Journal of Medical Education Development
e-ISSN: 2980-7670
www.https://edujournal.zums.ac.ir

Volume 18, Issue 3
December 2025
Pages 118-129

Review Article

Educational data mining in medical education: a scoping review

Meisam Dastani [1], Mostafa Kashani [2*]

[1] Social Determinants of Health Research Center, Gonabad University of Medical Sciences, Gonabad, Iran
[2] Sirjan School of Medical Sciences, Sirjan, Iran

Article info

Article history:
Received 20 Feb. 2025
Revised 16 Mar. 2025
Accepted 13 Jul. 2025
Published 1 Oct. 2025

***Corresponding author:**
Mostafa Kashani, Sirjan School of Medical Sciences, Sirjan, Iran.
Email: mostafa.kashani@sirums.ac.ir

How to cite this article:
Dastani M, Kashani M. Educational data mining in medical education: a scoping review. *J Med Edu Dev*. 2025;18(3):118-129.

Abstract

Background & Objective: The increasing complexity and volume of data in medical education highlight the importance of using advanced analytical techniques, such as data mining, to analyze educational data. This review aims to identify and assess the applications of educational data mining in medical education.

Materials & Methods: This research is a scoping review conducted based on the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR). Data was collected on January 8, 2025, utilizing search strategies specifically tailored for the Scopus, Web of Science (WOS), and PubMed databases. The inclusion criteria consisted of research articles related to medical education and data mining. In contrast, the exclusion criteria included non-research articles, articles written in languages other than English, and articles that had been retracted. The screening of articles was performed in three stages: titles, abstracts, and full texts. Finally, the selected articles were reviewed and reported based on data mining tools, algorithms, software, and results.

Results: The data mining applications identified were categorized into four main themes: predicting students' performance, identifying at-risk students, analyzing student interactions in online learning, and evaluating the quality of exams. Algorithms that are commonly used include Artificial Neural Networks (ANN), Naive Bayes, and K-means clustering.

Conclusion: Data mining is a powerful tool for analyzing educational data, particularly for planners in medical sciences. It can help improve the quality of educational systems and enhance student academic success through its various techniques. The intentional use of data mining can also support strategic decision-making within educational systems, leading to improved teaching quality and a reduction in socio-economic disparities among students.

Keywords: educational data mining, data mining, medical sciences, student

Introduction

Medical sciences education is critical because it is one of the fields that provides professional staff for health services. In contemporary times, modern tools such as artificial intelligence play a vital role in advancing and enhancing education in medical sciences, contributing significantly to improving learning quality and facilitating education [1]. Additionally, artificial intelligence has demonstrated its capabilities in the field of medical sciences [2]. An emerging and effective set of tools in the field is Educational Data Mining (EDM). Leveraging artificial intelligence and machine learning, EDM can effectively identify complex patterns and uncover hidden relationships in educational data [3].

EDM is an interdisciplinary field that combines education and computer science to extract meaningful patterns and actionable knowledge from large-scale educational data. Unlike traditional statistical techniques that test predefined hypotheses, EDM emphasizes a data-driven discovery approach, where hypotheses emerge from the data itself [4]. By employing various algorithms and tools, EDM enables researchers to examine student behavior, detect learning patterns, and optimize

educational settings [5, 6]. One of the key theoretical foundations of EDM is the learner-centered analysis framework, which emphasizes analyzing students' interactions within their learning environments. Additional frameworks, including the learning cycle, cognitive models, and motivational theories, provide conceptual bases for interpreting data within the context of medical education. Identifying and applying these frameworks allows for a deeper understanding of educational dynamics and supports more meaningful analyses [7, 8]. EDM encompasses a range of analytical tasks, including description, prediction, estimation, classification, clustering, and association analysis. These tasks are implemented through numerous data mining techniques, including regression, Naive Bayes, decision trees, neural networks, K-means clustering, Apriori, and FP-Growth algorithms [9]. In practice, EDM has been applied to predict student dropout rates using regression models [10, 11], forecast academic performance using techniques such as random forests, k-nearest neighbors, and support vector machines [12], and discover hidden relationships in educational datasets through association rule mining [13]. Clustering algorithms have also been used to group students based on performance indicators [14]. Several studies have demonstrated the practical value of EDM. For example, Rueangket et al. identified key factors affecting medical students' learning outcomes [15], while other research highlighted the predictive power of academic history and demographic features in determining student success [16]. Additionally, Yağcı employed a machine learning-based model to predict students' final grades using midterm results, highlighting how data-driven insights can inform decision-making in higher education [12]. Similarly, the study by Feng et al. combined clustering, discriminant analysis, and Convolutional Neural Networks (CNNs) to predict students' academic performance, highlighting the potential of hybrid EDM models to enhance prediction accuracy and early intervention in education [17]. Overall, EDM has demonstrated significant potential in improving educational quality, identifying at-risk students, and facilitating targeted interventions.

Despite the growing use of EDM in medical education, a comprehensive and structured review of its applications, tools, techniques, and outcomes has yet to be conducted. Moreover, it remains unclear which educational aspects have received the most attention in existing studies and what gaps persist in the application of these techniques within medical education. Considering the increasing complexity of medical education and the vast amount of data generated in the medical field, there is a growing need for advanced tools and techniques to analyze this data and extract meaningful insights. Therefore, the objective of this study is to conduct a scoping review to (1) identify the applications of data mining in medical education, (2) categorize the tools and techniques employed, and (3) summarize the key outcomes reported in the literature. By systematically reviewing relevant publications, this study aims to provide a comprehensive overview of how data mining contributes to the advancement of medical education. The data items extracted in this review were aligned with the PCC (Population, Concept, and Context) framework, which served as the foundation for defining the study's scope. The "Population" element referred to any studies focusing on learners, educators, or educational environments within the field of medical education. This included, but was not limited to, medical students, instructors, and academic institutions involved in various forms of medical teaching and learning. The "Concept" centered on the application of data mining or text mining techniques. This included studies that utilized algorithmic approaches or computational models to extract patterns, predict outcomes, or derive insights from educational data. Both structured data mining methods and unstructured text mining approaches were considered within this category, providing a broader perspective on the analytical tools used in the studies. The "Context" referred to any educational setting related to medicine. This encompassed a wide range of academic environments, including undergraduate medical education, postgraduate training, continuing medical education, and professional development programs. By including diverse contexts, the review aimed to capture the full breadth of how data mining techniques are employed across different stages and settings of medical education.

Materials & Methods

Design and setting(s)

This study was a scoping review conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines [18]. The objective was to identify and evaluate studies on the application of data mining in medical education.

Research framework

The research objectives were defined using the PCC (Population, Concept, and Context) framework. The

population included studies related to learners, educators, or educational environments in the field of medical education. This concept refers to the application of data mining or text mining techniques. Context included any medical education context, such as undergraduate, graduate, continuing medical education, or professional training programs.

Information sources and search strategy

The literature search was carried out on January 8, 2025, across three reputable databases: PubMed, Web of Science (WOS), and Scopus. The search was designed to retrieve all relevant articles related to data mining, education, and medical sciences. No time restrictions were applied, and the coverage extended to all articles published up to January 2025. The search strategy used the following queries: Scopus: TITLE-ABS-KEY ((education* OR "literacy program*" OR "training program*" OR workshop) AND ("data mining" OR "text mining") AND (medic*)); Web of Science (WOS): TS = ((education* OR "literacy program*" OR "training program*" OR workshop) AND ("data mining" OR "text mining") AND (medic*)); PubMed: (education*[Title/Abstract] OR "literacy program*" [Title/Abstract] OR "training program*" [Title/Abstract] OR workshop[Title/Abstract]) AND ("data mining" [Title/Abstract] OR "text mining" [Title/Abstract]) AND (medic*[Title/Abstract]) It is worth noting that the term "text mining" was included in the search strategy because it refers to the process of extracting patterns, concepts, or meaningful information from unstructured textual data such as narrative notes, open-ended responses, or written messages. Since text, mining focuses on extracting meaningful information from unstructured textual data—unlike data mining, which deals with structured databases [19]—its inclusion was intended to ensure comprehensive coverage of relevant literature.

Eligibility criteria and study selection

All search results were limited to English-language articles. The results were imported into EndNote, and duplicate entries were removed. The screening procedure was then carried out through multiple steps: First, title screening was performed, where research that did not pertain to the study goals was eliminated after examining the titles of the publications. Second, after selecting relevant papers for full-text review, the abstracts of the remaining publications were assessed. Third, to ensure all inclusion criteria were satisfied, the entire texts of the selected articles were reviewed. Two reviewers with expertise in information technology and medical education independently conducted screening. In cases of disagreement, conflicts were resolved through discussion and consensus; if needed, a third reviewer was consulted to reach a final decision. The inclusion criteria for this study comprised original research articles, publications related to medical education, and studies that employed text mining or data mining as a research methodology. Conversely, the exclusion criteria included non-research articles, such as reviews and letters to the editor, studies not written in English, publications without full-text availability, and articles that had been retracted.

Data collection methods & analysi

Lastly, the articles were reviewed and analyzed to address the objectives and research questions identified in this scoping review. The outcomes of this review are presented in a table that includes the authors and year of the study, objective, dataset, algorithms, software used, results, and conclusions.

Results

Figure 1 illustrates the process of finding, screening, and selecting relevant papers for this scoping review, as depicted in the PRISMA diagram. The first step includes searching three databases: PubMed, Web of Science (WOS), and Scopus. This step yielded a total of 2,122 studies. Among them, after removing duplicates and performing an initial screening based on titles and abstracts, a large number of articles were excluded. In the next stage, 78 studies were selected for full-text review. In the next step, 15 studies were excluded as irrelevant to the subject, 47 studies were excluded because they were not original articles, and two studies were excluded due to a lack of full-text access or being written in a language other than English, leaving a total of 78 studies for review. Finally, 14 studies were selected as eligible articles for analysis and conclusions in this study. This process clearly shows the systematic identification and selection of articles. **Table 1** presents the results of the selected studies in the area of data mining in medical education. Educational data include students' grades, demographics, exam scores, online learning environment behavior, and academic and behavioral characteristics (**Table 1**). In EDM, these data are analyzed mainly to predict student performance, identify at-risk students, evaluate exam quality, discover association rules, and analyze student interaction. The software used in this

analysis includes Python (with libraries such as Scikit-learn, TensorFlow, Keras, Pandas, and NumPy), R, SPSS, Weka, Orange, MATLAB, and Stata. These tools are among the leading options for processing educational data due to their advanced analytical capabilities [20-27, 29-33].

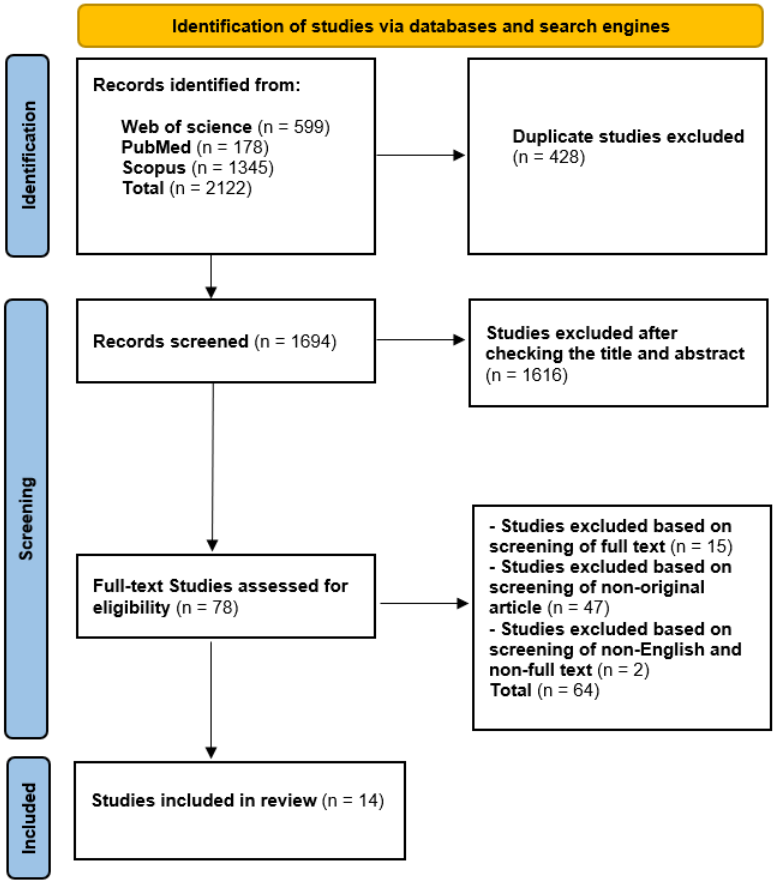

Figure 1. Flow chart of the study selection process according to the PRISMA guidelines

In the realm of EDM, various applications have emerged, including predicting student performance, identifying at-risk students, assessing the quality of online exams, discovering association rules, and analyzing student interactions in online learning environments. For instance, predicting student performance often involves algorithms, such as Artificial Neural Networks (ANN) and Naive Bayes, which utilize various admission criteria and demographic features to forecast the performance of medical students. The outcomes of these studies suggest that enhancing the prediction of students' performance [20, 23, 29] is achievable through the optimal allocation of weights for admission factors. For predicting at-risk students, ANN and Naive Bayes are used to predict first-year medical students who may be at risk. The results showed that students' prior knowledge is the most critical factor in predicting academic success [21, 33]. Additionally, EDM has successfully predicted the complexity level of exercises in online learning systems [30]. In the field of evaluating the quality of online exams, k-means clustering has been used to analyze the quality of online exams during the covid-19 pandemic. It was found that the experiences from the first semester influenced the characteristics of second-semester exams. Consequently, it was suggested that appropriate guidelines should be established and that more advanced classification questions should be included [22]. Apriori and Eclat algorithms were utilized to uncover association rules in the Iranian national medical entrance exam data within the domain of association rule discovery. The results indicated that students with high scores are accepted for the exam, while those with low scores are rejected, regardless of other factors [24]. The use of SNA to study student

interactions in online environments has garnered increasing interest, particularly in the context of online Problem-Based Learning (PBL) environments. The results showed that student-instructor interactions have a positive correlation with student performance, and that SNA indicators can predict low-performing students with an accuracy of 93.3% [27].

ANN, Naive Bayes, decision trees, random forests, logistic regression, KNN, and clustering algorithms such as K-means and PAM are some of the algorithms commonly used in these studies. These algorithms are widely utilized due to their ability to perform prediction, classification, and pattern recognition within educational data [20-25, 27, 31-33]. According to the findings obtained from the reviewed studies, the applications of educational data mining in medical education can be categorized into five main domains, as presented in **Table 2**.

Table 1. Comprehensive analysis of selected studies on educational data mining applications in medical education

No.	Authors	Objective	Dataset	Algorithms used	Software used	Results and conclusion
1	Investigation of the relationship between pre-admission criteria and medical students' performance and propose optimal admission criteria weights	Data on medical students at King Khalid University in Saudi Arabia (including high school scores, general aptitude tests, and standardized progress tests)	DT, NN, RF, NB, KNN	Python (SciPy.stats, scikit-learn, SciPy.optimize)	Optimal weights for pre-admission criteria: High school score (0.3), general aptitude test (0.2), and standardized progress test (0.5). NN and NB achieved highest performance in predicting student outcomes	Investigation of the relationship between pre-admission criteria and medical students' performance and propose optimal admission criteria weights
2	Predicting academic progress of first-year medical students and identifying at-risk students (categorized as regular or irregular, based on academic standing)	Data on 7,976 medical students at the National University of Mexico (including demographic information, academic records, and diagnostic test results)	ANN, NB	Python (Scikit-learn, TensorFlow, Keras), R	ANN model slightly outperformed NB in classifying students as regular or irregular. Prior academic knowledge emerged as the strongest predictor of academic success	Predicting academic progress of first-year medical students and identifying at-risk students (categorized as regular or irregular, based on academic standing)
3	Reviewing the quality of online exams conducted in two consecutive semesters during the covid-19 pandemic and analyzing data using k-means clustering	Data on 1,269 online multiple-choice exams at Birjand University of Medical Sciences (first semester: 535 exams, second semester: 734 exams)	K-means clustering	SPSS 19 and rattle package in R 3.6.3	Average percentage of correct answers: 69.97 ± 19.16. First semester: 43% very difficult, 16% difficult, 7% moderate. Second semester: 43% difficult, 16% moderate, 41% easy. First semester experience affected second semester exam characteristics	Reviewing the quality of online exams conducted in two consecutive semesters during the covid-19 pandemic and analyzing data using k-means clustering
4	Predicting the CMBSE using classical and hybrid machine learning models	Data on 1,005 medical students from top universities in Iran (including demographic information, basic science course grades, GPA, and CMBSE exam information)	LR, KNN, SVM, RF, ADA, XGB, Stacking	Python 3.9.1 and Scikit-learn, Pandas, NumPy	Hybrid ML models, particularly RF and Stacking, delivered highest predictive accuracy (83%) in classifying students' pass/fail status on CMBSE. Regression-based models closely aligned predicted scores with actual results	Predicting the CMBSE using classical and hybrid machine learning models
5	Discovering hidden rules and patterns from data in a national medical exam database using association rule mining algorithms (Apriori and Eclat) and DEA	Data on 7,723 participants in the national medical entrance exam of Iran (including demographic information, exam scores, and other relevant features)	Apriori, Eclat	Weka, R	Students with highest scores were admitted, while those with lowest were rejected, regardless of other features. This method effectively identified patterns that did not influence admission decisions	Discovering hidden rules and patterns from data in a national medical exam database using association rule mining algorithms (Apriori and Eclat) and DEA
6	Reviewing the factors affecting the continuous use of an adaptive learning platform (Osmosis) by medical students using EDM methods	Data on 6,787 medical students in the United States who used the Osmosis platform (responses to MCQs and flashcards, response accuracy, confidence level, time spent per item, and device type) between August 1, 2014, and July 31, 2015	Multivariate LR	Stata 13	Frequent use of mobile devices, participation in group activities, and holding a paid subscription were significantly associated with sustained platform usage. Regular users rated the platform more positively but response accuracy did not differ substantially from less frequent users	Reviewing the factors affecting the continuous use of an adaptive learning platform (Osmosis) by medical students using EDM methods

No.	Authors	Objective	Dataset	Algorithms used	Software used	Results and conclusion
7	Ting et al.	Examining the mediating role of academic performance in the relationship between internet use, gender, diet, illicit drug use, and delinquency among digital-era youth	Data from Wave III of the Add Health longitudinal study (including demographic, health, education, and crime-related behavior information)	Pearson correlation analysis, mediation analysis	IBM SPSS Statistics 26.0, PROCESS Macro 4.2	Academic performance mediates the relationship between internet usage, gender, dietary habits, drug use, and delinquent behavior. Results emphasize the importance of improving academic success to help reduce risk behaviors in youth
8	Saqr et al.	Examining the role of SNA in understanding online PBL interactions and predicting student performance	Data on student and instructor interactions from the Moodle learning management system (including information on posts, responses, and final student performance)	SNA, multivariate LR, LR	SPSS, Gephi, RapidMiner	SNA metrics revealed positive correlation between student–instructor interactions and academic performance in online PBL environments. The approach achieved 93.3% accuracy rate in identifying low-performing students
9	Sattari and Samouei	Predicting the performance of medical university professors in delivering virtual education during the covid-19 pandemic based on problem-solving methods and demographic characteristics	Data on 252 professors from medical universities in Iran (including demographic information, problem-solving methods, and performance in virtual education)	RF, CHAID, ID3	Not declared	Behavioral traits and personal capabilities significantly influenced professors' effectiveness in delivering virtual education during the pandemic. Key factors included self-regulation, persistence, interpersonal problem-solving, and tendency to seek challenges
10	Rakhmanov and Dane	Improving the prediction accuracy of first-year student performance using the Wonderlic Personnel Test and the ROCF test	Data on 111 computer science and medical students from the University of Nigeria (including IQ test results, ROCF test results, and academic performance)	LR, multiple regression	Not declared	The combination of IQ test scores and results from the ROCF test proved a valuable predictor of academic performance. This integrated approach is recommended as a supplementary tool for developing reliable predictive models in education
11	Fang et al.	Predicting the difficulty level of exercises in online education systems using the BEDP framework	Two datasets include 913,851 math exercises and 28,360 medical exercises	Deep Multimodal Feature Extraction, BSR-ADVI	Not declared	The BEDP framework, incorporating deep multimodal feature extraction and Bayesian Softmax Regression, achieved high accuracy in predicting exercise difficulty. It effectively handled diverse educational content and uncertainty in difficulty estimation
12	Hussain et al.	Comparing the performance of different data mining algorithms (including clustering, classification, and association rule extraction) in analyzing data of students applying for admission to medical colleges in Assam, India	Real data, including 666 samples with 11 features from the CEE for admission to medical colleges in Assam, India	K-means, Hierarchical Clustering, PAM, SOM, DT (J48), NN (MLP), NB, Apriori	Orange, Weka, R Studio	Among various classification and clustering algorithms tested, the MLP achieved the highest classification accuracy at 90.84%. For clustering, K-means and PAM outperformed hierarchical clustering, with a silhouette score of 0.54. Association rule mining using Apriori algorithm revealed functional patterns in admission data
13	Ebrahimzadeh et al.	Predicting academic failure among medical students at Lorestan University of Medical Sciences using a classification tree	Academic data of all medical students at Lorestan University of Medical Sciences (including student GPA, duration of study, academic probation, withdrawal, expulsion, and comprehensive exam results)	CART	SPSS 22	Academic failure occurred in 26.4% of the student population. Key predictors included enrollment in guest courses, class sizes of fewer than 40 students, regional quota admissions, and gender, all of which significantly influenced the likelihood of academic failure
14	Saadatdoost et al.	Knowledge discovery from higher education data using SOM and analyzing the Unified Distance Matrix (U-Matrix) and Component Planes	Academic data of students from Iranian medical universities (including year of registration, type of education, mode of study, and level of education)	SOM	MATLAB	From 1988 to 2005, enrollment at Tehran University of Medical Sciences declined. However, recent trends show an increase in postgraduate student numbers, along with a rise in part-time and evening courses, indicating these formats have become financially beneficial for universities

Abbreviations: DT, decision tree; NN, neural network; RF, random forest; NB, naive Bayes; KNN, k-nearest neighbor; ANN, artificial neural network; LR, logistic regression; SVM, support vector machine; ADA, AdaBoost; XGB, XGBoost; SNA, social network analysis; CHAID, chi-square automatic interaction detection; ID3, iterative dichotomiser 3; ROCF, Rey-Osterrieth Complex Figure; BSR-ADVI, Bayesian Softmax Regression with Automatic Differentiation Variational Inference; PAM, partitioning around medoids; SOM, self-organizing map; MLP, multi-layer perceptron; CART, classification and regression trees; GPA, grade point average; CMBSE, Comprehensive Medical Basic Sciences Examination; EDM, educational data mining; MCQ, multiple choice question; DEA, data envelopment analysis; PBL, problem-based learning; BEDP, Bayesian inference-based exercise difficulty prediction; CEE, common entrance exam.

Table 2. Main thematic categories of educational data mining in medical education

No.	Key thematic category	Description
1	Prediction of student academic performance	A significant portion of the reviewed studies [20, 23, 29, 31] focused on predicting students' academic success or failure. These studies demonstrated that algorithms such as ANNs, decision trees, and logistic regression perform well in prediction tasks. Most of these studies emphasize that combining individual-level variables, such as high school grades, GPA, or prior academic knowledge, plays a crucial role in enhancing the accuracy of predictive models.
2	Identification of at-risk students	Research works [21, 32] employed algorithms like Naive Bayes and ANNs to identify students at risk of academic failure or dropout. The findings indicate that prior knowledge and individual characteristics are among the most significant factors in determining the at-risk status of these students.
3	Analysis of exam quality and online education	Studies [22, 25, 26, 30] analyzed online examinations or instructor performance in virtual education to evaluate the quality of assessment processes during the pandemic. Algorithms such as K-means clustering and decision trees were employed to evaluate exam quality and identify behavioral patterns. A common theme in these studies is the impact of prior experience and individual psychological traits on the enhancement of virtual education.
4	Association rule mining and pattern analysis	In studies such as [24, 31, 33], algorithms like Apriori and Eclat were used to extract hidden rules from educational data. These rules typically provide insights into the combinations of variables and their impact on students' acceptance or failure. However, some of the findings emphasize that certain features, although present in the datasets, have no practical influence on final decision-making.
5	Analysis of interactions in online learning environments	Studies [25, 27] utilized SNA to examine student–instructor interactions within online learning management systems. The findings suggest that both the quantity and quality of these interactions significantly impact students' academic performance. Moreover, SNA proved effective in accurately identifying students who are low performing.

Note: This table presents a qualitative categorization of educational data mining applications in medical education based on systematic literature review methodology.
Abbreviations: ANNs, artificial neural networks; GPA, grade point average; SNA, social network analysis.

However, a review of the existing literature reveals that many studies in the field of EDM suffer from methodological limitations that can compromise the accuracy and generalizability of their findings. One standard limitation is the reliance on narrowly defined educational indicators—such as grades and standardized test scores—while overlooking critical variables, including students' cognitive abilities, motivational factors, social characteristics, and the contextual features of their learning environments, as well as teacher attributes [20, 21, 32]. Some studies have exclusively employed a limited range of cognitive assessments, such as the Rey-Osterrieth Complex Figure (ROCF) test, while neglecting other well-established psychometric instruments that could provide a more comprehensive view of learner characteristics [29]. Moreover, several investigations suffer from small sample sizes or biased data selection practices—for instance, excluding students with incomplete records—, which may distort the accuracy of predictive models [21, 25]. Another recurring issue is the absence of cross-validation or sensitivity analysis, which are essential for assessing the robustness and reliability of machine learning models [25]. Additional limitations observed include the unavailability of detailed demographic data [22], the narrow focus on a specific subgroup of users within an educational platform [25], and an overdependence on historical or context-dependent datasets [23, 26, 27]. Collectively, these shortcomings highlight the need for future research to incorporate more diverse and multidimensional datasets, adopt advanced hybrid modelling approaches, and account for the cultural, social, and technological aspects of modern educational systems to enhance research quality and applicability.

Discussion

The findings of this study suggest that EDM plays a crucial role as an effective and flexible tool in analyzing educational data and enhancing learning systems. Educational institutions store various types of student data, ranging from academic records to personal details such as parental income and educational background [4]. However, one of the key limitations observed in the reviewed studies is their heavy reliance on localized and context-specific datasets, which limits the generalizability of findings to broader educational settings [34]. In this regard, Baker, Martin, and Rossi also pointed out the challenges of implementing EDM systems on a larger scale, particularly in resource-constrained environments or when dealing with incomplete datasets [34].

Furthermore, the findings from this review underscore the importance of various types of educational data, including student grades, demographic characteristics, and online interaction records. Utilizing advanced data mining techniques enables the identification of hidden

patterns within the data. Admission criteria and demographic variables play a crucial role in predicting student success. In line with this, Waheed et al. achieved 85% prediction accuracy using demographic and geographic features [36]. Similarly, Costa-Mendes et al. predicted academic performance using variables such as income, age, employment status, cultural level, and place of residence [37]. Cruz et al. also utilized these socio-economic variables to predict student performance [38]. In another study, Alam enhanced educational outcomes by analyzing diverse data, including academic records, demographics, and external sources like social media platforms and online forums [39].

Nevertheless, the use of sensitive data, such as socio-economic and personal information, raises significant concerns regarding data privacy and ethics, which many of the reviewed studies failed to address [40] entirely. Romero and Ventura emphasized that these challenges can substantially hinder the real-world adoption of EDM in educational settings [41].

Furthermore, identifying at-risk students using data mining algorithms underscores the importance of prior data and its impact on academic success [21, 31]. For example, Ahmad and Shahzadi predicted poor academic performance with 85% accuracy by analyzing variables related to study habits, learning skills, and student-instructor interaction [42]. However, one major challenge in this area is the absence of a standardized definition for "at-risk" students across studies, which complicates cross-study comparisons [43]. Furthermore, implementing these predictive models in real-world scenarios necessitates a robust technological infrastructure and trained personnel [44].

In the evaluation of online exams, clustering-based techniques—such as the k-means algorithm—have demonstrated that multiple factors, including students' prior experiences, can influence assessment quality. These analyses suggest that incorporating higher-order questions and structured exam design can enhance the reliability of online evaluations [22]. However, many studies did not adequately consider external factors, such as internet accessibility or the quality of student devices, which can significantly impact assessment outcomes [45]. In the area of association rule mining, algorithms such as Apriori and Eclat have revealed functional patterns for determining student admission decisions based on test scores [24]. Moreover, analyses of student-instructor interactions in online learning environments suggest that positive engagement improves academic performance. The use of SNA indicators to identify low-performing students highlights the potential of EDM for designing interaction-focused learning models [27]. However, the reliance of SNA on dense interaction data makes it challenging to apply in low-resource educational environments (46). Additionally, inconsistencies found in the literature—such as the varying effects of online interaction on performance across different contexts—indicate the need for further empirical studies, as emphasized by Siemens [47].

Overall, this review highlights the substantial potential of EDM to transform the field of medical education. Nonetheless, challenges such as the lack of high-quality data, ethical and privacy concerns, and infrastructural limitations still impede its widespread application [41]. Future research should prioritize the development of generalizable EDM frameworks, the integration of ethical safeguards, and the creation of scalable solutions for settings with limited resources.

One of the limitations of the present study is that, as a scoping review, its primary objective was to map and categorize the scope of existing research rather than conduct a systematic evaluation of study quality or generalizability. Moreover, several reviewed studies demonstrated weaknesses, including small sample sizes, reliance on outdated datasets, the absence of validation methods (e.g., cross-validation), and limited analytical depth. For instance, studies relying on data from over a decade ago may no longer align with current medical education practices due to structural and technological changes in the field.

Conclusion

The current work illustrates the capability of EDM as a broad yet robust tool for identifying and improving the quality of learning systems. Advanced algorithms like ANN, decision trees, and random forests not only allow educational institutes to predict students' performance but can also help them identify latent patterns in the underlying data and use those insights for strategic decision-making. Additionally, it aids in identifying at-risk students, improving exam quality, and fostering communication between students and instructors by leveraging educational, demographic, and interaction data in online learning environments. The educational process can be optimized by the design of data-driven learning systems and predictive analyses, which will generate early intervention opportunities for students. The future of the educational system will be primarily driven by data-centricity and the adoption of sophisticated techniques, as this study demonstrates.

The findings of this research will provide valuable insights for researchers and policymakers on the application of data mining and machine learning algorithms in medical education. This, in turn, can lead to enhancements in the quality of teaching and student performance. These techniques can be used efficiently to analyze students' learning behaviour and predict their academic performance.

Additionally, these can be utilized to develop student-specific learning frameworks and identify areas of weakness, while offering remedial sessions to enhance learning outcomes.

The results of this study may help policymakers formulate evidence-based policies that can enhance constructive interactions between students and instructors, optimize online learning environments, and construct better-quality examinations. These strategies ultimately contribute to the advancement of the educational system and may lead to the training of more proficient and effective medical science professionals.

Ethical considerations

This is a review study, and all ethical principles in research have been observed. The study was approved by the Ethics Committee of Sirjan School of Medical Sciences under the ethics code IR.SIRUMS.REC.1403.049.

Artificial intelligence utilization for article writing

Some sentences and paragraphs were translated and linguistically enhanced using ChatGPT-4o (OpenAI) to improve the clarity and academic quality of the English writing.

Acknowledgment

Not applicable.

Conflict of interest statement

The authors declare that they have no conflicts of interest at any stage of the study.

Author contributions

MD contributed to the conception and design of the study, literature review, data extraction, writing the initial draft, and critical revision of the manuscript. MK supervised the research process, contributed to the methodological framework, provided critical feedback, and edited the final version of the manuscript.

Funding

The Sirjan School of Medical Sciences supported this study (Grant No: 403000047).

Data availability statement

All data related to this study are fully available within the article.

References

1. Dastani M, Mohseni M. Artificial intelligence chatbots in medical education: a literature review of potential benefits and challenges. *Int J Med Rev.* 2024;11(2):710–5. https://doi.org/10.30491/ijmr.2024.459251.1282

2. Dastani M, Kashani M, Mohseani M. Mapping the research on the application of artificial intelligence in cancer: a scientometric analysis. *Health Manag Inf Sci.* 2023;10(3):121–9. https://doi.org/10.30476/jhmi.2024.101181.1198

3. Dol SM, Jawandhiya PM. Classification technique and its combination with clustering and association rule mining in educational data mining—a survey. *Eng Appl Artif Intell.* 2023;122:106071. https://doi.org/10.1016/j.engappai.2023.106071

4. Dutt A, Ismail MA, Herawan T. A systematic review on educational data mining. *IEEE Access.* 2017;5:15991–6005. https://doi.org/10.1109/ACCESS.2017.2654247

5. Salloum SA, Alshurideh M, Elnagar A, Shaalan K. Mining in educational data: review and future directions. In: Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020); 2020. Cham: Springer International Publishing; 2020. p. 92–102. https://doi.org/10.1007/978-3-030-44289-7_9

6. Nahar K, Shova BI, Ria T, Rashid HB, Islam AS. Mining educational data to predict students performance: a comparative study of data mining techniques. *Educ Inf Technol.* 2021;26(5):6051–67. https://doi.org/10.1007/s10639-021-10575-3

7. Khine MS. Educational data mining and learning analytics. In: Khine MS, editor. *Artificial intelligence in education: a machine-generated literature overview.* Singapore: Springer Nature Singapore; 2024. p. 1–159. https://doi.org/10.1007/978-981-97-9350-1

8. Koedinger KR, D'Mello S, McLaughlin EA, Pardos ZA, Rosé CP. Data mining and education. *Wiley Interdiscip Rev Cogn Sci.* 2015;6(4):333–53. https://doi.org/10.1002/wcs.1350

9. Aleem A, Gore MM, editors. Educational data mining methods: a survey. In: 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT); 2020 Apr 10–12. New York: IEEE; 2020. https://doi.org/10.1109/CSNT48778.2020.9115734
10. do Nascimento RL, das Neves Junior RB, de Almeida Neto MA, de Araújo Fagundes RA. Educational data mining: an application of regressors in predicting school dropout. In: Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018; 2018 Jul 15–19; New York, NY, USA. Cham: Springer International Publishing; 2018. p. 246–57. https://doi.org/10.1007/978-3-319-96133-0_19
11. Sarra A, Fontanella L, Di Zio S. Identifying students at risk of academic failure within the educational data mining framework. *Soc Indic Res.* 2019;146(1):41–60. https://doi.org/10.1007/s11205-018-1901-8
12. Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn Environ.* 2022;9(1):11. https://doi.org/10.1186/s40561-022-00192-z
13. Rojanavasu P. Educational data analytics using association rule mining and classification. In: 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON); 2019. New York: IEEE; 2019. p. 142–5. https://doi.org/10.1109/ECTI-NCON.2019.8692274
14. Li Y, Gou J, Fan Z. Educational data mining for students' performance based on fuzzy C-means clustering. *J Eng.* 2019;2019(11):8245–50. https://doi.org/10.1049/joe.2019.0938
15. Rueangket P, Thaebanpakul C, Sakboonyarat B, Prayote A. Educational data mining: factors influencing medical student success and the exploration of visualization techniques. *Front Educ.* 2024;9. https://doi.org/10.3389/feduc.2024.1390892
16. Batool S, Rashid J, Nisar MW, Kim J, Kwon HY, Hussain A. Educational data mining to predict students' academic performance: a survey study. *Educ Inf Technol.* 2023;28(1):905–71. https://doi.org/10.1007/s10639-022-11152-y
17. Feng G, Fan M, Chen Y. Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access.* 2022;10:19558–71. https://doi.org/10.1109/ACCESS.2022.3151652
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169(7):467–73. https://doi.org/10.7326/M18-0850
19. Alwidian SA, Bani-Salameh HA, Alslaity AA. Text data mining: a proposed framework and future perspectives. *Int J Bus Inf Syst.* 2015;18(2):127–40. https://doi.org/10.1504/IJBIS.2015.067261
20. Qahmash A, Ahmad N, Algarni A. Investigating students' pre-university admission requirements and their correlation with academic performance for medical students: an educational data mining approach. *Brain Sci.* 2023;13(3):456. https://doi.org/10.3390/brainsci13030456
21. Monteverde-Suárez D, González-Flores P, Santos-Solórzano R, García-Minjares M, Zavala-Sierra I, de la Luz VL, et al. Predicting students' academic progress and related attributes in first-year medical students: an analysis with artificial neural networks and Naïve Bayes. *BMC Med Educ.* 2024;24(1):74. https://doi.org/10.1186/s12909-023-04918-6
22. Abedi F, Eghbali B, Akbari N, Sadr E, Salmani F. Online assessment in two consequent semesters during COVID-19 pandemic: K-means clustering using data mining approach. *J Educ Health Promot.* 2022;11:307. https://doi.org/10.4103/jehp.jehp_1466_21
23. Mastour H, Dehghani T, Moradi E, Eslami S. Early prediction of medical students' performance in high-stakes examinations using machine learning approaches. *Heliyon.* 2023;9(7). https://doi.org/10.1016/j.heliyon.2023.e18248
24. Zehtab Hashemi H, Abedian S, Parvasideh P, Bahrevar Z, Madani S. Discovering rules from a national exam repository: a use case for data analysis from Iranian medical schools entry exam. *Stud Health Technol Inform.* 2022;294:796–800. https://doi.org/10.3233/SHTI220586
25. Menon A, Gaglani S, Haynes MR, Tackett S. Using "big data" to guide implementation of a web and mobile adaptive learning platform for medical students. *Med Teach.* 2017;39(9):975–80. https://doi.org/10.1080/0142159x.2017.1324949
26. Ting TT, Lim ET, Lee J, Wong JS, Tan JH, Tam RC, et al. Educational big data mining: mediation of academic performance in crime among digital age young adults. *Online J Commun Media Technol.* 2024;14(1):e202403. https://doi.org/10.30935/ojcmt/14026

27. Saqr M, Fors U, Nouri J. Using social network analysis to understand online problem-based learning and predict performance. *PLoS One.* 2018;13(9):e0203590. https://doi.org/10.1371/journal.pone.0203590
28. Sattari M, Samouei R. Predicting the performance of faculty members of medical universities in providing virtual learning in the covid-19 pandemic in terms of problem-solving methods and their individual social characteristics: a data mining study. *Tehran Univ Med J.* 2023;80(12):986–91.
29. Rakhmanov O, Dane S, editors. Improving the prediction accuracy of academic performance of the freshman using Wonderlic Personnel Test and Rey-Osterrieth Complex Figure. In: Information and Communication Technology and Applications; 2021. Cham: Springer International Publishing; 2021. https://doi.org/10.1007/978-3-030-69143-1_5
30. Fang J, Zhao W, Jia D, editors. Exercise difficulty prediction in online education systems. In: 2019 International Conference on Data Mining Workshops (ICDMW); 2019 Nov 8–11. New York: IEEE; 2019. https://doi.org/10.1109/ICDMW.2019.00053
31. Hussain S, Atallah R, Kamsin A, Hazarika J, editors. Classification, clustering and association rule mining in educational datasets using data mining tools: a case study. In: Cybernetics and Algorithms in Intelligent Systems; 2019. Cham: Springer International Publishing; 2019. https://doi.org/10.1007/978-3-319-91192-2_21
32. Ebrahimzadeh F, Hajizadeh E, Birjandi M, Feli S, Ghazi S. Predicting the incidence of academic failure in medical students of Lorestan university of medical sciences using classification tree. *Iran J Epidemiol.* 2018;14(3):234–45. https://www.cabidigitallibrary.org/doi/full/10.5555/20193160242
33. Saadatdoost R, Alex Tze Hiang S, Jafarkarimi H, editors. Application of self organizing map for knowledge discovery based in higher education data. In: 2011 International Conference on Research and Innovation in Information Systems; 2011 Nov 23–24. New York: IEEE; 2011. https://doi.org/10.1109/ICRIIS.2011.6125693
34. Baker RS. Challenges for the future of educational data mining: the Baker learning analytics prizes. *J Educ Data Min.* 2019;11(1):1–7. https://doi.org/10.5281/zenodo.3554745
35. Baker RS, Martin T, Rossi LM. Educational data mining and learning analytics. In: Rupp AA, Leighton JP, editors. *The Wiley handbook of cognition and assessment: frameworks, methodologies, and applications.* Hoboken: Wiley; 2016. p. 379–96. https://doi.org/10.1002/9781118956588.ch16
36. Waheed H, Hassan SU, Aljohani NR, Hardman J, Alelyani S, Nawaz R. Predicting academic performance of students from VLE big data using deep learning models. *Comput Human Behav.* 2020;104:106189. https://doi.org/10.1016/j.chb.2019.106189
37. Costa-Mendes R, Oliveira T, Castelli M, Cruz-Jesus F. A machine learning approximation of the 2015 Portuguese high school student grades: a hybrid approach. *Educ Inf Technol.* 2021;26(2):1527–47. https://doi.org/10.1007/s10639-020-10316-y
38. Cruz-Jesus F, Castelli M, Oliveira T, Mendes R, Nunes C, Sa-Velho M, et al. Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon.* 2020;6(6). https://doi.org/10.1016/j.heliyon.2020.e04081
39. Alam A, editor. The secret sauce of student success: cracking the code by navigating the path to personalized learning with educational data mining. In: 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN); 2023 Apr 21–22. New York: IEEE; 2023. https://doi.org/10.1109/ICSTSN57873.2023.10151558
40. Pardo A, Siemens G. Ethical and privacy principles for learning analytics. *Br J Educ Technol.* 2014;45(3):438–50. https://doi.org/10.1111/bjet.12152
41. Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2020;10(3):e1355. https://doi.org/10.1002/widm.1355
42. Ahmad Z, Shahzadi E. Prediction of students' academic performance using artificial neural network. *Bull Educ Res.* 2018;40(3):157–64. https://eric.ed.gov/?id=EJ1209686
43. Lonn S, Aguilar SJ, Teasley SD. Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. *Comput Human Behav.* 2015;47:90–7. https://doi.org/10.1016/j.chb.2014.07.013
44. Aldowah H, Al-Samarraie H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: a review and synthesis.

45. *Telemat Inform.* 2019;37:13–49. https://doi.org/10.1016/j.tele.2019.01.007
46. Nguyen Q, Thorne S, Rienties B. How do students engage with computer-based assessments: impact of study breaks on intertemporal engagement and pass rates. *Behaviormetrika.* 2018;45(2):597–614. https://doi.org/10.1007/s41237-018-0060-1
47. Joksimović S, Gašević D, Loughin TM, Kovanović V, Hatala M. Learning at distance: effects of interaction traces on academic achievement. *Comput Educ.* 2015;87:204–17. https://doi.org/10.1016/j.compedu.2015.07.002
48. Siemens G. Learning analytics: the emergence of a discipline. *Am Behav Sci.* 2013;57(10):1380–400. https://doi.org/10.1177/0002764213498851