Discover Computing

Review

Educational data mining: a 10-year review

Emi Kalita[1] · Solomon Sunday Oyelere[2,3,4] · Silvia Gaftandzhieva[5] · Kandala N. V. P. S. Rajesh[6] · Senthil Kumar Jagatheesaperumal[7] · Asmaa Mohamed[8] · Yomna M. Elbarawy[8] · Abeer S. Desuky[8] · Sadiq Hussain[1] · Mehmet Akif Cifci[9,10] · Paraskevi Theodorou[11] · Slavoljub Hilčenko[12] · Jiten Hazarika[13] · Tazid Ali[1]

Received: 13 October 2024 / Accepted: 5 May 2025

Published online: 19 May 2025

Abstract
This systematic review comprehensively examines the application and impacts of Educational Data Mining (EDM) over the past decade. It explores the use of various data mining tools and techniques, statistics, and machine learning algorithms in education. The review discusses how EDM helps understand and improve the learning experience, educational strategies, and institutional efficiency. It highlights the iterative process of EDM, its applications, and the benefits it offers to different stakeholders, including students, teachers, and educational institutions. The paper also discusses the challenges related to data ethics, privacy, and security in EDM. Key sections include a methodology for conducting the systematic review, exploring different data mining techniques and learning styles, and using Artificial Intelligence in EDM. The review concludes with a discussion of findings, future research directions, and a summary of the study's contributions and limitations.

Keywords Education data mining · Multimodal learning analytics · Artificial intelligence in education · Explainability in education

✉ Solomon Sunday Oyelere, Solomon.Oyelere@ltu.se; S.Oyelere@exeter.ac.uk; soyelere@uj.ac.za; Emi Kalita, emikalita@ymail.com; Silvia Gaftandzhieva, sissiy88@uni-plovdiv.bg; Kandala N. V. P. S. Rajesh, kandala.rajesh2014@gmail.com; Senthil Kumar Jagatheesaperumal, senthilkumarj@mepcoeng.ac.in; Asmaa Mohamed, asmaamohamed89@azhar.edu.eg; Yomna M. Elbarawy, y.elbarawy@azhar.edu.eg; Abeer S. Desuky, abeerdesuky@azhar.edu.eg; Sadiq Hussain, sadiq@dibru.ac.in; Mehmet Akif Cifci, mcifci@bandirma.edu.tr; meh.cifci@kvk.lt; Paraskevi Theodorou, theodoroup@unipi.gr; Slavoljub Hilčenko, slavoljubhilcenko@gmail.com; Jiten Hazarika, jiten_stats@dibru.ac.in; Tazid Ali, tazid@dibru.ac.in | [1]Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh 786004, India. [2]Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, 931 87 Skellefteå, Sweden. [3]Department of Computer Science, Faculty of Environment, Science and Economy, University of Exeter, Exeter, England. [4]Research Group On Data, Artificial Intelligence, and Innovations for Digital Transformation, Johannesburg Business School (JBS), University of Johannesburg, Johannesburg, South Africa. [5]Faculty of Mathematics and Informatics, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria. [6]School of Electronics Engineering, VIT-AP University, Vijayawada, India. [7]Department of Electronics & Communication Engineering, Mepco Schlenk Engineering College, Sivakasi 626005, India. [8]Department of Mathematics, Faculty of Science, Al-Azhar University (Girls Branch), Cairo 11754, Egypt. [9]Department of Computer Engineering, Bandirma Onyedi Eylul University, Balikesir 10200, Türkiye. [10]Engineering and Informatics Department, Klaipėdos Valstybinė Kolegija/Higher Education Institution, 92294 Klaipeda, Lithuania. [11]Department of Digital Systems, University of Piraeus, Piraeus, Greece. [12]College for Vocational Education of Preschool Teachers and Coaches, Subotica, Serbia. [13]Department of Statistics, Dibrugarh University, Dibrugarh 786004, India.

1 Introduction

Over the last decade, organisations from all industries and areas of activity have taken action to digitise their core activities. The advent of digital technologies is also revolutionising the way educational institutions carry out their core activities and provide services to key stakeholder groups [1]. Since the provision of high-quality services is decisive for the rise of educational institutions in the rankings of various organisations, and hence a prerequisite for attracting more students, many educational institutions annually invest funds for the implementation of software tools for conducting various forms of online learning (b-learning, e-learning, distance learning, hybrid learning), tracking the learning process, providing online administrative services for students and staff, etc. These changes were catalysed in the spring of 2020 when COVID-19 forced educational institutions in all countries to rapidly switch entirely to distance learning and provide online services to students and faculty [2, 3]. The way educational institutions functioned was no longer the same after the return to traditional forms of learning. Many teachers continued to build on and use the developed online courses and e-learning materials as a supplement to face-to-face training, as well as the governing bodies of educational institutions sought and allocated financial resources to introduce innovations that would improve the student experience.

Using the software solutions deployed daily, users generate a large amount of data about student enrolment and attendance, exam results, demographics, usage of software systems, etc. [4]. Governing bodies must understand the importance of data to leverage its insights for improving all processes in educational institutions [5], from training and research to overall management.

Technology has fundamentally transformed the way of capturing information [6]. The gathered data motivates the rapidly growing interest of researchers and the increase of research on extracting hidden and valuable information and patterns from large data sets to allow a better understanding of the core processes in educational institutions and decision-making for their improvement [7]. Educational data mining (EDM) is an interdisciplinary field that emerged in the 1990s [8] and has spurred a significant impact on education in the last decade and continues to grow [6, 9]. EDM employs data mining tools and techniques (Classification, Regression, Clustering, Association Rule Mining, Outlier Detection, Optimization Techniques, Text Mining and Social Networks Analysis), statistics, and machine learning algorithms (Decision Trees, Neural Networks, Naive Bayes, K -Nearest Neighbour, etc.) to discover previously unknown information, relationships and patterns in large data repositories [10]. EDM analyses large datasets in educational institutions to build models that extract knowledge from educational data to improve the student learning experience and the educational process to explain educational strategies for better decision-making and institutional efficiency [11]. EDM process is an iterative process with four phases [10]: Problem definition, Data preparation and gathering, Modelling and evaluating, and Deployment. Building accurate models requires advanced data science skills for choosing the appropriate machine learning algorithm for a given problem formulation and configuring a specific model with the optimal parameters' values [12]. EDM aims to create and enhance methods for analysing educational data, which frequently contain several levels of meaningful structure, to uncover new insights into how students learn in such environments [13]. EDM works toward improving educational processes by introducing better and more effective learning practices [14]. By extracting valuable information from data, any educational institution can achieve the highest quality level of services and improve decision-making processes [5].

1.1 Purpose and utility of educational data mining

EDM aims to utilise data-driven strategies such as machine learning, deep learning, explainable AI, and multimodal learning analytics to identify trends that can improve teaching strategies, student participation, and learning results. By thoroughly analysing educational datasets, EDM allows teachers, administrators, and policymakers to make informed decisions that support personalised learning, prompt intervention tactics, and curriculum enhancements. EDM benefits various stakeholders in ensuring high-quality services in the education sector students, teachers, system administrators, and governing bodies of educational institutions. Students can use EDM tools to get recommendations for course selection [10], learning content and tasks tailored to their individual needs and progress in the learning process [15], the implementation of which would improve the results achieved to date, offering learning paths based on the results of the performance of tasks, receiving information that will help them achieve the set learning goals and alerts about the potential possibility of achieving poor results in the final exams and dropping out of the education system [16]. EDM gives teachers valuable information, which helps them better understand how students learn. Based on this information, teachers can make decisions to improve the quality of courses and take measures to increase student success. The information obtained during the educational process allows them to discover student learning habits and preferences in e-learning, schedule courses [16], evaluate the effectiveness of learning activities, and make changes in the study programs according to student characteristics to improve course adaptation and customization [10]. In addition, the information obtained about students' performance allows them to organize collections of patterns based on student behaviour similarity in using course materials [17], detect usual and unusual patterns in student learning and behaviour [18–21], detect common mistakes, identify priority learning needs and classify students in groups [9, 16, 18, 19] based on their needs in guidance and monitoring to provide appropriate advising, feedback and tasks [15]. Teachers can predict students' performance and their final course grades [6, 12, 15, 19, 21–27], including based on data from previous years and midterm exam grades, identify at-risk students who need special attention and make informed decisions that lead to a higher student success rate [6, 9, 28] and timely interventions at the earliest stage possible to achieve optimal student performance and decrease student dropout rate [6, 15, 25, 29]. In turn, system administrators can use EDM to understand problems in systems usage to improve their efficiency and optimize the resources used according to user needs (optimal server size, network traffic distribution, etc.).

EDM can help the governing bodies improve the activities of their educational institution in all aspects and achieve their goals to enhance the efficiency and quality of education [30, 31]. EDM is a valuable source of information to predict student enrolment [10] and dropout rate [29], finding problems leading to dropping out [10], predict and increase student performance and graduation rates [9, 19, 31, 32], evaluate curriculum based on the competitiveness of graduates [33] and optimize its renewal [9] to attract more students and maximize the use of campus resources (human and material) [9, 16], monitoring institutional performance efficiently, etc. In addition, EDM helps governing bodies identify honorary students in the early stages and offers them well-deserved opportunities such as scholarships, internships, and workshops [15]. In such a way, the governing bodies make accurate decisions [34] and develop a policy that can increase the quality of education as a critical driver for scientific and technological advancement and countries' economic and social development [31, 32].

Despite the many benefits for all stakeholder groups, EDM also has some disadvantages related to data ethics, privacy, and security [10, 35]. Based on in-depth analysis, Guan and his colleagues [35] found three core problems in educational data ethics (violation of privacy during data collection, storage, and sharing; the deprivation of the ability to make independent choices by prediction of educational data; the lack of "forgetting ability") and propose the corresponding learner-centred problem- solving strategy, combined with technologies such as blockchain, 5G technology, and federated learning to form targeted solutions from different levels.

There are many general data mining tools (such as WEKA, SPSS, R, Rapid Miner, Elvira, H2O, SAS, Watson Analytics, Azure Machine Learning, Matlab, Orange, Statistica, NeuralWorks Professional II/PLUS) used for extracting, processing, analysis, and visualization of educational data. Although these tools have proven effective for analysing educational data [29], they remain inaccessible to stakeholders without advanced data mining skills, limiting the benefits of EDM. This fact encourages researchers to design and develop software tools that will allow the non-expert stakeholders to gain insights from data visualized in different dashboards [36] and conduct experiments with appropriate automated parameter configurations [12] to take advantage of using EDM.

1.2 Contributions of the article

The key contributions of this article are:

- This paper provides a decade-long systematic review of EDM, examining the evolution of machine learning, deep learning, and explainable AI in education. Analysing rigorously selected studies highlights EDM's key trends, methodologies, and applications.
- It follows a PRISMA-compliant systematic review process, refining inclusion and exclusion criteria, search methodologies, and quality appraisal techniques. It also introduces structured thematic analysis and quantitative synthesis, ensuring the rigorous evaluation of EDM studies.
- This paper offers an in-depth analysis of AI-driven techniques in education, including multimodal learning, blockchain-based learning analytics, and chatbot-driven education. It also discusses the role of XAI in improving the transparency and interpretability of data-driven educational decisions.

- This paper provides practical insights and recommendations on selecting the most suitable data mining techniques for various educational contexts. It identifies existing gaps in EDM research, proposes future research directions, and suggests how institutions can effectively implement AI-driven educational analytics.

1.3 Organization of the article

This article provides a deep systematic review of the application of educational data mining in the last decade—from 2013 to 2023. Section 2 describes the criteria for selecting articles for conducting the survey. Section 3 explores different data mining techniques applied in EDM, such as Classification, Regression, Clustering, Association Rule Mining, Outlier Detection, Optimization Techniques, Text Mining, and Social Networks Analysis. Section 4 reviews different learning styles, such as e-learning, hybrid, and blended learning. Section 5 discusses the usage of data fusion in multimodal learning analytics and educational data mining. Section 6 explains the use of AI in educational data mining. Finally, the authors discuss the findings (Sect. 7), give some future research directions (Sect. 8), and summarize the contributions and limitations of the current study (Sect. 9).

2 Methodology

In this section, we describe our method for performing a systematic literature review (SLR), covering search methods, database choices, criteria for inclusion and exclusion, and quality assessment. We adhere to the PRISMA framework to maintain rigour and clarity. Figure 1 shows the updated PRISMA diagram, which offers a detailed view of the four-stage selection criteria and the systematic search strategy.

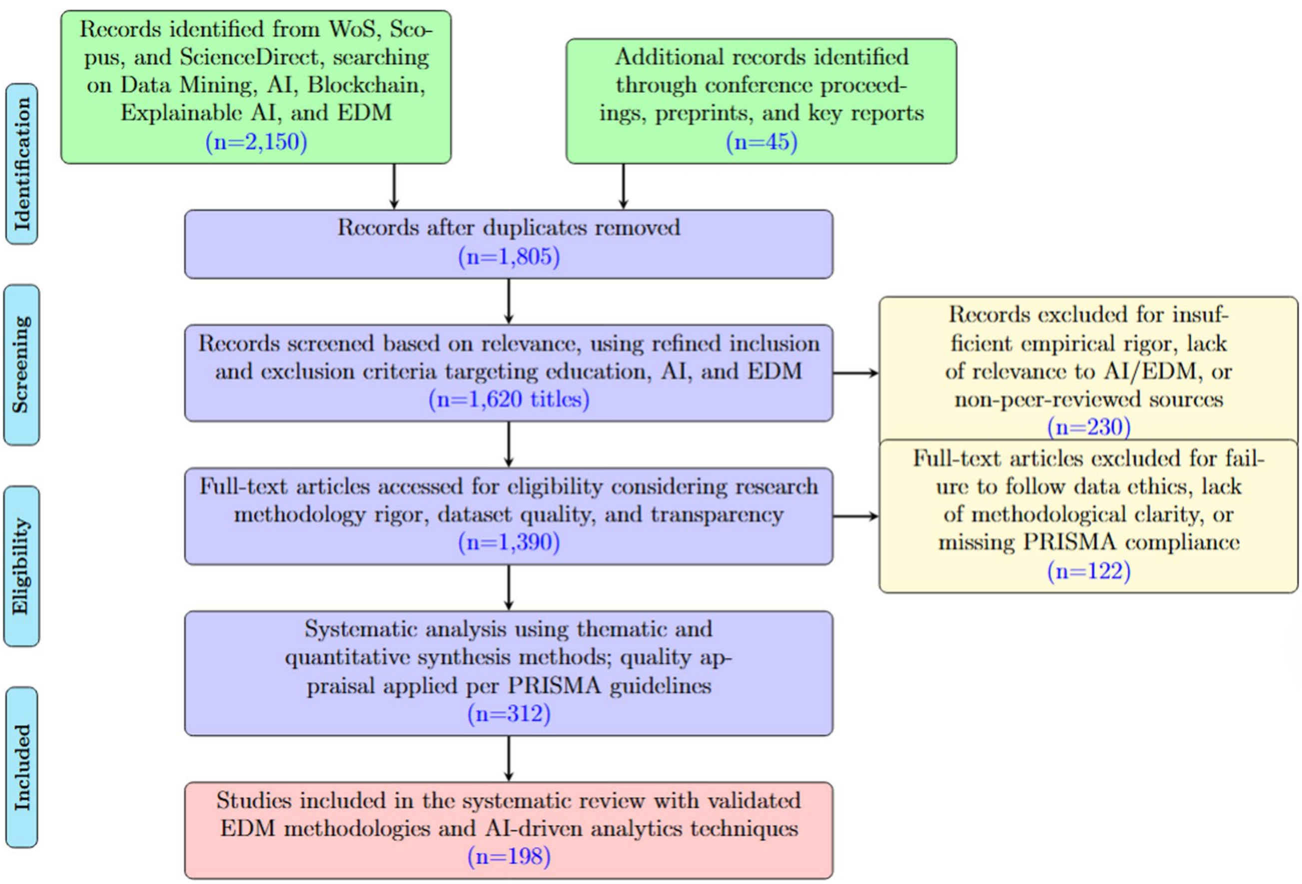

Fig. 1 Our methodology for systematic review: the PRISMA flowchart for literature search and selection

2.1 Search methods

To conduct a thorough and methodical review of Educational Data Mining (EDM), we opted for five primary databases: Scopus, Web of Science (WoS), IEEE Xplore Digital Library, Google Scholar, and ScienceDirect. We selected these databases due to their broad technical and educational research scope, comprehensive collection of peer-reviewed literature, and suitability for bibliometric and citation analysis.

We devised a comprehensive Boolean search approach to gather pertinent studies and systematically enhance our inquiries. The leading search terms consisted of "Education" AND "Data Mining", ensuring consideration of research intersecting these fields. To encompass works using artificial intelligence methodologies, we broadened the search with ("Machine Learning" OR "Deep Learning" OR "Explainable AI"), allowing the detection of articles exploring wider AI-inspired strategies in education. Furthermore, we incorporated ("Blockchain" OR "Chatbots" OR "Multimodal Learning") to cover nascent technologies used in EDM. Including terms such as Blockchain and Chatbots was warranted due to their rising significance in student engagement, learning analytics, and automated educational assistance systems.

2.2 Inclusion and exclusion criteria

We formulated specific inclusion and exclusion criteria to uphold methodological rigour and comply with PRISMA guidelines. We considered studies published as peer-reviewed journal articles and high-quality conference papers between 2013 and 2023. Eligible papers had to be written in English and primarily focus on Educational Data Mining, Machine Learning (ML), Deep Learning (DL), Explainable AI (XAI), and Multimodal Learning. The choices were narrowed down even more by adding studies that added to EDM techniques, methods, and uses through empirical, theoretical, or experimental means.

Articles not written in English or those lacking methodological rigour were excluded. Additionally, sources without peer review, patent documents, editorials, opinion pieces, and theoretical studies lacking empirical grounding were not considered. Unlike the former version of this methodology, conference papers were not automatically excluded; their scientific merit and methodological standards were evaluated. Furthermore, studies not aligned with EDM methodologies, AI-driven educational analytics, or themes pertinent to this research were excluded. These modifications respond to reviewer concerns about previously excluding significant conference papers and adopting an overly simplistic exclusion strategy.

3 Data extraction and study selection process

A dual-stage screening approach was implemented to guarantee the inclusion of high-quality and pertinent studies. We excluded articles failing to meet the inclusion criteria during the first phase, which consisted of a review of titles and abstracts. In the second phase, a comprehensive full-text review was conducted to evaluate the research's quality, relevance, methodological integrity, and compatibility with study goals. Google Sheets enhances transparency and rigor by meticulously documenting each screening decisions. A systematic thematic analysis was conducted to identify research categories, and a quantitative synthesis was utilized when appropriate. The authors worked together to clarify questions and keep inter-rater reliability high, ensuring that the studies chosen were always the same.

Using the PRISMA framework, our study selection process encompassed four crucial phases. During the identification phase, we sourced 2195 records from databases such as Scopus, WoS, IEEE Xplore, Google Scholar, and ScienceDirect to ensure comprehensive coverage of high-impact research. The Screening phase involved eliminating duplicate records, resulting in 1805 studies for further examination. During the eligibility phase, 1390 full-text articles were looked at using strict methodology and PRISMA-compliant criteria for what to include and what to leave out. Following a thorough quality assessment, we selected 198 high-quality studies for the systematic review in the inclusion phase. This method ensures that the PRISMA framework and the search strategy are clear and easy to copy, and the study selection process is thorough and well-documented. Incorporating the reviewers' input has significantly improved our systematic review's clarity, transparency, and methodological soundness, making substantial contributions to Educational Data Mining.

4 Explainable AI

The term "explainable AI" (XAI) describes an AI system's capacity to offer understandable explanations about its operation to foster confidence and transparency for the system. In education, XAI has the potential to fundamentally change how students view, absorb, and comprehend the material being taught. It can also foster a healthy knowledge transfer, resulting in a highly optimized classroom. Personalized learning experiences, adaptive assessments, and individualized feedback facilitated by XAI contribute to improved student comprehension and retention. Additionally, XAI may help learning analytics teachers identify areas where their students may be deficient and require more focused instruction and support [37]. The following paragraphs present various works of XAI in education. The authors in [38] present an XAI system that tries to enhance the accuracy of predictions about student performance. The system applies LSTM and Random Forest algorithms to a dataset including 305 male and 175 female students from various countries, including India, Jordan, Palestine, Saudi Arabia, Egypt, and the USA. Their approach to predicting student performance relies on a distinctive set of "behavioural features", such as hand-raising frequency, learning management system resource usage, parental survey feedback, and school satisfaction. The results suggest that both models surpassed conventional ones, with the Random Forest model achieving an accuracy of 96.9%, and the LSTM—89% accuracy. To provide meaningful explanations of the predicted results, they utilized the LIME and SHAP approaches, which are local model agnostic methods. The analysis highlighted significant variations in feature importance across the models. Parental surveys and behavioural features were most impactful for the LSTM model, while student absences and behavioural features were more significant for the Random Forest model.

The authors in [39] present a course-agnostic XAI system to predict early student performance, especially at-risk students, which facilitates the implementation of interventions to correct student performance. They explored various ML algorithms including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), XGBoost (XGB), LightGBM (LGBM), Voting Classifier (VTC), and Stacking Classifier (STC). The dataset consisted of data generated from undergraduate and graduate courses delivered online over 16 weeks using the Blackboard LMS during the academic year 2020–2021. The results of the prediction accuracy of the proposed system demonstrated that Logistic Regression showed the best overall performance. To provide better insights into the decision-making processes of the system, the authors employed the model agnostic method SHAP to offer both local and global explanations. The SHAP method's results indicate that it is possible to explain the prediction in terms of individual students' grades and group grades. However, the effectiveness of the results is somewhat limited due to the constrained number of courses, which were all of the same sort.

The work [40] tries to find ways to positively influence student growth in secondary education by examining and assessing students' results. Towards this direction, they proposed a prediction model exploring the Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost, and Naive Bayes methods. Various surveys and school reports from two Portuguese schools, balanced by using K-Means SMOTE (Synthetic Minority Oversampling Technique) before classification, were used for this process, with the results indicating that the method of Support Vector Machine (SVM) outperformed the rest with an accuracy value of 96.89%. The authors used the LIME model as their method of interpretation of the predictions and gave explanations in terms of a student's grades given to them (first and second-period grades and final grades).

The authors in [41] propose an XAI system to guide students in planning their education to achieve their desired career goals. They explored various White and Black Box models with the help of an educational dataset comprising academic and employability attributes that are important for jobs. The results of the analysis indicate that the Naive Bayes method outperforms the methods of Logistic Regression, Decision Tree, SVM, KNN, and Ensemble with Recall and F- measure values of 91.2% and 90.7%, respectively. Concerning the interpretation of prediction results, the authors utilised model-specific explanations. The authors in [42] propose an explainable AI solution supporting the students' self-regulation and improving their course performance. This is achieved by combining learning analytics and machine learning to provide automated, intelligent feedback and action recommendations. In more detail, it predicts student performance at the assignment level, identifies the key factors influencing student outcomes, and produces informative feedback, offering targeted insights for improvement. To facilitate the presentation of this information, the author built a dashboard. The proposed solution explored the use of the Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF), Multi-Layer perceptron (MLP), and the BayesNet methods. The input came from a university-level programming course delivered consecutively for two years (2019 and 2020) to distance-learning students through a Learning Management System (LMS) at Stockholm

University, Sweden. The results of the analysis indicate that the Random Forest method provides satisfying results with a value of 91% for both Recall and F-measure, while the overall system improved students' learning outcomes, assisted them in self-regulation, and had a positive effect on their motivation. A limitation of the dashboard is its lack of feedback on the effectiveness of recommendations in improving student knowledge. The goal of the proposed solution was to provide poor-performing students' insights into what could potentially help to improve and to show teachers how they could improve the course structure and resources. With the help of the local model agnostic method LIME, the authors identified five factors, namely active participation in the LMS, engagement in forum discussions, prior programming experience, quiz scores, and assignment grades, as influential contributors to students' academic performance.

5 Chatbots in educational research

Conversational agents, often known as chatbots, have become cutting-edge tools in educational research, dramatically supporting this platform for data gathering, analysis, and student participation. AI and NLP-powered chatbots offer vital resources for gathering, evaluating, and using data to improve the educational process. Chatbots play a promising role in providing individualised educational experiences. They could respond to each student's unique requirements with better communication interaction, providing resources, feedback, and help specifically catered to them [43, 44]. This helps to create a more encouraging learning environment and gives teachers insight into the strengths and weaknesses of individual students. Chatbots enable instant feedback on student performance. They enable prompt interventions by having the ability to give tests, grade answers, and offer immediate evaluation [45]. This real-time feedback loop makes the educational system more responsive by quickly correcting problems and encouraging a deeper comprehension of the material [46]. Chatbots' interactive features also help boost student engagement, including gamification features, such as challenges, awards, and quizzes, to make learning more robust and entertaining [47, 48]. This keeps students interested and encourages a promising attitude toward learning. Chatbots are also an effective tool for gathering data because they may provide important details about students' preferences, behaviour, and learning styles [46]. Applying EDM techniques to this data supplies researchers and educators with a comprehensive view of the learning environment, enabling evidence-based improvements to curriculum and instructional design [49].

6 Blockchain in education

Adopting blockchain technology revolutionises data management, security, and verification in education. Blockchain offers a secure, decentralised way to verify academic qualifications [50]. Educational institutions may safely issue and confirm degrees, diplomas, and transcripts by keeping academic records on a blockchain. This makes the employment process easier for firms and reduces the possibility of false credentials [51]. The decentralised structure of blockchain guarantees the safe storage and transmission of instructional data. By reducing the risk of data breaches and unauthorised access, this enhanced security addresses concerns about student data confidentiality and privacy [52]. Academic records can't be tampered with because of blockchain's immutability. Cerberus [53] sets itself apart by integrating seamlessly with current verification systems, tackling actual fraud situations, and employing on-chain smart contracts to enable revocation without necessitating users to handle crypto-graphic credentials or digital identities. It provides an auditable and transparent trail of a student's academic career provided by the time-stamped and linked entries of each entry on the blockchain to earlier records.

Education records have more integrity and credibility as a result of this transparency. Financial transactions and administrative chores like student registration and enrollment can be streamlined with blockchain technology. The initiative in [54] is established as an inclusive learning environment that uses appropriate authoring tools and technologies such as microsites, blockchain, and universal accessibility standards. The adoption of blockchain-based electronic lab notebooks for research data management is examined in [55]. Here, according to the findings, social norms have an impact on both lowering perceived risk and raising the intention to use, whereas perceived utility and ease of use both enhance perceived usefulness and decrease perceived danger. Institutions can also automate procedures with smart contracts, which lessens the administrative load and lowers the risk of data management errors.

7 Different data mining techniques applied in EDM

EDM is a specialised area of research that aims to extract important insights and knowledge from data within the education sector. EDM utilises Data Mining (DM) methodologies, including clustering and regression, to analyse data and generate forecasts that aid in addressing inquiries in the field of education. Various machine-learning methods have been utilised in this sector throughout the years. However, only recently has Deep Learning garnered significant interest in education. Deep Learning is a specific area within machine learning that uses neural networks to address complicated challenges in natural language processing (NLP) and computer vision. Deep learning is exceptionally proficient in analysing and manipulating textual and digital data, including photos, audio, and video. Deep learning can provide automated grading, attendance tracking, monitoring systems, or intelligent learning systems in education.

Nevertheless, the implementation of DL in education remains unresolved due to stringent education regulations and policies and the numerous problems associated with data collection and use. Moreover, education is susceptible to many ethical and legal disputes. Data mining and statistical tools to identify patterns in extensive educational data sets that would otherwise be unanalysable [56]. EDM utilizes e-learning platforms such as Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), and, more recently, Massive Open Online Courses (MOOC) to gather comprehensive and diverse information about students' learning activities in educational environments. These systems track several aspects of student activity, such as the frequency of accessing learning materials, the accuracy of exercise responses, and the duration spent reading or watching educational content. The data can be analysed to tackle many educational concerns, including producing suggestions, creating adaptive systems, and automating the grading of student tasks. While machine learning has long been used to analyse educational data, Deep Learning applications in EDM have emerged only recently. Data mining techniques have been widely studied and applied in educational settings over the past few decades. This research topic has gained significant appeal recently, largely due to the increased availability of online datasets and learning systems.

Several surveys on EDM have been published to date. As documented in the literature, the initial EDM survey was created by Romero and Ventura in 2007 [8]. Subsequent enhancements were made in 2010 [57, 58]. The authors examined over 300 studies conducted before 2010. They identified eleven distinct categories or tasks within the field of EDM: data analysis and visualisation, feedback provision for instructors, student recommendations, prediction of student performance, student modelling, detection of undesirable student behaviours, student grouping, social network analysis, development of concept maps, creation of course- wares, and planning and scheduling. The survey outlined the methodologies and strategies utilised in the field of EDM for each of these categories. Baker and Yacef [59] introduced a survey on EDM in 2009. This study examined the patterns and changes in research carried out by this community by comparing its present condition with the initial years of EDM. In this instance, the authors have highlighted four applications/tasks within this field: enhancing student models, enhancing domain models, examining the pedagogical support offered by learning software, and doing scientific research on learning and learners. A compilation of the most frequently referenced publications in EDM from 1995 to 2005 was provided, highlighting their impact on the EDM community. In 2014, PenaAyala conducted a comprehensive survey using data mining techniques on over 240 papers in EDM [60]. The statistical and clustering methods were executed successfully and identified a specific set of educational capabilities. Additionally, they revealed a pattern of EDM techniques and two distinct patterns of value instances that represent EDM approaches using descriptive and predictive models. This study primarily emphasised computational approaches rather than applications of EDM, in contrast to earlier literature surveys. Recently, four additional studies have been included in this compilation of surveys. Bakhshinategh et al. [61] conducted the initial research, which examined various EDM activities and applications, classifying them by objective. Romero and Ventura developed a hierarchy of thirteen types organised into five basic tasks [58]. These activities include Student Modelling, Decision Support Systems, Adaptive Systems, Evaluation, and Scientific Inquiry.

In 2019, Aldowah et al. [30] focused exclusively on studies conducted within higher education. The analysis was grounded on four dimensions: computer-supported learning, predictive, and behavioural analytics, and computer-supported visualization. According to earlier studies, the authors discovered that particular EDM techniques could provide the most effective solution for various learning difficulties. These techniques offer student-centred strategies and resources for educational institutions. The most recent review of EDM was conducted by Trung et al. [62], the authors presented a subjective perspective on the progress and utilization of EDM in combination with Deep Learning. It included several aspects, such as the individuals involved, the collected data, the applications, the technologies employed, and the challenges faced throughout the implementation phase.

The next section will thoroughly analyse relevant literature, focusing on research that has utilized different approaches in machine learning. This examination will be divided into two independent subsections, each focusing on a different aspect of research methods. The first subsection will examine academic publications that have effectively utilized traditional machine learning approaches, highlighting the effectiveness of well-established algorithms and methodologies in answering research questions. The second subsection will focus on a discussion of articles that have utilized deep learning approaches. This subsequent section will explore the complexities of neural networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other sophisticated deep learning structures. These two independent subsections provide a systematic examination to clarify the machine learning techniques and methodologies researchers employ. Table 1 contains a summary of different data mining techniques applied in EDM.

7.1 Classical machine learning approaches

The article [71] presents an online ensemble classifier for predicting student performance in distance education courses. The method uses educational data mining to identify students at risk of poor performance and trains supervised machine learning algorithms on student marks. The ensemble, which combines Naive Bayes, k-NN, and Winnow algorithms, outperforms individual algorithms and batch ensemble methods like AdaBoost, Random Forest, and Rotation Forest. The methodology can identify struggling students early, allowing tutors to provide targeted support. It also enables continuous model updating each term as new student data arrives, eliminating the need for periodic retraining. The high computational efficiency makes it suitable for large-scale implementation.

The paper [72] explores using data mining techniques to analyse historical student grades and activity data in e-learning systems. The authors propose building incremental logistic regression models for predicting student dropout in several courses, using grades on weekly activities as predictor variables. These models perform well, with classification accuracy reaching 100% by week 12. Predictive models identify potential dropouts early, allowing targeted tutoring interventions. When applied in a later academic year, this system helped reduce the dropout rate across courses by 14% compared to previous years.

The methodology is platform-agnostic and can be applied to any traditional online course with grade data. It is easy to automate for regular predictive monitoring once models are built. The approach helps reduce dropout rates when applied systematically across courses. Ethical use of predictive data is essential to avoid students' at-risk excessive monitoring. However, the methodology has weaknesses, such as no predictions in the early weeks until initial grades are available, less interpretability than other data mining models, and the need for custom predictive models for each course. Additionally, the method may be limited due to the distinct nature of MOOCs and may require technical expertise to implement data mining modelling and process automation.

The paper [73] presents a model for predicting student performance in large university courses using low-cost variables, focusing on large groups with over 50 students. The model uses data from the university's student information system and learning management system, including student demographic data, academic records, and LMS usage metrics. The model analyses data from 119,366 course enrolments across 811 courses over three semesters. It incorporates three scenarios: using only fixed student data, adding full-semester LMS data, and using LMS data up to midterm. The model compares three algorithms—linear regression, robust linear regression, and random forest—and finds that random forest has the best performance overall. The model aims to predict student grades early, allowing instructors to intervene and support at-risk students to prevent failure. The main techniques used in the paper include data collection and pre-processing, modelling algorithms, prediction scenarios, and evaluation metrics.

This paper [74] demonstrated how a genetic algorithm for feature selection (GAFS) in EDM can improve prediction accuracy for student academic performance. The study compares classifier performance on a student dataset with and without GAFS for feature selection, finding that GAFS improves accuracy for all classifiers. Random forest with GAFS performs best, increasing accuracy from 79.79% to 82.29%. The study highlights the importance of GAFS in selecting relevant features and overcoming issues with local search in other feature selection methods. It suggests the GAFS approach could be applied to other EDM tasks with high-dimensional data to improve model accuracy. The paper [75] uses four machine learning algorithms—SVM, naive Bayes, decision tree C5.0, and random forest—as baseline models. K-fold cross-validation and Principal Component Analysis (PCA) are used to improve these models. The hybrid models, tested on three datasets, showed a notable improvement in classification accuracy. The hybrid model using Random Forest achieved 99.72% accuracy, significantly higher than the baseline RF model. Cross-validation and PCA also improved model robustness and generalizability. The paper concludes that combining these algorithms with

Table 1 Summary of different data mining techniques applied in EDM

Ref	Key points	Data Type	Data	Features	Techniques	objective	Advantages	Shortcomings
[63]	Even–Odd Crossover, inspired by genetic algorithm crossover operators	Numerical data	4 real-world imbalanced student performance datasets	Dataset 4: 25 variables and 600 samples	RF, SVM, and KNN	Effective resampling technique for balancing data	A new crossover operator to handle imbalanced data	Only considers binary classification
[64]	SVM classifier with an MLP kernel	3 real-world educational datasets related to student performance	Portuguese course dataset	3 student performance datasets	SVM	Understanding students' performance and identifying factors that affect it	Optimizes SVM parameters through SA	SA optimization process is computationally expensive
[65]	GB, XGBoost, and LightGBM as base classifiers	Demographic, social/emotional, and school-related attributes	Portugal Alentejo region	Categorical and numerical features	HELA ensemble approach	Performance links to cognitive and learning	Integrating boosting and stacking techniques	Class imbalance in a multi-class setting is not addressed
[66]	Analyse slow learners in semester exams	Sample qualitative data set	97 data in the data set	Categorical and numerical features	C4.5, AODE, Naïve Bayes, and Multi-Label KNN	Evaluate different classification techniques	Compares classification like C4.5, Naïve Bayes, AODE	No variable importance analysis to identify the most predictive attributes
[67]	Machine learning algorithms	Categorical and numerical features	Multivariate educational dataset	Scholastic Features, Demographic Features, Emotional Features	Logistic regression, random forest, SVM, KNN, XGBoost, and decision tree	Identify factors influencing and predict performance	Utilizes a real-world educational dataset, and conducts exploratory analysis	Limited data from a single educational institute, lacks cross-validation
[68]	Data mining algorithms on educational data	Real-world, multivariate, structured, static educational data	Student dataset from a university in Indonesia	Demographic details: Gender, caste, parental occupation, Academic details	Decision Tree, Neural Networks, and Naive Bayes	Evaluating and comparing the algorithms' performance	Compares multiple algorithms to identify the optimal ones	High training scores do not highlight overfitting risks
[69]	Techniques like k-means clustering, association rule mining with Apriori, and decision trees	The data contains time- series records of student academic performance over multiple semesters	91 student records with 17 attributes containing academic performance	Academic and demographic features	k-means clustering, association rule mining with Apriori, and decision tree classification with ID3	Data mining togethercan maximize the ability toanalyse academic performance	Real student academic data from a university information system	Lacks rigorous validation, and arbitrarily sets association rule thresholds
[70]	ANN to predict academic performance	Educational, multivariate, temporal, real-world, categorical, and secondary data	Educational data includes academic records of engineering students	Features like Grade Points (GP) and Cumulative Grade Point Average (CGPA)	ANN regression models for predictive modelling of student performance	Predict engineering students' academic performance using early semester scores	Presents a methodology that uses actual grade data to predict performance	It is limited to a single institution and electrical engineering students

cross-validation and PCA results in highly accurate models for predicting student performance. However, the methodology has weaknesses, such as focusing only on predictive accuracy and not model interpretability. The hybrid approach can help identify students at risk of poor performance early, guide educational policies, and be applied to other learning outcomes. Further validation with larger datasets and more algorithms is needed.

The COVID-19 pandemic accelerated the need for improved online learning systems, driving the use of data mining to enhance student performance and retention. The paper [64] proposes using a support vector machine (SVM) classifier with a multilayer perceptron (MLP) kernel to predict student academic performance, optimized using a Simulated Annealing (SA) algorithm. SA is an optimization method that can escape local optima by accepting worse solutions probabilistically. Experiments were conducted on three real student performance datasets, and the proposed SA-SVM approach improved accuracy, sensitivity, precision, and F-measure compared to SVM with default parameters. The results were statistically significant compared to naive Bayes and decision trees used in previous studies. The findings support an increased focus on leveraging AI techniques to enhance understanding of student success factors and demonstrate the potential of using SVM and evolutionary algorithms like SA for optimizing educational data mining. The SA-SVM approach shows promise but needs more extensive testing and refinements to address limitations around model generalization, class imbalance, and practical application.

The research [67] paper uses educational data mining techniques to predict student academic performance. It uses 480 instances and 17 attributes from an educational institute and employs machine learning algorithms like logistic regression, random forest, SVM, KNN, XGBoost, and decision trees. The data is pre-processed, and exploratory analysis is performed to understand feature relationships. Feature engineering transforms the data into suitable inputs for the models. The results show that machine learning can predict student performance based on key attributes. Future work could consider additional features like financial, health, and food habits. The authors in [69] implemented an EDM system using student academic records from a university in Indonesia, using techniques like k-means clustering, association rule mining with Apriori, decision tree classification with ID3, and outlier detection with DBSCAN. The results showed that clustering identified two main groups of students based on academic performance, while association rule mining found strong associations between passing certain courses. The decision tree model predicted on-time graduation with high accuracy, and outlier detection identified anomalies in the data. The paper demonstrates that EDM can be an effective tool for knowledge discovery and data-driven decision-making in higher education institutions using available academic data. However, its weaknesses include a limited dataset, lack of rigorous validation methodology, arbitrary association rule thresholds, decision tree overfitting, and insufficient domain knowledge of educators. Key findings from the study include identifying two main groups of students based on academic performance, strong associations between passing certain courses and predicting on-time graduation with 93.41% accuracy using early semester grade attributes. The EDM methodology demonstrates how academic databases can be mined to extract useful knowledge to inform academic management and policies.

The HELA [65] is a hybrid ensemble learning algorithm designed to predict students' academic performance using gradient-boosting algorithms. It uses a stacking ensemble to integrate diverse base models, achieving higher accuracy than individual classifiers. The HELA is tested on student math and Portuguese course datasets and shows promising results. The methodology uses real student grade data, implements multiple classification algorithms, and provides binary and multi-class classification results. It achieves high accuracy results, with 96.6% for binary and 78% for 5-level grading. However, the small dataset of 395 students limits generalizability, and there is no train-test split or cross-validation for base classifiers. Class imbalance in multi-class settings is not addressed, and the predictive patterns learned are limited. The model also lacks feature selection or engineering to improve data quality and model generalization.

The paper [68] discusses the use of data mining in education, specifically the K- Means algorithm for performance analysis. The authors evaluate various clustering, classification, and association rule mining algorithms on a student dataset to identify the best methods. They use a dataset of 600 student records from a private university in Indonesia, which is cleaned and pre-processed before analysis. The Apriori algorithm in the Orange tool is used for association rule mining, which filters rules based on minimum support, confidence, and maximum antecedent/consequent values. The paper compares tools like Weka, Orange, and R-Studio to determine the most suitable tools for analysis tasks. The results from these techniques can help identify at-risk students and improve pedagogical planning to enhance student quality and reduce failure risks. The methodology's strengths include various clustering, classification, and association rule mining algorithms comprehensive evaluation, using real student datasets, comparing tools like Weka, Orange, and R-Studio, and proposing a potential framework to analyse student performance and generate predictive solutions. However, the weaknesses include the small dataset size, limited attributes considered, potential generalization of performance, limited detail on data pre-processing steps, negative silhouette values for K-means clustering, no validation of predicted results on new unseen student data, and no qualitative analysis of actual impact or usefulness for educational institutions.

7.2 Deep learning approaches

The authors in [76] presented a novel approach to personalize e-learning experiences by analysing user activity sequences. The authors propose a two-stage approach, modelling user activity sequences and then clustering these models to identify meaningful patterns and styles. They identify three levels of clustering: detecting predefined, well-established problem-solving styles, discovering new problem-solving styles along known learning dimensions, and discovering potentially interesting learning dimensions and associated problem-solving styles. The authors present a custom model that captures the activity sequences of learners involved with problem-solving in an Intelligent Tutoring System (ITS). The paper's findings have significant implications for e-learning and adaptive education, providing insights into learners' problem-solving styles, improving learning outcomes, and reducing the workload of human experts. However, the methodology's strengths include personalized learning, real-time feedback, adaptive assessments, continuous progress monitoring, and data-driven decision-making.

NNSPPM [70] is an ANN model developed to predict the academic performance of electrical engineering students at a Malaysian university. The model uses students' grades in fundamental courses from the first and third semesters as input variables, and the output variable is their cumulative GPA in semester 8. Tested on data from 391 matriculation students and 505 diploma students, the model achieved correlation coefficients of 0.9774 and 0.9245, respectively. The model's results indicate that performance in fundamental early courses strongly influences overall performance at graduation, allowing targeted support. Although limited to electrical engineering students, the model shows promising accuracy and could be expanded to other disciplines.

The research in [77] paper focuses on predicting student dropout in MOOCs using temporal models like recurrent neural networks (RNN) and input–output hidden Markov models (IOHMM). It proposes an RNN model with long short-term memory (LSTM) cells, which outperforms other models significantly. The LSTM model achieves an AUC score of 0.9 on test data for predicting dropout over weeks, beating baseline models like SVM and logistic regression. The study highlights the importance of LSTM in capturing temporal dynamics better and suggests that it can be applied to other student modelling tasks in education. However, the methodology lacks details on MOOC data, no ablation study, limited model interpretation, and no ethical issues. The study also does not consider contextual factors like demographics and location. The analysis would benefit from additional implementation details and a thorough discussion of ethics and limitations.

The authors in [78] proposed a systematic approach for developing a question-answering system for closed domains like education acts documents, using NLP techniques like POS tagging, keyword extraction, and similarity scoring. It uses indexing and keyword matching for efficient document retrieval and extracts answers by matching keywords and POS tags. However, it lacks large-scale evaluation, discusses handling complex questions, and lacks discussion on interpretation and explanation of system decisions. The methodology can be used for domain-specific QA systems and open-domain scenarios, but further analysis and evaluations are needed. Jayakodi et al. [79] presented a novel approach to classify exam questions based on Bloom's taxonomy using WordNet and Cosine similarity. The algorithm achieved an accuracy of 71% in classifying exam questions into the correct category of Bloom's taxonomy, including higher-level categories like Analysing and Evaluating. The study highlights the potential of NLP techniques, such as WordNet and Cosine similarity, to improve educational assessments and support student learning. The methodology used in the study has strengths, such as its novelty, improved accuracy, automation, adaptability, and potential for adaptability. However, it has weaknesses, such as limited domain knowledge, complexity of WordNet, limited interpretability, and potential for bias.

The research paper [80] has developed a method to improve student success prediction accuracy by training a Recurrent Neural Network (RNN) across all students at a university, regardless of their course. This approach addresses the lack of sufficient training data for individual courses and the inability to build predictive models for new courses without historical data. The RNN model outperformed baseline methods like random forests, explaining 13% of the variance in exam performance across all disciplines. The study demonstrates the feasibility of training deep learning models across all students rather than building individual course models. However, it also has weaknesses, such as limited interpretability, large volumes of historical log data, and the need for continuous monitoring and tuning over time.

The research paper [81] proposes using a Stacked Denoising Autoencoder (SDA) to build a deep neural network model for predicting student dropout. The model uses real student data from past years to train the neural network and predict dropout likelihood for current students. The model achieves 92.4% accuracy on test data, outperforming other machine learning methods like SVM and Random Forests. Pre-training with SDA improves model accuracy compared to no pre-training. The model can help schools provide timely counselling and intervention to students at high risk of dropping out, reducing resource wastage, and improving student retention. It can also help identify at-risk students early and prevent issues like disconnection from society or lack of future planning. The paper highlights the potential of SDA as an effective deep learning technique for student dropout prediction, offering educational, social, and economic benefits.

The ConRec [82] deep learning model is proposed to extract features from MOOC activity data for dropout prediction and reduce manual feature engineering. The model combines CNN and RNN layers and achieves performance comparable to traditional classification methods. It is more extensible and can be applied to other EDM problems involving sequential data, such as course completion prediction. Key strengths of the model include its ability to capture multi-scale temporal patterns, its independence from domain expertise in feature design, and its trainability via standard backpropagation. Its weaknesses include limited experimental validation, interpretability, and reliance on heuristics. Further rigorous testing is required to establish the generalizability of the approach. More research is needed to validate its impact on practical educational systems.

The research paper [83] presents a novel approach to multimodal emotion recognition using 3-dimensional convolutional neural networks (C3Ds) cascaded with multimodal deep-belief networks (DBNs). The method improves emotion detection by combining spatiotemporal information from audio and video streams and deep learning techniques for emotion recognition by discriminative features of visual appearance and audio. The findings suggest that incorporating additional sources of information like audio and gestures can improve the reliability of the recognition process. The methodology also uses deep learning techniques, such as C3Ds and DBNs, which have shown great performance in computer vision and NLP applications. However, the methodology has limitations, including a limited dataset, lack of comparison with other methods, and computational complexity. Further research and validation on larger datasets are needed to fully assess its strengths and weaknesses.

The paper [84] discussed using deep features in plagiarism detection in programming assignments. It argues that current methods based on n-gram techniques fail to capture long-term dependencies and non-contiguous interactions in source code. The proposed method uses deep features from a character-level Recurrent Neural Network (char-RNN) pre-trained on Linux Kernel source code, which captures non-contiguous interactions within n-grams. The experiments show that the deep features classify effectively assignment program submissions as copy, partial-copy, and non-copy, resulting in significant improvements in the f1-score for binary and three-way classification tasks.

The paper emphasized the importance of capturing non-consecutive interactions and obfuscated code, which are often overlooked by existing plagiarism detection tools. However, the paper lacks a detailed analysis of the limitations of using deep features and the char-RNN approach for plagiarism detection in programming assignments. The generalizability of the proposed method beyond programming assignments and the Linux Kernel source code is not explicitly addressed. The paper also lacks comparison with other state-of-the-art plagiarism detection methods and does not discuss the ethical implications of using deep features and plagiarism detection in the educational context.

The article [85] discusses the emergence and current state of deep learning techniques in EDM, focusing on four main tasks: predicting student performance, detecting undesirable behaviours, generating recommendations, and evaluating/assessing. It also discusses using popular datasets like ASSISTments, MOOC datasets, and AI tutoring system datasets. The article introduces DL, covering neural network basics, training, architectures like CNN and LSTM, hyper parameters, and frameworks like Keras. It notes that 67% of reviewed papers found DL outperformed baseline methods on EDM tasks, but some argue that DL models lack interpretability compared to traditional approaches. Future opportunities include applying DL advances like deep reinforcement learning and GANs to unexplored EDM tasks and improving model transparency.

Unequal representation of academic levels in educational datasets, known as imbalanced datasets, poses issues for classifiers. Handling imbalance datasets in EDM is illustrated as follows.

7.3 Handling imbalance datasets in educational data mining

The presence of class imbalance in educational statistics may impede the accuracy of predictive models, as these models are often built under the assumption of balanced class distribution. Upon examining the research on imbalance learning, the existing techniques can be broadly categorized into four groups: algorithm-based methods, data resampling-based methods, cost-sensitive methods, and classifier ensemble-based methods [86, 87]. Various sampling approaches can be employed to address the issue of class imbalance, which refers to situations where classification is either mildly or

Table 2 Summary of Different EDM Models with Imbalance Datasets Handling

Author	Method
Wongvorachan et al. [89]	Missing data imputation, outlier removal, and variable selection are used for data preprocessing. Random oversampling (ROS), Random undersampling (RUS), hybrid of SMOTE-NC and RUS used for resampling. Random Forest tuned using randomized grid search is the classification method
Shams et al. [63]	Random oversampling (ROS), Random undersampling (RUS), a hybrid of SMOTE-NC and RUS, and existing techniques like SMOTE, SLSMOTE, Cluster-SMOTE, and Bor-SMOTE are used for balancing the distribution
Al et al. [101]	Data-level approaches Random oversampling (ROS) and random undersampling (RUS). Hybrid approach: Synthetic minority over-sampling technique for nominal and continuous (SMOTE-NC) and RUS
Hassan et al. [100]	Random Over-sampling (ROS), SMOTEENN (SMOTE with Edited Nearest Neighbor), and Borderline-SMOTE (Synthetic Minority Over-sampling Technique). The classifiers used in this study include AdaBoost, Random Forest, Gradient Boosting, Extra Trees, and Bagging
Pristyanto et al. [99]	SMOTE and OSS to handle the imbalanced class distribution in the multiclass dataset. SVM is the classifier
Thai et al. [98]	Manipulating classifiers internally, one-class learning, and ensemble methods to address class imbalance. Random over-sampling, SMOTE, and cost-sensitive learning (CSL) modify the class distribution

excessively uneven. Three resampling techniques commonly used are random oversampling (ROS) [88], random undersampling (RUS) [89], and the hybrid resampling technique. Multi-class datasets can utilize 10 resampling algorithms, including SMOTE [90, 91], Distance SMOTE [90], BorderLineSMOTE [92], KmeansSMOTE, SVMSMOTE [93], LN SMOTE [94], MWSMOTE [95], Safe Level SMOTE [96], and SMOTETomek [97]. Table 2 summarises different EDM models which applied different imbalanced dataset handling techniques.

Thai-Nghe et al. [98] discussed the issue of class imbalance in predicting student academic performance, which degrades classifier performance. To address this issue, the paper investigated techniques such as oversampling with SMOTE and cost-sensitive learning. Experiments on university datasets from Vietnam and Thailand show that SMOTE consistently improves metrics like AUC, F-measure, and misclassification cost over baseline classifiers. Cost-sensitive learning reduces total misclassification costs by weighting false negatives more heavily than positives. Combining oversampling via SMOTE and cost-sensitive learning works better than either technique alone. The proposed methods can improve the prediction of minority class outcomes on real imbalanced educational data for student performance monitoring. The paper's strengths include using real-world university datasets, comparing multiple classifiers, using proper evaluation metrics, hyperparameter tuning, and cross-validation. However, the paper also lacks feature engineering, ablation studies, and statistical tests to verify the improvements.

Imbalanced datasets, characterized by the uneven distribution of cases across different classes, provide difficulties in constructing precise predictive models [99]. The study evaluated the performance of six classifiers (kNN, SVM, MLP, decision trees, Naive Bayes, and logistic regression) on an unbalanced dataset of pharmacy student grades. The dataset is manipulated using data-level sampling techniques such as SMOTE and random undersampling. Most classifiers demonstrated poor performance when applied to the imbalanced dataset. Ensemble approaches such as bagging, random forests, stacking, and XGBoost are highly efficient in dealing with class imbalance. Data-level sampling has its benefits, but ensemble approaches offer greater resilience. Decision trees have remarkable performance on imbalanced datasets as they effectively handle outliers. No statistically significant disparity is observed between data-level and algorithm-level methods. The main conclusion is that ensemble strategies regularly enhance the performance of classifiers on imbalanced EDM data and should be the preferred solution. However, the selection of the appropriate handling approach still relies on the specific scenario at hand.

The paper [100] explored the class imbalance problem in EDM, focusing on two datasets from a Malaysian university. The study evaluates ensemble learning methods and sampling techniques for handling multiclass imbalances on these datasets. The results show that combining Student Information System (SIS) and E-Learning (EL) data improves model performance compared to using them separately. Balancing the class distribution with sampling techniques further enhances the performance of ensemble models on imbalanced data. The random oversampling technique with the AdaBoost ensemble achieves the best overall performance, while the SMOTEENN hybrid over/undersampling technique consistently performs well across models. The study highlights the utility of sampling and ensembles for improving multiclass classification of imbalanced student performance data, with implications for how such data should be handled in educational data mining applications. The findings suggest that incorporating academic and behavioural data provides a more comprehensive view of students for prediction, and sampling methods should be used to handle an imbalance in student performance data before modelling.

Shams et al. [63] introduced the Even–Odd Crossover, a new resampling method inspired by genetic algorithm crossover operators, which exchanges attribute values between samples at even/odd indices to create new minority class instances. The method outperforms other oversampling methods on most datasets in recall, F-score, and G-mean, showing it improves handling of the minority class.

The paper demonstrated Even–Odd Crossover as an effective resampling technique for balancing data and enhancing imbalanced classification performance, especially for student success prediction tasks. It uses four real-world imbalanced datasets, compares multiple classification algorithms, considers multiple evaluation metrics beyond just accuracy, and validates datasets with different imbalance ratios, sizes, and features. However, the paper has some limitations, such as only considering binary classification of student performance, no cross-validation, and does not tune classifier hyperparameters for each dataset. Additionally, it does not consider feature selection or weighting of attributes, assumes balancing classes improve performance, but trade-offs exist, and requires additional compute and data transformations for resampling techniques.

Authors in [101] investigated techniques for handling imbalanced datasets in EDM systems. Imbalanced datasets, where some classes have more examples than others, pose challenges for building accurate predictive models. The study tests six classifiers (kNN, SVM, MLP, decision trees, Naive Bayes, and logistic regression) on an imbalanced dataset of pharmacy student grades before and after applying data-level sampling techniques like SMOTE and random undersampling. The analysis reveals that most classifiers exhibit suboptimal performance on imbalanced datasets. However, ensemble methods, specifically bagging, random forests, stacking, and XGBoost, are exceptionally adept at mitigating the effects of class imbalance. Data-level sampling helps in some cases but is less robust than ensemble methods. Decision trees perform well on the imbalanced dataset due to the handling of outliers. No statistically significant difference between data-level and algorithm-level techniques is found.

Wongvorachan et al. [89] compared different resampling techniques for handling class imbalance in EDM. The research finds that ROS performs best for moderately imbalanced data, while RUS underperforms. The hybrid approach and baseline are in between for moderate imbalance and ROS overfits in extreme imbalance cases.

The baseline model performs poorly. The findings suggest that ROS works well for moderately imbalanced data, while the SMOTE-NC + RUS hybrid approach is best for extremely imbalanced data. The methodology uses a real educational dataset, selects relevant predictors, and compares techniques across moderate and extreme imbalance levels. However, it only evaluates techniques on a single dataset and does not explore ensembles of resampling techniques. Proper class imbalance handling can improve predictive accuracy and promote fairness in EDM models. Ensemble techniques consistently improve the classifier's performance on imbalanced EDM data and should be a go-to solution, but picking the right handling approach still depends on the problem.

8 Data fusion in multimodal learning analytics and educational data mining

This section discusses the different types of educational data that can be used for fusion and analysis.

8.1 Educational learning systems

As mentioned in the previous sections, educational learning styles have evolved in the last decade, especially after the pandemic [102]. These learning systems are majorly classified as (i) blended-educational learning (BEL), (ii) hybrid educational learning (HEL), and (iii) smart educational learning (SEL). All these learning models follow different approaches to delivering instructions. Though the objectives of these models are the same, they are distinguishable by their characteristics [102].

In BEL, online learning components are integrated into conventional in-person instruction [103]. Here, learners interact face-to-face with the instruction while accessing the online components, like video lectures, discussion forms, quizzes, etc. In HEL, the instructor will teach the persons in the classroom and online participants at the same time using video conferencing methods [104]. SLE is the usage of digital technologies in educational learning. It sometimes involves smart gadgets, educational apps, or other digital technologies to improve the learning experience [105]. Analyzing the enormous data produced by these learning activities can significantly enhance educational practices [102].

8.2 Educational data mining (EDM)

It is a valuable tool that uses data mining and artificial intelligence techniques on the data collected from different educational environments to achieve the abovementioned objective. However, most EDM methods analyse educational data based on a single data source [106]. However, this limited data analysis may reflect something other than the overall educational process. A few limitations are highlighted below [107].

8.2.1 Limitations of using a single data source for educational data mining

Limited Knowledge A single data source will provide a partial perspective on student learning involvement. Adding various sources like questionnaires, assessments, and behavioural data can contribute to more knowledge enhancement.

Bias A single data source may produce skewed results and inaccurate conclusions.

Limited Generalizability The conclusions drawn from a single source cannot be adapted to different contexts and populations.

Therefore, by considering the above facts, combining several educational data sources will lead to a successful educational learning model, which is often called multimodal learning analytics (MLA) [102].

MLA The data for MLA depends on the source of it. The aforementioned different learning systems produce different data types. Broadly, this data type can be categorized into the following: Time series [108], multimedia (audio, video, and images) [109, 110], text [111], and numerical [112]. Further, various categories of data sources have been used under the abovementioned classification in the literature—examples of physical, physiological, and digital information [102]. Table 3 presents the data sources used for physical, physiological, and digital information [102].

Researchers have used the combination of the data mentioned in Table 3 for various objectives [102] like, analysis and prediction of student's learning processes, recognizing the student's, teacher behaviour modelling, forecasting of student's engaging in the classroom etc. Over the last decade, several data fusion methods have been developed by exploiting multiple data sources discussed above for various student and teacher-centric developmental objectives. The recent works on data fusion in multimodal learning analytics and educational data mining are presented in Table 4.

Table 3 Summary of data sources used for physical, physiological, and digital acquisition

Category	Data source	Data type
Physical	Heart rate Blood volume Student's head position Teacher movement in the class Hand movements Student's posture Eye movement Motivation from questionnaire	Time Series
Physiological	Electrocardiogram Electroencephalogram Skin Temperature Electro dermal activity	Time Series
Physical	Real-time video recordings Student attention using video Student audio Student Images Teacher eye tracking Teacher's presentation (audio and video) Interaction between students Teacher gestures Facial expressions of teachers and students	Multimedia
Physiological	Student eye tracking	Multimedia
Text	Cognitive attention of students using video of head movements Message exchange during activities Dialogues between students Course information Browser history Interviews with students	Multimedia
Digital	Online class session timings Evaluation records Multiple choice questions Student scores Exam results (classroom, online)	Numerical

Table 4 Summary of data fusion in multimodal learning analytics and educational data mining

Ref	Data type	Data source	Data	Learning environment	Features	Classifier	Objective	Advantages	Shortcomings
[108]	Time series, and Video	Physical and Physio-logical	EEG, electrodermal, heart rate, body temperature, blood volume, and gestures	Face-to-face	Correlation coefficient and Fourier coefficients of EEG, statistical movements of eye tracing data	Random Forest	Learning behaviour system	Eye features related to variations in learning,	European university's study may lead to biased results
[109]	Time series, Audio and Video	Physical, and multi-media	Teacher's eye-tracking, and movements in the classroom	In-Person classroom	10 features from eye-tracking, using Fourier analysis 140 features are extracted from the acetometer data, and over 7000 features from audio and video	Various machine learning models	Analysis of teaching and learning processes	Reported good accuracies that multimodal data fusion can analyse the teacher learning abilities	Adding more teacher activities can significantly improve the model's performance
[113]	Video and nu merical	Physical, and digital	Theory classes data from lectures, practicals, and final exam scores	BEL	Pre-processed the data using various normalization techniques, applied various fusion techniques	Six types of machine learning algorithms	Predict the student's status on the subject	Better result with the combination of ensemble attribute selection with REPTree classification algorithm	It may lead to biased results. The experiments are conducted in controlled environments
[114]	Video and time series	Physical	Teacher's gestures, Teacher behaviours including the way of teaching	Online classes	RGB information for human actions and skeleton information are formed from the videos	Support Vector Machine (SVM)	Construction of Teacher behaviour dataset	Useful tool to analyse the teacher's behaviour and self-expression in the classroom to adopt it	Accuracies are limited, and need various deep-learning methods
[115]	Time series, audio and videos	Physical and digital	Heartrate of the data, the teacher-student conversa-tion, and facial expressions	Online class environ-ment	Eye and mouth-related features from the face, the statistical feature from the heart rate, and the acoustic feature from the audio conversation	SVM and random forest, and MLP	Recognition of student's mental state	Good results in classifying the students' mental states into concentration	The experiment population is only four

9 Discussion

In this section, we delve into the nuanced implications of the presented research, exploring the multifaceted aspects of educational data mining and providing insightful analysis to elucidate the broader significance of the findings.

XAI in education revolutionizes learning by providing personalized experiences, adaptive assessments, and tailored feedback. It enhances learning outcomes, promotes self-regulation, and empowers educators, fostering an optimized and transparent educational environment. Studies utilizing machine learning algorithms and interpretability methods showcase XAI's effectiveness in predicting student performance, enabling early intervention, and providing insights for informed educational decisions. These advancements offer more transparent, effective, and student-centric educational practices. XAI provides understandable explanations of AI-based systems, altering student engagement with learning materials, and facilitating customized learning experiences, adaptive tests, and individualized feedback, thus enhancing student performance. Most proposed systems employ machine learning algorithms such as Logistic Regression, Decision Trees, Multilayer Perception, Random Forest, and Support Vector Machine, with a preference for model-agnostic methods like LIME and SHAP, offering valuable insights into decision-making processes.

Chatbots in education serve as advanced tools for data gathering, analysis, and student engagement. They offer personalized learning experiences by addressing individual student needs and providing tailored communication, feedback, and resources. The real-time feedback loop enhances comprehension, while gamification features boost student engagement. Chatbots also serve as effective data-gathering tools, offering insights into student preferences, behaviours, and learning styles. Applying EDM approaches to this data enables evidence-based instructional design and curriculum improvement decisions.

Blockchain in education revolutionizes data management, security, and verification by offering a decentralized, tamper-proof method for confirming academic qualifications. It enables secure issuance and verification of degrees, diplomas, and transcripts, reducing the risk of false credentials. The technology ensures the safe storage and transmission of instructional data, enhancing security and privacy. Immutability prevents tampering with academic records, and solutions like Cerberus provide auditable and transparent trails of students' journeys. This transparency enhances the integrity and credibility of education records. Blockchain streamlines financial transactions, administrative tasks, and research data management, offering benefits such as automation through smart contracts. Such means of using blockchain can transform the educational sector with increased security, transparency, and efficiency.

EDM includes a range of data-driven approaches, each tailored for specific analytical challenges and educational contexts. Selecting the optimal method hinges on the data characteristics, the precise research objective, and the intended usage. Machine learning models like decision trees, SVM, and ANNs are highly effective for predicting student performance and implementing early interventions. Decision trees provide clear decision-making pathways, which benefits educators needing transparent forecasts about at-risk students. On the other hand, ANNs and deep learning models are better at getting accurate results with large datasets, which makes them suitable for large learning management systems.

In personalized learning and adaptive education, clustering methods like k-means and hierarchical clustering assist in classifying students according to their learning patterns, enabling educators to customize lesson plans to meet individual needs. Adaptive learning platforms use reinforcement learning and recommender systems to provide dynamic suggestions for study resources based on students' interactions. For assessment and automatic grading, NLP techniques, sentiment analysis, and text mining play crucial roles in assessing students' written materials, discussion boards, and feedback. Intelligent grading and evaluation of student essays is made possible by NLP-driven models like BERT and GPT transformer architectures. This makes sure that the feedback process is both automated and personalized.

The evolution of educational learning systems, including blended, hybrid, and intelligent approaches, has generated diverse types of academic data. Recognizing the limitations of single data source analysis, integrating MLA emerges as a crucial strategy, leveraging various data types such as time series, multimedia, text, and numerical data. The application of educational data mining techniques and the development of data fusion methods in the last decade underscores the importance of a comprehensive and diverse approach to improve educational practices, encompassing student learning processes, teacher behaviour modelling, and engagement forecasting in the classroom.

10 Future research directions

This section critically examines the implications, limitations, and potential avenues for future research, fostering a comprehensive understanding of the subject matter. Authors in [116] demonstrated that very few studies had been conducted in areas such as intelligent tutor systems, student profiling, examination scheduling, timetabling, learner motivation, the augmentation of the semantic web in education and its useableness, implications of education affordability, the effect of classroom deco- ration to augment teaching and learning and learner annotation. These are the few negligible attributes that need to be explored in educational data mining. In [117], the authors suggested some of the crucial future research areas that need attention by the researchers in the domain. They found that most data sources focus only on student data, but only a few studies examined data on teachers. It will be fascinating to integrate the teacher and student data in a single platform to extract information on whether teacher characteristics influence student behaviour and similar such kinds of studies. Several open research questions exist regarding integrating environmental and psychometric data in EDM. Specifically, it is of interest to determine how the learning environment (e.g., lighting, humidity, temperature) influences student psychological processes. Furthermore, there is a need for more exploration of enhanced data fusion techniques in educational data, including possibilistic and probabilistic approaches, filters, and Dempster-Shafer evidence theory. Data fusion of multimodal data can be used in many unaddressed areas such as course organization and construction, learning strategy recommendations, and classroom planning.

In case of finding an impact on student performance, future research can aim at focusing on other influencing factors like domain knowledge and teaching quality before the commencement of the courses apart from established factors like quizzes, midterm examinations, performance in earlier attended courses, family background, etc. [118]. In place of grades or marks, the proper measure of value addition can aid in determining the impact of other influencing factors. Different aspects of learning can be exploited by mining unstructured educational data in examination question papers, course syllabi, answer scripts, etc. Social network analysis can be another tool for measuring the impact of student behaviour. Most of the studies predicted the performance of the students after course commencement. It will be challenging for the researchers in their future studies to predict before course commencement efficiently utilizing domain knowledge, student past performance and background, student behaviour in recent times, and previous teaching quality of the teacher. While most studies classify student success or grades, predicting final scores or student grades remains a promising avenue for future research. Hence, before course commencement enhancing the success prediction efficacy can be another research goal of the future.

eXplainable Artificial Intelligence (XAI) facilitates not only the predictions on academic performance but also necessitates the assessment of underlying features that play a crucial role in exhibiting such predictions. Accomplishing explainable EDM would be a cornerstone in this domain and should be among the substantial goals for a trustworthy and healthier EDM practice in the long horizon [41]. More predictive analytics with proper explainability and trustworthiness are expected on educational datasets. Such interpretable outcome leverages the students in career counselling, effective decision-making, and job placement [119].

Social media applications are being adopted and utilized in the University scenario [120]. Researchers explored the research model related to outcomes and experiences of social media. The Technology Acceptance Model (TAM) was exploited to understand the perceived usefulness, perceived playfulness, perceived ease of use, and behavioural intention of utilizing social media for e-learning in higher educational institutes of different regions. In future research, more in-depth study from the perspective of educators and students is needed. This will explore how social networking sites are implemented in education and what factors affect their use. Further research to expand our knowledge on the associations between the use patterns of social media and utility perceived in the educational resources' context opens new avenues in the domain [121]. Future research may shed more light on online learning networks, behavioural and content analysis, topics discussed among various learning communities, explore the relationship between community structures and topics, and examine the participant's perception across and within communities [122].

Opinion mining AKA sentiment analysis in education is utilized to explore student/teacher opinions and improve the pedagogically teaching–learning practices, reduce course abandonment, and students' performance [123]. Educational institutions use sentiment analysis tools to extract insights from student feedback in both traditional and MOOC courses. Opinion spam detection, negation and polysemous words, and multi-polarity are some challenges in this research space. Education-oriented sentiment annotation techniques are one of the hot topics for future research.

Education-oriented unsupervised annotation methods can overcome data labelling and lexical ambiguity issues. Reinforcement learning can be utilized based on education-oriented annotation to showcase enhanced support for goal-based learning tasks. Topic ontology can be exploited to construct an educational knowledge hub. Multimodal data from the student feedback can be examined to define relationships among them.

Smart education (integrating AI and IoT into education) positively impacts of learners' motivation, attendance, and engagement [124]. 5G, AI, and IoT integration led us to many smart solutions, but education institutions have not yet fully utilized these solutions. The transition to smart education faces challenges, including social and computational resistance. Researchers can delve deeper into these smart solutions (smart administration, smart classroom, smart assessment, smart pedagogy, etc.) and help educationists deliver their resources more smartly and plan accordingly.

Integration of video games/mobile video games in education catalysts the modern process of education digitalization [125]. Prospects for future research areas include the application of mobile video games in skill development such as enhancement of programming skills, cloud computing, and Big Data. The students who utilized such materials in the training process could be interviewed to conduct a regression analysis of such platforms. Potential risks and barriers associated with video games such as addiction, violence, and teacher and parental attitudes also need to be taken into account for the future development of such games [113].

Future research in blockchain within educational data mining could involve exploring decentralized and transparent systems for secure student data management, ensuring privacy, and verifying academic credentials. Additionally, investigating smart contracts to automate and enhance processes like accreditation, certification, and secure sharing of educational records could be a focus [126, 127]. Integrating blockchain with emerging technologies like AI for more advanced analytics and personalized learning experiences may also be an area of exploration. On the other hand, Chatbot can be used for data collection. It can provide real-time feedback and guide in finding the proper course materials and relevant information. Chatbots can be used for assessment and quizzing [128]. Provide hints for questions, and offer explanations. Research can be done on refining NLP capabilities to understand better and respond to students' queries [122].

11 Conclusion

The study reviews the trends and techniques applied in educational data mining for a decade (2013–2023). It demonstrated the current trends and strategies while showcasing the existing and prevailing methodologies in the domain. The study also focuses on the future trends of EDM. Based on the comprehensive exploration of EDM, explainable artificial intelligence, chatbots, and blockchain technology in education presented in the preceding discussion, it is evident that these advancements are reshaping the landscape of teaching and learning in profound ways. This discussion has underscored the multifaceted implications of leveraging these technologies, elucidating their potential to revolutionize various aspects of education, from personalized learning experiences to data management and security. XAI emerges as a transformative force in education, offering personalized experiences, adaptive assessments, and tailored feedback that enhance learning outcomes and empower students and educators. Through understandable AI explanations and customized learning, XAI increases student engagement and empowers educators with predictive analytics and early intervention strategies for informed decision-making. The preference for model-agnostic methods like LIME and SHAP highlights the importance of interpretability in ensuring transparency and trust in AI-driven educational practices. Similarly, chatbots are identified as advanced tools for data gathering, analysis, and student engagement, offering personalized learning experiences and real-time feedback loops that enhance comprehension and boost student engagement. By addressing individual student needs and providing tailored communication, feedback, and resources, chatbots contribute to a more dynamic and interactive learning environment while offering valuable insights into student preferences, behaviours, and learning styles that inform evidence-based decisions for instructional design and curriculum improvement.

Blockchain technology, on the other hand, revolutionizes data management, security, and verification in education by offering a decentralized, tamper-proof method for confirming academic qualifications and ensuring the secure issuance and verification of degrees, diplomas, and transcripts. Through its immutability and transparency, blockchain enhances the integrity and credibility of education records while streamlining financial transactions, administrative tasks, and research data management through automation via smart contracts. Moreover, the evolution of educational learning systems, including blended, hybrid, and intelligent approaches, has generated diverse educational data types, necessitating the integration of machine learning algorithms to leverage various data types such as time series, multimedia, text, and numerical data. This comprehensive and diverse approach, encompassing student learning processes, teacher behaviour modelling, and engagement forecasting in the classroom, underscores the importance of educational data mining techniques and the development of data fusion methods in improving educational practices.

In conclusion, the insights derived from this discussion emphasize the transformative potential of educational data mining, explainable artificial intelligence, chatbots, and blockchain technology in revolutionizing teaching and learning. By offering personalized, transparent, and efficient educational experiences, these advancements pave the way for a more dynamic, engaging, and effective educational environment that caters to the diverse needs of students and educators alike. As we continue to harness the power of these technologies, it is essential to prioritize ethical considerations, data privacy, and equitable access to ensure that education remains a catalyst for empowerment, innovation, and social progress in the digital age.

Acknowledgements The authors are thankful to all the individuals who helped directly or indirectly to complete this study. For open access, the author has applied a 'Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version.

Author contributions Emi Kalita, Solomon Sunday Oyelere, Silvia Gaftandzhieva designed the study. Emi Kalita, Kandala N V P S Rajesh 4, Senthil Kumar Jagatheesaperumal, Asmaa Mohamed, Yomna M. Elbarawy, Abeer Sayed Desouky, Sadiq Hussain performed conceptualization and Methodology. The software and analysis of data presentation have been performed by Mehmet Akif Cifci, Paraskevi Theodorou, Slavoljub Hilčenko, Jiten Hazarika, Tazid Ali. Furthermore, Emi Kalita, Jiten Hazarika, Tazid Ali and Sadiq Hussain wrote the original draft and visualized the figures. The final manuscript has been reviewed and edited by Solomon Sunday Oyelere, Silvia Gaftandzhieva. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by Lulea University of Technology. This paper is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0001-C01. The paper reflects only the author's and the European Union is not responsible for any use that may be made of the information it contains.

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate The author of this manuscript confirms that: (i) Approval and/or informed consent is not required as it is a review paper and no data is used in experimentation. (ii) All procedures followed were by the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964 and its later amendments.

Consent for publication Not Applicable.

Competing interests The authors declare no competing interests.

References

1. Lian B, Putra M. The role and effectiveness of education administration in developing institutions of higher education in the digital era. In: 4th International Conference on Education and Social Science Research (ICESRE). KnE Social Sciences, 2022; pp. 592–602
2. Younas M, Noor U, Zhou X, Menhas R, Qingyu X. COVID-19, students satisfaction about e-learning and academic achievement: mediating analysis of online influencing factors. Front Psychol. 2022;13:948061.
3. Hamdan KM, Al-Bashaireh AM, Zahran Z, Al-Daghestani A, Al- Habashneh S, Shaheen AM. University students' interaction, internet self-efficacy, self-regulation, and satisfaction with online education during pandemic crises of COVID-19 (SARS-CoV-2). Int J Educ Manag. 2021;35(3):713–25.
4. Ling NH, Chen CJ, Teh CS, John DS, Ch'ng LC, Lay YF. Global trends of educational data mining in online learning. Int J Technol Educ. 2023;6(4):656–80.
5. Gaftandzhieva S, Hussain S, Hilčenko S, Doneva R, Boykova K. Data- driven decision making in HEI: state-of-play. Int J Adv Comput Sci Appl (IJACSA). 2023;14(6):397–405.

6. Baek C, Doleck T. Educational data mining: a bibliometric analysis of an emerging field. IEEE Access. 2022;10:31289–96.
7. Hussain S, Gaftandzhieva S, Maniruzzaman M, Doneva R, Muhsin ZF. Regression analysis of student academic performance using deep learning. Educ Inf Technol. 2021;26(1):783–98.
8. Romero C, Ventura S. Educational data mining: a survey from 1995 to 2005. Expert Syst Appl. 2007;33(1):135–46.
9. Jagatheesaperumal SK, Ahmad K, Al-Fuqaha A, Qadir J. Advancing education through extended reality and internet of everything enabled metaverses: applications, challenges, and open issues. IEEE Trans Learn Technol. 2024;17:1120–39.
10. Zoric A. Benefits of educational data mining. J Int Bus Res Mark. 2020;6(1):12–6.
11. Silva C, Fonseca J. Educational data mining: a literature review. In: Advances in intelligent systems and computing. Advances in intelligent systems and computing. Cham: Springer; 2017. p. 87–94.
12. Tsiakmaki M, Kostopoulos G, Kotsiantis S, Ragos O. Implementing AutoML in educational data mining for prediction tasks. Appl Sci. 2019;10(1):90.
13. Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2020. https://doi.org/10.1002/widm.1355.
14. Aleem A, Gore MM. Educational data mining methods: A survey. In: 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT). IEEE, 2020
15. Alturki S, Alturki N. Using educational data mining to predict students' aca- demic performance for applying early interventions. J Inf Technol Educ Innov Pract. 2021;20:121–37.
16. Triayudi A, Widyarto WO. Educational data mining analysis using classification techniques. J Phys Conf Ser. 2021;1933(1):012061.
17. Kriˇzani´c S. Educational data mining using cluster analysis and decision tree technique: a case study. Int J Eng Bus Manag. 2020;12:184797902090867.
18. Vora DR, Iyer K. EDM-survey of performance factors and algorithms applied. Int J Eng Technol. 2018;7(2.6):93–7.
19. Alshareef F, Alhakami H, Alsubait T, Baz A. Educational data mining applications and techniques. Int J Adv Comput Sci Appl. 2020. https://doi.org/10.14569/IJACSA.2020.0110494.
20. Juhanˇ´ak L, Zounek J, Rohl´ıkov´a L. Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. Comput Human Behav. 2019;92:496–506.
21. Abdelmagid AS, Qahmash AIM. Utilizing the educational data mining techniques" orange technology" for detecting patterns and predicting academic performance of university students. Inf Sci Lett. 2023;12:1415–31.
22. Shahiri AM, Husain W, Rashid NA. A review on predicting student's performance using data mining techniques. Procedia Comput Sci. 2015;72:414–22.
23. Kaur H, Bathla EG. Student performance prediction using educational data mining techniques. Int J Future Revolut Comput Sci Commun Eng. 2018;4(12):93–7.
24. Nichat AA, Raut DAB. Predicting and analysis of student performance using decision tree technique. Int J. 2017;5:7319–28.
25. Bin Roslan MH, Chen CJ. Educational data mining for student performance prediction: a systematic literature review (2015–2021). Int J Emerg Technol Learn. 2022;17(05):147–79.
26. Zhang Y, Yun Y, An R, Cui J, Dai H, Shang X. Educational data mining techniques for student performance prediction: method review and comparison analysis. Front Psychol. 2021;12:698490.
27. Alamgir Z, Akram H, Karim S, Wali A. Enhancing student performance prediction via educational data mining on academic data. Inf Educ. 2023. https://doi.org/10.15388/infedu.2024.04.
28. Liu Q, Tong S, Liu C, Zhao H, Chen E, Ma H, Wang S. Exploiting cognitive structure for adaptive learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA. 2019
29. Agrusti F, Bonavolont`a G, Mezzini M. University dropout prediction through educational data mining techniques: a systematic review. J E-Learn Knowl Soc. 2019;15:161.
30. Aldowah H, Al-Samarraie H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: a review and synthesis. Telemat Inform. 2019;37:13–49.
31. Meghji AF, Mahoto NA, Asiri Y, Alshahrani H, Sulaiman A, Shaikh A. Early detection of student degree-level academic performance using educational data mining. PeerJ Comput Sci. 2023;9:1294.
32. Jacob D, Henriques R. Educational data mining to predict bachelors students' success. EmergingScience J. 2023;7:159–71.
33. Arifin M, Widowati Farikhin. Using education data mining (EDM) and tracer study (TS) data as materials for evaluating higher education curriculum and policies. KnE Soc Sci, 2022.
34. Inusah F, Missah YM, Najim U, Twum F. Integrating expert system in managing basic education: a survey in Ghana. Int J Inf Manag Data Insights. 2023;3(1):100166.
35. Guan X, Feng X, Islam AYMA. The dilemma and countermeasures of educational data ethics in the age of intelligence. Humanit Soc Sci Commun. 2023. https://doi.org/10.1057/s41599-023-01633-x.
36. Ban˜eres D, Rodr´ıguez ME, Guerrero-Rold´an AE, Karadeniz A. An early warning system to detect at-risk students in online higher education. Appl Sci. 2020;10(13):4427.
37. Rachha A, Seyam M. Explainable ai in education: current trends, challenges, and opportunities. SoutheastCon. 2023;2023:232–9.
38. Jagatheesaperumal SK, Pham QV, Ruby R, Yang Z, Xu C, Zhang Z. Explainable AI over the Internet of Things (IoT): overview, state-of-the-art and future directions. IEEE Open J Commun Soc. 2022;3:2106–36.
39. Jang Y, Choi S, Jung H, Kim H. Practical early prediction of students' performance using machine learning and explainable ai. Educ Inf Technol. 2022;27(9):12855–89.
40. Hasib KM, Rahman F, Hasnat R, Alam MGR. A machine learning and explainable ai approach for predicting secondary school student performance. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2022; pp. 0399–0405.
41. Guleria P, Sood M. Explainable ai and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. Educ Inf Technol. 2023;28(1):1081–116.
42. Afzaal M, Nouri J, Zia A, Papapetrou P, Fors U, Wu Y, Li X, Weegar R. Explainable ai for data-driven feedback and intelligent action recommenda- tions to support students self-regulation. Front Artif Intell. 2021;4:723447.

43. Riel J. Essential features and critical issues with educational chatbots: toward personalized learning via digital agents. In: Handbook of research on mod- ern educational technologies, applications, and management. Hershey: IGI Global; 2021. p. 246–62.
44. Haristiani N, Rifa'i MM. Combining chatbot and social media: enhancing personal learning environment (ple) in language learning. Indones J Sci Technol. 2020;5(3):487–506.
45. Kooli C. Chatbots in education and research: a critical examination of ethical implications and solutions. Sustainability. 2023;15(7):5614.
46. Hwang G-J, Chang C-Y. A review of opportunities and challenges of chatbots in education. Interact Learn Environ. 2023;31(7):4099–112.
47. Mun˜oz SAS, Gayoso GG, Huambo AC, Tapia RDC, Incaluque JL, Aguila OEP, Cajamarca JCR, Acevedo JER, Rivera HVH, Arias- Gonz´ales JL. Examining the impacts of chatgpt on student motivation and engagement. Soc Space. 2023;23(1):1–27.
48. Abbas N, Whitfield J, Atwell E, Bowman H, Pickard T, Walker A. Online chat and chatbots to enhance mature student engagement in higher education. Int J Lifelong Educ. 2022;41(3):308–26.
49. Chang DH, Lin MP-C, Hajian S, Wang QQ. Educational design principles of using ai chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. Sustainability. 2023;15(17):12921.
50. Ocheja P, Agbo FJ, Oyelere SS, Flanagan B, Ogata H. Blockchain in education: a systematic review and practical case studies. IEEE Access. 2022;10:99525–40.
51. Guustaaf E, Rahardja U, Aini Q, Maharani HW, Santoso NA. Blockchain-based education project. Aptisi Trans Manag (ATM). 2021;5(1):46–61.
52. Priya N, Ponnavaikko M, Aantonny R. An efficient system framework for managing identity in educational system based on blockchain technology. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE, 2020; pp. 1–5
53. Tariq A, Haq HB, Ali ST. Cerberus: a blockchain-based accreditation and degree verification system. IEEE Trans Comput Soc Syst. 2022. https://doi.org/10.1109/TCSS.2022.3188453.
54. Martins V, Oyelere SS, Tomczyk L, Barros G, Akyar O, Eliseo MA, Amato C, Silveira IF. A blockchain microsites-based ecosystem for learning and inclusion. In: Brazilian Symposium on Computers in Education (Simp´osio Brasileiro de Inform´atica na Educa¸c˜ao-SBIE), 2019; vol. 30: p. 229
55. Woo C, Yoo J. Exploring the determinants of blockchain acceptance for research data management. J Comput Inf Syst. 2023;63(1):216–27.
56. Churov´a V, Vyˇskovsky` R, Marˇs´alov´a K, Kudl´aˇcek D, Schwarz D, et al. Anomaly detection algorithm for real-world data and evidence in clinical research: implementation, evaluation, and validation study. JMIR Med Inf. 2021;9(5):27172.
57. Romero C, Ventura S. Educational data mining: a review of the state of the art. IEEE Trans Syst, Man, Cybern, Part C (Appl Rev). 2010;40(6):601–18.
58. Romero C, Ventura S. Data mining in education. Wiley Interdiscip Rev Data min Knowl Discov. 2013;3(1):12–27.
59. Baker RS, Yacef K, et al. The state of educational data mining in 2009: a review and future visions. J Educ Data Min. 2009;1(1):3–17.
60. Pen˜a-Ayala A. Educational data mining: a survey and a data mining-based analysis of recent works. Expert Syst Appl. 2014;41(4):1432–62.
61. Bakhshinategh B, Zaiane OR, ElAtia S, Ipperciel D. Educational data mining applications and tasks: a survey of the last 10 years. Educ Inf Technol. 2018;23:537–53.
62. Trung BD, Son NT, Tung ND, Son KA, Anh BN, Lam PT: Educational data mining: a systematic review on the applications of classical methods and deep learning until 2022. In: 2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA). IEEE, 2023; pp. 1–15
63. Shams SA, Omar AH, Desuky AS, Abou-Kreisha MT, Elsharawy GA. Even-odd crossover: a new crossover operator for improving the accuracy of students' performance prediction. Bull Electr Eng Inf. 2022;11(4):2292–302.
64. Mahareek EA, Desuky AS, El-Zhni HA. Simulated annealing for SVM parameters optimization in student's performance prediction. Bull Electr Eng Inf. 2021;10(3):1211–9.
65. Keser SB, Aghalarova S. Hela: a novel hybrid ensemble learning algorithm for predicting academic performance of students. Educ Inf Technol. 2022;27(4):4521–52.
66. Mayilvaganan M, Kalpanadevi D. Comparison of classification techniques for predicting the performance of students academic environment. In: 2014 Inter- national Conference on Communication and Network Technologies, IEEE, 2014; pp. 113–118.
67. Nirmala M, Selvi TS, Saravanan V. Student academic performance prediction under various machine learning classification algorithms. Int J Res Appl Sci Eng Technol. 2021;9(11):221–37.
68. Triayudi A, Haerani R, et al. Data mining k-means algorithm for performance analysis. J Phys Conf Ser. 2022;2394:012031.
69. Chrisnanto YH, Abdullah G. The uses of educational data mining in academic performance analysis at higher education institutions (case study at unjani). Matrix J Manaj Teknol Inf. 2021;11(1):26–35.
70. Arsad PM, Buniyamin N, et al. A neural network students' performance prediction model (nnsppm). In: 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), IEEE. 2013; pp. 1–5.
71. Kotsiantis S, Patriarcheas K, Xenos M. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowl-Based Syst. 2010;23(6):529–35.
72. Burgos C, Campanario ML, Pen˜a D, Lara JA, Lizcano D, Mart´ınez MA. Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout. Comput Electr Eng. 2018;66:541–56.
73. Sandoval A, Gonzalez C, Alarcon R, Pichara K, Montenegro M. Centralized student performance prediction in large courses based on low-cost variables in an institutional context. Internet High Educ. 2018;37:76–89.
74. Farissi A, Dahlan HM, Samsuryadi. Genetic algorithm based feature selection for predicting student's academic performance. In: Emerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing. Springer, 2020; 4, pp. 110–117.
75. Sokkhey P, Okazaki T. Hybrid machine learning algorithms for predicting academic performance. Int J Adv Comput Sci Appl. 2020;11(1):32–41.
76. K¨ock M, Paramythis A. Activity sequence modelling and dynamic clustering for personalized e-learning. User Model User-Adapt Interact. 2011;21:51–97.
77. Fei M, Yeung D-Y. Temporal models for predicting student dropout in massive open online courses. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), IEEE. 2015; pp. 256–263

78. Lende SP, Raghuwanshi M. Question answering system on education acts using NLP techniques. In: 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), IEEE. 2016; pp. 1–6
79. Jayakodi K, Bandara M, Meedeniya D. An automatic classifier for exam questions with wordnet and cosine similarity. In: 2016 Moratuwa Engineering Research Conference (MERCon), IEEE. 2016; pp. 12–17
80. Corrigan O, Smeaton AF. A course agnostic approach to predicting student success from VLE log data using recurrent neural networks. In: Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings. Springer, 2017; 12, pp. 545–548.
81. Kuo JY, Pan CW, Lei B. Using stacked denoising autoencoder for the student dropout prediction. In: 2017 IEEE International Symposium on Multimedia (ISM), IEEE. 2017; pp. 483–488.
82. Wang W, Yu H, Miao C. Deep model for dropout prediction in MOOCs. In: Proceedings of the 2nd International Conference on Crowd Science and Engineering, 2017; pp. 26–32
83. Nguyen D, Nguyen K, Sridharan S, Ghasemi A, Dean D, Fookes C. Deep spatiotemporal features for multimodal emotion recognition. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. 2017; pp. 1215–1223
84. Yasaswi J, Purini S, Jawahar C. Plagiarism detection in programming assignments using deep features. In: 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), IEEE. 2017; pp. 652–657
85. Herrera-Flores AH-BB, Tomas D, Navarro-Colorado B. A systematic review of deep learning approaches to educational data mining. Complexity. 2019. https://doi.org/10.1155/2019/1306039.
86. Fern´andez A, Garc´ıa S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets, vol. 10. Cham: Springer; 2018.
87. Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: experimental evaluation. Inf Sci. 2020;513:429–41.
88. Kamalov F, Leung H-H, Cherukuri AK. Keep it simple: random over- sampling for imbalanced data. In: 2023 Advances in Science and Engineering Technology International Conferences (ASET), IEEE. 2023; pp. 1–4
89. Wongvorachan T, He S, Bulut O. A comparison of undersampling, over- sampling, and smote methods for dealing with imbalanced classification in educational data mining. Information. 2023;14(1):54.
90. Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH. Radius-smote: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data. IEEE Access. 2021;9:74763–77.
91. He H, Bai Y, Garcia EA, Li S. Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE. 2008; pp. 1322–1328
92. Han H, Wang W-Y, Mao B-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing, Springer, 2005; pp. 878–887.
93. Gavrylenko S, Vladislav Z, Khatsko N. Methods for improving the quality of classification on imbalanced data. In: 2023 IEEE 4th KhPI Week on Advanced Technology (KhPIWeek), IEEE, 2023; pp. 1–5.
94. Maciejewski T, Stefanowski J. Local neighbourhood extension of smote for mining imbalanced data. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2011; pp. 104–111.
95. Barua S, Islam MM, Yao X, Murase K. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng. 2012;26(2):405–25.
96. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27–30, 2009 Proceedings, Springer. 2009; 13, pp. 475–482.
97. Wang Z, Wu C, Zheng K, Niu X, Wang X. Smotetomek-based resampling for personality recognition. IEEE access. 2019;7:129678–89.
98. Thai-Nghe N, Busche A, Schmidt-Thieme L. Improving academic performance prediction by dealing with class imbalance. In: 2009 Ninth International Conference on Intelligent Systems Design and Applications, IEEE. 2009; pp. 878–883
99. Pristyanto Y, Pratama I, Nugraha AF. Data level approach for imbalanced class handling on educational data mining multiclass classification. In: 2018 International Conference on Information and Communications Technology (ICOIACT), IEEE. 2018; pp. 310–314.
100. Hassan H, Ahmad NB, Anuar S. Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. J Phys Conf Ser. 2020;1529:052041.
101. Al-Ashoor A, Abdullah S. Examining techniques to solving imbalanced datasets in educational data mining systems. Int J Comput. 2022;21(2):205–13.
102. Chango W, Lara JA, Cerezo R, Romero C. A review on data fusion in multimodal learning analytics and educational data mining. Wiley Interdiscip Rev Data Min Knowl Discov. 2022;12(4):1458.
103. Hrastinski S. What do we mean by blended learning? TechTrends. 2019;63(5):564–9.
104. Raes A. Exploring student and teacher experiences in hybrid learning environments: does presence matter? Postdigit Sci Educ. 2022;4(1):138–59.
105. Chen X, Zou D, Xie H, Wang FL. Past, present, and future of smart learning: a topic-based bibliometric analysis. Int J Educ Technol High Educ. 2021;18:1–29.
106. Sharma K, Giannakos M. Multimodal data capabilities for learning: what can multimodal data tell us about learning? Br J Edu Technol. 2020;51(5):1450–84.
107. Liu Z, Yang S, Tang J, Heffernan N, Luckin R. Recent advances in multimodal educational data mining in k-12 education. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020; pp. 3549–3550
108. Giannakos MN, Sharma K, Pappas IO, Kostakos V, Velloso E. Multi- modal data as a means to understand the learning experience. Int J Inf Manage. 2019;48:108–19.
109. Prieto LP, Sharma K, Kidzinski L, Rodr´ıguez-Triana MJ, Dillenbourg P. Multimodal teaching analytics: automated extraction of orchestration graphs from wearable sensor data. J Comput Assist Learn. 2018;34(2):193–203.
110. Liu R, Stamper J, Davenport J, Crossley S, McNamara D, Nzinga K, Sherin B. Learning linkages: integrating data streams of multiple modalities and timescales. J Comput Assist Learn. 2019;35(1):99–109.
111. Gaftandzhieva S, Doneva R, Jagatheesaperumal SK. Approaches and tools for quality assurance in distance learning: state-of-play. Int J Adv Comput Sci Appl. 2023;14(8):1–13.

112. Yue J, Tian F, Chao K-M, Shah N, Li L, Chen Y, Zheng Q. Recognizing multidimensional engagement of e-learners based on multi-channel data in e- learning environment. IEEE Access. 2019;7:149554–67.
113. Chango W, Cerezo R, Romero C. Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. Comput Electr Eng. 2021;89:106908.
114. Wu D, Chen J, Deng W, Wei Y, Luo H, Wei Y. The recognition of teacher behavior based on multimodal information fusion. Math Probl Eng. 2020;2020:1–8.
115. Peng S, Nagao K. Recognition of students' mental states in discussion based on multimodal data and its application to educational support. IEEE Access. 2021;9:18235–50.
116. Chango W, Lara JA, Cerezo R, Romero C. A review on data fusion in multimodal learning analytics and educational data mining. Wiley Interdiscip Rev Data Min Knowl Discov. 2022;12: e1458.
117. Khan A, Ghosh SK. Student performance analysis and prediction in class- room learning: a review of educational data mining studies. Educ Inf Technol. 2021;26(1):205–40.
118. Tousside B, Dama Y, Frochte J. Towards explainability in modern educational data mining: A survey. In: Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. SCITEPRESS - Science and Technology Publications. 2022
119. Alshurideh MT, Kurdi BA. Factors affecting social networks acceptance: an extension to the technology acceptance model using PLS-SEM and machine learning approach. Int J Data Netw Sci. 2023;7(1):489–94.
120. G´omez-Garc´ıa M, Soto-Varela R, Boumadan M, Matosas-L´opez L. Can the use patterns of social networks in university students predict the utility perceived in digital educational resources? Interact Learn Environ. 2020;31:1–14.
121. Du H, Xing W, Zhu G. Mining teacher informal online learning networks: Insights from massive educational chat tweets. J Educ Comput Res. 2023;61(1):127–50.
122. Shaik T, Tao X, Dann C, Xie H, Li Y, Galligan L. Sentiment analysis and opinion mining on educational data: a survey. Nat Lang Process J. 2023;2:100003.
123. Badshah A, Ghani A, Daud A, Jalal A, Bilal M, Crowcroft J. Towards smart education through internet of things: a survey. ACM Comput Surv. 2023;56(2):1–33.
124. Quadar N, Rahouti M, Ayyash M, Jagatheesaperumal SK, Chehri A. IoT-AI/Machine learning experimental testbeds: the missing piece. IEEE Internet Things Mag. 2024;7(1):136–43.
125. Arias M. Using video games in education. J Mason Grad Res. 2014;1(2):49–69.
126. Rani P, Sachan RK, Kukreja S. A systematic study on blockchain technology in education: initiatives, products, applications, benefits, challenges and research direction. Computing. 2023;106:1–43.
127. Ocheja P, Flanagan B, Ogata H. Blockchain in education: Connecting learning records and contents through the blockchain. In: Proceedings of Blockchain in Kyoto 2021 (BCK21), 2021; p. 011006
128. Okonkwo CW, Ade-Ibijola A. Chatbots applications in education: a systematic review. Comput Educ Artif Intell. 2021;2:100033.