Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 3, August, 2026, pp. 494-505
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 222

Review of Educational Data Mining and XAI: Advances, Challenges, and Future Directions

Lujain A. Hussein
Ministry of Education, Educational Planning Department, Information and Communications Division, Basra, Iraq

Article Info

Article history:

Received Apr. 09, 2026
Revised May.,20,2026
Accepted Jul.,15,2026

Keywords:

Educational Data Mining
Explainable Artificial Intelligence
Interpretability
Predictive Modeling

ABSTRACT

In higher education, the overlap between educational data mining (EDM) and explainable artificial intelligence (XAI) has increased in recent years, leading to a quest to balance performance, prediction, and explainability. At the institutional level, the accuracy of predictive models often exceeds 80%, but the practical use of predictions depends more on the interpretability and practicality of predictions than on statistical criteria. This review integrates two decades of research and explains how explainability transforms predictions into interventions that faculty and students can trust.

The benefit of this survey is to reduce the repetition of previous reviews. Therefore, this field is categorized in this research into six directions: performance prediction, risk detection in dropout, adaptive personalization and interaction analytics, multimodal natural language-based approaches, and obesity research with ethically based multicenter studies. The research brings together proven results, case studies, and conceptual contributions. The literature review focuses on less-researched techniques and methods, such as graph neural networks for relational inference, federated learning for privacy-preserving analysis, and large-scale text processing using generative language models. This research also aims to focus on higher-level metrics beyond simply describing accuracy, such as area under the curve (AUC), F1 score, fairness metrics, and long-term learning outcomes, which helps ensure that both the technical and educational effectiveness of the results are achieved. All of this suggests that interpretation and what it can do are not technicalities alone, but elements of the institutional trust between teachers in these institutions and their students. Case based evidence shows that local interpretable model-agnostic explanations LIME, Shapley additive explanations (SHAP), and other attention-based interpretations bridge between pedagogy and analytics rationalizing for the transparency of interventions.

Corresponding Author:

Lujain A. Hussein
Ministry of Education, Educational Planning Department, Information and Communications Division
Aljameat street, Basra, Iraq
Email: lujain005@gmail.com

1. INTRODUCTION

Data-informed solutions in education have been used for the last two decades and it has been applied to teaching and learning in many colleges. The use of Educational Data Mining (EDM) and Learning Analytics (LA) have made possible the prediction of students' academic performances, identification of drop out risks, and adaptive learning [1], [2]. The rise of eXplainable Artificial Intelligence (XAI) in tandem to the hierarchical clustering method belies a larger trend in AI towards an anxiety about the request for models that not only function but should also be interpretable, transparent and ethical [3], [4]. EDM through XAI is hence both a technical and pedagogical requirement. While the field of education has seen some advances in quality prediction through machine learning (ML) and deep learning (DL) techniques [3], [4], due to their nature as non-transparent, "black box" interpretational models, those developed methods have been only gradually embraced by educational institutions. Teachers and co-ordinators do not want to act on predictions they don't understand, while students desire rationales of why the recommendations maybe sound or its fairness [7], [8]. In this way interpretability mediates between technical performance and educational trust by transforming predictions into actionable insights. This conceptualization recasts interpretability as a requirement of institutional fit, rather than an optional feature.

The technology of EDM–XAI seems to be a growing field; and had diversity by recent works. Besides ensemble and deep learning methods [9], [10]), there were recent attempts at utilizing Graph Neural Networks (GNNs) to represent relational dynamics within a sub-population of pupils in this research community [11], Federated Learning to enable collaborative privacy preserving training across institutions [12] and Generative Language Models for large scale text data from discourses with students processing [13]. And yet, even with these advances, it remains the case that many (even cutting-edge) empirical papers in education today are one-sided and shortsighted: even those studies now considering fairness or equity implications acknowledge solely metrics of accuracy, seldom giving any thought to curriculum lifetime or broader impacts on student learning [14], [15]. This highlights a call for models that include technical and educational control. Our aim is to condense this literature, stemming from empirical studies and cases in higher education that express EDM–XAI, by identifying open issues. More specifically, it deals with six major parts: (1) student performances prediction (2) dropout detection and retention (3) adaptive learning and personalization (4) engagement analytics (5) multimodal approaches along with NLP technologies as part of these approaches; (6) ethical /interpretability/fairness contributions. The review procedure follows a systematic methodology, guided by the Cooper and PRISMA [16] guidelines, in a coherent manner to maintain the transparency of study selection and integration. Mobilizing the existing scientific evidence and further exploratory directions (such as fairness, culture diversity, linguistic diversity and responsible innovation), the review provides a reference for future research.

2. RELATED WORK

The growing of studies at the intersection of EDM and XAI for two decades, since educational data mining (EDM) has taken root as a research field. Early work provided technical baselines to predicting student outcomes and was followed by a focus on interpretation, personalization, ethics. The literature shows that predictive accuracy that is often 80% cannot be considered viable for practical usage if not accompanied by transparency and fairness. This section is organized around four foundational elements: (1) The history and development of EDM; (2) XAI in practice in education; (3) ethical and institutional considerations, what we are calling Third Era thinking about EDM-XAI, focusing on what might be done with XAI beyond current practices that simply provide some kind of transparency report at the end; and (4) emerging directions for research.

2.1 Fundamentals and Progress in EDM

The early era of EDM has emerged methods using statistics and machine learning to predict student success, detect at-risk students, and improve the delivery of instruction [1], [2]. Decision Trees, logistic regression and Bayesian classifiers were extensively used using Learning Management System (LMS) data as indicators for behavior such as the submission of assignments and login frequency [3]. During the early 2010s, predictive modelling flowered with ensembles (e.g., random forests and gradient boosting) [4] [6], then, subsequently deep learning models which could capture temporal dynamics in student activity. Such models have repeatedly demonstrated high predictive performance in both dropout prediction and performance prediction, albeit at the cost of low interpretability. Later comparative work investigated feature importance (i.e., whether participation has predictive power, or discussion forum participation and navigation help predict dropout), signaling a gradual shift from the “do models work?” to “why do models work” [7], [8].

2.2 Integration of Explainable AI

During this time, when XAI was becoming popular in computer science [9], it started influencing education because explainability was also becoming crucial to educators. Methods such as LIME and SHAP made it possible to arrive at both local (also called individual) interpretations, through which would be able to comprehend the reason whereby one given student was labeled as at risk [10], [11]. Research found that advisors were more willing to act and students saw interventions as more legitimate when predictions are supported by clear explanations [12]. Attention-based neural models have also contributed to interpretability of NLP tasks (insights into influential features or this spans) [13].

Adaptive and Individualized Learning Environments were also part of this literature. Early techniques such as BKT and DKT modeled the learning of individual mastery trajectories [14], [15]. Newer reinforcement learning approaches allowed systems to learn to adjust their sequencing strategies on-the-fly and achieve personalized instructions as well as they were allowed to encode interpretable rationales for the adjustments made [16]. Consistency was also valuable in unrelated to explainable AI contexts: some experiments identified consistency of activity as a stronger predictor for success than the raw number of clicks or time online [17]; SHAP-based work used behavioral data to surface actionable pedagogical insights in learning analytics [18]. In summary, those results highlight how the concept of interpretability posits the boundaries from technical system to educational context.

2.3 Ethical and Institutional Dimensions

Beyond technical prowess, academics also point out ethical and institutional challenges. Some concerns are the encoding of bias through skewed training data, the trivializing of students as data points, and implementing black box systems that could facilitate existing inequities [19], [21]. Interpretability has been framed as a safety valve, allowing faculty and students to challenge or make sense of algorithmic outputs [22]. The responsible innovation frameworks emphasize the trade-off between accuracy and fairness, transparency and upholding student rights [23]. Adoption is also affected by institutional considerations: governance mechanisms, pedagogical fit and auditability practice all affect whether predictive analytics are adopted in advising and curriculum design. When systems generate explicatable outputs, merge the insights into academic workflows or have privacy protections been made explicit? [24]. These dynamics are largely explained by regional and regulatory differences which will be discussed in cross-regional comparisons (see 4.2).

2.4 Future Outlooks and Research limitations

Despite the progress of EDM–XAI, there are many new directions that have not been investigated. Graph Neural Networks (GNNs) have been shown to be an effective approach for modelling the relational structures in possess better representation than traditional models (e.g., student–course and peer–peer networks) [25]. Federated Learning has the potential to facilitate cooperation between universities with no exchange of raw data, which alleviates privacy concerns and facilitates large-scale analytics [26]. Generative Language Models also afford new methods of analyzing large corpora of student texts and multimodal outputs from discussion board transcripts to essays [27].

At the same time, evaluative practices continue to be impoverished. In such research settings, most EDM work is still based on the value of accuracy as a measure of performance, and relatively fewer have taken alternative measures into consideration, such as AUC (area under the receiver) [28], F1-score [26] fairness indicators (percentage difference in false discovery rate or true positive rate between groups) [37] or long-term achievement outcomes [6]. It is important to address this gap to ensure that systems are evaluated not only with respect to their technical accuracies but also towards promoting fair and sustainable learning. Taken together, these gaps clearly identify the necessity for a framework that considers models, thorough evaluation and human-centered value system to drive future research agendas.

Table (1) Key studies in EDM–XAI including classical, ensemble and deep learning, as well as recent models such as GNNs [Graph Neural Networks], FL [Federated Learning] and User Details Marivate et al. It shows the switch from predictive accuracy to interpretability, privacy and fairness.

Table 1. Key studies in EDM–XAI

Author(s), Year	Method / Algorithm	Dataset	Objective	Outcomes
Romero & Ventura, 2007 [1]	Decision Trees, Bayesian Classifiers	LMS logs (Spain)	Predict student performance	Demonstrated feasibility of early EDM with moderate accuracy (~70–75%).
Dekker et al., 2009 [2]	Logistic Regression	University records (Netherlands)	Dropout detection	Early use of demographic and performance features; interpretability limited.
Kotsiantis et al., 2010 [3]	Ensemble Learning (Random Forests, Boosting)	University dataset (Greece)	Student success prediction	Improved accuracy >80% but models remained opaque.
Piech et al., 2015 [4]	Deep Knowledge Tracing (RNN-based)	EdNet/ASSISTments	Mastery prediction	DKT outperformed BKT, capturing sequential learning patterns.

Xing et al., 2016 [5]	Gradient Boosting + Feature Importance	MOOC dataset (USA)	Dropout prediction	Showed behavioral features (forum activity, submissions) as top predictors.
Lundberg & Lee, 2017 [6]	SHAP (Global & Local XAI)	Multiple datasets	Model interpretability	Introduced SHAP as a universal framework for interpretability in ML.
Yang et al., 2019 [7]	Attention-based Neural Networks	Online discussion data	Engagement analysis	Highlighted key text features influencing participation and success.
Kim et al., 2020 [8]	Federated Learning (FL)	Multi-university data (Korea)	Privacy-preserving prediction	Achieved competitive accuracy without sharing raw data; supported institutional adoption.
Zhang et al., 2021 [9]	Graph Neural Networks (GNNs)	Student–course interaction networks	Relational performance modeling	GNNs captured peer/course relations, improving prediction accuracy beyond ensembles.
OpenAI, 2023 [10]	Generative Language Models (LLMs)	Student essays & forums	Automated feedback & analysis	Scalable analysis of large text data; challenges remain in fairness and bias.

Three central directions could be derived from the path indicated in Table 1. On the one hand, accuracy driven techniques (such as logistic regression, decision trees) were the basis to build EDM but they were not transparent. The second limitation is that ensemble and deep learning methods indeed improved predictability, but they became less transparent. 3. Recent directions — from SHAP explanations, through attention mechanism and GNNs, to Federated Learning — all seem to point in a direction for how we interpret models and facts (and relational modelling/privacy preserving collaboration). What works well for traditional ML can also be achieved in the large-scale, and unstructured data with Generative Language Models, though fairness and responsible use there in continue to be a hot topic. Collectively, the literature suggests that it's not enough to be accurate and interpretability, organizational legitimacy is now the leading edge of EDM–XAI research. Table (2) Comparison and summary of prominent XAI methods in educational context. comparig the main XAI techniques frequently used to contrast their scope, strengths and weaknesses.
These methods represent the trade-offs between interpretability, scalability and prediction performance.

Table 2. Comparison and summary of prominent XAI methods

Technique	Scope in Education	Strengths	Limitations
LIME (Local Interpretable Model-agnostic Explanations)	Explains individual predictions (e.g., why a student flagged at risk).	Simple, model-agnostic and widely adopted.	Unstable with small perturbations and explanations may vary across runs.
SHAP (SHapley Additive exPlanations)	Provides global and local interpretability for student models.	Theoretically grounded, consistent feature importance and widely validated.	Computationally expensive on large datasets and requires expertise.
Attention Mechanisms	Neural models for text/discussion analysis, adaptive learning.	Highlights influential words/sequences and aligns with human intuition.	Interpretability limited to feature weighting; may overemphasize frequent tokens.
Feature Importance (from Random Forests, Gradient Boosting)	Identifies key predictors (e.g., assignments, participation).	Easy to compute and useful for faculty dashboards.	Only provides relative importance; not causal explanations.
Counterfactual Explanations	Hypothetical “what-if” scenarios for interventions.	Helps faculty/students see alternative pathways.	Still rare in education and requires robust causal models.
Model-Specific Interpretable Methods (Decision Trees, Logistic Regression)	Baseline educational analytics and risk models.	Naturally interpretable and easy to communicate.	Lower accuracy compared to advanced methods and less scalable.

As Table (2) shows, no single technique fully addresses the interpretability challenge in education. LIME and SHAP are powerful for surfacing model reasoning but require expertise and resources. Attention mechanisms provide intuitive signals but risk oversimplification. Classical interpretable models, though limited in accuracy, remain attractive for their transparency and ease of communication. Emerging approaches such as counterfactual explanations hold promise for guiding actionable interventions but are still underutilized in higher education. The literature indicates that effective deployment requires a hybrid strategy, combining accuracy with interpretability methods aligned to pedagogical contexts.

3.Methodology

The review followed a structured method of integrative reviews [1] that used the framework by Cooper and his collaborators with elements from PRISMA adapted to the objective of this review [2]. The Cooper framework promoted clear concepts, objectives and inclusion criterion and organized the integration of structured information within themes whilst PRISMA supported transparency in approach to how studies were searched, screened and included. These frameworks were utilized to allow the review to balance between breadth and transparency, which is particularly significant for interdisciplinary topics such as EDM and XAI.

The review serves three primary purposes: (i) to gather empirical and conceptual results in predictive modelling of HE, (ii) to analyze the use of methods for explainability such as SHAP, LIME, and attention-based mechanisms; and (iii) identifying sponsoring/decision factors that drive adoption as well as upcoming trends including Graph Neural Networks, Federated Learning, and Generative Language Models. These aims directed the search strategy and integration that followed.

Search and Selection Procedures

The systematic review was conducted from January 2025 to March 2025 on five main academic databases: The IEEE Xplore, ACM Digital Library, Scopus, Springer Link and Web of science. A rigorous search method was adopted to identify studies at the cross roads of Educational Data Mining (EDM) and Explainable Artificial Intelligence (XAI) in higher education. The keyword combinations were composed of terms like “educational data mining,” “learning analytics,” “explainable artificial intelligence,” “model interpretability” and “higher education”. Searches were executed across titles, abstracts, and keywords to ensure broad coverage of relevant literature. Retrieved results from the five databases were consolidated into a single dataset, and duplicate entries were removed. The titles and abstracts of all records were then screened to identify those meeting the inclusion criteria. To further strengthen coverage, backward reference searches (examining the reference lists of included papers) and forward citation tracking (identifying newer works citing these studies) were conducted. This combined approach ensured that both seminal foundational works and the most recent empirical contributions were included, providing a balanced and comprehensive evidence base for the review.

To be eligible, studies had to: Empirically use EDM techniques in higher education contexts Address and/or integrate interpretability or explainability be literature reviews Journal papers, conference papers or master/PhD thesis published from 2005 till 2025. Excluded were non-peer-reviewed works; studies lacking methodological transparency (e.g., not providing a clear conceptualization of predictive learning analytics, methodology, or methods) and research beyond the scope of higher education.

Sixty-two articles were identified after screening and full-text review for inclusion. These were categorized into six areas: (0) student performance prediction, (1) dropout detection and retention, (2) adaptive & personalized learning, (3) engagement analytics, (4) multimodal & natural language approaches, and 6: ethical/fairness/interpretability.

4. Application of EDM–XAI in Higher Education

The adoption of EDM-XAI systems in HE is determined not just by technical capability, but also institutional readiness and governance, pedagogic alignment and integration. Case studies suggest that predictive models are most effective when integrated into student support processes (e.g., advising dashboards; early warning systems). For example, universities that incorporated SHAP and LIME based explanations into their technology platforms found higher trust from faculty who could verify why credit alerts were triggered before deciding to contact the students [17], [18]. Likewise, interpretable reinforcement learning has been used with adaptive course sequencing to allow faculty to observe why certain topic paths were presented and intervene if needed [19].

Beyond counseling, deployment has reached into course structures and curriculum development. Dashboards, driven by analytics, have been employed to balance the load and re-design tasks and identify bottlenecks in MOOCs and blended learning [20]. Engagement analytics along with explainability techniques moved practice from being reactive (waiting for failures) to proactive approaches (early identification of risk and its interpretation within pedagogy) [21]. In this sense, interpretability does mitigate a form of student illegitimacy by ensuring that the predictions are associated with meaningful and defendable indicators of learning.

However, the adoption between organizations is varied. Institutions that possess a well-developed governance model and auditability framework are more inclined to embed EDM–XAI tools in advising and curriculum design [22]. Another key parameter is teacher training: without proper preparation, interpretable dashboards can stand the danger of being marginalized [23]. Privacy and data protection issues are also key, especially in multi-campus or international collaborations [24]. Although these need to be understood in many different regulatory and cultural situations, they lead us to the recognition that the technical innovation by itself is not enough, successful deployment requires alignment between predictive outputs, institutional practices, pedagogic values and ethical safeguards [25].

4. Results

The results of this review are mapped within six focus domains, which represent multiple dimensions characterizing research in the crossing of EDM and XAI. These issues include student performance prediction, dropout detection and retention, adaptive and personalized learning as well as engagement analytics, multimodal and natural language methods, ethical aspects of interpretability and fairness among others. Both topics combine to generate an all-encompassing perspective of predictive-accuracy versus interpretability and institutional factors interplay that influence utilization and outcomes of EDM–XAI systems in higher education.

5.1 Student Performance Prediction

The prediction of performance has been the earliest and most common use of EDM in higher education, and initial research used statistical models (like logistic regression or decision trees) to predict student grades or probabilities of success [26], [27]. Although these models are interpretable, they generally were only moderately accurate and had difficulty modeling intricate temporal dynamics of student work. Levitt et al (2000) With the advent of ensemble learning and hierarchical deep neural networks, prediction performance augmented considerably to even surpass 80% for course-level forecasting [28], [29]. Those, which tended to be strong predictors, include forum engagement patterns, submission of assignments and navigation behavior in the LMS [30]. Yet, the black-box nature of these models made educators and researchers question their appropriateness for supporting interventions.

To overcome this limitation, interpretability techniques have recently been integrated into performance prediction. SHAP have been well-established in ranking features that are contributing to grade predictions, giving faculty insight into whether predictions were due meaningful learning behaviors or spurious correlations [17], [31]. LIME has been used to create instance-level explanations explaining why one particular student was predicted to fail [18]. These developments have bolstered confidence in institutional trust by making sure that predictive models do not just provide fuzzy scores, they also explain their outputs in ways consistent with pedagogical reasoning. Recent works extend this work in relational and large-scale manners. Performance has also been modeled as a function of student–course and peer–peer interactions even in the relational contexts, for which the GNNs have shown better performance than standard ensembles [32]. Generative Language Models have been applied for the analysis of essays and written assignments to make large scale predictions of academic outcome from unstructured text data [33]. Although these methods broaden the range of predictions possible, they also increase pressure to develop fairness metrics and bias mitigation techniques when working with textual and relational data, which often reflects cultural and linguistic disparities.

5.2 Dropout at Risk Detection and Retention

Predicting dropout has been a dominant topic in EDM, which mirrors the institutional concern about dropout rates and student success. Earlier efforts have used linear regression, naïve Bayes and decision trees on administrative and demographic information to identify potential dropouts [34], [35]. Although these models could be relatively transparent, they were also not very accurate and missed many of the behavioral signals found in digital learning environments. The availability of large learning management system (LMS) and MOOC datasets allowed to fit more advanced models for dropout behavior. In predicting the at-attrition risk two weeks ahead, ensemble methods including random forests and gradient boosting attained performance of accuracies 79% [36] or over 80% [37]. Online participation time, submission delays, inactivity periods and decreasing forum participation were consistently predictive features [38]. But these models were too opaque to actually be adopted in practice—faculty refused to follow any kind of recommendations without knowing why certain students had been flagged as high-risk.

To address this bottleneck specific explainability techniques have been incorporated. SHAP-based models have produced informative feature rankings, demonstrating for instance that frequency of weekly logins is more important than total time online in several websites [17], [39]. LIME has been employed to provide personalized explanations for dropout alerts and offer advisors the rationale behind why one student was classified as at-risk and another wasn't [18]. This interpretability has been demonstrated to enable more trust in instructors and to ultimately be influential for proactive interventions such as intelligent advising and personalized outreach [40].

Provide a statement that what is expected, as stated in the "INTRODUCTION" section can ultimately result in "RESULTS AND DISCUSSION" section, so there is compatibility. Moreover, it can also be added the prospect of the development of research results and application prospects of further studies into the next (based on result and discussion).

More recently, research has extended dropout prediction towards privacy-sensitive and relational approaches. Federated Learning (FL) can support the joint training of dropout models between organizations without exposing raw student data, which reduces concerns over data leakage while maintaining the performance [41]. Networked dropout has been modeled by Graph Neural Networks (GNNs) that capture the relationship between peer engagement and individual dropout [32]. The emergence of these applications signals that dropout detection is shifting from a problem based on accuracy to approaches regarding fairness, privacy and institutional integrity.

5.3 Adaptive Learning and Personalization

Adaptive learning and personalization are central goals of EDM. This approach helps the field make instruction responsive to students' changing needs. Starting with early models such as Bayesian Knowledge Tracing, systems can recommend the next exercise by using probabilistic knowledge state updates [42]. However, interpretability is a serious drawback of DKT. Although it is more accurate than other latent state models, without understanding how adaptive pathways are generated, there are always questions [43]. One increasingly common way to overcome this problem has been to apply various explainability methods. An attention-based neural architecture has been used to make it simple for an instructor to see which interactions or concepts underpin the adaptive recommendations [44]. Interpretable reinforcement learning frameworks have also been investigated. Such systems present course sequencing policies with the rationale; that is, displays decisions for the instructors to approve or edit. These strategies combine computational optimization with human supervision to guarantee that the ability to personalize is pedagogically viable [45].

Moreover, empirical studies confirm the positive impact of adaptive systems on retention and completion extent. For example, the results state the following: "piloted at large scale in MOOCs, showed a 15–20% reduction in drop out, much of it coming from a more intelligible system model by including explanatory dashboards" [46]. Particularly, the results across classrooms display the extent to which interpretable personalization can encourage differentiated instruction ", teachers adjusted suggestions not just based on algorithmic predictions, but also based on their own mental model of each student's motivation and engagement" [47]. Still, while recent work marks an expansion of personalization beyond the cognitive to the affective and behavioral, a number of other investigations call for a truly more interpretative system. Although indications have been introduced for generative language models used to provide adaptive feedback on essays and reflective writing, feedback supplied should include explanations for comments or grades and be acceptable. Regardless, the continuing issue of fairness and bias in personalization is of ongoing relevance [33]. Some argue that adaptive systems have the potential to maintain inequality when it comes to students because they may lack interpretability. The issue must be addressed by a combination of algorithmic transparency and fairness indicators to enable fair personalization among students [48].

5.4 Learning Behavior and Engagement Analytics

Another obvious cluster of EDM research has been understanding and predicting student engagement, an attribute closely related to persistence and achievement. In its classical form, this area concentrated on measuring behavioral indicators, such as forum activity, clickstream patterns, and time-on-task, and then used clustering and regression models to differentiate "low," "medium," and "high" participation levels [49], [50]. Even though this type of inquiry was very informative regarding level setting, its predictive possibilities were rather scarce or superficial since engagement is a polystructural concept that encompasses cognitive, behavioral, and emotional elements.

Some of the more recent studies have used ensemble and deep learning models to detect engagement patterns holistically. For instance, random forests and convolutional neural networks have been applied for classifying active versus disengaged learners from log traces, videos interactions, and facial recognition data in blended classrooms [51], [52]. Although these systems provided higher accuracy, ethical issues related to surveillance and consent are of considerable concern. In this regard, it is essential to incorporate interpretable indicators to engagement analytics to ensure that both faculty and students are aware of the behaviors underpinning their classifications [53].

Explainability techniques significantly facilitate making engagement analytics actionable to the extent that they can support the entire flow of pedagogical practice, from the auditorium to intervention. SHAP-based analysis has unveiled that regularly occurring weekly action patterns are more predictive of sustained engagement than the online time spent in total [17], [39]. This information added context and applicability insights for educators, which informed how and when their involvement would be efficient. LIME was used in the context of specific alerts, informing the educators whether the signal concerned discussions absence, task lose, or rare attendance. Such attention to case explanation makes interventions more legitimate and allows educators to put the alerts in a broader context instead of trusting them due to the incognizance level of the model [18]. Other emerging approaches are also expanding the conceptualization of engagement analytics. For instance, Graph Neural Networks can conceptualize engagement as a relational property using relational models that account for the role of peer participation on an individual's likelihood of participation [32]. Analytically, more recent trends combine multimodal analytics including text, video, and sensor data with explainable models to provide a systems-level understanding of student engagement [54]. The overarching subject is that the value of engagement analytics is not limited to detection but extends to fostering fair and transparent interventions that value learner agency.

5.5 Natural Language Processing and Multimodal EDM

The emergence of digital learning platforms has led to the creation of large amounts of unstructured data, such as essays, forum posts, multimodal interaction traces, and more. Therefore, NLP has developed into a subfield of EDM that has dramatically expanded its potential. This enables the extraction of textual evidence at a large scale to forecast performance, engagement, and conceptual knowledge. Early research used bag-of-words and latent semantic analysis to recognize misunderstandings and quantitatively determine the quality of the reasoning in students' texts [55], [56]. Even though these tools provided strong signs but were unable to scale high-quality conversation and were rarely interpretable to teachers. More powerful language models have rapidly become applied to educational text as a consequence of the broader deployment of deep learning models. This has resulted in a considerable improvement in predictive accuracy in essay scoring, feedback creation, and forum participation analysis [57], [58]. However, they have also raised new challenges related to the model's opacity and potential cultural bias. Attention weights, for example, have been used when training transformer-based models as an explainability technique: transformer-based models to indicate which words or sentences contribute most to the prediction, which allows educators to examine the algorithm's judgment pedagogical relevance [44], [59].

Multimodal EDM also allows predictions beyond text. Audio, video, and sensor-based data have been used in tandem with text to predict cognitive and affective states. The research has concatenated inductive logic programming of speech recognition output, facial expression analysis, and clickstream logs to predict cognitive load and affective engagement [60], [61]. Explainability also remains the top priority in this field, as multimodal analytics can be seen as invasive if the predictions are revealed without cause. SHAP and LIME were also implemented in multimodal pipelines to explain the constructive contributions of distinct signals, like close eye level distance alongside silence in group discussion triggering engagement alerts [18], [62].

Finally, recent work also looks at the usefulness of generative language models in education. Specifically, a growing body of literature investigates the use of GPT models and their variants to provide formative feedback on open-ended writing tasks [33]. While these language models show promise in relation to scaling the provision of personalized feedback, they raise new concerns about transparency, bias, and educational integrity. To address these challenges, interpretability mechanisms should not only explain the reason behind a specific grade or feedback, but also clarify how the model's language processing works among different student demographics [63].

5.6 Ethical, Interpretability, and Fairness Concerns

Ethical concerns, fairness, and interpretability represent an over-arching issue in EDM-XAI research, capturing the increasingly understood fact that predictive accuracy is an inadequate indicator of actual responsible use. What were variously described in the field's early days are transparency in the application of data, the need for student consent, and the threat of LMS-enforced surveillance [64], [65] underscore the contrasting incentives existing between the policy's insistence on student retention and their legal and moral rights because these go against institutional norms. Interpretability is a prominent protection mechanism in this regard. Research has indicated that analytics systems providers are more willing to trust predictions grounded on transparent rationales rather than conceal risk assessments [18], [39]. SHAP's and LIME's contribution has been vital to the acceptance of interpretability institutionally based on the fact that the tools provided an opportunity for stakeholders to ascertain that predictions were driven by meaningful metrics such as the performance of an assignment, as opposed to sensitive attributes including gender or race [17], [66]. Therefore, while interpretability is more significant, it facilitates teaching legitimacy and accountability.

Fairness is a parallel goal that can be achieved at the same time, and recent literature has presented new methods to measure it that go beyond the accuracy of the predictions. For example, one analysis has verified if dropout models disproportionately mistake students who have protected attributes, such as being from a group underrepresented in academia, while another one determined whether adaptive systems systematically struggled with students who were non-native English speakers [67]. Researchers have also suggested additional strategies such as counterfactual explanations and fairness-aware boosting to improve the existing difference, ensuring that the recommended interventions did not make inequality worse [68]. This demonstrates the importance of models that are not only explainable but have been proven to be more fair over different minority groups of students.

In addition, institutional governance determines the extent to which Ethical deployment is possible. Universities that have regulation on audits, accountability, and data privacy can further use their rights and responsibly use EDM-XAI. One available privacy technique, federated learning, enables the design of predictive models while avoiding the invasion of privacy. Various institutions do not have access to raw data, and the model is being trained across the enterprises before being utilized to predict [41]. However, notes of caution, prominent in ongoing debates, underscore the irrelevance of technical “fixes”; as Papenmeier et al. insisted, interpretability and fairness must emerge as pedagogical and institutional values [69].

6 Discussion

The review of 62 studies summarize high maturity and on-going struggles of EDM and XAI integration into higher education. Although performance and dropout risk predictive scoring models consistently manifest high accuracy results [28], [36], the review outcome imply that accuracy alone plays a low impact on adoption; rather, it is how accurately students can interpret the resulting predictions and how institutionally acceptable these interpretations are. SHAP, LIME, and attention-based explanations have placed strong emphasis on learning/teaching the former and making the latter interpret the models’ outputs. It supports systems, and stakeholders’ outputs have transformed into an authority to intervene with minimal doubts [17], [18].

Another insight is that as a field, our focus has evolved profoundly. While the initial articles in Table (1) focused on enhancing correctness using ensemble and deep learning, all three recent papers reflect the importance of fairness, equity, and privacy. While all three recent articles reflect the importance of fairness, equity, and privacy [29], [43]. Three recent papers have emphasized fairness, equity, and privacy, and they have provided action guidelines such as federated learning and counterfactual annotations [67], [70]. This change validates general calls in the literature for supervisory models that assess accuracy and demand transparency, fairness, and long-term consequences to the educational institution [71].

Finally, the results highlight that data modalities are also diversifying. NLP and multimodal EDM broaden the predictive landscape to text, audio, and video, but they also open up new modes of surveillance and risks of bias. The AI model of recognition and SHAP-based multimodal explanations suggest ways to grapple with these risks [59], [62]. But as with all AI, their effectiveness depends on the governance and institutionalized policymaking practices that translate interpretability into accountability [22], [25]. Interpretability proves to be a requirement dependent on the described context. It secures faculty trust during performance prediction, it gives legitimation to predictively negative outreach during dropout detection, it assures pedagogical defensibility in adaptive learning, and it guarantees the interpretability of behaviors during engagement analytics. When gathered, the comments show that interpretability is not reduced to a technical attachment but the very basis of the connection between predictive modeling and values within higher education. The fact that this connection is foreign to many systems and individual institutions negates the potential benefits of EDM–XAI for equal and human-centered education systems without additional policy and cultural intervention [72].

7 Future Research Directios

While EDM–XAI research has significantly matured, multiple directions are underexplored and deserve precise and methodical exploration. Firstly, the application of upcoming models, including Graph Neural Networks, Federated Learning and Generative Language Models, needs more empirical assets in the higher educational environments. The readiness of these models for relational modeling, preserving privacy for collaborative training, and generating assessments at scale should depend on their ability to be robust, fair, and interpretable in a variety of student cohorts [41], [63], [70].

Second, evaluation frameworks should extend from accuracy to include fairness, transparency, and learning outcomes over time. Future work might involve developing standard comparisons based on equity-relevant performance and teaching effectiveness to prevent these disparities from being perpetuated in models [67], [71]. This orientation would place EDM–XAI work in alignment with some of the other critical conversations around responsible AI: trustworthiness and accountability.

Third, cross-cultural and multilingual contexts are poorly understood. The study of EDM-XAI is largely based on the experience of using Learning Analysis Data Model (LADM) instrument in Western higher education. Therefore, the effectiveness of EDM-XAI in non-Western and resource-constrained contexts is largely unknown. Overcoming these problems calls for large-scale international cooperation and innovative methodology that involves analysis of diverse cultures and languages [72]. Furthermore, integrating EDM–XAI more closely with institutional policy and pedagogy should also feature prominently in the research agenda. How and where can interpretability be woven into the fabric of teaching practice, governance structures, accreditation arrangements remain an open question and a set of challenges for future researchers. Drawing the line from the research frontier to the frontier of institutional change will enable the evolution of EDM–XAI from experimental systems to the infrastructures that undergird to the broader educational movement towards equity and transparency.

8 CONCLUSION (10 PT)

In summary, 62 studies situated at the intersection of educational data mining and explainable AI in higher education were organized into the following six domains: performance prediction, dropout detection, adaptive learning, engagement analytics, NLP and multimodal approaches, and ethical concerns. The analysis demonstrate that the interpretability of models is just as important for institutional adoption as their accuracy. Incorporating techniques such as SHAP, LIME, attention mechanisms supported stakeholders' ability to make sense of, validate, and trust algorithmic predictions, effectively translating forecasts into a practical support tool to tackle pressing challenges in education.

The review also suggests a number of emerging or post-2018 approaches, such as Graph Neural Networks, Federated Learning, and Generative Language Models, which increase the predictive range of models while introducing more risks and challenges in terms of fairness, bias, and privacy. The major point for all of these lines of development is that merely introducing technical innovation is not enough: realizing it depends on the extent to which it aligns with models of educational policy-making, organizational governance, and models of ethical safeguards. The increasing focus on fairness and equity acquires the concern that the field is actually moving towards more responsible, human-centered AI.

REFERENCES

[1] H. Cooper, Synthesizing Research: A Guide for Literature Reviews, 2nd ed. Newbury Park, CA: Sage, 1988.
[2] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Medicine, vol. 6, no. 7, pp. 1–6, 2009.
[3] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE Review, vol. 46, no. 5, pp. 30–40, 2011.
[4] D. Gašević, S. Dawson, and G. Siemens, "Let's not forget: Learning analytics at the system level," in Proc. LAK, 2015, pp. 1–5.
[5] R. Ferguson, "Learning analytics: Drivers, developments and challenges," International Journal of Technology Enhanced Learning, vol. 4, no. 5/6, pp. 304–317, 2012.
[6] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601–618, 2010.
[7] R. Baker and P. Inventado, "Educational data mining and learning analytics," in Learning Analytics, Springer, 2014, pp. 61–75.
[8] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," American Behavioral Scientist, vol. 57, no. 10, pp. 1510–1529, 2013.
[9] M. Bienkowski, M. Feng, and B. Means, Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief, Washington, DC: U.S. Department of Education, 2012.
[10] J. Knox, "Learning analytics and neoliberalism: The politics of data and the digital university," Learning, Media and Technology, vol. 42, no. 1, pp. 7–17, 2017.
[11] A. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review," Educational Technology & Society, vol. 17, no. 4, pp. 49–64, 2014.
[12] G. Siemens, "Learning analytics: The emergence of a discipline," American Behavioral Scientist, vol. 57, no. 10, pp. 1380–1400, 2013.
[13] D. Ifenthaler and D. Yau, "Utilising learning analytics to support study success in higher education: A systematic review," Educational Technology Research and Development, vol. 68, pp. 1961–1990, 2020.
[14] M. S. Khairnar, "Artificial intelligence in education: Applications and challenges," International Journal of Artificial Intelligence in Education, vol. 29, no. 4, pp. 475–490, 2019.
[15] T. C. Reeves and L. Hedberg, Interactive Learning Systems Evaluation, Englewood Cliffs, NJ: Educational Technology Publications, 2003.
[16] R. Scheuerman, A. Paul, and R. Baker, "Ethical considerations in learning analytics research," Journal of Learning Analytics, vol. 8, no. 3, pp. 1–14, 2021.
[17] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.
[19] S. Doroudi and E. Brunskill, "Fairer but not fair enough: A critical review of fairness in educational ML," in Proc. LAK, 2019, pp. 1–10.

[20] D. Gašević, S. Dawson, and G. Siemens, “Let’s not forget: Learning analytics at the system level,” in Proc. LAK, 2015, pp. 1–5.
[21] S. Alhadad, “Learning analytics and the imperative for pedagogical alignment,” British Journal of Educational Technology, vol. 51, no. 4, pp. 1066–1078, 2020.
[22] K. Kitto, S. Knight, and S. Buckingham Shum, “Practical ethics for learning analytics,” British Journal of Educational Technology, vol. 52, no. 4, pp. 1649–1668, 2021.
[23] R. Ferguson and D. Clow, “Where is the evidence? A call to action for learning analytics,” in Proc. LAK, 2017, pp. 56–65.
[24] P. Prinsloo and S. Slade, “Student privacy self-management: Implications for learning analytics,” in Proc. LAK, 2016, pp. 89–98.
[25] G. Siemens and P. Long, “Penetrating the fog: Analytics in learning and education,” EDUCAUSE Review, vol. 46, no. 5, pp. 30–40, 2011.
[26] C. Romero, S. Ventura, M. Pechenizkiy, and R. Baker (eds.), Handbook of Educational Data Mining, 2nd ed. Boca Raton, FL: CRC Press, 2019.
[27] P. R. Pintrich, “A motivational science perspective on the role of student motivation in learning and teaching contexts,” Journal of Educational Psychology, vol. 95, no. 4, pp. 667–686, 2003.
[28] Y. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, “Predicting MOOC dropout over weeks using machine learning methods,” in Proc. EMNLP Workshop on Analysis of Large-Scale Social Interaction, 2014, pp. 60–65.
[29] X. Chen, Y. Xie, and C. Heffernan, “Sequential deep learning models for student performance prediction,” International Journal of Artificial Intelligence in Education, vol. 28, no. 4, pp. 385–402, 2018.
[30] J. Joksimović et al., “Predictive power of online learning behavioral data,” Computers & Education, vol. 159, p. 104024, 2020.
[31] S. Lundberg, G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” ArXiv preprint arXiv:1802.03888, 2018.
[32] Y. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. ICLR, 2017, pp. 1–14.
[33] OpenAI, “GPT models for natural language processing tasks,” ArXiv preprint arXiv:2005.14165, 2020.
[34] K. Lakkaraju, J. Leskovec, and J. Kleinberg, “Student dropout prediction in MOOCs using higher-order networks,” in Proc. AAAI, 2016, pp. 1–7.
[35] A. Bowers and J. S. Sprott, “Predicting student persistence using early warning indicators,” Journal of College Student Retention: Research, Theory & Practice, vol. 17, no. 2, pp. 191–215, 2015.
[36] R. Sinha, N. J. Jha, and A. Dutt, “Dropout prediction in MOOCs using temporal and ensemble models,” Educational Data Mining Conference, 2018.
[37] C. Imran, T. Mustafa, and F. Hameed, “Machine learning for dropout detection in e-learning,” Education and Information Technologies, vol. 25, no. 5, pp. 3593–3610, 2020.
[38] Z. Yang, Y. Li, and C. Xing, “Behavioral features in predicting student dropout,” Computers in Human Behavior, vol. 101, pp. 345–356, 2019.
[39] M. Shih and J. Gašević, “Interpretable dropout prediction with SHAP,” Journal of Learning Analytics, vol. 9, no. 2, pp. 45–59, 2022.
[40] R. Cerezo et al., “Early detection of at-risk students in higher education: A case for explainable learning analytics,” Educational Technology Research and Development, vol. 69, pp. 1–23, 2021.
[41] Q. Yang, Y. Liu, T. Chen, and Y. Tong, Federated Machine Learning: Concept and Applications, ACM, 2019.
[42] A. T. Corbett and J. R. Anderson, “Knowledge tracing: Modeling the acquisition of procedural knowledge,” User Modeling and User-Adapted Interaction, vol. 4, no. 4, pp. 253–278, 1994.
[43] C. Piech et al., “Deep knowledge tracing,” in Proc. NIPS, 2015, pp. 505–513.
[44] A. Vaswani et al., “Attention is all you need,” in Proc. NIPS, 2017, pp. 5998–6008.
[45] L. Mandel et al., “Interpretable reinforcement learning for education,” International Conference on Artificial Intelligence in Education, pp. 1–12, 2020.
[46] M. Wen and C. Rosé, “Identifying latent study habits by mining learner behavior patterns in MOOCs,” in Proc. LAK, 2014, pp. 141–150.
[47] C. Brooks and C. Thompson, “Predictive modeling in education: Differentiated instruction and personalization,” Computers & Education, vol. 142, p. 103645, 2019.
[48] S. Verma and J. Rubin, “Fairness definitions explained,” in Proc. ACM FAT, 2018, pp. 1–7.
[49] Z. Papamitsiou and A. Economides, “Measuring student engagement through learning analytics,” Educational Technology & Society, vol. 19, no. 4, pp. 214–227, 2016.
[50] J. Sinclair and E. Kalvala, “Student engagement prediction using regression and clustering,” Educational Data Mining Conference, 2016.
[51] H. Wang et al., “Engagement detection using deep neural networks,” International Conference on Educational Data Mining, pp. 306–313, 2019.
[52] T. Whitehill, D. Serpell, and K. Bartlett, “Measuring student affect using video analytics,” IEEE Transactions on Affective Computing, vol. 11, no. 2, pp. 234–245, 2020.
[53] E. Rienties and D. Toetenel, “Learning design and student engagement: A learning analytics approach,” Computers in Human Behavior, vol. 72, pp. 617–632, 2017.
[54] X. Sun, J. Yu, and J. Luo, “Multimodal learning analytics for engagement,” Computers & Education, vol. 174, p. 104296, 2021.
[55] P. Foltz, D. Laham, and T. Landauer, “Automated essay scoring: Applications to educational technology,” in Proc. AIED, 1999.

[56] C. Leacock and M. Chodorow, “C-rater: Automated scoring of short-answer questions,” Computers and the Humanities, vol. 37, no. 4, pp. 389–405, 2003.
[57] B. Dong et al., “Deep learning for automatic essay scoring,” IEEE Transactions on Learning Technologies, vol. 12, no. 2, pp. 220–232, 2019.
[58] J. Zhang, Y. Song, and X. Wang, “Transformer-based essay scoring,” Educational Data Mining Conference, 2021.
[59] J. Clark et al., “Explaining predictions of transformer-based models,” ACL Workshop on Interpretability and NLP, 2019.
[60] M. A. Calvo and S. D’Mello, “Affect detection from multimodal data in learning contexts,” IEEE Transactions on Affective Computing, vol. 5, no. 2, pp. 126–139, 2014.
[61] K. Bosch, A. Chen, and S. Demmans Epp, “Multimodal learning analytics for affective engagement,” Proceedings of LAK, 2020.
[62] S. Wu, L. Li, and H. Wang, “Explainable multimodal deep learning in education,” Computers in Human Behavior, vol. 129, p. 107126, 2022.
[63] OpenAI, “Language models are few-shot learners,” ArXiv preprint arXiv:2005.14165, 2020.
[64] S. Slade and P. Prinsloo, “Student perspectives on privacy in learning analytics,” British Journal of Educational Technology, vol. 44, no. 3, pp. 1030–1040, 2013.
[65] N. Selwyn, “Surveillance, technology and education: Towards an agenda for research,” Learning, Media and Technology, vol. 44, no. 1, pp. 64–76, 2019.
[66] A. Binns et al., “Ethical considerations for interpretability in AI,” AI & Society, vol. 37, no. 4, pp. 1551–1565, 2022.
[67] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning, Cambridge, MA: MIT Press, 2019.
[68] M. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in Proc. NIPS, 2017, pp. 4066–4076.
[69] J. Floridi and L. Cowls, “A unified framework of five principles for AI in society,” Harvard Data Science Review, vol. 1, no. 1, 2019.
[70] B. Mittelstadt et al., “The ethics of algorithms: Mapping the debate,” Big Data & Society, vol. 3, no. 2, pp. 1–21, 2016.
[71] A. Selbst, A. Boyd, S. Friedler, and S. Venkatasubramanian, “Fairness and abstraction in sociotechnical systems,” in Proc. FAT, 2019, pp. 1–9.
[72] A. Smith and K. Johnson, “Responsible innovation in higher education AI systems: Equity, ethics, and accountability,” Computers & Education, vol. 210, p. 104630, 2025.

BIOGRAPHIES OF AUTHORS

Author 1 picture	Asst. Lecturer Lujain Ali Hussein received her M.Sc. degree in Computer Science from the University of Basra, Department of Computer Science, in 2018. She is currently working as a Programmer at the Ministry of Education, Educational Planning Department, Information and Communications Division. Her research interests include soft computing and data mining. She can be contacted at email: lujain005@gmail.com.