IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Systematic Literature Review on Explainable Learning Analytics and Educational Data Mining

SACHINI GUNASEKARA[1], MIRKA SAARELA[2]
[1]Faculty of Information Technology, BP.O. Box 35, FI-40014, University of Jyvaskyla, Finland (e-mail: savisama@jyu.fi)
[2]Faculty of Information Technology, BP.O. Box 35, FI-40014, University of Jyvaskyla, Finland (e-mail: mirka.saarela@jyu.fi)

Corresponding author: Sachini Gunasekara (e-mail: savisama@jyu.fi).

This work was supported by the Academy of Finland (project no. 356314).

ABSTRACT Artificial intelligence (AI) is gaining traction for its ability to extract valuable information from vast amounts of data, despite concerns about a lack of transparency in decision-making processes. The rise of explainable AI (XAI) has improved human comprehension and handling of AI systems by providing clear explanations for their decisions. This study aims to present a comprehensive overview of recent research on explainability in Learning Analytics (LA) and Educational Data Mining (EDM) from 2009 to July 2025. Through a detailed evaluation of experimental studies and adherence to the PRISMA guidelines, we initially discovered 1,656 studies, which were subsequently narrowed down to a final corpus of 212 studies for in-depth systematic analysis. Six databases were examined using a systematic keyword search: IEEE Xplore Digital Library, ACM Digital Library, Springer, Web of Science, ScienceDirect, and SCOPUS. This methodology provides a comprehensive summary of the present status of explainability studies regarding educational models and provides details about the results, methods, techniques, and efficacy of explainability uses in the realm of education. In particular, researchers have increasingly adopted post-hoc methods—especially SHAP—from 2024 to 2025, signaling a shift toward interpretable, model-agnostic tools. It also addresses the influence of metrics, models, and data types on explainability. Interestingly, only a few papers in our collection included quantified explanations of prediction models that used metrics related to explainability, such as sensitivity and stability. Nevertheless, within our corpus, several articles focused on measures related to model performance and fairness in educational settings. At the end of the review, significant findings about explainability in EDM and LA are summarized, along with a discussion of the research's limitations and future research.

INDEX TERMS Educational data mining, Explainability, Learning analytics, Metrics, Systematic review

I. INTRODUCTION

OVER the past decades, increasingly complex data have transformed the way machine learning is applied across domains. In medicine, for example, early models in the 1980s and 1990s relied primarily on structured, tabular data such as blood pressure, cholesterol, or glucose levels to predict health risks. With the digitization of healthcare and advances in sensing technologies, richer data sources became available, including neuroimaging (MRI, EEG, MEG), high-resolution X-ray scans, and other multimodal data streams. These new modalities demanded more sophisticated non-linear models, often based on deep learning, that could capture complex structures in the data. At the same time, the interpretability of earlier statistical models was lost, raising concerns about transparency and trust in high-stakes contexts like medical decision-making [1]–[4]. A similar dynamic can be observed in education. The earliest predictive models relied on relatively simple tabular data, such as grades, demographics, or attendance records, to estimate outcomes like dropout risk or course performance. With the rise of digital learning environments in the late 2000s, data sources became increasingly diverse and sophisticated, extending to log files, clickstream data, and interaction traces, and more recently to multimodal streams such as video, eye-tracking, physiological signals (e.g., heart rate or EEG), and rich interaction data from online platforms [5]–[14]. These data require powerful nonlinear models capable of handling high dimensionality [15]–[19], but as in medicine, this shift has come at the cost of transparency and interpretability.

Explainability is therefore not a luxury but a necessity. In education, stakeholders ranging from students and teachers to policymakers and EdTech developers need to understand and trust the models guiding decisions. Students are entitled to explanations of why they were denied admission or how they might improve performance; educators may seek insight into why an algorithm flags a student as "at risk"; and policymakers require transparency to ensure fairness and accountability. The legal context reinforces this demand: the European Union's General Data Protection Regulation (GDPR) grants individuals a right to explanation in automated decisions [20], while ethical guidelines issued by the European Commission emphasize accountability, fairness, and transparency as central principles of responsible AI [21].

Explainable artificial intelligence (XAI) has thus emerged as a key research area to address these challenges [22]. Within education, the demand for XAI spans multiple stakeholders, including students, educators, policymakers, parents, school administrators, developers, EdTech companies, and data scientists (Table 1). Trust in AI is the common denominator across these groups [23].

In this review, we focus on how XAI has been applied in the fields of educational data mining (EDM) and learning analytics (LA). Since EDM and LA emerged as distinct fields around 2009 (see Section II-A), our review covers research published from that point up to the present (July 2025). Building on this foundation, our review synthesizes 212 studies, making it the most comprehensive review to date dedicated to explainability in EDM and LA.

Specifically, we address the following research questions::

1) *How do different data types affect the explainability of EDM and LA models?*
2) *Which XAI approaches and explanation techniques are employed across educational contexts?*
3) *Which metrics are used to evaluate explainability, and how are they suited to the educational domain?*

By systematically bringing together educational data types, AI models, explainability techniques, and evaluation practices, our review moves beyond fragmented perspectives to present a unified map of the research landscape. This integrated view enables us to identify methodological gaps, reveal underexplored connections, and highlight opportunities for advancing XAI in EDM and LA.

II. BACKGROUND

A. EDUCATIONAL DATA MINING AND LEARNING ANALYTICS

The 21st century is seeing a major change in the education sector due to technological progress. With the methodical examination of educational data, LA and EDM emerge as important tools that transform teaching and learning. These two closely connected "sister" disciplines emphasize data interpretation and analysis in educational contexts to enhance student learning [24], [25]. The purpose of EDM is to offer methods for analyzing all the types of data collected from educational settings. Its primary goal is to examine different types of data to discover solutions to concerns in educational studies. It emerged from the domains of data mining and artificial intelligence [26].

The International Educational Data Mining Society defined EDM as a "discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students and the settings which they learn in" [27]. A growing amount of data recorded from various computer-based education environments (such as learning management systems and intelligent tutoring systems) is analyzed in most EDM case studies [28].

In comparison, LA is articulated as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [29]. LA works to understand and enhance learning and the settings in which it is conducted with the assistance of stakeholders like educators and administrators. With an emphasis on enhancing learning and teaching via the practical use of data analysis, LA frequently incorporates insights from educational research and pedagogy.

B. WHAT IS EXPLAINABILITY?

Explainability is becoming increasingly crucial as AI systems are used to make judgements that significantly affect people's lives, including those in the criminal justice, healthcare, financial, and educational fields. In the 1970s, the concept of explainability in AI was initially studied in expert systems like DARPA's Explainable Expert Systems (EES), which aimed to provide clear explanations for decisions [30]. As advanced machine learning (ML) methods became available, the emphasis of AI research shifted to improving descriptive and prediction abilities. Different terms in different research groups refer to the fundamental ideas influencing this explainability landscape; as a result, explainability in ML lacks a common and clear definition [31], [32].

Some authors (e.g., [33]) attempt to distinguish between explainability and interpretability—often describing the former as post-hoc methods to make model predictions understandable and the latter as the inherent transparency or simplicity of a model. However, this distinction is not consistently applied across the literature. Many recent studies use the terms interchangeably, such as Miller [34], who equates the two and defines interpretability as "the degree to which an observer can understand the cause of a decision," and Zini et al. [35], who explicitly state that "explainability and interpretability are used interchangeably" in their work. In this review, we follow this broader and pragmatic usage, treating both concepts under the umbrella of explainable AI.

These explanations should also offer practical and actionable information to support decision-making processes

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3643645

TABLE 1: XAI in Education: Who Needs It, Why, and For What Purposes?

Who	Why	For what
Educator	Helps to understand AI model decisions, improving their guidance and support for students, and ensuring transparency and trust in AI systems.	To interpret model predictions, provide explanations for AI-driven decisions, and integrate AI insights into their teaching strategies.
Student	Making AI processes easier to understand, enhancing critical thinking and problem-solving skills.	To identify the reasons behind AI decisions and apply this knowledge to their learning processes.
Model developer	Ensures the models they create are transparent, interpretable, and accountable.	To provide clear explanations for outputs and ensure ethical and transparent AI practices.
Government	Ensures that the application of AI in education fits with legal requirements, ethical standards, and accountability standards.	Ensuring fairness, accountability, and policy compliance through the monitoring and regulation of AI applications, particularly in EDM.
School administrators	Helps in the adoption and application of AI in educational settings by providing information to support decision-making.	Assessing AI tools, making sure they are used responsibly, and enhancing institutional policies in accordance with AI findings.
AI ethics researchers	Ensure that AI systems are created with accountability, transparency, and fairness in consideration.	To evaluate and enhance AI models for ethical compliance and establish guidelines for ethical AI usage in education, especially in EDM.
Parents	Concerned about the possible effects of AI on their children's education, privacy, and opportunities in the future.	To assure fairness and explainability in EDM generated recommendations and to comprehend how AI-driven selections impact children's education.
EdTech companies	It is necessary to create interpretable and user-friendly educational tools driven by AI.	To ensure that EDM models are credible and responsible while creating AI solutions that are clear, explainable, and in line with educational needs.
Educational Data scientists	Require explainability to validate AI-driven insights in education.	To ensure that educators and policymakers can rely on the trustworthy, comprehensible, and accountable outcomes that EDM models produce.
Policymakers	Responsibility for establishing regulations and rules regarding AI in education.	To develop regulations that ensure the responsible use of AI and prevent it from creating bias or inequality in EDM results.

effectively [36]. Besides, these notions produce explanations in various formats, such as rules, numerical data, text, visual information, or combinations thereof. Explanations can be theoretically assessed using a set of notions that can be formalized into metrics. The explainability and performance of a particular ML model are frequently trade-offs [37]. Users may find it challenging to comprehend how highly accurate models—like deep neural networks—make decisions due to their complexity and difficulty in interpretation. These types of models are also referred to as "black boxes". Conversely, simpler models like decision trees are easier to explain but may not achieve the same level of performance. These types of models are also referred to as "white boxes".

C. TAXONOMY OF EXPLAINABILITY METHODS

ML models deemed interpretable because of their basic structure, such as sparse linear models or short decision trees, are referred to as "intrinsically" interpretable models. Interpretation techniques applied after model training are referred to as "post-hoc" interpretable. One post-hoc interpretation technique is the relevance of permutation features. In addition, these techniques can be used with models that are inherently interpretable [32]. These interpretation techniques can be used with models that are inherently interpretable. They can also be categorized as "global" or "local" depending on their explainability scope. A comprehensive insight is offered by modular global explanations, which explain the entire model. For example, feature importance across the whole dataset or global surrogate models are modular global explanations. In contrast, local explanations give insights into a specific prediction, such as why a particular student is predicted to succeed or fail. Techniques like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are examples of local explanations [32], [37].

Additionally, explanation techniques can also be either "model-specific", altered to a particular type of model, or "model-agnostic", applicable to any model. Early explainable approaches were mostly model-specific, including techniques like fuzzy rule algorithms, decision trees, and linear regression. These models were straightforward and easily interpretable. Deep learning and ensemble techniques are examples of non-linear black-box models that evolved as ML progressed to get more accuracy on intricate datasets. For these sophisticated models, post-hoc explainability techniques, such as SHAP values or LIME, are often used to provide insights into their predictions.

D. RELATED REVIEWS ON EXPLAINABILITY IN EDM AND LA

Several reviews related to explainability in EDM and LA have been published, but all have limitations in scope, focus, or time coverage compared to ours. For example, some reviews concentrate on predictive modeling for student performance using relatively small evidence bases, such as 12 studies [38], 14 studies [39], and 62 studies [40]. Others are broader mappings of EDM or LA research but do not focus specifically on explainability, for example, 402 articles on EDM/LA in higher education [41] or 50 studies on dropout prevention [42]. Several systematic reviews also target adjacent topics, including AI in higher education (66 studies [43], 33 studies [44], and 92 studies [45]) or XAI across domains more generally (91 studies [46]), but none of these provide a dedicated synthesis of explainability in EDM and LA.

Two very recent surveys published in 2025 come closer to our scope but remain distinct. Kalita et al. [47] provide a decade-long systematic review of EDM (2013–2023) that briefly mentions explainable AI as one theme, but their purpose is a broad mapping of EDM techniques and applications. Prentzas and Binopoulou [48], in contrast, review XAI exclusively in primary education, covering 23 studies and developing a categorization tailored to elementary settings.

Our review is novel in that it systematically covers all identified explainability-focused work in EDM and LA from 2009 (the emergence of these disciplines) to July 2025. Unlike prior surveys, we did not restrict our scope to a particular educational level or learner population; studies from primary, secondary, and higher education are included. Overall, we adopt an explainability-first lens and synthesize a substantially larger evidence base (212 studies) than most related reviews, systematizing evaluation practices and design choices for explanations and incorporating the most recent work beyond the temporal and topical scope of earlier surveys.

III. METHODOLOGY

A Systematic Literature Review (SLR) tries to integrate research by bringing together factors necessary for comprehending it. Literature reviews serve as a commonly employed method to consolidate existing research findings within a particular field of study. The approach outlined in [49], [50] was applied in this systematic review, which complied with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

A. SYSTEMATIC REVIEW PLANNING

As far as we are aware, no prior study attempted to offer a systematic and full review of the empirical work that is currently available on explainable LA and EDM. Thus, this study aims to compile the findings from the review, systematize, and summarise the empirical work done in the field over time. This study offers an in-depth review that might be helpful to many stakeholders, particularly educators, students, and model developers, in understanding areas such as enhanced learning outcomes, approaches to promote equality in the educational domain, and already explored and implemented approaches.

Search strategies: To ensure that the research included in the review study was of an acceptable quality for finding primary studies, we decided to limit our inclusion criteria to empirical, peer-reviewed works. Six databases were examined using a systematic keyword search: IEEE Xplore Digital Library, ACM Digital Library, Springer, Web of Science, SCOPUS, and ScienceDirect. These databases have been selected on an international scale due to their broad multidisciplinary reach and reliability. They serve as vast knowledge bases that ensure broad coverage of citation indexing and provide researchers with access to high-quality data from scientific publications. The second cycle involved using the snowball approach to search each selected paper's reference section for further related publications.

A checklist was developed to ensure the credibility of the chosen articles. Eight of questions on the list have been altered from Kitchenham et al's [51] SLR guidelines. Additionally, the questions were categorised according to design, conduct, analysis, and conclusion in order to make the validation process easier.

B. SEARCHING STRINGS

Three key terms—explainability, educational data mining, and learning analytics—that must show up in the possibly significant papers are covered by the search string used. The three primary words together could cover a broad range of possible studies at the intersection of explainability and LA or EDM. All of the search terms for articles concerning explainability and EDM/LA were found using the boolean operators "OR" and "AND" in the search.

The following provides an overview of the search strings that were created and modified to meet the advanced search requirements of each database (different spellings and plural are indicated by the wildcard *): ((intelligib*, transparen*, interpretab*, explanation*, explainab*, XAI) AND ("learning analytics" OR LA OR "educational data mining" OR EDM)). The same search string was utilized across all databases within the constraints of each search tool. Among the items that are being searched for are conference papers and journal articles. The SLR followed the analytical process in accordance with PRISMA guidelines [49], [50]. Figure 1 outlines the main phases of our complete systematic review.

After generating the search string, the search queries must be executed in the selected databases, resulting in a total of 1,656 articles. The synthesis included a list of 1,602 articles, which were determined after eliminating duplicate articles and those identified as ineligible by the automation tool.

Selection criteria: After the previous phase, all of the papers' titles, abstracts, and keywords were examined by the reviewers to see if they were relevant for the systematic review. Research on LA, EDM, and explainability was not included in the total of 1,166 papers, and 34 articles were not retrieved, which the researchers omitted at this point. As a result, 402 publications were further examined in the review study. Out of the 68 papers that were selected for further evaluation in the review research, 66 new articles were discovered in the second cycle through searches conducted in the referencing section of each selected paper.

Initially, we considered the four criteria outlined in Table 2 for inclusion and exclusion. Peer-reviewed English-language articles on explainable EDM or LA published in peer-reviewed international journals and conference proceedings met the requirements to be included in the SLR. Articles that did not meet the requirements for the research topics were excluded. In order to ensure the selection of superior research that significantly advances our understanding of

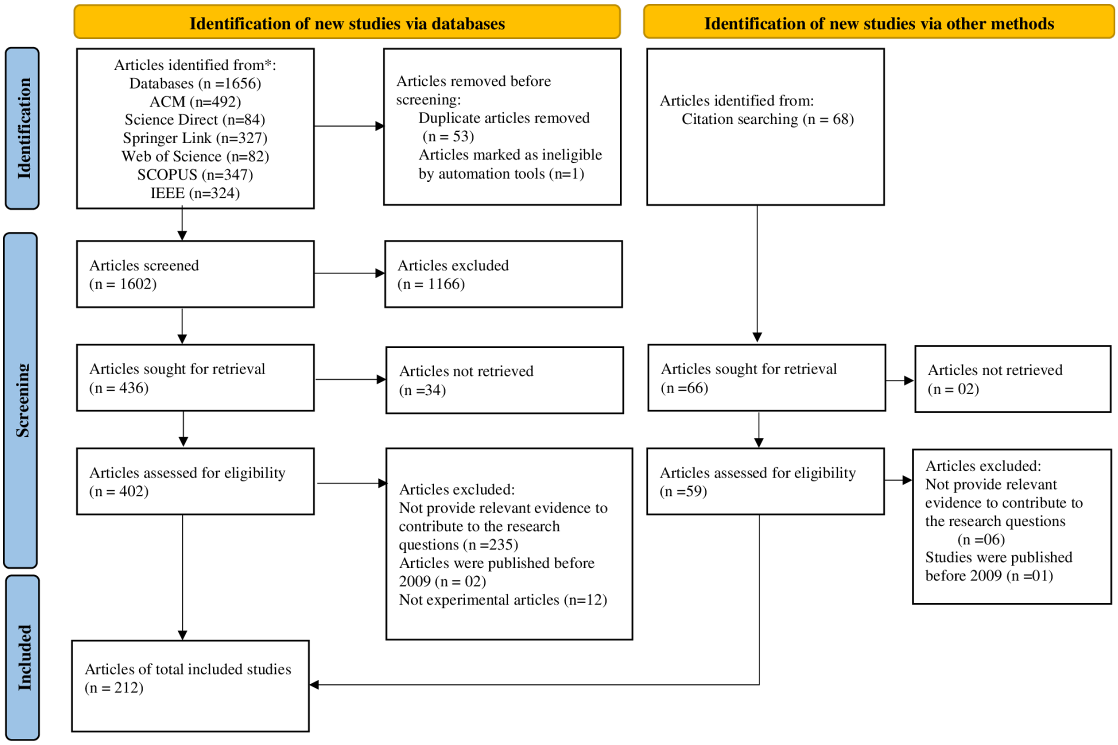
Identification of new studies via databases
Identification of new studies via other methods
Identification
Screening
Included
Articles identified from*: Databases (n =1656) ACM (n=492) Science Direct (n=84) Springer Link (n=327) Web of Science (n=82) SCOPUS (n=347) IEEE (n=324)
Articles removed before screening: Duplicate articles removed (n = 53) Articles marked as ineligible by automation tools (n=1)
Articles identified from: Citation searching (n = 68)
Articles screened (n = 1602)
Articles excluded (n = 1166)
Articles sought for retrieval (n = 436)
Articles not retrieved (n =34)
Articles sought for retrieval (n =66)
Articles not retrieved (n = 02)
Articles assessed for eligibility (n = 402)
Articles excluded: Not provide relevant evidence to contribute to the research questions (n =235) Articles were published before 2009 (n = 02) Not experimental articles (n=12)
Articles assessed for eligibility (n =59)
Articles excluded: Not provide relevant evidence to contribute to the research questions (n =06) Studies were published before 2009 (n =01)
Articles of total included studies (n = 212)

FIGURE 1: Systematic literature review PRISMA chart.

explainability within these fields, the inclusion criteria were especially designed.

Consequently, the review process systematically removed articles that did not meet the requirements of the research topics. We used exclusion criteria to eliminate preprints, duplicate studies, and research without empirical support in order to preserve the integrity of our analysis. We also limited our search to research that was published between 2009 and July 2025, since this is when explainability in EDM and LA has become fairly common. This ensures that the included research is relevant to recent developments in these fields. Moreover, two separate reviewers evaluated study eligibility in order to reduce selection bias, and disagreements were settled by discussion.

C. SYSTEMATIC REVIEW EXECUTION

Throughout the extraction process, 235 papers were removed since neither of the reviewers could discover enough information in them to answer the research questions. Also, two articles that were published before 2009 and nine articles that were not experimental studies were identified and eliminated. As a result, 1,503 articles were excluded based on the exclusion criteria, and 153 articles were determined to be relevant to the EDM and LA models' explainability. Moreover, we looked at 59 more publications that generally discuss the concept of explainability in EDM and LA after applying inclusion and exclusion criteria based on the following references, which were gathered using the snowball approach. Consequently, a final sample of 212 articles (released through recognised channels of publishing [52]) was then incorporated into our review.

TABLE 2: Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Primary studies published in peer-reviewed journals or conference proceedings	Reviews, book chapters, editorials, magazines, grey literature, pre-prints, and duplicates
Published between 2009 and July 2025	Published before 2009
Written in English	Written in languages other than English
Experimental or empirical research relevant to XAI in education	Not relevant to the purpose of this review

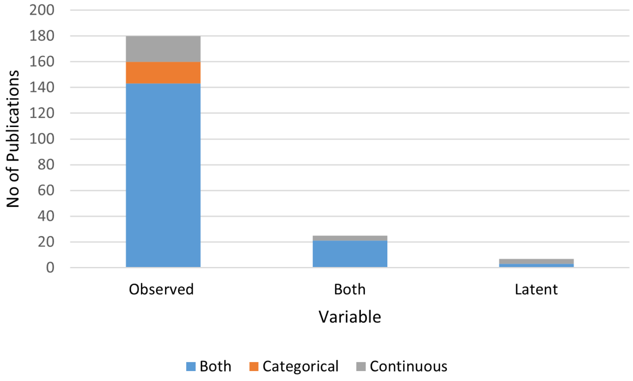

FIGURE 2: The number of articles for each variable type in our dataset during 2009–July 2025. *According to related works, researchers used basic types of data, such as categorical and continuous.*

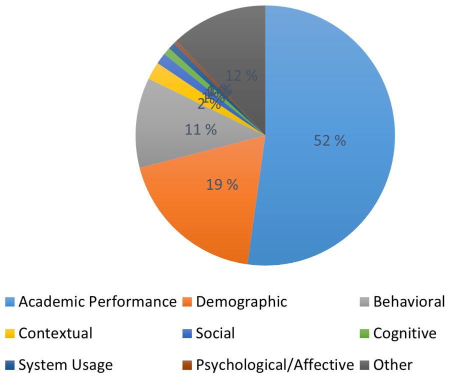

FIGURE 3: Distribution of feature types used in our dataset during 2009–July 2025. *Academic performance and demographic features are the most frequently used, while behavioral, contextual, and affective types are less common.*

IV. RESULTS

A. RELATIONSHIP BETWEEN VARIABLES AND DATA TYPES

Figure 2 displays statistics from previously published studies based on observed and latent variables. This figure outlines the variable types that are commonly observed, the usage of both, and the latent variables that are employed in the educational data analysis works included in this systematic review. Additionally, according to related works, researchers used basic types of data, such as categorical and continuous. It is evident that both categorical and continuous data types in the same dataset have been extensively employed, not only to forecast performance but also to address explainability and fairness concerns.

B. TYPES OF FEATURES

Figure 3 presents the distribution of feature types used in educational data analysis across the reviewed studies. The analysis reveals that academic performance features are by far the most dominant, according for 52.2% of all features, followed by demographic (18.8%) and behavioral (11.2%) features. Less commonly used were cotextual (2.2%), social (1.4%), cognitive (0.9%), system usage (0.9%), and psychological/affective (0.4%) features. This distribution suggests that researchers predominantly rely on direct academic and student background attributes, while behavioral, contextual, and emotional features are used less often.

C. TYPE OF LA MODELS EMPLOYED

Due to these basic needs for AI systems' ethics, fairness, and explainability, there has been a notable increase in research interest in explainability techniques for comprehending and interpreting black-box automated decisions. As a result, several directing model explanation algorithms have been developed, especially in the field of supervised learning models (e.g., [53]–[55]). The statistics of the articles that have been published using the four different types of data-driven AI methods—unsupervised learning, supervised learning, both supervised and unsupervised learning, and semi-supervised learning—are displayed in Figure 4. Algorithms from all four AI methods have been used in datasets from 2009 to July 2025. The most commonly used type of learning model is supervised learning, which is followed by both supervised and unsupervised learning, semi-supervised learning, and then the unsupervised learning model (Figure 4).

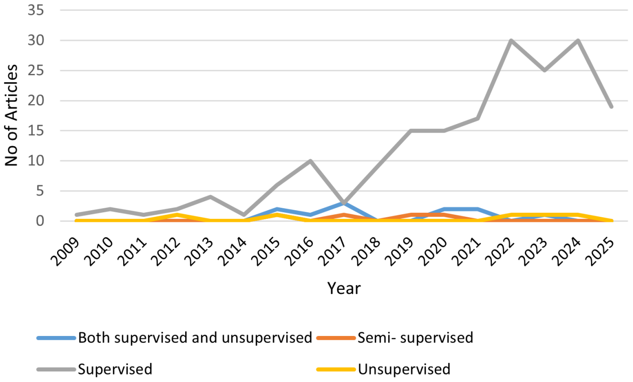

FIGURE 4: AI models with the number of articles in our dataset during 2009–July 2025. *The most commonly used type of learning model is supervised learning, which is followed by both supervised and unsupervised learning, semi-supervised learning, and unsupervised learning model.*

D. RELATIONSHIPS BETWEEN APPLICATIONS AND LA/EDM MODELS

It is important to give readers a summary of the explainability of the EDM and LA models used in various education-related applications. We used data-driven AI methodology to offer statistics on the number of publications in various applications for each of the four main categories of AI techniques in Figure 5. As can be seen, supervised models have been

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3643645

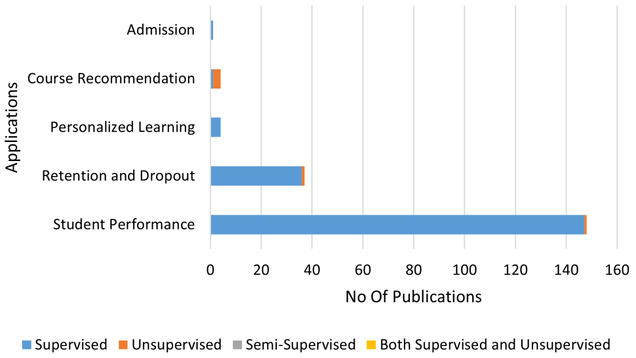

FIGURE 5: Relationships Between Applications and LA models.

used in the majority of applications, indicating that annotated data is available in most cases (e.g., [8], [54]–[57]). In two applications, such as student performance and dropout, semi-supervised models have been used (e.g., [58]–[60]).

E. UTILIZED ALGORITHMS IN PREDICTION MODELS

The classification ML model task is the most often utilized algorithm in our reviewed EDM and LA papers (Figure 6). More precisely, among these algorithms, decision trees, support vector machines, random forest, and naive Bayes algorithms are frequently applied to build the classification models.

V. DISCUSSION

An extensive qualitative analysis of the studies was conducted from the following viewpoints in order to answer the research questions: (i) data types used in models; (ii) LA models, XAI approaches, and XAI/explanation techniques can be employed to ensure explainability; and (iii) metrics used to evaluate the explainability, performance, and fairness of the models.

A. ANALYSIS OF RESULTS REGARDING THE FIRST RESEARCH QUESTION

For educators, it is crucial to consider and comprehend the many variables that influence prediction models in the application area. Additionally, the educational dataset is dynamic and contains a variety of variables for each student, including demographic data, courses taken, behavioral aspects, past performance records, and much more (see Figure 3).

This distribution suggests that while researchers predominantly rely on direct academic indicators and student background attributes, there is still limited integration of richer behavioral, contextual, and emotional features. As such, this is not a simple operation. Based on its format and structure, data can be classified into several types, including latent and observed data.

Recent researchers have shown that LA techniques can access not only demographic information (e.g., family size, gender, age), environmental variables (e.g., urban/rural setting, home location), and familial variables (e.g., family support, parental education level) (e.g., [22], [55], [61]–[67]), but also a number of features targeted towards students, such as “engagement”, “achievement”, “satisfaction”, “participation”, “reflection” and “motivation”. Observed variables, which are usually more straightforward to understand, have the potential to promptly enhance the explainability of models (e.g., [68]–[70]).

Figure 2 illustrates the variable types that are commonly observed. Latent variables related to education, such as satisfaction, learning motivation, and reflection, are challenging to explain (e.g., [71]–[73]). Moreover, as pointed out by Susnjak et al. [74], many latent variables significantly influencing student outcomes are not directly measurable to extract meaningful insights; more processing and interpretation are required to extract insights from that type of data.

The above-mentioned variable types are categorized into continuous and categorical data. Figure 2 makes it clear that categorical and continuous data types have been used extensively in the same dataset. For example, in the study by Mingyu et al. [75], the two primary feature categories found in university big data are numerical features (e.g., exercise score, exercise time, lab score, lab attendance score, age, GPA, etc.) and categorical features (e.g., gender, status, school type, etc.) and in modeling, the two kinds of features are meant to be handled differently.

Since LA and EDM studies usually work with vast collections of data and complicated data types, such as audio, images, text, structured data with many features, and so on, Sahlaoui et al. [76] said that the explainability of ML models is becoming more and more crucial. Romero et al. [77] suggested that the highest accuracy results are achieved with both numerical and categorical type data. They said, “in general, models derived from categorical data are easier to understand than those obtained from numerical data because teachers can more easily interpret categorical values than precise magnitudes and ranges”. In summary, in order to maintain data integrity and validate a model’s ethical decisions, explainability should be used in EDM and LA to identify the important variables that might directly affect any AI application.

B. ANALYSIS OF RESULTS REGARDING THE SECOND RESEARCH QUESTION

Figure 7 summarizes how many papers in our corpus used ante- versus post-hoc methods. The figure clearly shows that most of the EDM/LA models use ante-hoc or white-box methods since the learned model itself is explained. However, this statistic seems to be changing. Figure 8 shows the distribution of ante-hoc, post-hoc, and both model stages over the years.

While ante-hoc techniques initially dominated the field, post-hoc methods have seen a sharp rise, particularly from 2021 onwards. This trend becomes more marked in 2024 and 2025, during which post-hoc approaches—especially

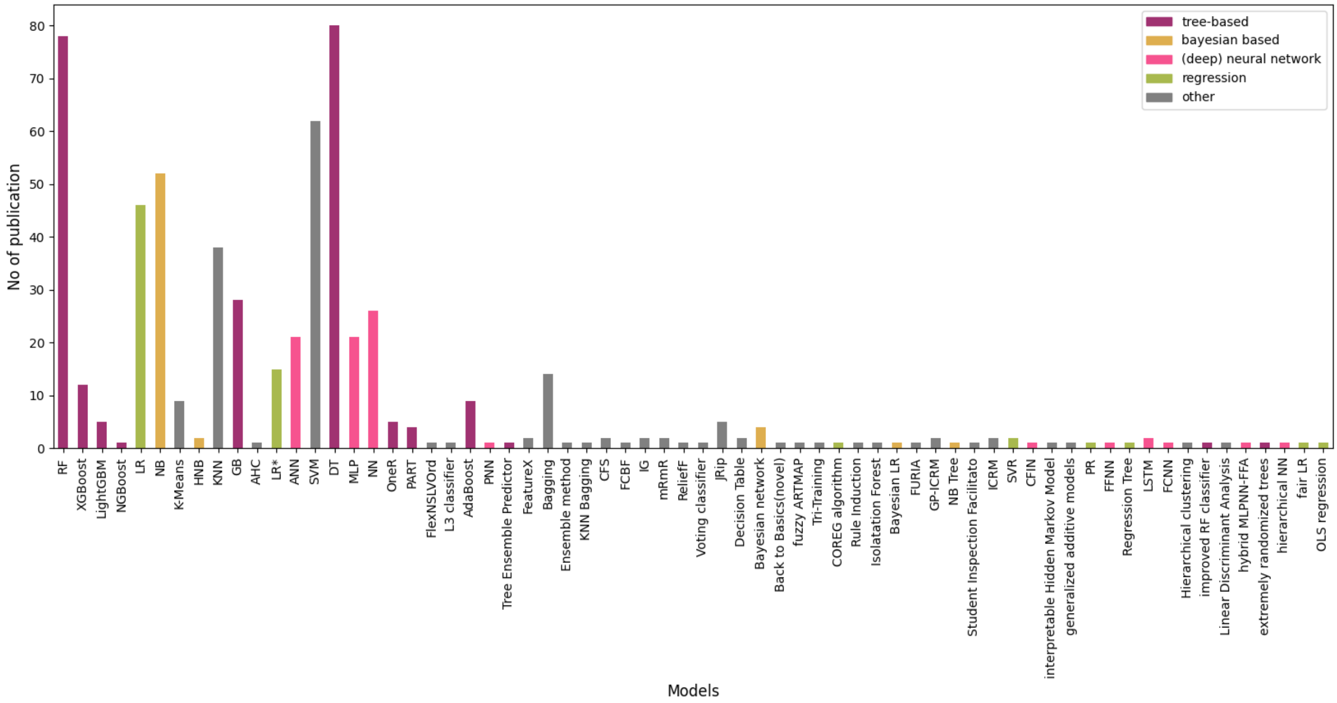

FIGURE 6: Frequency of applied ML algorithms in our dataset during 2009–July 2025.

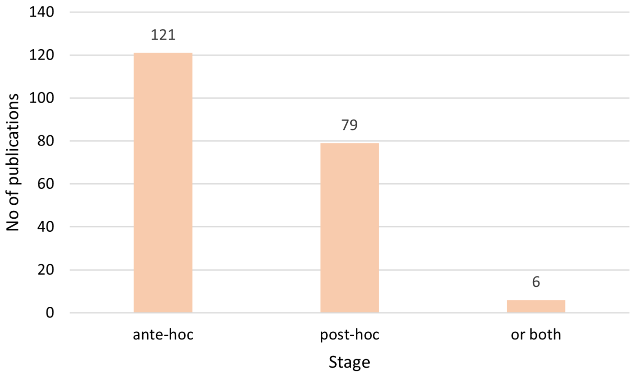

FIGURE 7: Various explainable model stages correlated with the number of articles in our dataset during 2009–July 2025. *Most of the models in this field use the white-box method.*

SHAP (e.g., [78]–[90])— have been increasingly adopted by researchers. Although post-hoc usage has significantly narrowed in recent years. This observation suggests a growing preference for post-hoc explainability techniques, and it is likely that their adoption will continue to expand in upcoming studies.

Most of the examined research studies used tree-based approaches, followed by deep neural networks, Bayesian-based-based, regression, and others. As shown in Figure 6, across the evaluated literature, 67 different models were used. More specifically, tree-based models such as conventional decision trees (e.g., [91]–[93]), random forests (e.g., [94]–[97]), XGBoost (e.g., [74], [96], [98], [99]), and gradient boosting (e.g., [68], [100], [101]) are the most frequently employed models. Other frequently employed models are K-nearest neighbours (e.g., [54], [102]–[104]), naïve bayes (e.g., [65], [105], [106]), support vector machine (e.g., [55], [64], [107]), and logistic regression (e.g., [70], [97], [100]).

Besides, Melo et al. [63] and Duan et al. [108] indicated that decision rules, decision trees, and linear models are some of the most explainable models in the literature. Romero et al. [77] stated "we recommend using decision trees, rule induction, and fuzzy rule algorithms because they are white-box models that provide comprehensible results, allow an interpretation to be made of the model obtained, and can be used for making decisions". Moreover, many researchers have examined the common use of decision tree models, which are most likely due to their accuracy and interpretability combined [105], [109]. Moreover, Ghimire et al. [110] stated that "EML models, such as decision trees or interpretable neural networks, play a crucial role in educational settings by allowing educators to understand and trust the model's predictions and recommendations."

As reported by Rachha et al. [111], there have been a lot of logistic regressions employed in the education domain. As demonstrated by Rangone et al. [112], supervised ML systems may be used to address quite a few educational issues, particularly those involving regression and classification. Additionally, Queiroga et al. [94] indicated that while the random forest is conceived of as a "black-box" model, it is simple to convert its models into ones that can be interpreted. Conversely, Matetic et al. [6] found that while decision tree

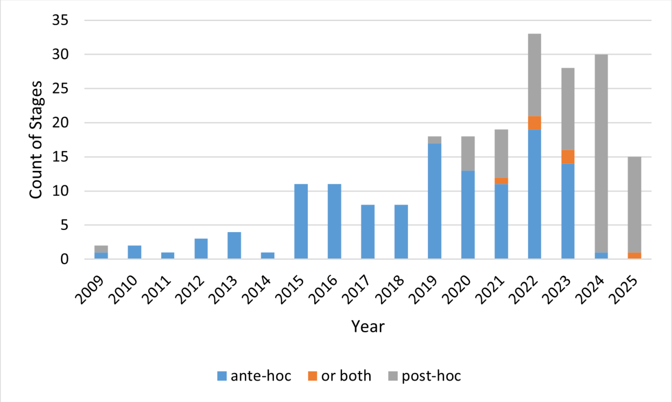

FIGURE 8: Distribution of different stages of explainable model with published year in our dataset during 2009–July 2025. *Post-hoc model adoption will likely expand over the next few years.*

models are easy to comprehend, their prediction performance is not up to level with the most effective supervised learning methods, like ANN.

The following are some examples of regression techniques: CO-training REGressors (COREG) (e.g., [60]), ordinary least squares regression (e.g., [113]), polynomial regression (e.g., [114]), support vector regression(e.g., [115], [116]), and linear regression (e.g., [107], [117]–[120]).

Deep learning approaches contain artificial neural network (e.g., [121]–[123]), multi-layer perceptron (e.g., [63], [94], [102], [105]),neural network (e.g., [64], [103], [124]), and long-short term memory (e.g., [115], [125]). Although there are some promising early applications of deep neural network learning in EDM and LA, Waheed et al. [7] pointed out that the volume of research on the subject is typically limited. Matetic et al. [6] noted that because ANN models are black-box models that do not support generating a predicted interpretation, they are not widely used in the field of EDM. Besides, Swamy et al. [126] highlighted "explainability in educational deep learning models can lead to better-informed personalised interventions, curriculum personalisation, and informed course design."

Although several basic ML techniques, such as linear regression, decision trees, and decision rules are easily comprehensible, Pereira et al. [127] pointed out that these techniques frequently lack predictive ability. This is most likely because complex datasets commonly provide greater accuracy for non-linear black-box models like ensembles and deep learning.

Clustering approaches includes K-Means clustering (e.g., [109], [128], [129]), hierarchical clustering (e.g., [130]) and agglomerative hierarchical clustering (e.g., [62]).

XAI Techniques: On the basis of the architecture of existing AI systems, researchers have investigated a number of methods to produce model outputs that are comprehensible and interpretable. Initially, this was achieved by using intrinsically explainable methods, such as Bayesian networks, decision trees, and linear regression. To extend interpretation techniques to all ML models, more recently, post-hoc approaches have been developed. These techniques do not rely on intrinsically explainable models, but can be applied after any model to explain the model's decisions in hindsight. The two most popular post-hoc explainability techniques are SHAP and LIME [131], [132].

SHAP, a game theory-based approach, has been applied across several domains to assess the relative value of features, offering both local and global explanations. Unlike Gini coefficients, which are limited to tree-based algorithms for global explanation, SHAP can be applied to any model [7], [60], [112], [133], [134]. SHAP assigns each feature a Shapley value, indicating the model's performance without that feature. However, in educational contexts, SHAP faces several limitations. First, it is computationally expensive, making frequent or real-time deployment challenging for large-scale student datasets [135]. Second, SHAP assumes feature independence, which is often violated in educational data where features tend to be interrelated, potentially leading to misleading interpretations [136]. Third, SHAP explanations can be harder to interpret in complex models or large datasets, making it challenging for educators and students to clearly understand them [137]. Finally, SHAP explanations may vary in reliability depending on the model type, which can limit their consistent use in educational settings [138]. These challenges highlight the need for careful consideration when applying SHAP in education and suggest opportunities for methodological refinement to improve explainability in this domain.

A local explanation framework called LIME applies agnostic techniques to the model, which require model execution for each explanation and provide the decision route (e.g., [6], [63], [66], [93], [139]). Its transparency and flexibility allow for simple comprehension of the estimation of feature importances and their significance in the final prediction. Also, it is an efficient way of providing explanations. LIME's drawback is that it does not offer a comprehensive understanding of the model's behaviour or insights into how the model decides over the whole dataset. It must thus be used together with other methods to ensure the explainability and transparency of the model. Tousside et al. [140] further noted that while LIME has been widely applied to image and text datasets, its application in EDM remains limited, partly because educational data are often contextual, temporal, and interdependent, which can reduce the reliability of perturbation-based explanations.

In explainable AI, counterfactual instances are a technique that illustrates "what-if" scenarios—how altering specific input features would change a model's prediction. This approach's primary benefit is that it helps users comprehend the causal impact of particular features and offers clear, actionable insights. It is particularly helpful for individualized decision- making and feedback. Counterfactuals, on the other hand, are difficult to create, have the potential to oversimplify intricate models, and usually only provide one explanation at

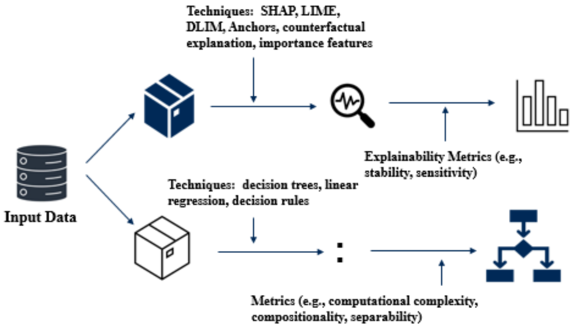
Techniques: SHAP, LIME, DLIM, Anchors, counterfactual explanation, importance features
Explainability Metrics (e.g., stability, sensitivity)
Input Data
Techniques: decision trees, linear regression, decision rules
Metrics (e.g., computational complexity, compositionality, separability)

FIGURE 9: Depending on if the LA/EDM model is black- or white-box, different explainability metrics can be used. For a black box model, we typically need additional tools (visualized by the magnifying glass) to look into the model's internal and retrospectively explain its decision-making process. Thus, explanation metrics are computed from the post-hoc explanations but may also use quantities of the original model (such as its predictions). For white box models, we can get the explanations directly from the model and also compute the explanation metrics using only quantities from the original model and its explanations.

a time rather than providing a comprehensive view of model behaviour (e.g., [56], [74]).

C. ANALYSIS OF RESULTS REGARDING THE THIRD RESEARCH QUESTION

Figure 9 provides a summary of the predictive model that includes the three important metrics (explainability, performance, and fairness) that need to be addressed while developing and implementing educational models. Reaching this balance is essential for building trust among all stakeholders, including students, educators, and model developers. It makes sure that various parties' needs and interests are sufficiently addressed of, which improves learning results and improves fairness in the educational system. By maintaining a harmonious integration of these metrics, the model not only performs effectively but also remains transparent and unbiased, fostering an inclusive and supportive educational environment.

As mentioned above, explainability explains the how and why of an AI model's decision, but it takes out a lot of information, indicating how much trust may be placed in that decision. Therefore, it is still unclear how explainability should be measured [126], [141]. A technique that is frequently used to measure the explainability of ML models is the analysis of the models' observed explainability using qualitative approaches on human subjects. However, these approaches might not be practicable or appropriate given the variation in application domains and prediction types. Consequently, as quantitative methods allow for the objective comparison of various ML models without the involvement of human subjects, they should be employed to assess explainability.

Figure 10 provides the distribution of explainability metrics, presenting only a few studies that have been done on explainability in applications (e.g., student performance, retention, and dropout). These studies relate to the development or utilization of specialized tools that employ more interpretable techniques, such as SHAP, anchors, and LIME. With a focus on explainability, the following part offers a comprehensive overview of a few articles in the field of LA.

Baranyi et al. investigated the early detection of students who are at risk of dropping out [142]. Budapes University's preenrollment educational data was employed in a deep neural network learning model. They employed the SHAP values and permutation importance techniques to analyse and comprehend the model's decisions.

Alwarthan et al. [143] showed that random forest, a tree-based technique, outperformed other data mining approaches in detecting students at risk early on. The global surrogate model, SHAP, LIME, and other XAI approaches were used in this study to increase the predictability of the prediction model. These methods were helpful in understanding the complex prediction models, explaining the forecast results, and pinpointing the reasons behind failures.

A timely evaluation of students' academic performance and suitable intervention require an interpretable model analysis. Each of the seven features (learning progression for objective practice questions, in-class discussion participation, learning progress, number of posts, number of replies, text vector, and learning progression for subjective practice questions) in the unified feature vector's contribution to the random forest classifier was evaluated by Qu et al. [144] using the SHAP approach. They discovered that text features had a greater impact on the classification model. This demonstrated even further how combining text features may enhance classification model performance.

Pereira et al. [127] used the SHAP method to identify the successful and ineffective behaviours that explain prediction paths, collective behaviors, and individual predictions in order to predict student performance instantly.

Sargsya et al. [145] used the LIME approach to extract academic performance indicators from the longitudinal dataset made available for the Fragile Families Challenge (FFC) competition, including gender, financial class, and family background. LIME was selectively applied to student occurrences that the random forest classifier identified as "top" or "low" in order to build feature weights particular to each subject. Following that, the results were clustered using the k-means approach. They draw attention to how deeply insightful the suggested pipeline may be. For example, test results correlated with the features that appeared to have a significant effect on future academic performance in two of the first clusters they discovered.

Using the LIME technique, Vultureanu-Albişi et al. [66] not only compared the effectiveness of various ensemble tree-based models on two datasets but also interpreted the models that were produced and provided significant explanations for

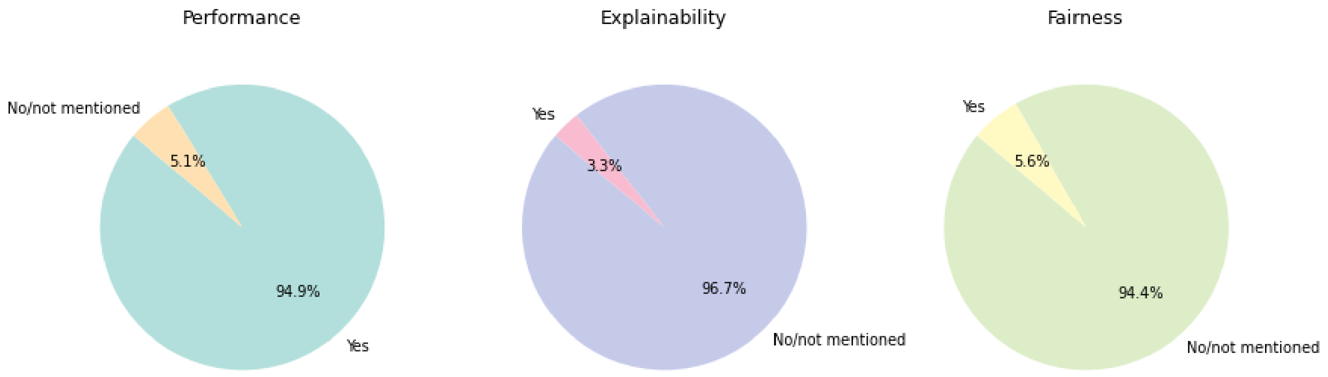

FIGURE 10: Distribution of metric usage in our corpus of reviewed studies (2009–July 2025): The pie charts illustrate the proportion of studies that reported performance, explainability, and fairness metrics. While nearly all studies included performance metrics (94.9%), only a small fraction reported metrics related to explainability (3.3%) or fairness (5.6%).

the predictions made to enhance students' performance. The number of hands raised, the number of prior failures, the number of absent days, and the student's travel time to school were recorded as variables of the explanation features.

Implementing a novel explainable framework, Karacapilidis et al. [53] provided results that are accurate, reliable, and comprehensible when predicting students' performance. To develop an effective prediction model, it utilizes the recently published NGBoost algorithm together with the LIME and SHAP techniques for providing local and global explanations, respectively. The numerical studies show that NGBoost has the best overall classification performance based on all performance metrics.

Several approaches for estimating students' time-on-task based on LMS trace data and widely used analytical models were examined in the research presented by Kovanovic et al. [117]. Regarding how effectively they take into consideration student differences in the five performance measures, they contrasted time on task and count measures. Overall, the results showed that time-on-task estimations performed better than count data.

With the use of prediction tools and techniques, Veluri et al. [55] introduced a unique ML model for conducting EDM and LA on university student data to determine which variables impact college students' selection of their major. It was possible to predict the attitudes, behavior, and performance of students. Support vector machines, naive bayes, and artificial neural networks are a few of the ML techniques under study. Regarding the student data prediction, the experimental results suggest the artificial neural network technique produces more accurate outcomes.

To facilitate predicting students' performance in four ranking classes (withdrawn $<$ fail $<$ pass $<$ distinction), Gamez-Granados et al. [105] proposed to use an optimized ordinal classification algorithm, FlexNSLVOrd. An experimental study with the Open University Learning Analytics Dataset (OULAD) datasets is carried out by developing highly understandable models. Based on statistical analysis, the findings demonstrate that FlexNSLVOrd performs better than other models like random forest, J48, and PART.

Using a blended learning model, Cagliero et al. [102] offered an explainable LA method to analyze it. The use of associative classifiers at different times and a review of the model characteristics that resulted in the selection of pass or fail success rates were suggested. Associative models showed the same level of efficiency as the top-performing classifiers when applied to real student data obtained from the provided dataset. Additionally, the models enable domain experts to identify successful and at-risk students.

A personalized knowledge tracing framework with an attention-based mechanism was presented by Zhao et al. [146] to highlight the key elements from the learner's interactions and features in the dataset that was employed. In addition to improving prediction performance, the mechanism recommended future learning activities and offered explanations for the existing state of knowledge acquisition. Thus, they noticed that the learner's attributes were the main predictors of the model in the early stages, while the learning activity attributes were more significant in the latter stages.

The above-mentioned research presents prediction models for explainability, personalized learning, and student performance; however, none of them take into consideration the quantification and evaluation of explainability [69], [95], [141], [147]. Therefore, some metrics for evaluating explainability systems have been presented in the last few years. The literature listed below offers quantifiable prediction models that quantify explainability metrics, including the *stability*, *sensitivity*, and increasing fidelity.

The stability of the models has a major effect on the quality of ML applications when they are applied to new and unknown data. This stability has traditionally been evaluated using distributional modifications to data and model predictions, shifts in model performance, and changes to model parameters. Predictive model stability has been built on model coefficients and performance ratings, especially in the LA domain.

Tiukhova et al. [148] examined the stability metrics for assessing the explainability of the proposed model in order to predict student performance. They examined the stability of significant student performance indicators using SHAP, a widely recognised feature attribution approach. They also looked at the feature impacts and feature importance rankings of the factors that demonstrate stability, which are crucial for advising students. Out of eight common ML algorithms, the NB method was determined to be the most stable based on agreement metrics in this study.

Using both logistic regression and neural networks as AI models, Li et al. [149] predicted high-risk students. Following that, students' learning outcomes are assessed, and explanations are provided using explanation AI approaches such as SHAP and LIME. To evaluate these explanations' accuracy and stability, six evaluation indicators are used. Quantifying the greatest change in the explanation produced by XAI with respect to the predicted input and result probabilities, respectively, are Relative Input Stability (RIS) and Relative Output Stability (ROS). The explanation's stability is assessed using these parameters. Additionally, post-hoc models such as SHAP or LIME generate explanations that are evaluated for consistency and fidelity using other quantified indicators such as Feature Agreement (FA) and Rank Agreement (RA).

Nnadi et al. [150] mentioned a variety of XAI approaches to understand student adaptability prediction models in a comprehensive way, including LIME, anchors, SHAP, Accumulated Local Effects (ALE), and counterfactual. LIME and Anchors elucidate local, instance-specific explanations; counterfactuals show how sensitive predictions are to feature perturbations; and SHAP and ALE provide an overall view of feature contributions.

Sohail et al. [141] presented a new framework based on hybrid statistical fuzzy theory that addressed the adaptability and explainability limitations of the ML models currently in use in the educational domain. It also provided explainability in the form of rules outlining the reasoning behind a certain outcome. The fuzzy index, a measure of explainability, demonstrates how highly interpretable the model is. A fuzzy system explainability metric called the fuzzy index is introduced, which draws inspiration from the Nauck index. The fuzzy index is assessed using a hierarchical fuzzy system that has six input variables and one output variable.

Li et al. [114] evaluated the performance and explainability of 17 ML models by comparing them to the test set using four widely used metrics and by using "statistical principles" to assess the models' explainability. They provide detailed textual information regarding the explainability of different models, even if it is unclear which principles were used to forecast the models.

Beyond stability and sensitivity, fidelity has emerged as a crucial explainability metrics in education. Fidelity measures how accurately an explanation reflects the model's true decision process. High fidelity ensures explanations truly represent the model's behavior, while low fidelity can mislead users. It can be evaluated locally (per instance) or globally (across data). Gunasekara and Saarela [151] highlighted that fidelity is a key metric for evaluating explanations in educational prediction tasks. In their study, ANNs combined with SHAP and LIME achieved higher fidelity than DTs, indicating that explanations more closely reflected the model's actual reasoning. They further noted that while high fidelity can improve trust in educational AI, complex models with high fidelity may still pose challenges for interpretability, making it essential to balance both dimensions when designing explainable systems for learning contexts.

Applying these metrics in education faces challenges such as data heterogeneity, high computational cost, and lack of contextual relevance for educators [23]. Moreover, there is no standardized framework for evaluating explainability in LA. We recommend developing domain-specific benchmarks, combining global and local metrics, and incorporating user evaluations to ensure explanations are both technically faithful and pedagogically meaningful.

However, Alamriet al. [147] discovered that the primary studies included in their SLR did not use any evaluation metrics to assess the explainability of the model. Additionally, according to Swamy et al. [126], there is currently a lack of standard metrics within the community for quantifying explainability methods. Sohail et al. [141] verified this and stated that it is not easy to conduct any sort of comparison or evaluation because none of the methods evaluate the explainability of student prediction systems.

Several major studies included in this systematic literature analysis revealed that most researchers did not use any assessment metrics to determine how explainable the model was [95], [126], [141], [147]. Specifically, we noted that explainability is not properly addressed in LA and EDM studies, according to multiple authors [124], [140], [147]. In the educational domain, where human-AI system interaction is a primary issue, Alonso et al. [22] and Tousside et al. [140] noted that in order to close this gap, new XAI systems must be developed that can convey data analysis findings in a way that is understandable to individuals.

Besides, in the domain of education, it is frequently preferable to achieve explainable results rather than simply highly accurate predictions [68]. In support of this, Cagliero et al. [102] noted that although a lot of data mining and ML techniques have been created to provide accurate predictions from past learner data, the most effective models frequently have limitations with explainability [102]. Likewise, as stated in several papers, there is a trade-off between explainability and model performance [75], [95], [127], [152]. To increase confidence and trust in the model's adoption, accuracy and explainability are equally important.

Explainable ML models may effectively give educators as well as learners insightful knowledge about the many programming behaviors and approaches to problem-solving that students use, which may be related to good or poor performance [64], [153]–[155]. Fuzzy logic and IF-THEN classification have been demonstrated to be effective in designing efficient tools for explanations when taking into consideration the prediction algorithm and XAI/explanation approaches utilized to evaluate explainability in education.

Also, because learning behavior is inherently unpredictable, it has been widely applied in the educational field [11], [124]. In intrinsically explainable models, feature selection is another common XAI/explanation technique. It improves both prediction performance and the model's explainability [69]. Although post-hoc model adoption will expand over the next few years, they frequently offer little insight concerning how the models arrived at their predictions, making it challenging for educators to figure out why a particular student may be performing well or poorly. Matetic and colleagues [6], for instance, emphasized that the artificial neural network model is a black-box model type that is unable to produce an interpretation of the expected outcome. ANN is not commonly utilized in EDM and LA because of this limitation. In Table 3, we list highlight statements on how and what researchers have pointed out regarding explainability in EDM or LA.

D. FAIRNESS IN EDM AND LA

In the domain of educational AI models, explainability and fairness are interconnected since decisions have a big impact on students' academic opportunities and their future success. Fairness is defined as preserving the reasonable and equitable distribution of benefits and costs as well as the freedom of people and groups from discrimination, unfair bias, and social stigma [162]. While fairness in LA has recently received notice that ML's fairness has garnered a lot of attention over a decade.

Furthermore, it's still difficult to decide which unfairness mitigation method or metric to apply in a given situation [96], [163] and by confirming this statement, Fenu et al. [164] mentioned "the rapidly growing area of fairness in educational AI presents many challenges and needs". We have clustered a limited amount of research works that have focused on fairness metrics in applications, as illustrated in Figure 10.

XAI is essential in addressing these challenges since it makes the educational model's decision-making process general, which helps stakeholders in identifying and comprehending possible biases or unfair results. Predictive models, for example, may unintentionally favour or disadvantage particular demographic groups—such as students based on economic status, gender, or ethnicity—leading to unfair learning experiences [8], [94], [95], [113], [155], [165], [166]. Dropout prediction is a well-known example, where research has demonstrated that models trained on historical enrolment and demographic data consistently over-predict dropout for under-represented groups, resulting in inefficient interventions [167]. Similarly, it has been demonstrated that using institutional records for student progress monitoring introduces more bias than using learning activity or assessment data, especially unfairly students from non-traditional backgrounds or students with disabilities [168]. Automated grading systems are another important example. Error rates vary significantly across languages and writing styles, according to research on multilingual and short-answer grading. This raises issues about fairness for students whose linguistic backgrounds differ from the training data [169]. Large-scale assessments have also revealed systematic disadvantages: for instance, automated essay scoring in language proficiency tests such as TOEFL has been found to assign lower scores to certain language groups compared to human raters, highlighting how algorithmic scoring can unintentionally reinforce inequalities [170]. Finally, issues of data preprocessing—such as how missing values are imputed—have been shown to affect techniques producing unequal impacts across student subgroups. Explainability techniques can provide insight into the causes of these biases, the features that influence predictions, and whether or not sensitive attributes are inappropriately affecting results.

1) Fairness metrics

There are particular features that might lead to bias in the predictive models, leading to incorrect assumptions throughout the learning process and in the models' output [94]. Several studies reveal that specific demographic groups (such as gender or ethnicity) of students may be the focus of bias in LA models [8], [94], [95], [155], [165], [166]. Several types of metrics have been proposed to quantify fairness. One of the metrics is demographic parity. For all groups having sensitive attributes (like race), it specifies that the sensitive attribute must be independent of the forecast and that the overall probability of a positive forecast of an occurrence must be the same (e.g., [94], [96], [155]). Equal opportunity (e.g., [171] and equalised odds (e.g., [163], [172]) are the other two metrics.

However, using fairness metrics alone isn't enough unless we can clearly understand how the AI makes decisions. Thus, integrating explainability helps us better identify and mitigate bias, promotes ethical AI development, and builds trust among educators, students, and decision-makers. Figure 11 displays the metrics used in previous research works of our SLR to evaluate fairness.

E. PERFORMANCE IN EDM AND LA

Model performance evaluation, which measures the model's precision in predicting the performance of different applications using the relevant dataset, was conducted using a variety of metrics. The performance metrics distribution for each of the studies selected was shown in Figure 10.

1) Performance metrics

Predictive classification accuracy (such as f-measure, precision, accuracy, and recall), error measurement (such as Mean Square Error (MSE), Ordinal Mean Absolute Error (OMAE),

TABLE 3: The statements were listed on how and what researchers have pointed out regarding explainability in EDM or LA in our dataset during 2009–July 2025.

Reference	Explanation of Explainability in EDM and LA	Reference	Explanation of Explainability in EDM and LA
[141]	"Moreover, the literature on quantification or measurement of explainability of AI models used for learning analytics is even more limited."	[155]	" Without a basic understanding of AI, students may struggle to respond appropriately to questions about 'explainability' or 'fairness' (data bias)."
[154]	"In education, the prediction models obtained should be comprehensible/interpretable for instructors in order that these models could be used directly for decision making and provide an explanation for the classification."	[156]	"The usage of the so-called explainable Artificial Intelligence, even if desirable, is still limited, especially whenever we consider educational datasets","Fuzzy logic has been proven to be useful to design effective tools for explanations.Moreover, it has been widely used in the educational domain since the learning behavior is inherently uncertain."
[6]	"The ANN model is the black-box model type and does not support generating the predicted outcome interpretation. Because of this deficiency, ANN is not frequently used in EDM and LA."	[64]	"Despite calls to increase the focus on explainability and interpretability in EDM and, in particular, student success prediction, so that it becomes useful for personalized intervention systems, only few efforts have been undertaken in that direction so far."
[22]	"In the educational field, where the interaction between humans and AI systems is a main concern, there is a need of developing new XAI systems,that are able to communicate, in a human understandable way, the data analysis results."	[68]	"To some extent, the pursuit of explainable outcomes in education context is preferred contrast to prediction results with high accuracy."
[102]	"Although several machine learning and data mining solutions have been proposed to learn accurate predictors from past learner-related data, the interpretability and explainability of the best performing models is often limited."	[140]	"In particular, we noticed that explainability is not well addressed in EDM although it would be of great importance in satisfying regulations about artificial intelligent systems as well as making EDM more trustworthy."
[126]	"The community does not yet have a set of standard metrics for evaluating explainability methods."	[111]	"The increasing reliance on AI in education raises important questions about explainability and accountability"
[147]	"None of the primary studies that were included in this systematic literature review has utilized any evaluation metric to assess the explainability of the model.State of the art explainable models have been utilized in many domains. However, they are yet to be explored in educational data mining."	[153]	"Explainable Machine Learning models can effectively help students and instructors gain insights into students' different programming behaviours and problem-solving strategies that can lead to good or poor performance."
[153]	"Explainable Machine Learning models can effectively help students and instructors gain insights into students' different programming behaviours and problem-solving strategies that can lead to good or poor performance."	[155]	" Without a basic understanding of AI, students may struggle to respond appropriately to questions about 'explainability' or 'fairness' (data bias)."
[157]	"The rising adoption of learning analytics and academic performance prediction technologies in higher education highlights the urgent need for transparency and explainability."	[158]	"There remains no universally accepted metric to assess explanation quality, as explainability is context-dependent and subjective. This lack of standardization complicates comparing methods across applications."
[159]	"This implies that the field of educational data mining seeks an interpretable artificial intelligence model. Such models prioritize transparency, interpretability, accountability, and a clear explanation of model decisions and outcomes."	[110]	"By using explainable AI (xAI) models, educational institutions can promote accountability and equity in their support strategies, ensuring that interventions are fairly distributed and that all students receive assistance based on transparent, evidence-based criteria."
[160]	"In Learning Analytics, a key challenge is designing feedback indicators that are both interpretable and actionable, enabling students to adjust their behaviors effectively to improve learning outcomes."	[161]	"However, beyond explain ability, it is essential that indicators are also action able: that is, they enable students to take concrete steps to improve their learning."
[97]	"The application of AI in education is significant and desirable, it is important to expand the discussion on the explainability of results obtained by ML models, so that the explainability of AI systems in the educational domain provides reliability and better understanding for stakeholders."	[160]	"Explainability is seen as a critical area for exploration and development of solutions aimed at enhancing the transparency of AI-based educational systems. However, beyond explain ability, it is essential that indicators are also action able: that is, they enable students to take concrete steps to improve their learning."

Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), and regression quality assessment (such as R-squared) can all be used to assess how well the algorithms perform. The different metrics that were used in each research study for predicting performance are summarized in Figure 12.

We summarized the common measures used in the comparison study to evaluate the performance of different algorithms, including overall performance metrics. Metrics based on the confusion matrix are the most commonly used for assessing models: in 106 studies, the authors evaluated accuracy (e.g., [73], [96], [115], [122], [130], [172]–[176]); 61 research looked into precision (e.g., [72], [74], [96], [107], [109], [127]), 63 used recall (e.g., [65], [96], [124], [125], [139], [163], [177]–[179]); 44 used some F-measure (e.g., [116], [125], [133], [163], [172], [174], [180]); 13 used sensitivity (true positive rate) (e.g., [93], [101], [181]–[183]), 12 used specificity (true negative rate) (e.g., [10], [57], [59], [96], [125]). In addition, 53 studies used the area under the curve (AUC) or the receiving operation characteristic (ROC) itself to analyse the data. The following metrics were also employed: variance, complexity, Mean Absolute Error

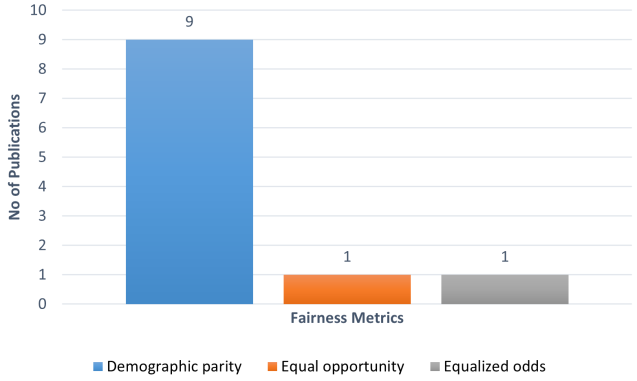

FIGURE 11: Fairness metrics that were used in our dataset during 2009–July 2025.

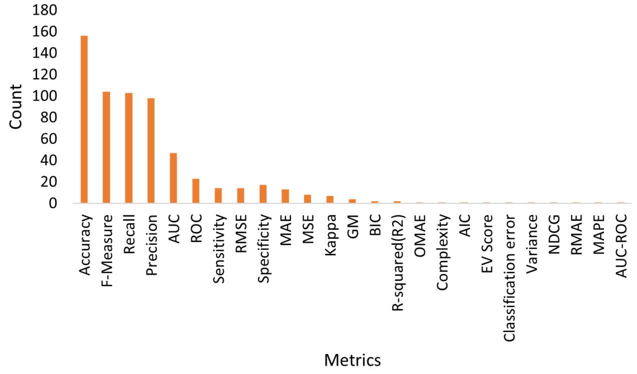

FIGURE 12: Performance metrics that were used in our dataset during 2009–July 2025.

(MAE), Root Mean Square Error (RMSE), Geometric Mean (GM), R-squared (R2), Cohen's kappa, and Ordinal Mean Absolute Error (OMAE).

From 2009 to July 2025, the number of articles focusing on different metrics—explainability metrics, fairness measures, and performance metrics—is shown in Figure 13. The figure highlights that, of the 161 publications in the literature, performance metrics have attracted the greatest attention—roughly 155 of them are devoted to the subject. In contrast, out of the 161 papers in our dataset, explainability metrics are addressed in only three articles and fairness measures in just 6, highlighting a significant disparity. This review highlights the need for a more balanced approach to research that incorporates explainability, performance, and fairness. For predictive models to foster ethical use and build trust, especially in educational settings where EDM and LA are increasingly important, they must be fair, comprehensible, and accurate.

A model's internal workings are made explicit through the use of several metrics, including separability, compositionality, and computational complexity. Computational complexity measures how well a model can process and generate

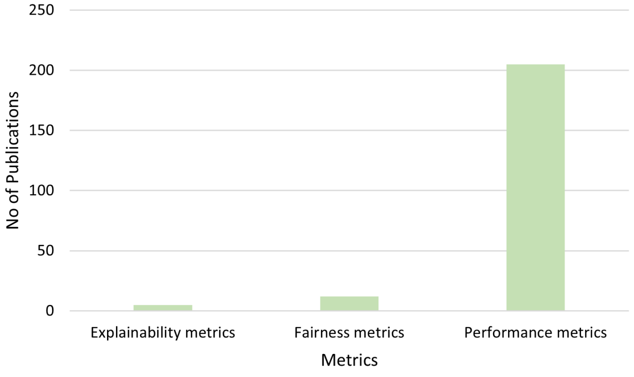

FIGURE 13: Number of publications using different metrics—explainability metrics, fairness measures, and performance metrics—that were used in our dataset during 2009–July 2025.

predictions; simpler models tend to be easier to understand because of their more basic functions. Compositionality is the ordering of feature importance values, or any threshold used to constrain the explanation, and separability refers to the ability to distinguish between different classes or categories based on the features or representations within a model. It is often associated with the clarity and distinctness of the boundaries between different groups in the feature space [184].

Conversely, the robustness and consistency of a model's predictions are assessed by several post-hoc explanation metrics, such as stability and sensitivity, which indirectly impact explainability. More stable models are simpler to understand and more reliable. Stability measures how much a model's predictions alter in response to small changes in input data or model parameters [148], [149]. Sensitivity measures the model's prediction sensitivity to variations in input characteristics, which affects how consistent and intelligible the model's outputs are under various conditions [150]. When combined, these metrics offer a thorough framework for assessing and enhancing the explainability of ML models, ensuring that they perform effectively and continue to be reliable and understandable in practical applications.

F. RECENT ADVANCES AND EMERGING TRENDS IN THE FIELD

Since the original submission, several key studies have advanced our understanding of ML in educational contexts, reflecting trends identified in our review. For instance, Pan et al. [185] review various ML models used in academic performance prediction, with a focus on algorithm comparison and dataset evaluation. This study underscores the growing role of ensemble methods, a trend we also identified, and highlights improvements in prediction accuracy and transparency in EDM.

Complementing this, recent advancements in deep learning have introduced even more complex models, such as the deep knowledge tracking model proposed by Zhonghua et al. [186]. This model incorporates an attention mechanism to address challenges in explainability and long sequence dependencies, significantly improving predictive performance in tracking students' knowledge acquisition. Both of these trends—ensemble techniques and deep learning—reflect a shift toward more complex models in educational ML, driving enhancements in accuracy, interpretability, and personalized learning.

Another important development is the shift away from traditional aggregated data explanations, advocating for individualized, instance-level predictions. For example, Saqr et al. [187] highlight the challenges posed by mispredictions and emphasize the importance of offering personalized feedback to enhance decision-making processes in education. These findings underscore the need for human-AI collaboration to improve the practical utility of AI systems.

These recent contributions reflect an ongoing shift toward more personalized, interpretable, and fair ML models in education, highlighting the evolving intersection of explainability, fairness, and algorithmic transparency.

G. STUDENT MODELS AND AUTOMATIC GRADING IN INTELLIGENT TUTORING SYSTEMS

Intelligent Tutoring Systems, which use AI to improve scalability and personalize education, depend significantly on student models and automatic grading systems. Through the analysis of several data points, including attendance, performance, behaviour, and wellness, student models monitor the progress of learners in order to predict outcomes and tailor learning paths accordingly. This personalization enables educators to provide timely interventions and better address the unique needs of diverse learners. Additionally, automatic grading systems improve scalability by reducing assessment time and effort requirements while promoting evaluation consistency [188].

The fairness of educational assessment has also been significantly impacted by automatic grading systems. To make grading fairer, it's important that AI systems are clear about how they work and why they give certain grades [189], [190]. However, since poorly designed models may inadvertently reinforce existing disparities, the fairness of these systems still depends on the quality of the algorithms and the data used to train them.

Typically, decision trees are inherently interpretable, making them favorable when clarity and explanation through feature importance metrics are priorities. However, they tend to underperform compared to more complex models like ANNs, which generally offer better predictive accuracy but at the cost of reduced transparency and explainability [151]. As previously stated, techniques like SHAP and LIME have been applied to bridge the explainability gap in neural networks, improving the transparency of intricate models. In the field of automatic grading, Condor et al. [191] introduced Neural Additive Models, which combine the explainability of additive models with the predictive power of neural networks. This enables educators to determine which aspects of a student's response affected the grade.

VI. CONCLUSION

In this study, we reviewed explainability in EDM and LA studies. In particular, we looked at the data types, variables, and metrics utilized in these studies, along with their explainability findings. On the one hand, we highlight an in-depth review of the current XAI approaches and techniques developed for evaluating explainability. Conversely, we emphasized the research trends in the field during the last 14 years, as well as the existing limitations and future directions of explainability in EDM and LA.

To sum up, we could bring attention to the often observed variable type and challenges in elucidating latent variables in education, such as satisfaction, reflection, and learning motivation. These variable types are categorized as continuous and categorical data. Our dataset reveals that both categorical and continuous data types have been extensively used, achieving the highest accuracy results with a combination of two data types.

Most models in the educational field use white-box models since the learned model itself is explained. According to our review, although post-hoc models are used less frequently than white-box models, their adoption has increased over time. This suggests that post-hoc model adoption will likely expand in the coming years. The evaluated literature primarily used tree-based, deep neural networks, regression, and Bayesian-based methods. Most of the examined research studies used classification approaches, followed by deep learning, regression, and clustering approaches. Decision trees, decision rules, and linear models were shown to have benefits in regard to explainability since they are able to explain and make explanations to domain experts on an intrinsic level, and they also have adequate prediction performance.

Deep neural networks are among the most accurate models that are intrinsically difficult to understand. Ongoing initiatives, however, may utilize post-hoc explainability approaches used in recent research to enhance the interpretability of such intricate models without compromising the way they perform. Model-agnostic strategies have been developed to extend interpretation techniques to any ML model. XAI approaches, such as anchors, SHAP, LIME, DLIME, feature importance, and counterfactual explanations, are utilized to analyze learner-generated data collected by different educational applications.

Many of the studies in our dataset include prediction models for explainability in different applications, and a few of them address quantifying and evaluating explainability [114], [141], [148]–[150]. Despite numerous studies mentioning the importance of explainability, none of them address the quantification and evaluation of it [69], [95], [141], [147]. Specifically, it was observed that explainability is not sufficiently addressed in EDM and LA, despite the fact that it would be crucial to both improving EDM's reliability and adhering to rules concerning AI systems.

As a result, in recent years, a few metrics for assessing explainability systems have been introduced. The mentioned literature provides quantitative prediction models that measure metrics related to explainability, such as sensitivity and stability. Although model performance, explainability, and fairness are interdependent, there is currently a lack of a systematic way to find the best algorithms or optimize for these objectives simultaneously in EDM and LA studies. In order to improve prediction performance, ensure transparency, user understanding, and explainability of the model and results, and address ethical issues in educational contexts, it becomes essential to develop explainable EDM and LA systems. Metrics are also required for explainability in EDM and LA models, in addition to being able to quantify performance and fairness with reliability.

One significant limitation of this systematic literature review is that we excluded grey literature and focused only on peer-reviewed papers. This approach may have the drawback of omitting certain important information, but it was selected to ensure the quality of the data used in the research. The use of databases and search engines to find relevant research is another limitation of this. All publications that pertain may not have been found using the search terms and filters that were applied, particularly those that convey similar concepts in explainable learning analytics using different terminology or keywords.

This might result in a partial overview of the field, potentially overlooking crucial studies that do not fit the predefined search criteria. Moreover, useful research from non-English-speaking regions could not be taken into account due to the removal of non-English publications, which can skew the results in favour of an English-centric perspective. Another limitation arises from the fact that gathering and summarising data in a systematic review involves personal judgement. This can lead to bias, especially when the research is complicated or the results are unclear.

Moreover, explainability often involves a trade-off with predictive performance or model complexity, which raises the significant issue of how to achieve a balance between efficacy and transparency in educational decision-making. Also, many current explainability techniques, including SHAP and LIME, were not created with the unique characteristics of educational data—which are often contextual, learner-specific, and temporal. This disparity may limit the accuracy of the explanations provided.

Advancing the field of explainable EDM and LA requires a well-defined, comprehensive, multi-dimensional roadmap. First, there is a pressing need to focus on developing standardized, reliable, and widely applicable metrics to assess the explainability of EDM and LA models. These metrics must be reliable and broadly applicable in order to enable consistent evaluation and comparison among various models and methodologies.

Second, research should be looking at how explainability interacts with other important elements like security and privacy. The proper application of learning analytics in educational contexts requires that models be not only explainable but also secure and privacy-preserving. Future studies can more effectively meet the intricate needs of modern educational environments by including data security and user privacy issues in the evaluation system.

Third, researchers should focus on the simultaneous optimization of explainability and fairness in EDM and LA. Although attempts have been made to improve model transparency, it is equally crucial to make sure that biases and inequality are not maintained by these models. Research should work towards developing methods that find a compromise between the requirement for equal consideration of all student groups and the necessity for clear, understandable explanations. This involves investigating techniques for identifying and reducing biases in algorithms and data.

By prioritizing both fairness and explainability, future studies can contribute to the development of more equitable and comprehensible LA systems, ultimately supporting more effective and inclusive educational outcomes. Addressing these dual goals will help create educational technologies that not only provide valuable insights but also promote fairness and equity in educational opportunities.

In summary, this review advances the field by offering a comprehensive and novel body of knowledge that systematically combines educational data, AI models, explainability, and fairness considerations. By integrating these dimensions into one systematic perspective, this study moves beyond reviews and delivers a holistic view that can inform and guide future research. To improve prediction performance, ensure transparency, enhance user understanding, and address ethical issues in educational contexts, it becomes essential to develop explainable EDM and LA systems equipped with reliable metrics for explainability.

ACKNOWLEDGMENT

This work was supported by the Academy of Finland (project no. 356314).

REFERENCES

[1] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, “Explainable ai (xai): Core ideas, techniques, and solutions,” *ACM computing surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[2] A. Gerdes, “The role of explainability in AI-supported medical decision-making,” *Discover Artificial Intelligence*, vol. 4, no. 1, p. 29, 2024.

[3] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, “Explainable AI for healthcare 5.0: opportunities and challenges,” *IEEe Access*, vol. 10, pp. 84 486–84 517, 2022.

[4] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[5] M. Afzaal, A. Zia, J. Nouri, and U. Fors, “Informative feedback and explainable ai-based recommendations to support students’ self-regulation,” *Technology, Knowledge and Learning*, vol. 29, no. 1, pp. 331–354, 2024.

[6] M. Matetic, “Mining learning management system data using interpretable neural networks,” in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 1282–1287.

[7] H. Waheed, S.-U. Hassan, R. Nawaz, N. R. Aljohani, G. Chen, and D. Gasevic, "Early prediction of learners at risk in self-paced education: A neural network approach," *Expert Systems with Applications*, vol. 213, p. 118868, 2023.

[8] B. Pei and W. Xing, "An interpretable pipeline for identifying at-risk students," *Journal of Educational Computing Research*, vol. 60, no. 2, pp. 380–405, 2022.

[9] W. Xing, D. Du, A. Bakhshi, K.-C. Chiu, and H. Du, "Designing a transferable predictive model for online learning using a bayesian updating approach," *IEEE Transactions on Learning Technologies*, vol. 14, no. 4, pp. 474–485, 2021.

[10] S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, and L. Porter, "A robust machine learning technique to predict low-performing students," *ACM transactions on computing education (TOCE)*, vol. 19, no. 3, pp. 1–19, 2019.

[11] A. Qin and M. Boicu, "Eduboost: An interpretable grey-box model approach to identify and prevent student failure and dropout," in *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2023, pp. 01–07.

[12] L. Yan, R. Martinez-Maldonado, L. Zhao, X. Li, and D. Gašević, "Physiological synchrony and arousal as indicators of stress and learning performance in embodied collaborative learning," in *International conference on artificial intelligence in education*. Springer, 2023, pp. 602–614.

[13] S. Tang and Z. Li, "Eeg complexity measures for detecting mind wandering during video-based learning," *Scientific Reports*, vol. 14, no. 1, p. 8209, 2024.

[14] Y. Şekerci, M. U. Kahraman, Ö. Özturan, E. Çelik, and S. Ş. Ayan, "Neurocognitive responses to spatial design behaviors and tools among interior architecture students: a pilot study," *Scientific Reports*, vol. 14, no. 1, p. 4454, 2024.

[15] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, no. 2, pp. 1–12, 2021.

[16] M. Saarela, V. Heilala, P. Jääskelä, A. Rantakaulio, and T. Kärkkäinen, "Explainable student agency analytics," *IEEE Access*, vol. 9, pp. 137 444–137 459, 2021.

[17] M. Saarela and T. Kärkkäinen, "Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator," *Journal of Informetrics*, vol. 14, no. 2, p. 101008, 2020.

[18] M. Saarela, O.-P. Ryynänen, and S. Äyrämö, "Predicting hospital associated disability from imbalanced data using supervised learning," *Artificial Intelligence in Medicine*, vol. 95, pp. 88–95, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365718303063

[19] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLOS ONE*, vol. 12, no. 4, p. e0174944, 2017.

[20] European Parliament and Council of the European Union, "European union general data protection regulation, articles 13–15," 2018. [Online]. Available: https://www.privacy-regulation.eu/en/13.htm

[21] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević, "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, 2022.

[22] J. M. Alonso and G. Casalino, "Explainable artificial intelligence for human-centric data analysis in virtual learning environments," in *International Workshop on Higher Education Learning Methodologies and Technologies Online*. Springer, 2019, pp. 125–138.

[23] S. Gunasekara and M. Saarela, "Explainability in educational data mining and learning analytics: An umbrella review," in *Proceedings of the International Conference on Educational Data Mining (EDM)*. International Educational Data Mining Society (IEDMS), 2024.

[24] R. Baker and P. S. Inventado, "Chapter 4: Educational data mining and learning analytics," in *Learning Analytics*. Springer, 2014.

[25] M. Saarela, *Automatic knowledge discovery from sparse and large-scale educational data: case Finland*. University of Jyväskylä, 2017, no. 262.

[26] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 6, pp. 601–618, 2010.

[27] International Educational Data Mining Society, "International educational data mining society," International Educational Data Mining Society, 2011, accessed June 26, 2017. [Online]. Available: http://www.educationaldatamining.org

[28] M. Saarela and T. Kärkkäinen, "Analysing student performance using sparse data of core bachelor courses," *Journal of Educational Data Mining*, vol. 7, no. 1, pp. 3–32, 2015.

[29] R. Ferguson, "Learning analytics: drivers, developments and challenges," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5/6, p. 304, 2012.

[30] W. Swartout, C. Paris, and J. Moore, "Explanations in knowledge systems: design for explainable expert systems," *IEEE Expert*, vol. 6, no. 3, pp. 58–64, 1991.

[31] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[32] C. Molnar, *Interpretable Machine Learning*. Lean Publishing, 2019.

[33] P. J. G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann, "The coming of age of interpretable and explainable machine learning models," *Neurocomputing*, vol. 535, pp. 25–39, 2023.

[34] T. Miller, "Explanation in artificial intelligence: insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[35] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–31, 2022.

[36] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[37] M. Saarela, V. Heilala, P. Jäskelä, A. Rantakaulio, and T. Kärkkäinen, "Explainable student agency analytics," *IEEE Access*, vol. 9, pp. 137 444–137 459, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3116664

[38] C. Li, M. Li, C.-L. Huang, Y.-T. Tseng, S.-H. Kim, and S. Yeom, "Educational data mining in prediction of students' learning performance: A scoping review," in *IFIP World Conference on Computers in Education*. Springer, 2022, pp. 361–372.

[39] S. U. Masruroh, D. Rosyada, N. A. R. Vitalaya *et al.*, "Adaptive recommendation system in education data mining using knowledge discovery for academic predictive analysis: Systematic literature review," in *2021 9th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2021, pp. 1–6.

[40] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, p. 237, 2020.

[41] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019.

[42] C. F. de Oliveira, S. R. Sobral, M. J. Ferreira, and F. Moreira, "How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 64, 2021.

[43] M. Bond, H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley, P. Pham, S. W. Chong, and G. Siemens, "A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour," *International journal of educational technology in higher education*, vol. 21, no. 1, p. 4, 2024.

[44] B. Memarian and T. Doleck, "Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100152, 2023.

[45] T. K. Chiu, Q. Xia, X. Zhou, C. S. Chai, and M. Cheng, "Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100118, 2023.

[46] A. Saranya and R. Subhashini, "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decision analytics journal*, vol. 7, p. 100230, 2023.

[47] E. Kalita, S. S. Oyelere, S. Gaftandzhieva, K. N. Rajesh, S. K. Jagatheesaperumal, A. Mohamed, Y. M. Elbarawy, A. S. Desuky, S. Hussain, M. A. Cifci *et al.*, "Educational data mining: a 10-year review," *Discover Computing*, vol. 28, no. 1, p. 81, 2025.

[48] J. Prentzas and A. Binopoulou, "Explainable artificial intelligence approaches in primary education: A review." *Electronics (2079-9292)*, vol. 14, no. 11, 2025.

[49] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *International journal of surgery*, vol. 8, no. 5, pp. 336–341, 2010.

[50] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *International journal of surgery*, vol. 88, p. 105906, 2021.

[51] B. Kitchenham, S. Charters *et al.*, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[52] M. Saarela and T. Kärkkäinen, "Can we automate expert-based journal rankings? analysis of the finnish publication indicator," *Journal of Informetrics*, vol. 14, no. 2, p. 101008, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1751157719302305

[53] I. E. Livieris, N. Karacapilidis, G. Domalis, and D. Tsakalidis, "An advanced explainable and interpretable ml-based framework for educational data mining," in *International Conference on Methodologies and Intelligent Systems for Technology Enhanced Learning*. Springer, 2023, pp. 87–96.

[54] G. Ramaswami, T. Susnjak, and A. Mathrani, "On developing generic models for predicting student outcomes in educational data mining," *Big Data and Cognitive Computing*, vol. 6, no. 1, p. 6, 2022.

[55] R. K. Veluri, I. Patra, M. Naved, V. V. Prasad, M. M. Arcinas, S. M. Beram, and A. Raghuvanshi, "Learning analytics using deep learning techniques for efficiently managing educational institutes," *Materials Today: Proceedings*, vol. 51, pp. 2317–2320, 2022.

[56] M. Tsiakmaki and O. Ragos, "A case study of interpretable counterfactual explanations for the task of predicting student academic performance," in *2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC)*. IEEE, 2021, pp. 120–125.

[57] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471–1483, 2020.

[58] G. Kostopoulos, S. Kotsiantis, and P. Pintelas, "Estimating student dropout in distance higher education using semi-supervised techniques," in *Proceedings of the 19th Panhellenic conference on informatics*, 2015, pp. 38–43.

[59] G. Kostopoulos, A.-D. Lipitakis, S. Kotsiantis, and G. Gravvanis, "Predicting student performance in distance higher education using active learning," in *International conference on engineering applications of neural networks*. Springer, 2017, pp. 75–86.

[60] S. Karlos, G. Kostopoulos, and S. Kotsiantis, "Predicting and interpreting students' grades in distance higher education through a semi-regression method," *Applied Sciences*, vol. 10, no. 23, p. 8413, 2020.

[61] G. Ramaswami, T. Susnjak, and A. Mathrani, "Supporting students' academic performance using explainable machine learning with automated prescriptive analytics," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 105, 2022.

[62] G. Cobo, D. García-Solórzano, J. A. Morán, E. Santamaría, C. Monzo, and J. Melenchón, "Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums," in *Proceedings of the 2nd International conference on learning analytics and knowledge*, 2012, pp. 248–251.

[63] E. Melo, I. Silva, D. G. Costa, C. M. Viegas, and T. M. Barros, "On the use of explainable artificial intelligence to evaluate school dropout," *Education Sciences*, vol. 12, no. 12, p. 845, 2022.

[64] L. Cohausz, "Towards real interpretability of student success prediction combining methods of xai and social science." *International Educational Data Mining Society*, 2022.

[65] S. Malik and K. Jothimani, "Enhancing student success prediction with featurex: A fusion voting classifier algorithm with hybrid feature selection," *Education and Information Technologies*, vol. 29, no. 7, pp. 8741–8791, 2024.

[66] A. Vultureanu-Albisi and C. Badica, "Improving students' performance by interpretable explanations using ensemble tree-based approaches," in *Proceedings of the 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2021, pp. 215–220.

[67] N. N. Sánchez-Pozo, J. S. Mejía-Ordóñez, D. C. Chamorro, D. Mayorca-Torres, and D. H. Peluffo-Ordóñez, "Predicting high school students' academic performance: A comparative study of supervised machine learning techniques," in *2021 machine learning-driven digital technologies for educational innovation workshop*. IEEE, 2021, pp. 1–6.

[68] J. Lu, J. Mou, and P. Li, "Interpretive analyses of learner dropout prediction in online stem courses," in *2023 5th International Conference on Computer Science and Technologies in Education (CSTE)*. IEEE, 2023, pp. 1–9.

[69] N. Kondo, T. Matsuda, Y. Hayashi, H. Matsukawa, M. Tsubakimoto, Y. Watanabe, S. Tateishi, and H. Yamashita, "Academic success prediction based on important student data selected via multi-objective evolutionary computation," in *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2020, pp. 370–373.

[70] V. T. N. Chau and N. H. Phung, "A cumulative increasing kemelized nearest-neighbor bagging method for early course-level study performance prediction," in *2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*. IEEE, 2021, pp. 91–96.

[71] M. F. A. Khan, J. Edwards, P. Bodily, and H. Karimi, "Deciphering student coding behavior: Interpretable keystroke features and ensemble strategies for grade prediction," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 5799–5808.

[72] S. E. Sorour and T. Mine, "Building an interpretable model of predicting student performance using comment data mining," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2016, pp. 285–291.

[73] C. Watson, F. W. Li, and J. L. Godwin, "Predicting performance in an introductory programming course by logging and analyzing student programming behavior," in *2013 IEEE 13th international conference on advanced learning technologies*. IEEE, 2013, pp. 319–323.

[74] T. Susnjak, "Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and chatgpt," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 452–482, 2024.

[75] Z. Mingyu, W. Sutong, W. Yanzhang, and W. Dujuan, "An interpretable prediction method for university student academic crisis warning," *Complex & Intelligent Systems*, vol. 8, no. 1, pp. 323–336, 2022.

[76] H. Sahlaoui, E. A. A. Alaoui, S. Agoujil, and A. Nayyar, "An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models," *Education and Information Technologies*, vol. 29, no. 5, pp. 5447–5483, 2024.

[77] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use moodle courses," *Computer Applications in Engineering Education*, vol. 21, no. 1, pp. 135–146, 2013.

[78] A. G. R. Sandeepa and S. Mohottala, "Evaluation of machine learning models in student academic performance prediction," in *2025 5th International Conference on Advanced Research in Computing (ICARC)*. IEEE, 2025, pp. 1–6.

[79] E. Kalita, H. El Aouifi, A. Kukkar, S. Hussain, T. Ali, and S. Gaftandzhieva, "Lstm-shap based academic performance prediction for disabled learners in virtual learning environments: a statistical analysis approach," *Social Network Analysis and Mining*, vol. 15, no. 1, pp. 1–23, 2025.

[80] M. N. Gul, W. Abbasi, M. Z. Babar, A. Aljohani, and M. Arif, "Data driven decisions in education using a comprehensive machine learning framework for student performance prediction," *Discover Computing*, vol. 28, no. 1, pp. 1–34, 2025.

[81] R. Katarya *et al.*, "Shapley explainable deep learning based knowledge distillation framework for student's performance prediction," in *2024 Second International Conference on Data Science and Information System (ICDSIS)*. IEEE, 2024, pp. 1–8.

[82] B. Alnasyan, M. Basheri, M. Alassafi, and K. Alnasyan, "Kanformer: an attention-enhanced deep learning model for predicting student performance in virtual learning environments," *Social Network Analysis and Mining*, vol. 15, no. 1, p. 25, 2025.

[83] M. Abdalkareem and N. Min-Allah, "Explainable models for predicting academic pathways for high school students in saudi arabia," *IEEE Access*, vol. 12, pp. 30 604–30 626, 2024.

[84] W.-C. Choi, C.-T. Lam, and A. J. Mendes, "Enhance learning performance predictions with explainable machine learning," in *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2024, pp. 1–9.

[85] B. Ujkani, D. Minkovska, and N. Hinov, "Course success prediction and early identification of at-risk students using explainable artificial intelligence," *Electronics*, vol. 13, no. 21, p. 4157, 2024.

[86] Y. Huang, Y. Zhou, and D. Wu, "Exploring factors causing the mathematics performance gaps of different genders using an explainable machine learning," *Computer Applications in Engineering Education*, vol. 33, no. 3, p. e70014, 2025.

[87] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Balancing performance and explainability in academic dropout prediction," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 2086–2099, 2024.

[88] W.-C. Choi, C.-T. Lam, and A. J. Mendes, "Analyzing the interpretability of machine learning prediction on student performance using shapley additive explanations," in *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*. IEEE, 2024, pp. 1–8.

[89] M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, and M. Rahman, "An explainable educational data mining system for predicting student academic performance," in *2024 IEEE International Conference on Signal Processing, Information, Communication and Systems (SPICSCON)*. IEEE, 2024, pp. 1–5.

[90] D. Yang, L. Cao, L. Pan, and S. Xie, "Interpretable grade prediction based on ecoc and shapley theory," in *Proceedings of the 2024 8th International Conference on Electronic Information Technology and Computer Engineering*. IEEE, 2024, pp. 49–54.

[91] A. Cano and J. D. Leonard, "Interpretable multiview early warning system adapted to underrepresented student populations," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 198–211, 2019.

[92] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, 2019.

[93] P. Kumar and M. Sharma, "Predicting academic performance of international students using machine learning techniques and human interpretable explanations using lime—case study of an indian university," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2019, Volume 1*. Springer, 2020, pp. 289–303.

[94] E. M. Queiroga, M. F. Batista Machado, V. R. Paragarino, T. T. Primo, and C. Cechinel, "Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay," *Information*, vol. 13, no. 9, p. 401, 2022.

[95] N. Capuano, D. Rossi, V. Ströele, and S. Caballé, "Explainable prediction of student performance in online courses," in *The Learning Ideas Conference*. Springer, 2023, pp. 639–652.

[96] O. B. Deho, S. Joksimovic, J. Li, C. Zhan, J. Liu, and L. Liu, "Should learning analytics models include sensitive attributes? explaining the why," *IEEE Transactions on Learning Technologies*, vol. 16, no. 4, pp. 560–572, 2022.

[97] F. da Conceição Silva, A. M. Santana, and R. M. Feitosa, "An investigation into dropout indicators in secondary technical education using explainable artificial intelligence," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 2025.

[98] H. Singh, B. Kaur, A. Sharma, and A. Singh, "Framework for suggesting corrective actions to help students intended at risk of low performance based on experimental study of college students using explainable machine learning model," *Education and Information Technologies*, vol. 29, no. 7, pp. 7997–8034, 2024.

[99] Q. Liu and M. Khalil, "Explainable ai in learning analytics: Improving predictive models and advancing transparency trust," in *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2024, pp. 1–7.

[100] J. Lin, W. Dai, L.-A. Lim, Y.-S. Tsai, R. F. Mello, H. Khosravi, D. Gasevic, and G. Chen, "Learner-centred analytics of feedback content in higher education," in *LAK23: 13th international learning analytics and knowledge conference*, 2023, pp. 100–110.

[101] B. Albreiki, T. Habuza, and N. Zaki, "Framework for automatically suggesting remedial actions to help students at risk based on explainable ml and rule-based models," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 1, p. 49, 2022.

[102] L. Cagliero, L. Canale, L. Farinetti, E. Baralis, and E. Venuto, "Predicting student academic performance by means of associative classification," *Applied Sciences*, vol. 11, no. 4, p. 1420, 2021.

[103] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[104] N. Nnamoko, J. Barrowclough, B. Onikoyi, and M. Liptrott, "Predicting student academic outcome using online behavioural statistics and neighbourhood influences: A machine learning approach," *SN Computer Science*, vol. 6, no. 6, p. 661, 2025.

[105] J. C. Gámez Granados, A. Esteban Toscano, F. J. Rodríguez Lozano, and A. Zafra Gómez, "An algorithm based on fuzzy ordinal classification to predict students' academic performance," 2023.

[106] A. A. Saa, "Educational data mining and students' performance prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2016.

[107] P. Guleria and M. Sood, "Explainable ai and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling," *Education and Information Technologies*, vol. 28, no. 1, pp. 1081–1116, 2023.

[108] X. Duan, B. Pei, G. A. Ambrose, A. Hershkovitz, Y. Cheng, and C. Wang, "Towards transparent and trustworthy prediction of student learning achievement by including instructors as co-designers: A case study," *Education and Information Technologies*, vol. 29, no. 3, pp. 3075–3096, 2024.

[109] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & education*, vol. 113, pp. 177–194, 2017.

[110] S. Ghimire, S. Abdulla, L. P. Joseph, S. Prasad, A. Murphy, A. Devi, P. D. Barua, R. C. Deo, R. Acharya, and Z. M. Yaseen, "Explainable artificial intelligence–machine learning models to estimate overall scores in tertiary preparatory general science course," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100331, 2024.

[111] A. Rachha and M. Seyam, "Explainable ai in education: Current trends, challenges, and opportunities," *SoutheastCon 2023*, pp. 232–239, 2023.

[112] G. N. Rangone, G. A. Montejano, A. G. Garis, C. A. Pizarro, and W. R. Molina, "An educational data mining model based on auto machine learning and interpretable machine learning," in *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)*. IEEE, 2022, pp. 1–6.

[113] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich, "Implications of ai (un-) fairness in higher education admissions: the effects of perceived ai (un-) fairness on exit, voice and organizational reputation," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 122–130.

[114] Z. Li and Z. Zhan, "Model selection and evaluation for learning analytics via interpretable machine learning," in *2023 5th International Conference on Computer Science and Technologies in Education (CSTE)*. IEEE, 2023, pp. 130–140.

[115] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in *Proceedings of the 2019 8th international conference on educational and information technology*, 2019, pp. 7–11.

[116] H. Li, C. F. Lynch, and T. Barnes, "Early prediction of course grades: Models and feature selection," in *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*, 2018, pp. 492–495.

[117] V. Kovanović, D. Gašević, S. Dawson, S. Joksimović, R. S. Baker, and M. Hatala, "Penetrating the black box of time-on-task estimation," in *Proceedings of the fifth international conference on learning analytics and knowledge*, 2015, pp. 184–193.

[118] M. Tsiakmaki, G. Kostopoulos, G. Koutsonikos, C. Pierrakeas, S. Kotsiantis, and O. Ragos, "Predicting university students' grades based on previous academic achievements," in *2018 9th international conference on information, Intelligence, Systems and Applications (IISA)*. IEEE, 2018, pp. 1–6.

[119] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, "Next-term student performance prediction: A recommender systems approach," *Journal of Educational Data Mining*, vol. 8, no. 1, p. 22–51, Sep. 2016.

[120] J. Xu, K. H. Moon, and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, 2017.

[121] D. Delen, "Predicting student attrition with data mining methods," *Journal of College Student Retention: Research, Theory & Practice*, vol. 13, no. 1, pp. 17–35, 2011.

[122] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," *Computers in Human behavior*, vol. 104, p. 106189, 2020.

[123] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100018, 2021.

[124] G. Casalino, P. Ducange, M. Fazzolari, and R. Pecori, "Incremental and interpretable learning analytics through fuzzy hoeffding decision trees," in *International Workshop on Higher Education Learning Methodologies and Technologies Online*. Springer, 2022, pp. 674–690.

[125] C.-H. Liao and J.-Y. Wu, "Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning

on student performance," *Computers & Education*, vol. 190, p. 104599, 2022.

[126] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser, "Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs," in *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. Durham, United Kingdom: International Educational Data Mining Society, July 2022, pp. 98–109.

[127] F. D. Pereira, S. C. Fonseca, E. H. Oliveira, A. I. Cristea, H. Bellhäuser, L. Rodrigues, D. B. Oliveira, S. Isotani, and L. S. Carvalho, "Explaining individual and collective programming students' behavior by interpreting a black-box predictive model," *IEEE Access*, vol. 9, pp. 117 097–117 119, 2021.

[128] W. Xing and S. Goggins, "Learning analytics in outer space: a hidden naïve bayes model for automatic student off-task behavior detection," in *Proceedings of the Fifth International Conference on learning analytics and knowledge*, 2015, pp. 176–183.

[129] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach," in *Proceedings of the fifth international conference on learning analytics and knowledge*, 2015, pp. 146–150.

[130] V. U. Kumar, A. Krishna, P. Neelakanteswara, and C. Z. Basha, "Advanced prediction of performance of a student in an university using machine learning techniques," in *2020 international conference on electronics and sustainable communication systems (ICESC)*. IEEE, 2020, pp. 121–126.

[131] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, 2025.

[132] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju, "Reliable post hoc explanations: Modeling uncertainty in explainability," *Advances in neural information processing systems*, vol. 34, pp. 9391–9404, 2021.

[133] Y. Jang, S. Choi, H. Jung, and H. Kim, "Practical early prediction of students' performance using machine learning and explainable ai," *Education and information technologies*, vol. 27, no. 9, pp. 12 855–12 889, 2022.

[134] R. L. C. Silva Filho, K. Brito, and P. J. L. Adeodato, "A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement," *Expert Systems with Applications*, vol. 221, p. 119729, 2023.

[135] L. Qin, Y. Zhu, S. Liu, X. Zhang, and Y. Zhao, "The shapley value in data science: Advances in computation, extensions, and applications," *Mathematics*, vol. 13, no. 10, p. 1581, 2025.

[136] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *International conference on machine learning*. PMLR, 2020, pp. 5491–5500.

[137] X. Huang and J. Marques-Silva, "On the failings of shapley values for explainability," *International Journal of Approximate Reasoning*, vol. 171, p. 109112, 2024.

[138] Y. Takefuji, "Reevaluating feature importance in machine learning: concerns regarding shap interpretations in the context of the eu artificial intelligence act," *Water Research*, vol. 280, p. 123514, 2025.

[139] K. M. Hasib, F. Rahman, R. Hasnat, and M. G. R. Alam, "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.

[140] B. Tousside, Y. Dama, and J. Frochte, "Towards explainability in modern educational data mining: A survey." in *KDIR*, 2022, pp. 212–220.

[141] S. Sohail, A. Alvi, and A. Khanum, "Interpretable and adaptable early warning learning analytics model," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3211–3225, 2022. [Online]. Available: https://doi.org/10.32604/cmc.2022.023560

[142] M. Baranyi, M. Nagy, and R. Molontay, "Interpretable deep learning for university dropout prediction," in *Proceedings of the 21st annual conference on information technology education*, 2020, pp. 13–19.

[143] S. Alwarthan, N. Aslam, and I. U. Khan, "An explainable model for identifying at-risk student at higher education," *IEEE Access*, vol. 10, pp. 107 649–107 668, 2022.

[144] Y. Qu, F. Li, L. Li, X. Dou, and H. Wang, "Can we predict student performance based on tabular and textual data?" *IEEE Access*, vol. 10, pp. 86 008–86 019, 2022.

[145] A. Sargsyan, A. Karapetyan, W. L. Woon, and A. Alshamsi, "Explainable ai as a social microscope: A case study on academic performance," in *International Conference on Machine Learning, Optimization, and Data Science*. Springer, 2020, pp. 257–268.

[146] J. Zhao, S. Bhatt, C. Thille, D. Zimmaro, and N. Gattani, "Interpretable personalized knowledge tracing and next learning activity recommendation," in *Proceedings of the seventh ACM conference on learning@ scale*, 2020, pp. 325–328.

[147] R. Alamri and B. Alharbi, "Explainable student performance prediction models: a systematic review," *Ieee Access*, vol. 9, pp. 33 132–33 143, 2021.

[148] E. Tiukhova, P. Vemuri, N. L. Flores, A. S. Islind, M. Oskarsdottir, S. Poelmans, B. Baesens, and M. Snoeck, "Explainable learning analytics: assessing the stability of student success prediction models by means of explainable ai," *Decision Support Systems*, vol. 182, p. 114229, 2024.

[149] M.-J. Li, S.-T. Li, A. C. Yang, A. Y. Huang, and S. J. Yang, "Trustworthy and explainable ai for learning analytics." in *LAK Workshops*, 2024, pp. 3–12.

[150] L. C. Nnadi, Y. Watanobe, M. M. Rahman, and A. M. John-Otumu, "Prediction of students' adaptability using explainable ai in educational machine learning models," *Applied Sciences*, vol. 14, no. 12, p. 5141, 2024.

[151] S. Gunasekara and M. Saarela, "Explainable ai in education: Techniques and qualitative assessment," *Applied Sciences*, no. 3, 2025.

[152] T. Doleck, D. J. Lemay, R. B. Basnet, and P. Bazelais, "Predictive analytics in education: a comparison of deep learning frameworks," *Education and Information Technologies*, vol. 25, no. 3, pp. 1951–1963, 2020.

[153] M. Hoq, P. Brusilovsky, and B. Akram, "Analysis of an explainable student performance prediction model in an introductory programming course." *International Educational Data Mining Society*, 2023.

[154] J. L. Olmo, C. Romero, E. Gibaja, and S. Ventura, "Improving meta-learning for algorithm selection by using multi-label classification: A case of study with educational data sets," *International Journal of Computational Intelligence Systems*, vol. 8, no. 6, pp. 1144–1164, 2015.

[155] Y. Jang, S. Choi, and H. Kim, "Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (at-eai) and analysis of its difference by gender and experience of ai education," *Education and Information Technologies*, vol. 27, no. 8, pp. 11 635–11 667, 2022.

[156] G. Casalino, G. Castellano, and G. Zaza, "Neuro-fuzzy systems for learning analytics," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2021, pp. 1341–1350.

[157] M. Lünich and B. Keller, "Explainable artificial intelligence for academic performance prediction. an experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1031–1042.

[158] S. Gunasekara and M. Saarela, "Quantitative assessment of explainability in machine learning models: A study on the oula dataset," in *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 2025, pp. 101–103.

[159] J. Zhou, Y. Cheng, and C. Dai, "From the perspective of explainable machine learning: A student feature selection strategy based on the geometric mean of feature importance and robustness," in *Proceedings of the 2024 International Conference on Computer and Multimedia Technology*, 2024, pp. 499–503.

[160] E. Félix, E. De Oliveira, I. Ramos, M. Perez-Sanagustin, E. Villalobos, I. Hilliger, R. Mello, and J. Broisin, "Designing actionable and interpretable analytics indicators for improving feedback in ai-based systems," in *17th International Conference on Computer Supported Education*. SCITEPRESS-Science and Technology Publications, 2025, pp. 428–435.

[161] M. J. Gomez, A. Armada Sanchez, M. Albaladejo-González, F. J. Garcia Clemente, and J. A. Ruipérez-Valiente, "Utilizing explainable ai to enhance real-time student performance prediction in educational serious games," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 2024, pp. 93–94.

[162] European Commission. (2019) Ethics guidelines for trustworthy ai. European Commission Digital Strategy. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[163] W. Jiang and Z. A. Pardos, "Towards equity and algorithmic fairness in student grade prediction," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 608–617.

[164] G. Fenu, R. Galici, and M. Marras, "Experts' view on challenges and needs for fairness in artificial intelligence for education," in *International Conference on Artificial Intelligence in Education*. Springer, 2022, pp. 243–255.

[165] D. Litman, H. Zhang, R. Correnti, L. C. Matsumura, and E. Wang, "A fairness evaluation of automated methods for scoring text evidence usage in writing," in *International Conference on Artificial Intelligence in Education*. Springer, 2021, pp. 255–267.

[166] M. Saarela, "On the relation of causality-versus correlation-based feature selection on model fairness," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 2024, pp. 56–64.

[167] R. Yu, D. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?" in *Proceedings of the 11th International Conference on Learning Analytics & Knowledge*. ACM, 2021, pp. 509–520.

[168] J. A. Idowu, A. S. Koshiyama, and P. Treleaven, "Investigating algorithmic bias in student progress monitoring," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100267, 2024.

[169] D. Ramalingam, H. Suresh, L. Paquette, and R. Singh, "Towards trustworthy auto-grading of short, multi-lingual, multi-type answers," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 919–946, 2022.

[170] R. S. Baker, A. Hawn, and S. Lee, "Algorithmic bias: The state of the situation and policy recommendations," 2023.

[171] K. Simbeck, "They shall be fair, transparent, and robust: auditing learning analytics systems," *AI and Ethics*, vol. 4, no. 2, pp. 555–571, 2024.

[172] C. Li, W. Xing, and W. Leite, "Using fair ai to predict students' math learning outcomes in an online platform," *Interactive Learning Environments*, vol. 32, no. 3, pp. 1117–1136, 2024.

[173] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in moocs," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 517–524.

[174] L. Wang, X. Li, Z. Luo, Z. Hu, and Q. Yan, "Multivariate cognitive response framework for student performance prediction on mooc," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 3, pp. 1221–1233, 2023.

[175] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Computers in Human Behavior*, vol. 98, pp. 166–173, 2019.

[176] F. Afrin, M. Hamilton, and C. Thevathyan, "On the explanation of ai-based student success prediction," in *International Conference on Computational Science*. Springer, 2022, pp. 252–258.

[177] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500–508, 2015.

[178] M. F. Musso, C. F. R. Hernández, and E. C. Cascallar, "Predicting key educational outcomes in academic trajectories: a machine-learning approach," *Higher education*, vol. 80, no. 5, pp. 875–894, 2020.

[179] X. Ma, Y. Yang, and Z. Zhou, "Using machine learning algorithm to predict student pass rates in online education," in *proceedings of the 3rd international conference on multimedia systems and signal processing*, 2018, pp. 156–161.

[180] J. Hardman, A. Paucar-Caceres, and A. Fielding, "Predicting students' progression in higher education by using the random forest algorithm," *Systems Research and Behavioral Science*, vol. 30, no. 2, pp. 194–203, 2013.

[181] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory," *Computers in human behavior*, vol. 47, pp. 168–181, 2015.

[182] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010.

[183] A.-S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decision Support Systems*, vol. 101, pp. 1–11, 2017.

[184] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/8/832

[185] J. Pan, Z. Zhao, and D. Han, "Academic performance prediction using machine learning approaches: A survey," *IEEE Transactions on Learning Technologies*, 2025.

[186] Z. Zhonghua, S. Tongping, and W. Jiaming, "Predicting student performance based on knowledge characteristics and learning ability," *IEEE Access*, 2025.

[187] M. Saqr and S. López-Pernas, "Why explainable ai may not be enough: predictions and mispredictions in decision making in education," *Smart Learning Environments*, vol. 11, no. 1, p. 52, 2024.

[188] S. Vetrivel, V. Arun, R. Ambikapathi, and T. Saravanan, "Automated grading systems: Enhancing efficiency and consistency in student assessments," in *Adopting Artificial Intelligence Tools in Higher Education*. CRC Press, 2025, pp. 41–61.

[189] F. Chai, J. Ma, Y. Wang, J. Zhu, and T. Han, "Grading by ai makes me feel fairer? how different evaluators affect college students' perception of fairness," *Frontiers in Psychology*, vol. 15, p. 1221177, 2024.

[190] M. S. Johnson and D. F. McCaffrey, "Evaluating fairness of automated scoring in educational measurement," *Advancing natural language processing in educational assessment*, pp. 142–164, 2023.

[191] A. Condor and Z. Pardos, "Explainable automatic grading with neural additive models," in *International Conference on Artificial Intelligence in Education*. Springer, 2024, pp. 18–31.

SACHINI GUNASEKARA received her MSc degree in Information Technology from the Sri Lanka Institute of Information Technology (SLIIT) in 2018. She is currently a Ph.D. student in Educational Technology and Cognitive Science in the Faculty of Information Technology at the University of Jyväskylä (JYU/IT), Finland. Her research interests include machine learning, learning analytics, and data mining.

MIRKA SAARELA a tenure-track assistant professor of Educational Technology and an Academy of Finland research fellow in the Faculty of Information Technology at the University of Jyväskylä (JYU/IT), Finland. She holds a doctoral degree in Mathematical Information Technology (JYU/IT, 2017). Her research lies at the intersection of learning analytics, machine learning, education, and artificial intelligence. She is especially interested in explainability and fairness in algorithmic decision-making. Her research on explainable AI was supported and awarded by several foundations, such as the Otto A. Malm Foundation, the Finnish Foundation for Share Promotion, the K. H. Renlund Foundation, and the Ulla Tuomisen Foundation.