

Explainable Machine Learning for Early Detection of Mathematics Learning Difficulties: A Systematic Review of Multimodal Educational Data and Predictive Modeling

Narges Hosseinzadeh

nhoss8@uic.edu

University of Illinois at Chicago <https://orcid.org/0000-0002-9712-9294>

Tannaz Goodarzvand Chegini

Montana State university <https://orcid.org/0009-0008-3194-9776>

Faraz Dadgostari

Montana State University

Neda Nazemi

Montana State University

Systematic Review

Keywords: Developmental dyscalculia, early detection, educational data mining, explainable artificial intelligence, eye-tracking, machine learning, mathematics learning difficulties, multimodal learning analytics, predictive modeling, systematic review, wheel-spinning

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11098732/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Explainable Machine Learning for Early Detection of Mathematics Learning Difficulties: A Systematic Review of Multimodal Educational Data and Predictive Modeling

Narges Hosseinzadeh¹, Tannaz Goodarzvand Chegini², Faraz Dadgostari³, Neda Nazemi⁴

¹College of Education, University of Illinois Chicago, Chicago, Illinois, 60607, USA

²Department of Mathematical Sciences, Montana State University, Bozeman, Montana, 59717, USA

³Department of Mechanical and Industrial Engineering, Montana State University, Bozeman, Montana, 59717, USA

⁴Gianforte School of Computing, Montana State University, Bozeman, Montana, 59717, USA

Corresponding Author: Narges Hosseinzadeh (e-mail: nhoss8@uic.edu).

This research was partially supported by the National Science Foundation under Award No. 2418359.

ABSTRACT Mathematics learning difficulties, including developmental dyscalculia, affect about 3–6% of school-age children and, when unaddressed, can lead to long-term academic underperformance, reduced career opportunities in science and technology, anxiety, low self-efficacy, and high costs to society. Conventional identification relies largely on post-test standardized scores, which rarely reveal specific misconceptions, maladaptive behaviors such as wheel-spinning, or inefficient problem-solving strategies, delaying intervention. This systematic review, conducted in accordance with the PRISMA 2020 guidelines, synthesizes research on explainable machine learning methods that use multimodal educational data, including interaction histories, clickstream data, eye-tracking, audiovisual classroom sensing, physiological cues, and contextual information, for the early detection and transparent explanation of mathematics learning difficulties. A March 2026 search of major databases and conference proceedings identified 18 sources. Interaction-log and clickstream data dominate early-warning systems, with decision-tree and gradient-boosting models commonly used to detect wheel-spinning, whereas eye-tracking provides process-level evidence of misconceptions and strategy use. Multimodal fusion approaches (early, late, hybrid/intermediate, and co-learning) combined with explainability methods (intrinsic glass-box models, Shapley additive explanations, local interpretable model-agnostic explanations, and counterfactual explanations) substantially improve predictive performance and interpretability, supporting educator trust and ethical deployment. Persistent challenges include reliance on proxy labels rather than clinical diagnoses, data alignment and synchronization, temporal and group leakage, limited cross-platform generalizability, and fairness. We propose a research agenda centered on privacy-preserving multimodal datasets, standardized evaluation of explanation usefulness for stakeholders, and integration with classroom interventions, and we discuss implications of the European Union Artificial Intelligence Act and U.S. and European data-protection laws.

INDEX TERMS Developmental dyscalculia, early detection, educational data mining, explainable artificial intelligence, eye-tracking, machine learning, mathematics learning difficulties, multimodal learning analytics, predictive modeling, systematic review, wheel-spinning.

I. INTRODUCTION

MATHEMATICS learning difficulties (MLD) are defined as ongoing problems with performing mathematical tasks at the appropriate age [1]. MLD is a major issue among students, teachers, and the entire community. Students with MLD face significant challenges when undertaking mathematical tasks at their developmental level. These obstacles may be associated with developmental dyscalculia or a mix of more general factors, such as inadequate teaching, language obstacles, attention or memory disorders, and socioeconomic disadvantages [1], [2].

Developmental dyscalculia is specifically defined by the diagnostic criteria of the DSM-5 and ICD-11 as significant and chronic mathematical learning impairments that cannot be attributed only to intellectual disability or the absence of

educational opportunities. The prevalence of dyscalculia is estimated at 3–6% among school-aged children, though this figure varies depending on the methods and criteria used to identify it [3]. In practice, MLD manifests as difficulty mastering number sense, number facts, and calculation, including poor understanding of numerical magnitude and relationships, reliance on immature strategies such as finger-counting for facts peers recall automatically, and difficulty applying mathematical concepts and procedures to solve problems, which together constitute the diagnostic criteria for impairment in mathematics [1]. These manifestations are considered the hallmark symptoms of math difficulties in clinical research, though the field's cognitive profile of MLD remains incomplete, making it difficult to cleanly differentiate MLD from general low math achievement in

typical school or clinical settings [4]. Conventional assessment relies primarily on standardized, norm-referenced math achievement tests and curriculum-based measurement administered at fixed intervals, supplemented by clinical batteries when a formal diagnosis is sought. These tools are well suited to confirming that a gap exists relative to grade-level peers, but by design they are administered only periodically and after difficulties have already become apparent in the classroom, offering little insight into the specific misconceptions or strategies underlying a given student’s struggle.

When unaddressed, MLD is linked to a range of negative consequences, including low academic performance, a lack of interest in science, technology, engineering, and mathematics (STEM) fields driven by math anxiety [5], decreased self-efficacy, and high societal costs due to remedial education and wasted productivity [6], [7]. The concept of early intervention has been widely recognized as the most efficient and cost-effective solution, with long-term better results, consistently outperforming remediation efforts in later years.

Existing school-based identification practices, however, are not sufficient. The reliance on late standardized test scores or classroom performance often leads to delayed identification, often after repeated failures [8]. These testing approaches, to a degree, give minimal information on particular misconceptions, maladaptive behaviors such as wheel-spinning [9], a pattern in which students persistently practice without making meaningful progress toward mastery, or inefficient problem-solving tactics that are the root of student challenges. In turn, this creates issues in identifying real MLD, rather than temporary difficulties, and long-term postponements can promote stigmatization and deprive educators of timely, specific support mechanisms [10].

Throughout this review, we use MLD as a broad term which includes diagnosable neurodevelopmental conditions and non-clinical difficulties, since formal diagnosis is rarely available in the large-scale educational datasets this review draws on. We have also reviewed articles that worked on observable indicators for MLD, and we can categorize our findings into three groups, summarized in Table 1: a) Behavioral and Persistence Patterns, visible in interaction logs, showing how a student practices over time; b) Strategic and Conceptual Patterns, visible in eye-tracking or think-aloud data and showing how a student is actually reasoning through a problem; c) Underlying Risk Factors and Diagnostic Categories, not observed directly in the moment, but instead assessed through standardized tools or defined clinically. Table 1 draws on the studies we included in this review, which are described in Section II.

TABLE 1
TAXONOMY OF OBSERVABLE MLD INDICATORS ACROSS INCLUDED STUDIES

Category	Problem Type	Definition
(a) Behavioral / Persistence Patterns	Wheel-spinning	Persistent, repeated practice on a skill without meaningful progress toward mastery [9], [11], [12]

	Non-persistence	Premature disengagement from a skill or task before mastery is achieved [11]
	Excessive hint-seeking / hint overreliance	A maladaptive pattern of dependence on system-provided hints or scaffolds rather than independent problem-solving attempts, often masking an underlying misunderstanding [13]
	Repeated errors / fast guesses	Rapid, low-latency incorrect responses indicating guessing rather than genuine problem-solving attempts [14]
	Guessing	Responses characterized by minimal engagement with the problem content, evidenced by very short response times or absence of a coherent solution strategy [14]
(b) Strategic / Conceptual Patterns	Misconceptions	Systematic, incorrect conceptual understandings that produce predictable, recurring error patterns across related problems [14]
	Inefficient problem-solving strategies	The use of valid but suboptimal procedural approaches that increase cognitive load or time-to-solution without necessarily yielding an incorrect answer [15]
	Number-line estimation strategy differences	Variation in the underlying representational strategy used to estimate numerical magnitude (e.g., logarithmic versus linear compression) [16]
	Cognitive strategy patterns (general)	Broader patterns of reasoning approach adopted during problem-solving that are not tied to a specific, identifiable error type [15]
(c) Underlying Risk Factors / Diagnostic Categories	Working memory / number sense deficits	Foundational cognitive capacities whose impairment is associated with, though not exclusive to, mathematical difficulty, typically measured through standardized cognitive assessments [4]
	Developmental dyscalculia (and subtypes)	A clinically diagnosable neurodevelopmental condition, per DSM-5 and ICD-11 criteria, marked by persistent impairment in the acquisition of arithmetic and number-sense skills not attributable to intellectual disability or inadequate instruction [3]

Note. Productive persistence; sustained, effective practice that leads to mastery, is not itself a difficulty, but is included in several studies [11], [12] as the contrastive category against which wheel-spinning is defined and detected.

Understanding how the field has arrived at its current state requires tracing the key developments that have shaped this intersection, as the combined use of early prediction methods, multimodal data analytics, and explainable artificial intelligence in support of mathematics learning has undergone considerable evolution through a series of milestones. In 1994, Corbett and Anderson developed Bayesian Knowledge Tracing (BKT), which is a probabilistic model that represents the knowledge state of learners over time [17]. As of 2013, the discovery of wheel-spinning

behavior indicated the utmost importance of mastery-learning paradigms in the early stages of development [9]. The year 2013 witnessed an increasing focus on the ethical aspects of learning analytics, especially regarding privacy, transparency, and informed consent [18]. In 2015, Deep Knowledge Tracing (DKT) was presented: recurrent neural networks were trained on large, aggregated datasets spanning many students to learn general patterns of skill acquisition; the resulting model was then used to generate individualized predictions of each student's future performance on specific skills [19]. In 2016, Local Interpretable Model-agnostic Explanations (LIME) was introduced [20]; the following year, SHAP (SHapley Additive exPlanations) [21] and more detailed taxonomy of multimodal machine learning challenges were also presented [22]. Scalable applications of early wheel-spinning detection were accompanied by a substantial improvement in the rigor and transparency of systematic reviews, as mandated by the PRISMA 2020 guidelines [23]. In 2021, the National Institute of Standards and Technology (NIST) identified four principles of explainable artificial intelligence [24], which were implemented in educational settings in 2022 and are characterized by stakeholder engagement and actionable insights [25]. Also, the European Union AI Act of 2024 included educational AI applications in the high-risk systems category. In 2025, time-dependent gaze-based prediction models appeared, which allowed recognizing certain mathematical problems when solving the problem [14]. For example, Kennel et al. tracked students' eye fixations while solving graphical differentiation problems and showed that gaze patterns alone could flag a specific misconception within the first 15–45 seconds of attempting the problem [14]. Together, these accomplishments clarify how knowledge-tracing methodologies, early-detection frameworks, systematic review requirements, and modern approaches are progressively integrated to describe and explain explainable, multimodal machine learning within the framework of mathematics education (Fig. 1).

Building on these developments, and in response to the identification gaps outlined above, scholars are increasingly advocating the use of rich, multimodal data streams that are frequently produced in online learning settings. Multimodal educational data refers to the process of gathering and combining data from various sources, including digital interaction logs, assessments, eye-tracking information, audiovisual recordings, physiological data, and contextual variables, and converting them through Multimodal Learning Analytics (MMLA) into valuable features, predictions, and feedback [26]. These data include logs and clickstreams from intelligent tutoring systems, eye-tracking data during

problem-solving tasks, audiovisual classroom videos, physiological data (e.g., electroencephalography or wearable sensors), and context information. All these varied sources of data allow identifying MLD indicators earlier and more accurately, as well as providing a deeper insight into the processes behind learning. For example, intelligent tutoring systems have shed light on the journey to mastery and patterns of persistence [11], [12], and eye-tracking research has revealed real-time misunderstandings and strategic reorientations in solving mathematical problems [14], [16]. However, these data modalities are typically analyzed separately and have not been integrated with explainable artificial intelligence (XAI) methods, which convert complex model outputs into clear, actionable insights for educators. These insights help teachers, students, and families understand, trust, and act on predictions about individual students in ways that are pedagogically sound and avoid stigmatizing language [27], [28].

In addition to technical performance, explainable machine learning should be adopted for pedagogical and ethical reasons. Teachers, administrators, learners, and families need explanations that are understandable enough to build trust, support informed decision-making, reduce negative labeling, and enable adherence to core educational values. This need is reflected in regulatory frameworks: the AI Act of the European Union defines high-stakes educational uses as high-risk systems, which must be highly accountable, and U.S. laws, including the Family Educational Rights and Privacy Act (FERPA) and the Children Online Privacy Protection Act (COPPA), as well as the principles of the General Data Protection Regulation (GDPR) in Europe, highlight the value of privacy, and fairness. Broader educational scholarship on AI integration similarly warns that continuous multimodal and biometric monitoring, while powerful for detecting moments of struggle, introduces critical vulnerabilities surrounding algorithmic opacity, data privacy, and the inadvertent reinforcement of demographic biases, underscoring that predictive tools must prioritize teacher agency and transparent, localized pedagogical decision-making over fully automated, black-box deployment [29]. Although considerable progress has been made in specific areas, there have been very few integrated solutions that combine multimodal educational information, early MLD-specific predictive modeling, and education-focused XAI. The bulk of current studies on learning analytics and education data mining focus on overall performance or general at-risk measures rather than math-specific difficulties that may be addressed through targeted educator interventions [28].

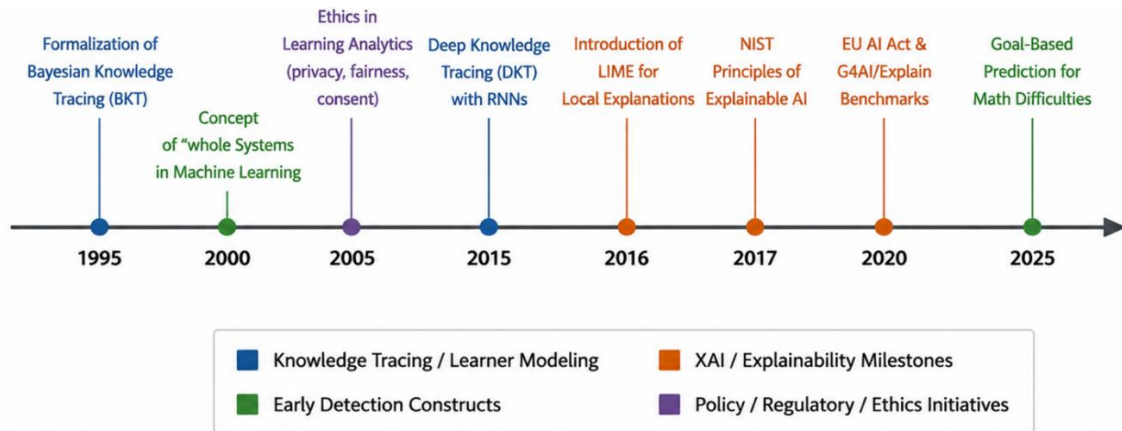


FIGURE 1. Timeline of key milestones linking early prediction, multimodal data, and explainability in mathematics learning support (1994–2025).

A. SCOPE AND TARGET AUDIENCE

The review addresses K–12 students and early university learners, developmental and educational stages in which mathematics learning difficulties and related risk indicators can substantially affect subsequent learning. The review aims to guide educators, educational technologists, curriculum developers, and policymakers to focus on evidence and data in decision-making to detect, at an early stage, learners at risk of developing mathematical learning difficulties and to take the required action.

B. SIGNIFICANCE STATEMENT

This review integrates three domains that have largely been examined separately: multimodal learning analytics, early predictive modeling for mathematics learning difficulties, and explainable artificial intelligence. Although recent reviews have examined multimodal learning analytics broadly [26], [28], [30], [51], they have not jointly synthesized mathematics-specific early-risk indicators, predictive earliness, multimodal evidence, and explainability within a single systematic framework. The present review addresses this gap by focusing specifically on mathematics learning difficulties and by examining not only predictive methods and data modalities but also the interpretability, methodological validity, and educational applicability of these systems.

C. PURPOSE AND RESEARCH QUESTIONS

This systematic review aims to gather and integrate current research on multimodal data and explainable machine learning methodologies. It aims to identify deficits in early mathematics detection and to offer suggestions for future research. The review is organized along the following research questions:

RQ1: What are the most frequently used types of multimodal data and learning analytics to support early identification of mathematics learning difficulties?

RQ2: What is the effectiveness of the use of early predictive models of mathematics learning difficulties to identify at-risk students relative to traditional assessment methods?

RQ3: What is the level of integration of XAI frameworks in these predictive models to enhance educators' understanding of the outcomes?

RQ4: What are the general pedagogical, technical, and ethical issues with the use of such predictive systems in the educational context in the real world?

II. METHODOLOGY

The literature search and selection process followed rigorous systematic review protocols to enhance transparency, reproducibility, and comprehensiveness. The methodology enabled the determination of relevant empirical research and theoretical models that overlap with mathematics learning challenges, predictive modeling, multimodal data, and explainable artificial intelligence. The review adopted transparent methods for study identification, screening, eligibility evaluation, and data synthesis, in line with the PRISMA 2020 guidelines.

A. SEARCH STRATEGY

1) DATABASES AND VENUES

Major digital libraries, including the ACM Digital Library, IEEE Xplore, SpringerLink, ScienceDirect, PubMed Central, ERIC, and arXiv, and proceedings of major conferences on learning analytics, such as Learning Analytics and Knowledge (LAK), Educational Data Mining (EDM), and Multimodal Learning Analytics (MMLA), were searched in March 2026. All resulting articles were exported and deduplicated using reference management software, then verified by hand. Backward and forward citation tracking were also conducted to identify additional papers in the area [30].

2) BOOLEAN SEARCH STRINGS

The searches were based on Boolean strings focused on mathematics learning challenges, early detection and prediction, multimodal data sources, and explainable or interpretable methods, and were complemented with keywords such as wheel-spinning, productive persistence,

and mastery learning. The following Boolean string template was used as the core search strategy:

*(math OR mathematics OR numeracy OR dyscalculia OR
"mathematics learning difficulties" OR "math learning
disabilities")
AND (early detection OR screening OR prediction OR "early
warning" OR "risk model" OR "knowledge tracing")
AND (multimodal OR "learning analytics" OR clickstream OR
log OR "eye tracking" OR video OR audio OR EEG OR
neuroimaging)
AND (explainable OR interpretable OR XAI OR SHAP OR
LIME OR counterfactual OR "feature importance")*

These include specific terms (e.g., "wheel-spinning," "productive persistence," or "mastery learning") reported or identified in early identification of learning problems using intelligent tutoring system (ITS) research. The search and synthesis drew on foundational definitions of mathematics learning difficulties (MLD) and dyscalculia (per DSM-5 and ICD-11 criteria), multimodal learning analytics frameworks, predictive modeling approaches for early detection (e.g., interaction logs and gaze data), educational explainability techniques, and key evaluation and governance considerations. The literature review also encompassed frameworks for multimodal machine learning methods, including representation-based, alignment-based, data-fusion-based, and co-learning-based approaches, as well as research on interpretable decision trees for detecting problems such as wheel-spinning, gaze-based forecasting of math difficulty, and clickstream-based math performance prediction.

B. STUDY SELECTION

1) INCLUSION AND EXCLUSION CRITERIA

This review was limited to K-12 students and early university learners, with the focus on mathematical performance, misconceptions, and related challenges. The studies included predictive modeling to address mathematics learning problems or early warning signs (including persistence problems, mastery deficits, or common misconceptions) and used explainable or interpretable methods.

We included the studies that fulfilled these conditions:

- 1) Data from empirical studies or studies on mathematics performance or difficulties, or high-quality literature and frameworks on XAI in education.
- 2) Use of at least one educational method (such as logs or assessments) and possibly more (such as eye tracking, video/audio, physiological signals, or neuroimaging).
- 3) A clearly described assessment method applied (for example, train/test splits, cross-validation, or temporal evaluation), or a detailed account of the evaluation strategy used.

We excluded studies in these categories:

- 1) Studies not related to math outcomes or difficulties (unless they provide evidence that XAI in education generally).

- 2) Purely correlational analyses without a predictive focus (unless the study validates a measurement method for MLD).
- 3) Opinion pieces without significant technical or methodological insights (outside of well-known articles on interpretability challenges).

For synthesis, included sources were distinguished by evidence type. Primary empirical studies provided the basis for conclusions regarding predictive performance and earliness, whereas review or framework-oriented sources were used to contextualize multimodal methods, explainability approaches, and broader methodological developments. These source types were therefore not treated as equivalent evidence for predictive performance.

2) SCREENING PROCESS

We used Covidence, a web-based systematic review management platform commonly used to support PRISMA-compliant screening workflows, to organize our references and systematically screened the records. First, we deleted the duplicate entries. Afterward, we performed the title and abstract screening and then fully reviewed the articles, using defined inclusion and exclusion criteria. To assure objectivity, the title and abstract screening was carried out by two reviewers. When disagreements arose, discussions were held, and when consensus was not reached, a third reviewer was consulted to resolve them, underscoring our reliance on a balanced, fair review process.

The literature search resulted in 1,256 records from the chosen databases and conference proceedings. A total of 937 records remained after removing 319 duplicates. During the title and abstract screening, 809 records were excluded as clearly irrelevant, leaving 128 articles eligible for full-text review. We included 18 articles that met all our inclusion and exclusion criteria after excluding 110, and used them to inform our evidence synthesis. The selection process is described in detail in the following PRISMA flow diagram (Fig. 2).

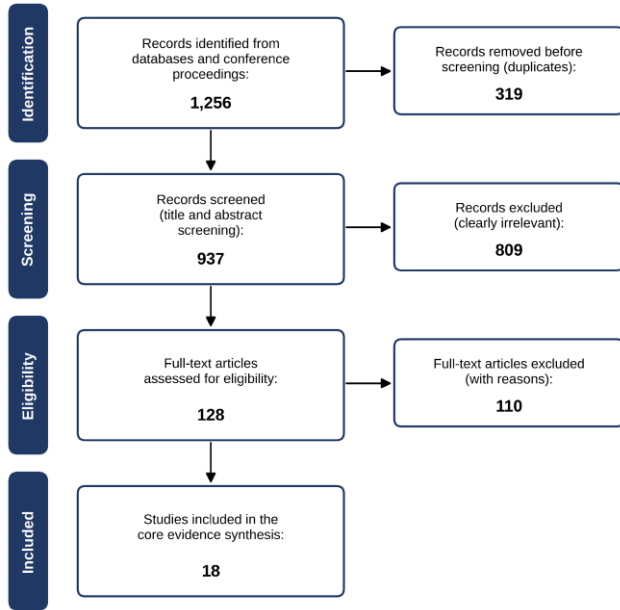


FIGURE 2. PRISMA 2020 flow diagram of the study selection process.

3) SCOPE AND REPORTING FRAMEWORK

Consistent with PRISMA 2020 guidelines, which encourage clear and honest procedures in systematic reviews and provide instructions for searching for studies, filtering them, setting inclusion criteria for study design, and analyzing data [23], the review’s scope focuses on the following:

Population/Setting: Learners we seek include K to 12 level students and early university students, particularly focusing on their mathematical achievement, misconceptions, and related problems.

Intervention/Phenomenon: We attempt to forecast MLD or early indicators (e.g., continued issues, lack of mastery, and widespread misconceptions) from a variety of educational datasets.

Explainability: These can be techniques that provide explanations (e.g., decision trees, additive models) after model analysis (e.g., SHAP, LIME, counterfactual systems) and education-specific methods for AI explanation (XAI).

C. DATA EXTRACTION AND QUALITY ASSESSMENT

Population characteristics, sample size, target/label definitions, methods, preprocessing, model types, fusion (where relevant), explainability, evaluation metrics, validation, and limitations were included in data extraction. Data were independently extracted by two reviewers using a standardized extraction framework; discrepancies were resolved through discussion, with a third reviewer consulted when necessary, and the final extracted information was independently verified by a fourth reviewer against the source articles. We used TRIPOD+AI to guide the assessment of reporting completeness and PROBAST+AI to structure the evaluation of methodological quality, risk of bias, and applicability of the predictive models. Together, these frameworks emphasize clear outcome definitions, appropriate handling of missing data, avoidance of data

leakage, and adequate model validation, ideally including external validation. We drew on the TRIPOD+AI and PROBAST+AI frameworks [31], which flag three risks as common threats to validity in educational prediction models: temporal leakage, where information from future time points leaks into predictive features; group leakage, where the same student appears in both the training and test sets; and proxy-label fragility, where an outcome label such as “struggle” is defined inconsistently across platforms rather than through a validated clinical measure. Of these, reliance on proxy labels rather than clinical diagnoses was the most consistent concern across the included studies, present in 14 of the 18 sources, as detailed in Section III.

III. RESULTS

A. MULTIMODAL DATA SOURCES AND FEATURE ENGINEERING

The current state of the art on the topic can be further subdivided into a few broad categories of modalities, enabling multimodal collaborations. The existing developments in the early identification of MLD may be divided into six main data modalities that collaborate to create multimodal partnerships: interaction logs and clickstream data, assessments and psychometric data, eye-tracking, audiovisual modalities and classroom sensing, physiological measures and neuroimaging, and socio-demographic and contextual factors (Fig. 3).

1) DATA MODALITIES

The most common and most scalable sources are interaction logs and clickstream data, which record data on item accuracy, number of attempts, use of hints, time taken on the task, progress, and navigation. Persistence and wheel-spinning in mathematics practice have been studied with educational systems, like ASSISTments [9], with later work comparing detection methods across multiple large-scale tutoring datasets [32].

The high-validity outcome measures are assessments and psychometric data in the form of standardized tests, curriculum-based assessments, timed arithmetic tasks, and cognitive evaluations of the working memory and number sense [4].

Eye-tracking technology provides in-depth process evidence by monitoring fixation durations, saccadic movements, and gaze transitions within areas of interest, thereby revealing specific misconceptions and strategic differences in the problem-solving process [14],[16].

By analyzing video and audio data, audiovisual modalities, and classroom sensing represent emotional states and the degree of engagement along with collaborative interactions. Moreover, physiological measures, such as electroencephalography and wearable devices, provide useful, though less commonly used, data to investigate numeracy development [33]. Neuroimaging methods, in particular, have been used to examine the neural correlates of numerical skill development in children with and without developmental dyscalculia [34]. There are, however,

limitations on their use in routine screening due to cost, privacy, and ethical concerns.

Lastly, socio-demographic and contextual factors, including socioeconomic status, language background, and educational history, may also add predictive power, but they should be handled with care to ensure fairness and minimal risk [35].

Interaction logs and clickstream data were the most commonly used modality, appearing in 11 of the 18 included

studies to capture user-system interactions. In 4 studies, eye-tracking was used; in 2 studies, physiological/sensor data were used. The results show that log-based and eye-tracking approaches are the most common ways of measuring. Interestingly, only 2 of the 18 sources used a combination of 3 or more of the six modality categories identified in this review [36], [37].

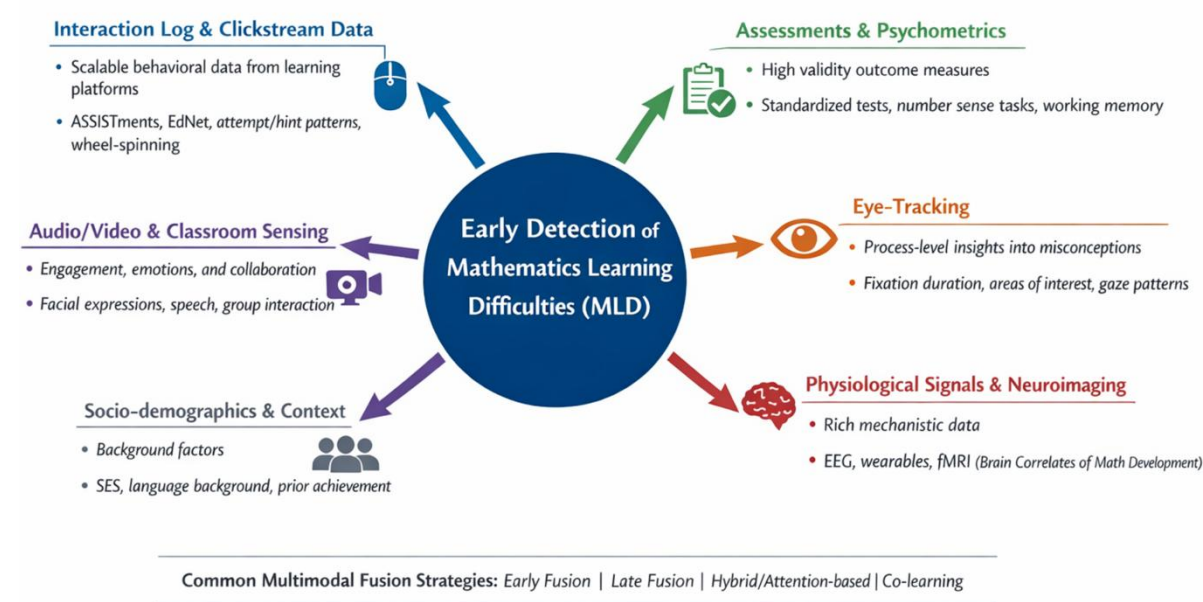


FIGURE 3. Taxonomy of multimodal data sources for early detection of mathematics learning difficulties (MLD).

2) PREPROCESSING AND FEATURE ENGINEERING

The practices of data preprocessing and feature engineering follow a systematic variability across modalities but have the common goal of converting raw input data into a representation that can be used in early predictive modeling [38].

a) Log/Clickstream Data

For interaction logs and clickstream data, common preprocessing tasks include session structuring, time normalization, normalization of attempt and hint rates, and the creation of early aggregate features based on initial interactions [13]. The resultant features often include mastery indicators (e.g., correct streaks and retention-test performance), behavioral indicators (e.g., repeated errors and fast guesses), and hierarchical representations between problem-level and student-level aggregates. Beyond static correctness metrics, process-level learning behaviors extracted from interaction logs can reveal critical distinctions in students’ underlying problem-solving strategies, such as differentiating systematic reasoning from unreflective trial-and-error, and their direct relationship with learning performance [39], [40].

b) Eye-Tracking Data

Analogous, though modality-specific processing is done to eye-tracking data, including fixation detection, converting

gaze coordinates to areas of interest (e.g., textual features or answer choices), and deriving measures of dwell time, saccade magnitude, and inter-region transitions, and temporal alignment of these measures to decision points. The features of gaze are especially helpful for tracking strategy changes and misconceptions in real time, and they can be fed into classical classifiers to issue real-time warnings of potential pitfalls in solving mathematical problems.

3) MULTIMODAL FUSION AND ALIGNMENT

Additionally, multimodal fusion involves aligning event-driven logs, high-frequency gaze streams, and frame-based audiovisual data. Successful multimodal fusion requires a close synchronization and alignment of heterogeneous streams of data, which is a daunting challenge considering the very different time properties of each modality. Interaction logs have an inherently event-based nature; eye-tracking data is recorded at a high rate (usually 60–120 Hz); and audiovisual data is frame-based. The multimodal machine learning architectures address these complexities by aligning representations, maintaining temporal dependencies via temporal synchronization modules, and employing fusion strategies that preserve temporal dependencies. In the context of Multimodal Learning Analytics (MMLA), defining end-to-end pipelines is essential to both predictive modeling and to produce timely and interpretable signals that can be used in real-time to support instructional interventions. Failure to

address alignment issues risks noise concatenation, thereby reducing model generalizability and practical usefulness in education [22].

B. PREDICTIVE MODELING METHODS FOR EARLY DETECTION

The studies selected for review used a variety of methods to develop predictive models. Nine studies used traditional machine learning techniques, three used deep learning techniques, and five used a hybrid of both. One study, by van Viersen et al. [16], relied on exploratory quantitative/qualitative coding rather than a formal machine learning classifier.

1) CLASSICAL AND GLASS-BOX MODELS

Glass-box models are classical machine-learning baselines and intrinsically interpretable models that are still commonly used in educational prediction tasks because of their auditability and ability to work with tabular features. Methods such as logistic regression, support vector machines, k-nearest neighbors, random forests, and gradient boosting have shown good performance in early-detection settings, especially when short time windows of log or gaze data are used. Decision trees can be especially useful for detecting wheel-spinning behavior, with interpretable splits based on variable measures of hint use, correctness, and time-on-task.

Five of the nine traditional machine learning studies used transparent models (glass-box models), most commonly decision trees, while the remaining four paired traditional

methods such as gradient boosting or support vector machines with post-hoc explanation techniques such as SHAP or LIME. Decision-tree models in particular have been used to detect wheel-spinning, as illustrated in the studies by Wang et al. [11] and Kai et al. [12].

2) DEEP LEARNING MODELS

Deep learning models, in contrast, are well-suited to handling unstructured, high-frequency sequences, such as raw event logs, gaze time series, and audiovisual representations. Bayesian Knowledge Tracing (and its deep-learning variants) is one of the key choices in temporal models, and the richness of features, as well as the time structure of the data, plays a vital role in its performance. Predicted probabilities must be calibrated to support meaningful intervention decisions. Moreover, it is suggested that evaluation metrics such as the Matthews correlation coefficient should be used to effectively address class imbalance in early-detection tasks [41].

Deep learning methods were employed in 3 of the 18 sources, demonstrating strong performance in classifying unstructured time series data.

3) SUMMARY OF INCLUDED STUDIES

Sample sizes across the included studies ranged from 11 to 419,832 students, as shown in Table 2. Notably, 14 of the 18 sources relied on proxy definitions of MLD, such as behavioral or performance-based indicators, rather than clinically validated diagnostic criteria.

TABLE 2
CHARACTERISTICS AND FINDINGS OF THE 18 INCLUDED SOURCES

Study (Year)	Modalities	Sample Size	Target / Label Definition	Models Used	Explainability Approach	Performance & Earliness	Key Limitations
Wang et al. (2020) [11]	ITS log data (clickstream: attempts, correctness, hints, time)	26,522 students; 1,088 problem sets	3-class persistence: non-persistence vs. productive persistence vs. wheel-spinning	Decision trees (early-window models)	Intrinsically interpretable (tree splits + feature inspection)	Macro AUC ~0.62 (first 9 opportunities); Wheel-spinning vs. productive persistence detectable by first 3 opportunities	Strong platform dependence; limited modalities; requires cross-platform validation
Kai et al. (2018) [12]	ASSISTments Skill Builder logs	Large-scale (thousands of students)	Wheel-spinning vs. productive persistence	Decision trees	Intrinsic (decision tree paths)	High accuracy in distinguishing states within early opportunities	Platform-specific rules; proxy labels may not generalize
Kennel et al. (2024/2025) [14]	Eye-tracking (+ retrospective think-aloud)	148 students (143 in analysis)	Multiclass math difficulties (e.g., graphical differentiation misconceptions)	KNN, Random Forest, SVM	Feature-based (engineered gaze metrics + statistical tests)	MCC > 0.7 at ~45s (multiclass); some binary splits >0.7 at 15–35s	Requires eye-tracking hardware; guessing category poorly predicted; needs more data for rare difficulties
van Viersen et al. (2013) [16]	Eye-tracking + number line tasks + working memory	1 dyscalculia case + 10 controls	Strategy differences in number line estimation (dyscalculia vs. typical)	Quantitative/qualitative coding (not full ML)	Strategy-level interpretability (fixation patterns)	Discriminative process signals (logarithmic vs. linear fit)	Exploratory case study; small sample; not scalable ML classifier
Djekourmane et al. (2026) [42]	Survey-based features including student demographics, socioeconomic	26,718 15-year-old students.	Mathematics performance based on the mean of ten plausible values.	Random Forest, LightGBM, XGBoost, AdaBoost, Elastic Net, and	SHAP (SHapley Additive exPlanations) values and SHAP	XGBoost achieved the best performance with an R-squared of 0.5703 and an	Cross-sectional design limits causal inference. The focus on East Asian systems limits external validity.

	status, instructional quality, and school experiences.			Linear Regression.	interaction analyses.	root mean square error of 65.06.	Relies heavily on student self-reports, which can be subject to bias.
Rohani et al. (2024) [43]	Clickstream/log features from math assignments	Competition dataset (large)	Correctness prediction & inferred difficulty patterns	CatBoost trees (“ClickTree”)	Intrinsic tree paths + SHAP-compatible feature importance	AUC $\approx$ 0.788 (EDM Cup 2023, 2nd place)	Proxy outcome (item correctness); generalization depends on task design
El Hmimdi et al. (2024) [44]	Eye movement time series	4,243 time series (children)	Learning disorders, including dyscalculia	Deep learning on time series	Post-hoc (feature importance from model)	Strong classification across learning disorder groups	Broad pathology grouping; privacy constraints; needs classroom validation
Swamy et al. (2022) [45]	Multimodal (logs + assessment data)	Massive open online course / online math learners	Student performance / at-risk prediction	LSTM + classical ML	Post-hoc (LIME, SHAP, counterfactuals)	High predictive accuracy with local/global explanations	Focus on general performance rather than math-specific misconceptions; potential data leakage risks
Hornos-Arias et al. (2025) [3]	Serious game logs (interaction + performance)	Children 5–12 (multiple datasets)	Developmental dyscalculia subtypes	ML/DL models on game data	Feature importance from best models	Promising accuracy for developmental dyscalculia screening	Game-specific; needs broader classroom validation
Odoom et al. (2025) [37]	Multimodal (logs, assessments, basic sensors)	Elementary/basic education students	Early learning difficulties (math component)	3-Tier ML framework (hybrid classical + ensemble)	Post-hoc interpretability layers	Scalable early risk prediction with good accuracy	Broad learning difficulties focus; limited math-specific metrics reported
Patel et al. (2025) [36]	Multimodal (handwriting, EEG, speech, behavior)	Multiple datasets	Learning disabilities including dyscalculia	ANN, SVM, CNN, hybrid models	Hybrid model interpretability	Superior accuracy with multimodal hybrids	Review-heavy; some datasets not purely math-focused
Sharada et al. (2017) [46]	ITS log data covering student performance, response speed, and general learning information.	623,904 data records from 14,512 unique students.	Successful retention versus wheel-spinning.	Deep Learning (Feed Forward Neural Networks), Random Forests, and Logistic Regression.	Feature importance ranking, highlighting variables like average standard response time and prior problem count.	Random Forest was the best model, achieving an AUC of 0.87 and a log loss of 0.235. Identifies wheel-spinning cases at an early stage in the enhanced mastery cycle.	Suffers from a class imbalance problem because wheel-spinning cases are the minority class.
Maniktala et al. (2020) [13]	ITS logs + hint usage patterns + help-seeking behavior	Multiple semesters of college students	Wheel-spinning detection & help-need prediction	Gradient boosting (XGBoost) + classical ML	Feature importance & tree visualization (SHAP-compatible)	Strong early detection with actionable intervention suggestions	Domain is formal logic; proxy labels; needs validation on math-specific platforms
Zhang et al. (2022) [15]	Log data + cognitive features	Large online learning dataset	Cognitive strategies in math	Logistic Regression, Random Forest	Intrinsic (trees) + feature coefficients	Good detection of cognitive patterns	Focus on strategies rather than clinical MLD
Owen et al. (2019) [47]	Game-based learning logs	Large-scale adaptive game dataset	Wheel-spinning in early math skills	Decision trees / ensemble classifiers	Intrinsic (tree paths)	High early separation of persistence states	Game-specific; limited to young learners
Kennel and Ruzika (2024) [48]	Raw eye-tracking (gaze) data	139 students, solving a math problem	Multiclass classification of math difficulties during problem-solving	Temporal Convolutional Network on sequential raw gaze data	Gaze feature importance	~80% accuracy; real-time dynamic classification	Requires eye-tracking hardware; unclear explainability; single-task context, needs further validation
Psyridou et al. (2024) [49]	Tabular/psychometric, 49 kindergarten-age (age 6) variables: arithmetic skills,	2000 Grade 6 students	Grade 6 math difficulties	Neural network, two-step: full 49-variable model, then a reduced model refit on only the top	Post-hoc variable importance from the trained multi layered perceptron	AUC = .818 (full model); AUC = .776 (reduced, top-predictors model). Top predictors: arithmetic skills,	Method inferred not confirmed; Global-only, no case-level explanation; No validation of explanation usefulness

	cognitive skills, home learning environment, parental measures, motivation, behavioral problems, gender			predictors the first model surfaced		number concepts, spatial skills, counting, attention-orientation.	
Zhang et al. (2019) [32]	ITS log data (Algebra and Geometry tutoring systems)	3 datasets; 132,551 + 419,832 + 8,175 student knowledge component pairs	Wheel-spinning vs. mastery, compared under two competing definitions (three-correct-in-a-row vs. Predictive Stability)	Logistic Regression (single-feature and full-feature), Random Forest, 5-layer ANN, 3-layer LSTM	Feature importance (Random Forest, Gini-based)	Random Forest most consistent across datasets; 77–91% precision by step 4; simple single-feature model competitive with complex ones	Two wheel-spinning definitions disagree substantially (14–42% agreement); highlights proxy-label inconsistency

4) MULTIMODAL FUSION STRATEGIES

When multiple types of data are available through multimodal fusion strategies, the methods generally fall into four known categories [50].

Early fusion takes features of various modalities, concatenated together before model training, allowing the algorithm to learn joint representations initially. Late fusion, on the other hand, involves training separate models on each modality and combining their predictions so that modality-specific strengths are retained at the cost of simpler integration. Intermediate or hybrid fusion is based on the idea of learning joint representations by using mechanisms like co-attention, which balances early and late fusion. Co-learning uses the insights of one modality to improve learning in another, transferring learned representations from a data-rich modality to guide a sparser or noisier one. The role of appropriate fusion strategies to handle the issues of representation, alignment, and data fusion is continuously emphasized in surveys of multimodal machine learning and MMLA [50]. Inappropriate fusion can lead to noise concatenation and a decrease in generalization ability [51].

a) Practical Applications

- 1) *Sliding-Window Classification*: In this method, models learn from data collected during the first (k) attempts or the first (t) seconds. They are then tested over longer time frames. This method is widely used to identify when students fall behind or anticipate hurdles based on gaze data.
- 2) *Time-to-event/Survival Prediction*: This method predicts how long it will take for an event to be realized (for instance, when a student will fail or succeed). Though this is primarily used in healthcare, it can also become useful in education, as it shows prompt decision-making.

b) Metrics on Evaluation and Early Detection Validation

In early detection in education, evaluation metrics need to accurately reflect how well the model classifies and the impact of its decisions. Discrimination metrics such as Area Under the Curve (AUC), Precision-Recall Area Under the Curve (PR-AUC), accuracy, and F1 scores are widely used.

However, they might fail under conditions of a class imbalance and changing trends over time. The Matthews correlation coefficient (MCC) is useful for predicting math problems as they emerge.

C. EXPLAINABILITY METHODOLOGY, HOW IT WORKS, AND HOW IT IS USED IN EDUCATION

This section proceeds in three parts: we first summarize the general categories of explainability methods used across the included studies, then examine what makes an explanation genuinely useful in an educational context specifically, and finally present a four-stage framework synthesizing how explainability can be applied across an early-intervention pipeline.

1) METHODS OF EXPLAINABILITY

Explainability methods can be intrinsic, encompassing glass-box models such as decision trees and generalized additive models, or post hoc. Local and global feature attributions provided by methods such as SHAP and LIME are effective, especially for tabular log and clickstream data. Gradient-based and saliency methods (including Integrated Gradients and Gradient-weighted Class Activation Mapping) are designed to assist deep learning models trained on gaze or image data. Moreover, the concept-based approach, along with counterfactual explanations, allows exploring the what-if situation and, consequently, connects the model's decisions to educationally relevant concepts.

When it comes to education, explanations should not only meet the demands of various stakeholders but also align with pedagogical goals and promote a sense of collective responsibility, rather than replacing the teacher's judgment. According to recent reviews, SHAP and LIME are among the most commonly used methods in educational studies, as they have been shown to promote transparency and build trust [25], [52]. The conceptualization of early intervention explainability can be implemented through four stages, including detection, diagnosis, decision support, and monitoring, which can be clearly and practically explained based on multimodal models. Indicatively, eye-tracking studies generate strategy signatures that can be interpreted to

improve the prediction and diagnosis of specific learning impairments. Simultaneously, log-based detectors indicate tendencies like excessive hint-seeking and repeated errors, thus informing the changes in the instructional design [53].

2) EDUCATION-SPECIFIC AND USABLE REQUIREMENTS FOR USEFUL EXPLANATIONS

Machine-learning predictions in education are rarely used as final, autonomous decisions. Instead, they typically guide teaching, provide feedback, or inform resource allocation. One major concept in XAI in education is that explanations should accommodate multiple parties, education objectives, and behavior. In other words, explanations ought to be about shared responsibility, not substituting teacher judgment. Furthermore, cultivating critical AI literacy and human-centered AI (HCAI) frameworks within teacher preparation is essential, as educators cannot effectively exercise human-in-the-loop oversight or navigate the pedagogical trade-offs of predictive models without structured professional learning that builds their interpretive and critical capacities [54]. The results of a review of XAI employed in educational research from 2019–2024 suggest that tools such as SHAP and LIME rank among the most popular. The goal is for these tools to help promote transparency and trust in predictive models. Nonetheless, the field also lacks the standard benchmarks for judging the effectiveness of these explanations for teachers and students: their comprehension and trust, as well as the steps taken in response to them [25], [52].

3) FOUR-STAGE EXPLAINABILITY FRAMEWORK FOR EARLY INTERVENTION

Early-intervention machine learning explainability can be deconstructed into four different stages:

- 1) *Detection*: Rapidly detect risk or hindrance.
- 2) *Diagnosis*: What the problem is, or what is the particular misconception or obstacle contributing to it.
- 3) *Decision Support*: Suggesting how best to use the available interventions.
- 4) *Monitoring*: Measuring how beneficial the interventions are and adapting risk assessments as needed.

An illustration of this technique would be to use eye-tracking to predict math difficulty. Studies have found that eye movements can inform predictive strategies and provide clear “strategy signatures” that increase predictive accuracy. This also aids early diagnosis of specific learning difficulties. Log-based interpretable detectors, for instance, are easily detectable and may help identify patterns (e.g., when hints are used too heavily and lead to mistakes), indicating that the design or instructional support may need to be strengthened.

While this framework outlines how explainability can function across an early-intervention pipeline, its real-world deployment raises pedagogical, technical, and ethical questions, which are examined in the Discussion in relation to RQ4.

IV. DISCUSSION

This review explores the possibility of using new educational data and XAI to detect mathematical learning problems early on. The research evidence shows a strong trend: single-method evaluation tools are increasingly being supplemented, and in some cases replaced, by methodologies that combine multiple data types. For example, interaction logs and eye tracking data can be used together, in real-time, to give insights into students’ misunderstandings and behaviors that affect their learning, for example if they struggle with a problem-solving process. This improvement provides a glimmer of hope for more effective early interventions and motivates teachers and researchers to find new solutions.

However, there are still some problems to be addressed. While tools like SHAP and LIME help interpret complex predictions, there is currently no uniformity in how these tools are tested to ensure educators are confident and comfortable in their understanding and trust. Furthermore, the widely varied research methods used in the studies make it difficult to make broad platform conclusions. This means there is a need for high standards of validation for specific use in education, and for a guarantee that predictive tools do not compound current issues of bias in identifying mathematical learning difficulties.

A. RQ1: WHAT ARE THE MOST FREQUENTLY USED TYPES OF MULTIMODAL DATA AND LEARNING ANALYTICS TO SUPPORT EARLY IDENTIFICATION OF MATHEMATICS LEARNING DIFFICULTIES?

The evidence suggests that richer, multimodal data can offer insights beyond what conventional assessments alone capture. The incorporation of multimodal data, including interaction logs and eye-tracking data, has been promising for real-time insights into students’ learning challenges. An eye-tracking system has been introduced alongside the log, helping identify distinguishing “strategy signatures” that further improved predictive accuracy and allowed the identification of individual mathematical difficulties at an earlier stage.

Educators and researchers can better understand the implications of these types of data and how they can be used at the beginning of a detection, thus providing some support for existing researchers who have suggested a multidimensional perspective on understanding student behaviors. This review highlights the importance of synchronizing across time to better understand the origins of learning difficulties and the potential of this process to develop novel opportunities for targeted early interventions to prevent or reduce mathematical learning difficulties.

B. RQ2: WHAT PREDICTIVE PERFORMANCE AND EVIDENCE OF EARLINESS HAVE BEEN REPORTED FOR MODELS USED TO IDENTIFY MATHEMATICS LEARNING DIFFICULTIES OR RELATED RISK INDICATORS?

When addressing the second research question about predictive modeling, the studies reviewed indicate that machine learning models can be used to detect learning

failures, such as overreliance on hints. This pattern of excessive hint reliance parallels a related concern in the broader AI-in-education literature, sometimes described as cognitive offloading, where reliance on automated support may reduce the mental effort students invest in constructing their own understanding [55]. However, most earlier studies remain centered on traditional machine learning methods and prioritize interpretability and simplicity over complexity.

These results are consistent with existing studies, but a problem remains: model performance differs considerably across learning platforms. Cross-platform generalizability is a critical problem to solve to ensure trust in AI tools that can be relied upon across different educational settings and to build future trust in novel innovations. A contribution to this challenge can spur a sense of purpose and motivation to create more dependable solutions.

C. RQ3: WHAT IS THE LEVEL OF INTEGRATION OF XAI FRAMEWORKS IN THESE PREDICTIVE MODELS TO ENHANCE EDUCATORS' UNDERSTANDING OF THE OUTCOMES?

In terms of explainability, the need for transparency of AI applications in educational contexts is rising. There are practical challenges to using these tools in mathematics education, including how to integrate them into existing instructional frameworks and how usable they are for teachers.

Regarding the third research question, how XAI can contribute to the role, this review underlines the necessity of empirical testing of the XAI framework with end users, such as teachers and students. Demonstrating to educators, especially those working with special education students, that are interested in understanding and enhancing explainability, can promote trust and confidence. It is crucial to understand these challenges to respond to stakeholders' concerns and to highlight the need to prioritize tackling real-world hurdles for future progress in explainability.

D. RQ4: WHAT ARE THE GENERAL PEDAGOGICAL, TECHNICAL, AND ETHICAL ISSUES WITH THE USE OF SUCH PREDICTIVE SYSTEMS IN THE EDUCATIONAL CONTEXT IN THE REAL WORLD?

Addressing the fourth research question reveals some serious pedagogical, technical, and ethical issues that will hinder the actual application of predictive analytics in education. It is worth clarifying that the included studies do not all examine the same population: some, such as van Viersen et al. [16], compare a small clinically diagnosed group against typical peers, while others, such as Wang et al. [11] and Kai et al. [12], screen large general student populations using a tutoring platform, without pre-selecting by diagnosis. This distinction matters for the ethical and equity issues discussed below, since a screening tool applied broadly to an entire student population carries different fairness and false-positive risks than a diagnostic comparison within an already-identified clinical group.

Under the European Union AI Act [56], certain educational AI systems may qualify as high-risk depending on their intended use. Their deployment must also account for

applicable requirements under FERPA [57], COPPA [58], and GDPR [59] concerning student records, children's personal information, privacy, transparency, and responsible data handling.

Currently, the field lacks sufficient formal clinical validation for diagnosing MLD or dyscalculia, and instead relies on proxy indicators, such as low scores on tasks or high error rates. This dependence makes it difficult to establish a clear connection between predictions and known cognitive disorders. To mitigate this label opacity, researchers are beginning to employ generative AI agents not as automated coders, but as interactive refinement tools to systematically surface, formalize, and document the implicit qualitative criteria human experts use to classify the cognitive demand levels of mathematical tasks [60].

Pedagogical concerns about fairness and equity are also introduced by these systems. Algorithmic identification tools are vulnerable to label bias, measurement bias arising from limited access to appropriate diagnostic instruments, and unequal access to diagnosis by class, language, and disability accommodations. This risk is particularly pronounced when interpreting process-level indicators such as eye-tracking and gaze dynamics; as empirical evidence in multimodal task environments demonstrates, demographic factors such as gender can significantly modulate gaze patterns, visual search behaviors, and problem-solving strategies without reflecting differences in actual learning ability [61]. Consequently, predictive models that rely on rigid gaze metrics risk misinterpreting legitimate strategic variations as cognitive deficits or persistence failures, thereby exacerbating algorithmic bias. A related risk is intervention bias, in which a risk score alters educators' behavior in ways that may worsen existing inequalities. These biases can negatively impact teachers' interactions with students, increase the likelihood of stigmatizing students, and unintentionally contribute to inequities in the classroom. This concern is further compounded across STEM educational AI applications, where unexamined algorithmic bias and a lack of model transparency risk reinforcing gender and demographic stereotypes, underscoring the necessity of explainable AI architectures paired with critical AI literacy [62]. Reflecting these concerns, empirical work in educational data mining has begun shifting toward fair-prediction methods for large-scale online learning data [63], [64]. However, resolving bias at the algorithmic level is only half the battle. To successfully navigate these socio-pedagogical tensions on the ground, researchers advocate for anchoring classroom implementation in the modified AI-TPACK framework [65], which equips STEM educators with the specialized technological knowledge and ethical competencies needed to critically evaluate, adapt, and co-design AI-driven diagnostic tools without compromising student equity. Applied to this context, ethical competence is treated not as a supplementary add-on but as a core part of how teachers interpret and act on the outputs of AI-based early-detection tools.

E. LIMITATIONS

This review provides an overview of the state of the field on multimodal data and XAI for the identification of early MLD. It does, however, recognize that there are some limitations:

- 1) The results are influenced by the methods used to search the databases, the time period analyzed, and the language used (only publications in English are included). Therefore, some literature in other languages, regional journals, or publications outside the specified timeframe may not have been included.
- 2) *Study Selection and Data Extraction by Reviewers:* Screening was performed by two primary reviewers, with a third consulted to resolve disagreements. While this is a standard approach, a small reviewer team may still introduce selection bias compared to larger review teams.
- 3) *Heterogeneity of Included Studies:* The included studies in this review have significantly different methods, which limits the extent to which findings can be directly compared across studies.
- 4) A major drawback in the literature is the reliance on proxy labels (e.g., low task scores, high error rates) rather than a formal, valid clinical diagnosis of MLD or dyscalculia. This dependency makes it difficult to establish a relationship between predictive models and well-defined cognitive learning disorders.
- 5) XAI and educational data mining are emerging fields that are rapidly developing. New deep learning methodologies and changing regulatory landscapes, like the European Union AI Act, mean that some of the latest methods or governance approaches may not be well captured in this review.

V. GAPS, CHALLENGES, AND FUTURE RESEARCH AGENDA

Despite significant research on multimodal learning analytics (MMLA) and explainable artificial intelligence (XAI) in education, there is limited evidence on how to effectively integrate “multimodal + explainable + math difficulty” approaches. Few studies have combined diverse data-sensing methods with explanations that teachers can use to identify math learning deficits early. Existing systems are often more concerned with predicting things like grade correctness or course grades, than with the specific cognitive mechanisms underlying math learning challenges.

The study of issues such as wheel-spinning and persistence in learning may be helpful, but they vary significantly between platforms and studies. In terms of generalizability and wider applicability, there was no cross-platform replication, so the same model doesn’t always translate to tutoring systems, schools, or other educational contexts. Furthermore, gaze models that follow attention over time may be overly task-specific unless deliberately designed for generalization. Evaluations frequently lack a clear mechanism to control for problems such as data leakage.

Relying solely on normative metrics is insufficient for informing intervention choices.

Generalizing from broader machine learning research, leakage can lead to unreliable performance claims; thus, stricter education-specific evaluation standards must be adopted. To build a strong research agenda, we suggest establishing privacy-sensitive, multimodal datasets that future researchers can use to develop and validate MLD early-detection models, rather than proposing any single diagnostic tool for immediate classroom use. In addition, scalable log datasets can be used with targeted, more accurate high-resolution methods (e.g., gaze tracking during tasks) and proper labeling to ensure they include diagnostics. More broadly, parallel work in other educational domains, for example, explainable ML frameworks for pedagogical feedback in language pronunciation instruction [66], suggests that interpretable, feedback-oriented model architectures developed outside mathematics education may offer transferable design patterns for future MLD-specific systems. Similar questions about balancing automation with educator judgment have emerged in other assessment contexts; for instance, comparisons of teacher-created, AI-generated, and AI-assisted test design in foreign language education have found that AI functions best as a support tool for educators rather than a replacement, particularly for those with less experience in a given domain [67]. It is also advisable to use available open datasets, including ASSISTments logs and math-related neuroimaging datasets, while adding multimodal elements that satisfy the governance requirements such as those discussed in Section IV-D.

Equally critical for real-world adoption is bridging predictive analytics with participatory classroom design; engaging educators in the co-design and iterative refinement of instructional tasks ensures that algorithmic early warnings translate into ecologically valid, actionable interventions that fit standard curricular routines [68].

Earliness should be considered as the primary outcome for measurement. Performance information should include the time (in seconds and opportunities) required for observations and the lead time (time needed to intervene), or the pace at which interventions are available. A common practice for evaluating bias in standard reporting and evaluation should also be grounded in best practices for predictive modeling.

Tools such as TRIPOD+AI for reporting and PROCAST+AI for bias assessment are recommended, including disaggregating results by subgroup to confirm that model validity holds for disadvantaged groups and not just the majority group, checking time series for outliers, and testing for leakage. We should value the validity of these explanations rather than just their plausibility, focusing on whether they are genuinely useful to educators and learners. The quality of explanations ought to be measurable, including checking for user comprehension, trust calibration to avoid trust arising from excessive confidence [64], action quality, whether appropriate interventions follow, and possible harms such as stigma or unfairness. This need for active trust calibration aligns with research in university

mathematics, which shows that learners must cultivate epistemic vigilance through continuous coherence-checking to prevent automation bias and maintain cognitive agency over AI-mediated problem-solving [69]. For high-stakes decisions in particular, the explanations provided must be interpretable.

Less transparent or complex models should be reserved for very low-stakes situations, unless there is a strong reason to avoid them even there. Where AI has implications for access, placement, or labeling, transparency in modeling is necessary, and the use of such models should be approached conservatively, consistent with the regulatory emphasis on transparency in high-stakes contexts discussed in Sections III-C and IV-D. As regulatory frameworks for educational AI continue to evolve, sustained attention to these emerging requirements will be essential for ensuring that future MLD detection systems remain both effective and accountable.

VI. CONCLUSION

This systematic review examined 18 sources at the intersection of multimodal educational data, predictive modeling, and explainable machine learning for the early identification of mathematics learning difficulties. In answer to RQ1 and RQ2, the evidence shows that interaction logs and clickstream data remain the most widely used and most scalable modality, particularly for detecting wheel-spinning and other persistence-related behaviors, while eye-tracking offers complementary, process-level evidence of misconceptions and strategic differences that conventional post-hoc assessments cannot capture. Predictive models built on these data sources, including interpretable tree-based methods and more complex classifiers, demonstrate the potential for identifying risk indicators from relatively early stages of student interaction; however, direct head-to-head comparisons with traditional standardized assessments remain limited.

With respect to RQ3, explainability methods, particularly SHAP and LIME, are increasingly paired with these predictive models to support educator trust and transparency, and a four stage framework of detection, diagnosis, decision support, and monitoring offers a practical structure for translating model outputs into classroom action. However, the field still lacks standardized ways of evaluating whether these explanations are actually comprehensible, trusted, and acted upon by teachers and students, and few studies have tested their frameworks directly with these end users. Turning to RQ4, the review identifies substantive pedagogical, technical, and ethical barriers to real-world deployment. Included studies vary considerably in the populations they examine, from small clinically diagnosed samples to large, unselected student populations, a distinction with direct consequences for fairness and false-positive risk when tools are deployed at scale. The field's continued reliance on proxy labels rather than clinical diagnoses, combined with risks of temporal and group leakage, further limits confidence in reported performance.

Under the European Union AI Act [56], certain educational AI systems may be classified as high-risk depending on their intended use; separately, FERPA [57], COPPA [58], and GDPR [59] establish applicable protections concerning education records, children’s personal information, privacy, transparency, and lawful data handling.

Taken together, these findings indicate that while multimodal, explainable predictive modeling holds real promise for the early identification of mathematics learning difficulties, realizing that promise will require closing three gaps: building privacy-preserving multimodal datasets that support cross-platform validation, establishing standardized methods for evaluating explanation usefulness to educators and families rather than assuming interpretability equates to utility, and grounding future systems in participatory design with teachers so that algorithmic outputs translate into pedagogically sound, ecologically valid classroom practice. Addressing these gaps all at once, rather than in isolation, is the most direct path toward predictive tools that are not only accurate but trustworthy, equitable, and genuinely usable in mathematics classrooms.

APPENDIX. LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
BKT	Bayesian Knowledge Tracing
CNN	Convolutional Neural Network
COPPA	Children’s Online Privacy Protection Act
DKT	Deep Knowledge Tracing
DL	Deep Learning
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
EDM	Educational Data Mining
EEG	Electroencephalography
FERPA	Family Educational Rights and Privacy Act
GDPR	General Data Protection Regulation
HCAI	Human Centered Artificial Intelligence
ICD-11	International Classification of Diseases, 11th Revision
ITS	Intelligent Tutoring System
KNN	K-Nearest Neighbors
LAK	Learning Analytics and Knowledge
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
MCC	Matthews correlation coefficient
ML	Machine Learning
MLD	Mathematics Learning Difficulties
MMLA	Multimodal Learning Analytics
NIST	National Institute of Standards and Technology
PR-AUC	Precision-Recall Area Under the Curve
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST	Prediction Model Risk of Bias Assessment Tool
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XAI	Explainable Artificial Intelligence

REFERENCES

- [1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA, USA: American Psychiatric Association, 2013, doi: 10.1176/appi.books.9780890425596.
- [2] World Health Organization, *International Statistical Classification of Diseases and Related Health Problems*, 11th ed. Geneva, Switzerland: World Health Organization, 2022. [Online]. Available: <https://icd.who.int/>
- [3] J. Hornos-Arias, S. Grau, and J. M. Serra-Grabulosa, "Early detection and intervention of developmental dyscalculia using serious game-based digital tools: A systematic review," *Information*, vol. 16, no. 9, Sep. 2025, Art. no. 787, doi: 10.3390/info16090787.
- [4] S. Haberstroh and G. Schulte-Körne, "The cognitive profile of math difficulties: A meta-analysis based on clinical criteria," *Front. Psychol.*, vol. 13, Mar. 2022, Art. no. 842391, doi: 10.3389/fpsyg.2022.842391.
- [5] R. J. Daker, S. U. Gattas, H. M. Sokolowski, A. E. Green, and I. M. Lyons, "First-year students' math anxiety predicts STEM avoidance and underperformance throughout university, independently of math ability," *npj Sci. Learn.*, vol. 6, no. 1, Jun. 2021, Art. no. 17, doi: 10.1038/s41539-021-00095-7.
- [6] T. Jungert and U. Andersson, "Self-efficacy beliefs in mathematics, native language literacy and foreign language amongst boys and girls with and without mathematic difficulties," *Scand. J. Educ. Res.*, vol. 57, no. 1, pp. 1–15, Feb. 2013, doi: 10.1080/00313831.2011.621140.
- [7] P. Atwell, D. Lavin, T. Domina, and T. Levey, "New evidence on college remediation," *J. Higher Educ.*, vol. 77, no. 5, pp. 886–924, Sep. 2006, doi: 10.1080/00221546.2006.11778948.
- [8] President's Commission on Excellence in Special Education, "A new era: Revitalizing special education for children and their families," U.S. Dept. Educ., Office Special Educ. Rehabil. Services, Washington, DC, USA, Jul. 2002.
- [9] J. E. Beck and Y. Gong, "Wheel-spinning: Students who fail to master a skill," in *Artificial Intelligence in Education* (Lecture Notes in Computer Science), vol. 7926, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Berlin, Germany: Springer, 2013, pp. 431–440, doi: 10.1007/978-3-642-39112-5_44.
- [10] R. Gersten, N. C. Jordan, and J. R. Flojo, "Early Identification and Interventions for Students With Mathematics Difficulties," *J Learn Disabil.*, vol. 38, no. 4, pp. 293–304, Jul. 2005, doi: 10.1177/00222194050380040301.
- [11] Y. Wang, S. Kai, and R. S. Baker, "Early detection of wheel-spinning in ASSISTments," in *Artificial Intelligence in Education* (Lecture Notes in Computer Science), vol. 12163, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Cham, Switzerland: Springer, 2020, pp. 574–585, doi: 10.1007/978-3-030-52237-7_46.
- [12] S. Kai, V. Almeda, R. S. Baker, C. Heffernan, and N. Heffernan, "Decision tree modeling of wheel-spinning and productive persistence in skill builders," Jun. 2018, doi: 10.5281/zenodo.3344810.
- [13] M. Manikala, C. Cody, A. Isvik, N. Lytle, M. Chi, and T. Barnes, "Extending the Hint Factory for the assistance dilemma: A novel, data-driven HelpNeed predictor for proactive problem-solving help," *J. Educ. Data Mining*, vol. 12, no. 4, pp. 24–65, Dec. 2020, doi: 10.5281/zenodo.4399683.
- [14] K. Kennel, S. Ishimaru, S. Küchemann, S. Steinert, J. Kuhn, and S. Ruzika, "Gaze-based prediction of students' math difficulties—A time dynamic machine learning approach to enable early individual assistance," *Int. J. Artif. Intell. Educ.*, vol. 35, no. 4, pp. 1724–1748, Dec. 2025, doi: 10.1007/s40593-024-00447-5.
- [15] J. Zhang *et al.*, "Using machine learning to detect SMART model cognitive operations in mathematical problem-solving process," Dec. 2022, doi: 10.5281/zenodo.7304763.
- [16] S. van Viersen, E. M. Slot, E. H. Kroesbergen, J. E. van't Noordende, and P. P. M. Leseman, "The added value of eye-tracking in diagnosing dyscalculia: A case study," *Front. Psychol.*, vol. 4, 2013, Art. no. 679, doi: 10.3389/fpsyg.2013.00679.
- [17] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994, doi: 10.1007/BF01099821.
- [18] S. Slade and P. Prinsloo, "Learning analytics: Ethical issues and dilemmas," *Amer. Behav. Sci.*, vol. 57, no. 10, pp. 1510–1529, Oct. 2013, doi: 10.1177/0002764213479366.
- [19] C. Piech *et al.*, "Deep knowledge tracing," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 28, 2015, pp. 505–513.
- [20] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 4765–4774.
- [22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," 2017, *arXiv:1705.09406*.
- [23] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Mar. 2021, Art. no. n71, doi: 10.1136/bmj.n71.
- [24] P. J. Phillips *et al.*, "Four principles of explainable artificial intelligence," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Rep. NISTIR 8312, Sep. 2021, doi: 10.6028/NIST.IR.8312.
- [25] H. Khosravi *et al.*, "Explainable artificial intelligence in education," *Comput. Educ.: Artif. Intell.*, vol. 3, 2022, Art. no. 100074, doi: 10.1016/j.caeai.2022.100074.
- [26] J. D. T. Guerrero-Sosa, F. P. Romero, V. H. Menéndez-Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. A. Olivas, "A comprehensive review of multimodal analysis in education," *Appl. Sci.*, vol. 15, no. 11, May 2025, Art. no. 5896, doi: 10.3390/app15115896.
- [27] C. Conati, K. Porayska-Pomsta, and M. Mavrikis, "AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling," 2018, *arXiv*, doi: 10.48550/ARXIV.1807.00154.
- [28] M. Mohammadi, E. Tajik, R. Martinez-Maldonado, S. Sadiq, W. Tomaszewski, and H. Khosravi, "Artificial intelligence in multimodal learning analytics: A systematic literature review," *Comput. Educ.: Artif. Intell.*, vol. 8, Jun. 2025, Art. no. 100426, doi: 10.1016/j.caeai.2025.100426.
- [29] S. E. Israel, L. C. Assaf, and P. Abedini, "Where to From Here?," in *Handbook of Research on Reading Comprehension*, 3rd ed., New York: Routledge, 2026, pp. 532–542, doi: 10.4324/9781003624509-32.
- [30] H. Ouhaichi, D. Spikol, and B. Vogel, "Research trends in multimodal learning analytics: A systematic mapping study," *Comput. Educ.: Artif. Intell.*, vol. 4, 2023, Art. no. 100136, doi: 10.1016/j.caeai.2023.100136.
- [31] K. G. M. Moons *et al.*, "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, vol. 388, Mar. 2025, Art. no. e082505, doi: 10.1136/bmj-2024-082505.
- [32] C. Zhang *et al.*, "Early detection of wheel-spinning: Comparison across tutors, models, features, and operationalizations," in *Proc. 12th Int. Conf. Educ. Data Mining (EDM)*, Montreal, QC, Canada, 2019.
- [33] S. Torres-Ramos, R. A. Salido-Ruiz, A. Espinoza-Valdez, F. R. Gómez-Velázquez, A. A. González-Garrido, and I. Román-Godínez, "A brain connectivity characterization of children with different levels of mathematical achievement based on graph metrics," *PLoS ONE*, vol. 15, no. 1, p. e0227613, Jan. 2020, doi: 10.1371/journal.pone.0227613.
- [34] U. McCaskey, M. Von Aster, U. Maurer, E. Martin, R. O'Gorman Tuura, and K. Kucian, "Longitudinal Brain Development of Numerical Skills in Typically Developing Children and Children with Developmental Dyscalculia," *Front. Hum. Neurosci.*, vol. 11, p. 629, Jan. 2018, doi: 10.3389/fnhum.2017.00629.
- [35] R. S. Baker and A. Hawn, "Algorithmic Bias in Education," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, Dec. 2022, doi: 10.1007/s40593-021-00285-9.
- [36] H. Patel, V. Sheth, and M. Shah, "An early detection of learning disabilities using machine learning: A comparative study of models, datasets, and diagnostic strategies," *Intelligent Hospital*, vol. 1, no. 2, Dec. 2025, Art. no. 100022, doi: 10.1016/j.inhs.2025.100022.
- [37] S. Odom, E. O. Osei, E. Q. Effah, V. Bofo, S. Dusu, and A. K. Appiah, "A 3-tier machine learning framework for early detection of learning difficulties in basic setting," *Research Square*, preprint, Nov. 2025, doi: 10.21203/rs.3.rs-7876870/v1.
- [38] J. Makhoul and T. Mine, "Analysis of click-stream data to predict STEM careers from student usage of an intelligent tutoring system," *J. Educ. Data Mining*, vol. 12, no. 2, pp. 1–18, Aug. 2020, doi: 10.5281/zenodo.4008050.
- [39] M. Askarnia, J. Cheon, A. Abbaszadeh, M. Huang, and B. Thapa, "Analyzing debugging patterns using an AI-generated debugging framework in block-based programming," in *Proc. Soc. Inf. Technol. Teacher Educ. Int. Conf. (SITE)*, Waynesville, NC, USA: AACE, 2025,

- pp. 9–14. [Online]. Available: <https://www.learntechlib.org/primary/p/225985/>
- [40] M. Askarnia, J. Cheon, M. Huang, C. Cusson, and A. Abbaszadeh, “The impact of debugging strategies on student learning performance in block-based programming,” *J. Appl. Instructional Des.*, vol. 15, no. 2, 2026, doi: 10.59668/2579.26932.
 - [41] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.
 - [42] D. Djekourmane, W. Zhang, M. Aziku, D. Wu, and Y. Zhang, “Using machine learning to predict student mathematics performance in six East Asian countries: Evidence from PISA 2022,” *Front. Psychol.*, vol. 17, Jun. 2026, Art. no. 1840049, doi: 10.3389/fpsyg.2026.1840049.
 - [43] N. Rohani, B. Rohani, and A. Manataki, “ClickTree: A tree-based method for predicting math students’ performance based on clickstream data,” Sep. 2024, doi: 10.5281/zenodo.13627655.
 - [44] A. E. El Hmimdi, Z. Kapoula, and V. Sainte Fare Garnot, “Deep learning-based detection of learning disorders on a large scale dataset of eye movement records,” *BioMedInformatics*, vol. 4, no. 1, pp. 519–541, Feb. 2024, doi: 10.3390/biomedinformatics4010029.
 - [45] 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,” 2022, *arXiv:2207.00551*.
 - [46] N. Sharada, M. Shashi, and X. Xiong, “Identification of wheel spinning cases while learning and retaining a skill in intelligent tutoring systems,” *Int. J. Comput. Appl.*, vol. 179, no. 3, pp. 27–31, Dec. 2017, doi: 10.5120/ijca2017915906.
 - [47] V. E. Owen *et al.*, “Detecting wheel-spinning and productive persistence in educational games,” in *Proc. 12th Int. Conf. Educ. Data Mining (EDM)*, 2019.
 - [48] K. Kennel and S. Ruzika, “Real-time prediction of students’ math difficulties using raw data from eye tracking and neural networks,” *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. ETRA, pp. 1–19, May 2024, doi: 10.1145/3655607.
 - [49] M. Psyridou *et al.*, “Early prediction of math difficulties with the use of a neural networks model,” *J. Educ. Psychol.*, vol. 116, no. 2, pp. 212–232, Feb. 2024, doi: 10.1037/edu0000835.
 - [50] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, “A review on data fusion in multimodal learning analytics and educational data mining,” 2025, *arXiv*. doi: 10.48550/ARXIV.2511.20871.
 - [51] G. Kostopoulos, S. Kotsiantis, T. Panagiotakopoulos, and A. Kameas, “A survey of multimodal learning analytics: Data, methods, systems, and responsible deployment,” *Future Internet*, vol. 18, no. 3, Feb. 2026, Art. no. 115, doi: 10.3390/fi18030115.
 - [52] G. Türkmen, “The review of studies on explainable artificial intelligence in educational research,” *J. Educ. Comput. Res.*, vol. 63, no. 2, pp. 277–310, Apr. 2025, doi: 10.1177/07356331241310915.
 - [53] V. Hassija *et al.*, “Interpreting black-box models: A review on explainable artificial intelligence,” *Cogn. Comput.*, vol. 16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/s12559-023-10179-8.
 - [54] E. Baran, M. Dilek, M. Ziba, and X. Xiao, “Human-centered AI for teacher educators: Designing professional learning for critical AI literacy,” *Comput. Educ. Open*, vol. 11, Dec. 2026, Art. no. 100399, doi: 10.1016/j.cao.2026.100399.
 - [55] L. B. Hsin, A. Hutchison, P. Abedini, J. Amayou, and V. Mousavi, “Teaching adolescents to read purposefully with generative AI,” *J. Literacy Technol.*, vol. 27, no. 1, pp. 37–60, 2026.
 - [56] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
 - [57] Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g, 1974. [Online]. Available: <https://www.ecfr.gov/current/title-34/subtitle-A/part-99>
 - [58] Children’s Online Privacy Protection Act of 1998, 15 U.S.C. §§ 6501–6506, 1998. [Online]. Available: <https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa>
 - [59] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
 - [60] B. Lee and S. Azimi Asmaroud, “AI agents as a refinement tool for operationalizing coding process in the context of textbook analysis,” in *Proc. 20th Int. Conf. Learn. Sci. (ICLS)*, 2026, pp. 1347–1351.
 - [61] P. Abedini, F. Luo, R. Liu, and A. Hutchison, “Gender differences in gaze patterns during block-based programming: A multimodal literacy perspective,” in *Proc. 57th ACM Tech. Symp. Comput. Sci. Educ. V.2*, St. Louis, MO, USA, Feb. 2026, pp. 1217–1218, doi: 10.1145/3770761.3777332.
 - [62] C. Denson and N. Bayati, “How GenAI Mentoring Can Support Underserved Students in STEM,” *AIEL*, vol. 1, no. 1, pp. 189–202, 2025, doi: 10.70725/904313hkkwwg.
 - [63] J. Gardner, C. Brooks, and R. Baker, “Evaluating the Fairness of Predictive Student Models Through Slicing Analysis,” in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, Tempe AZ USA: ACM, Mar. 2019, pp. 225–234, doi: 10.1145/3303772.3303791.
 - [64] Y. Feldman-Maggor, M. Cukurova, C. Kent, and G. Alexandron, “The impact of explainable AI on teachers’ trust and acceptance of AI EdTech recommendations: The power of domain-specific explanations,” *Int. J. Artif. Intell. Educ.*, vol. 35, no. 5, pp. 2889–2922, Dec. 2025, doi: 10.1007/s40593-025-00486-6.
 - [65] A. Ataran and L. Shojaei, “Bridging AI-TPACK and STEM education: Navigating facilitators, barriers, and ethical dimensions through a systematic review,” *STEM*, vol. 24, pp. 43–72, Jun. 2026, doi: 10.32674/b2pss783.
 - [66] P. Jahanbin, “Modeling L1 influence on L2 pronunciation: An MFCC-based framework for explainable machine learning and pedagogical feedback,” 2025, *arXiv:2504.13765*.
 - [67] A. Abe, L.-T. Chen, E. Maleki, E. Teye, S. Motohashi, and Y. Kadowaki, “Integrating AI in Foreign Language Assessment: Practicality and Validity in University-level Japanese Education,” in *Proc. Soc. Inf. Technol. Teacher Educ. Int. Conf. (SITE)*, 2025, pp. 3367–3375.
 - [68] F. Luo *et al.*, “Promoting elementary students’ computer science conceptual knowledge and coding performance: A mixed methods case study,” *J. Res. Technol. Educ.*, pp. 1–21, 2026, doi: 10.1080/15391523.2026.2661645.
 - [69] T. G. Chegini, B. Cummins, M. H. Wickstrom, and N. Hosseinzadeh, “From Artifact to Instrument: Upper-Level University Student-GenAI Interactions During Problem Solving in an Advanced Mathematics Course,” Jul. 15, 2026, In Review. doi: 10.21203/rs.3.rs-10297565/v1.



NARGES HOSSEINZADEH (M’26) received the B.S. degree in mathematics and its applications from Shahid Chamran University of Ahvaz, Ahvaz, Iran, in 2014, the M.S. degree in pure mathematics from Isfahan University of Technology, Isfahan, Iran, in 2016, and the M.S. degree in applied mathematics from Illinois Institute of Technology, Chicago, IL, USA, in 2023. She is currently pursuing the Ph.D. degree in curriculum and instruction, with a focus on mathematics and science education, at the University of Illinois Chicago, Chicago, IL, USA.

From 2023 to 2024, she was a Data Analytics Developer with PERSOWN Inc. From 2024 to 2026, she was a Lecturer with City Colleges of Chicago and National Louis University. She is currently a Graduate Research Assistant with the University of Illinois Chicago, Chicago, IL, USA. Her research interests include applied mathematics, mathematical modeling, data analysis, mathematics education, and learning analytics. Ms. Hosseinzadeh led the teams that won first place in the Insurance Track at the 2023 SAS Hackathon and first place overall in the Americas Region at the 2024 SAS Hackathon.

TANNAZ GOODARZVAND received the B.S. and M.S. degrees in pure mathematics from Imam Khomeini International University, Qazvin, Iran, in 2009 and 2013, respectively. She is currently pursuing the Ph.D. degree in mathematics education and the M.S. degree in data science at Montana State University, Bozeman, MT, USA. Since 2023, she has held Graduate Teaching Assistant positions at Montana State University, first with the Department of Education and, since 2024, with the Department of



Mathematical Sciences, where she also currently serves as Lead Instructor. Her research interests include computational mathematics, data science, artificial intelligence in mathematics education, and mathematical modeling.



FARAZ DADGOSTARI received the Ph.D. degree in systems and Information engineering from the University of Virginia, Charlottesville, VA, USA, in 2021. He was subsequently a Postdoctoral Research Fellow with its Biocomplexity Institute's Network Systems and Advanced Computing Division. He is currently an Assistant Professor with the Department of Mechanical and Industrial Engineering, Montana State University, Bozeman, MT, USA. His

research focuses on the mathematical foundations of optimization and machine learning, including causal machine learning, (inverse) reinforcement learning, decision-making under uncertainty, network optimization, and game theory, with applications to digital twins, multiagent systems, and human-artificial intelligence systems.



NEDA NAZEMI received the Ph.D. degree in Systems and Information Engineering from the University of Virginia, Charlottesville, VA, USA, in 2023. She is currently an Assistant Professor with the Gianforte School of Computing at Montana State University, Bozeman, MT, USA. Her research interests include data mining, machine learning, and artificial intelligence for decision making and resilience in sociotechnical systems, with applications in environmental sustainability, natural hazard management, health,

education, and food systems