International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(12s) 2026 pp. 1242-1252
https://doi.org/10.70917/ijcisim-2026-4027
Article

Machine Learning and Deep Learning Approaches for Cognitive Reasoning Assessment in Personalized Education: A Systematic Literature Review

Alka Patel[1], Rajesh Patel[2]

[1]Department of Computer Engineering, Faculty of Engineering and Technology, Sakalchand Patel University, Visnagar, Gujarat, India. Information Technology, L. D. College of Engineering, Ahmedabad, Gujarat, India.
Email: alkapatel227@gmail.com
[2]Department of Computer Engineering, Faculty of Engineering and Technology, Sakalchand Patel University, Visnagar, Gujarat, India.
Email: drrppatelce_spce@spu.ac.in

Abstract: The significance of cognitive reasoning evaluation is on the rise in contemporary education, where not only the students' acquired knowledge but also the process of information analysis, problem solving, and the application of knowledge to new contexts is considered essential for the evaluation. The rapid development of Artificial Intelligence (AI), especially in the area of machine learning (ML) and deep learning (DL), has opened up new opportunities for building intelligent systems that can help automate reasoning assessment and make it more personalized. New methods like Bidirectional Long Short-Term Memory (BiLSTM) networks, transformer language models, multimodal learning, Explainable Artificial Intelligence (XAI), and Large Language Models (LLMs) have made educational assessment more advanced and multifunctional by processing various kinds of learning data and improving model performance and explainability. This paper discusses a systematic review of 48 peer-reviewed articles published between 2016 and 2026 regarding recent developments in AI-based cognitive reasoning evaluation. The selected studies are discussed in terms of six fundamental research topics, which include traditional machine learning algorithms, sequential deep learning and knowledge tracing, transformer networks, multimodal learning frameworks, explainable AI, and cognitive diagnosis models. A structured review approach is applied in order to analyze and contrast different learning models, methods of explainability, educational data sets, and performance metrics, as well as detect trends in research and existing challenges. The conducted review reveals the evident shift from traditional prediction models to explainable and multimodal, as well as LLM-powered cognitive diagnostic systems. Despite the fact that recent methods demonstrated significant advances in predicting students' performance and offering personalized learning opportunities, there are still many challenges, such as a lack of benchmark data sets, annotation costs, domain generalization, fairness issues, transparency problems, and implementation of explainable AI in real-world settings. Based on these results, the paper proposes possible future research directions for creating reliable, interpretable, and learner-centered cognitive reasoning assessment frameworks.
Keywords: Cognitive Reasoning Assessment; Machine Learning; Deep Learning; Multimodal Fusion; Explainable AI; Personalized Education; Literature Review

1. Introduction

To know how students think, not just how correctly they respond, has emerged as a crucial aim in current educational evaluation. The existing means of assessment, such as exams, multiple-choice questions, and grading by the use of criteria, serve well in the evaluation of outcomes achieved but fail to reveal the thinking process behind achieving those outcomes. In this way, a right answer does not necessarily show the actual knowledge of the topic because it can be achieved through luck, while the wrong answer does not indicate a lack of knowledge of the subject since it can be partially understood and incorrectly applied [1], [2].

However, the rapid digitization process in the education sector offers opportunities for overcoming some of these weaknesses. The continuous production of huge amounts of learner interaction data from Learning Management Systems (LMS), online assessment systems, MOOCs, and intelligent tutoring systems is one way through which such digitization has taken place. Some of the digital footprints that can be captured include test scores, explanations provided by learners, time to answer questions, click-stream data, engagement in discussion forums, among others. EDM and LA studies have shown that such data can be analyzed to gain insights on

Copyright: © 2026 by the authors

learning and prediction of learner performance [12], [13].

Progresses made in the area of AI and specifically ML and DL have contributed significantly to the shift from traditional assessment based on outcomes to cognitive assessment based on processes. Early educational prediction models utilized traditional machine learning techniques, such as Decision Trees, Random Forests, Support Vector Machines, and Logistic Regression, in order to detect patterns related to students' performance. While these methods gave quite good results in terms of predicting accuracy, they lacked the ability to capture all the complexity of relations typical of human cognition [9], [10].

In more recent times, however, deep learning methods have helped in extracting features from different types of educational data sources. The use of sequential models like Long Short Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) has shown success in modeling temporal learning behavior, while transformer models such as BERT have greatly enhanced semantic analysis of students' written answers and explanations [18], [20]. Such advances have enhanced the capability of AI to evaluate cognitive reasoning skills by learning complex relationships between textual, numeric, and behavioral educational data.

An additional trend is that of Multimodal Learning Analytics (MMLA). This approach is characterized by the usage of information coming from different sources, thereby offering a more holistic representation of the students’ learning process. In contrast to using only grades and written answers in their assessments, this approach utilizes explanations, solutions, interactions, behavior, and other data coming from the learner [26], [29]. Through incorporating complementary views of the process of learning, multimodal models have shown better predictive power than unimodal models [26], [29].

However, while these innovations greatly enhanced the performance of predictive models, they gave rise to new problems. Modern state-of-the-art deep learning algorithms are often black boxes and, therefore, educators cannot comprehend the mechanism of predictions. In the educational context, AI-supported decisions may be responsible for assessing a learner, developing personalized educational pathway, and conducting instructional interventions, thus, it is important to ensure that they are clear and accountable. As a result, there is rising interest in Explainable Artificial Intelligence (XAI) aimed at generating human-understandable explanations of AI predictions. For example, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) methods help educators to determine the set of features which are used to predict a reasoning level of a particular learner [24], [25], [31].

Recently, fast growth of Large Language Models (LLMs) allowed to introduce new opportunities for intelligent educational systems. The key difference between modern LLMs and previous AI models which were created only for prediction purposes is that LLMs can conduct an interactive reasoning assessment, generate personalized feedback, answer questions of learners, and modify instructional material according to their individual needs. Natural language processing capabilities of LLMs make them perspective tools for creating next-generation educational technologies [37], [39].

Regardless of these technological advances, the research on cognitive reasoning assessment using AI is still scattered. Research in this domain involves many related areas such as knowledge tracing, cognitive diagnosis, educational data mining, multimodal learning analytics, explainable AI, and LLMs for education. Each community has developed its own datasets, learning algorithms, evaluation metrics, and application fields, and thus, it becomes very hard to have an overview of the present achievements and future research directions in this area. Moreover, very few review papers study these research developments altogether from the perspective of cognitive reasoning assessment.

In order to fill this gap, this paper provides a systematic review of 48 peer-reviewed studies published from 2016 to 2026. The selected studies are classified into six main categories which include (i) traditional statistical and machine learning algorithms, (ii) sequential deep learning and knowledge tracing methods, (iii) transformers and attention mechanisms, (iv) multimodal learning approaches, (v) explainable AI, and (vi) cognitive diagnosis models. The review analyzes the above-mentioned methods in terms of predictive performance, interpretability, multimodal capability, scalability, and data requirements.

These are the main contributions of this review paper:

- A systematic review of machine learning and deep learning methods used for cognitive reasoning assessment through a review methodology inspired by PRISMA guidelines.
- A detailed taxonomy that organizes the available literature into six main research categories and demonstrates the development of the field of AI-based cognitive reasoning assessment.
- A comparative study of the existing approaches based on their predictive power, interpretability, multimodal capabilities, scalability, and educational applicability.

- An evidence-based identification of research gaps and future research directions for the development of interpretable, personalized, and trustworthy cognitive reasoning assessment systems.

This paper is organized as follows. In Section 2, the methodology of the review is described, which includes the literature search strategy, study selection approach, and inclusion and exclusion criteria. The review results are presented in Section 3 with a thematic presentation of the selected papers in terms of the six identified research areas. In Section 4, the comparison of the approaches discussed in this review is provided. Section 5 discusses the main research gaps and presents possible directions for future research.

2. REVIEW METHODOLOGY

The review adheres to a systematic, PRISMA-inspired methodology that entails four steps, namely identification, screening, eligibility determination, and inclusion, all tailored specifically for a method-based systematic review.

2.1. Literature Search Strategy

A systematic search was performed in six databases utilizing various combinations of the following keywords: “cognitive reasoning assessment,” “machine learning education,” “deep learning student performance,” “multimodal learning analytics,” “explainable AI education,” “knowledge tracing,” “cognitive diagnosis model,” and “large language models personalized learning.” A summary of the databases used can be found in Table 1.

Table 1. Databases Searched and Rationale

Database	Rationale for Inclusion
IEEE Xplore	Primary venue for engineering, computing, and educational-technology conference and journal papers
ScienceDirect / Elsevier	Broad coverage of Computers & Education and Expert Systems with Applications journals
SpringerLink	Coverage of AIED, EDM, and learning-analytics conference proceedings
ACM Digital Library	Coverage of LAK, ICML, and multimodal interaction conference proceedings
arXiv	Rapid dissemination of recent transformer-, LLM-, and knowledge-tracing research
MDPI (Applied Sciences, Behavioral Sciences)	Open-access coverage of interdisciplinary explainable-AI and cognitive-diagnosis studies

2.2. Inclusion and Exclusion Criteria

Table 2 provides the inclusion and exclusion criteria utilized at the time of screening and selection phases in order to maintain the quality of technically relevant studies.

Table 2. Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Publication type	Peer-reviewed journal or conference paper, or archival preprint with technical contribution	Blog posts, editorials, non-technical opinion pieces
Topical focus	Applies ML/DL/XAI to reasoning, performance, or cognitive-state assessment in education	General AI papers with no educational-assessment application
Publication window	2016 – 2026	Pre-2016 studies unless foundational (e.g., BERT, LSTM, SHAP)
Language	English	Non-English full text with no translation available
Reporting quality	Describes dataset, method, and at least one quantitative outcome	Insufficient methodological detail to assess contribution

2.3. Study Selection Process

First, 612 papers were retrieved through database searching. Upon removal of 125 duplicate records, 487 unique records were found for title-abstract screening. Through this, 329 off-topic or non-technical records were removed, leaving 158 records for assessing eligibility. Finally, 110 papers were removed due to lacking sufficient technical details, being not applicable in the field of educational assessment, and/or being redundant to other comprehensive papers. Thus, 48 papers formed the final dataset which is categorized into thematic groups as follows: classical statistical/ML approaches (7), sequential DL and knowledge tracing approaches (10), transformer and attention based approach (8), multimodal fusion techniques (8), explainable AI techniques (8), and cognitive diagnostic models (7). Figure 1 depicts the whole selection process.

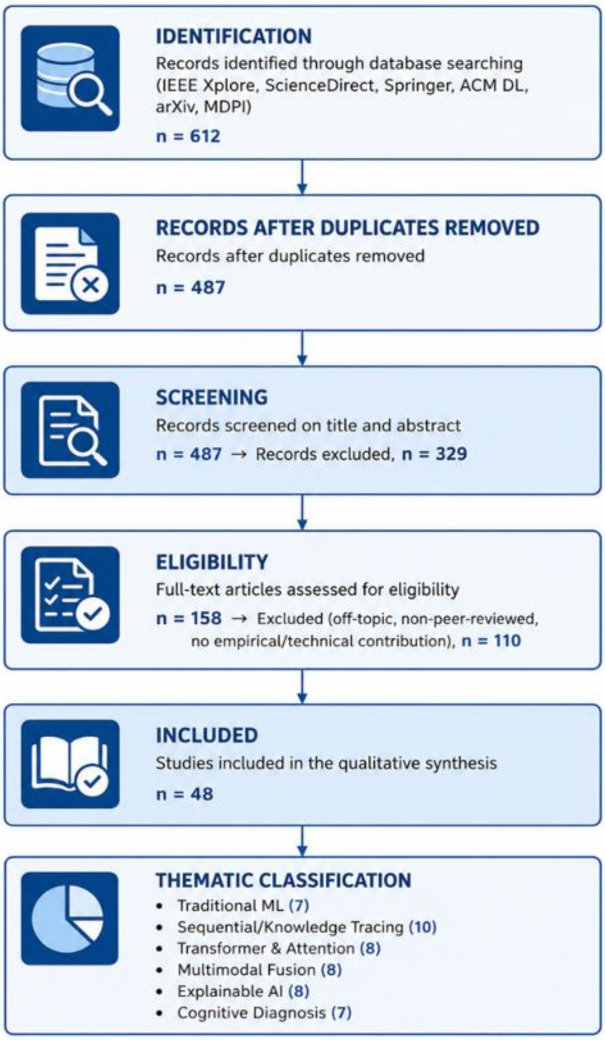

Figure 1. PRISMA-Informed Flow Diagram of the Study Selection Process

2.4. Data Extraction and Synthesis

The following data have been extracted from each included study: year of publication, family of techniques, characteristics of datasets used, evaluation metrics, performance results obtained, and limitations of the research work. After that, the studies were synthesized narratively for each thematic category (Section 3), and comparative analysis and gaps synthesis (Section 4 & Section 5) follow.

3. THEMATIC LITERATURE REVIEW

In this section, a synthesis of 48 reviewed papers is made, and it includes six thematic categories according to the historical and technological evolution of ML/DL methods applied to reasoning and learning analytics assessment problems. The taxonomy of the six categories is presented in Fig. 2.

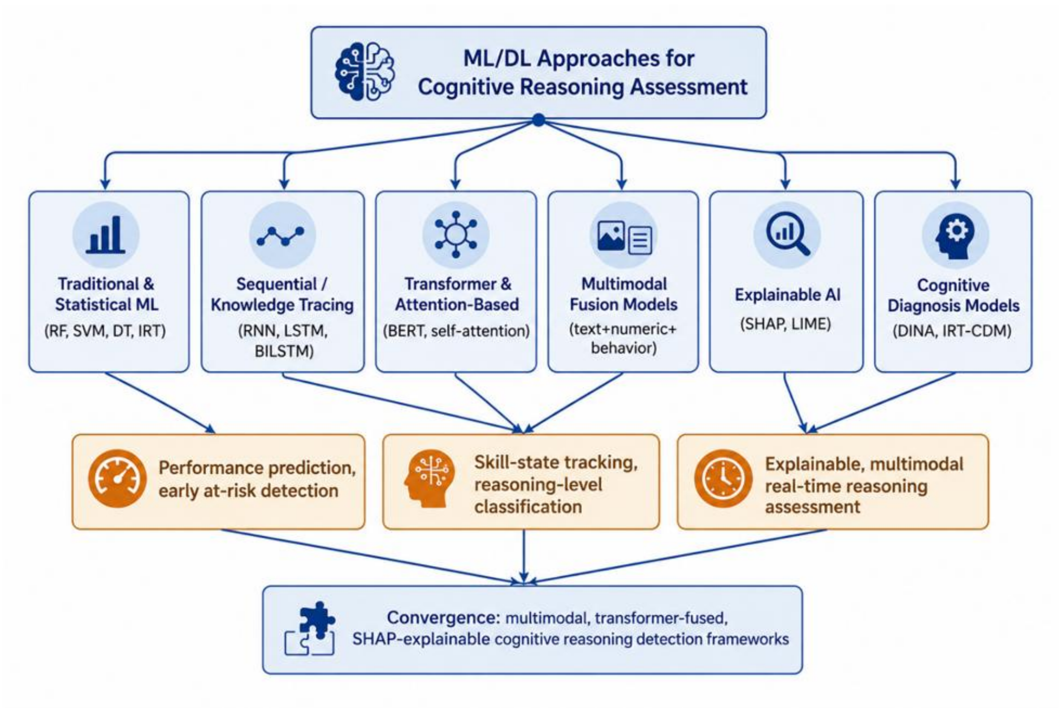

Figure 2. Taxonomy of ML/DL Approaches for Cognitive Reasoning Assessment

3.1. *Traditional Statistical and Machine-Learning Approaches*

Initial work in educational data mining used classical statistical and machine learning classifiers like Decision Trees, Random Forests, Support Vector Machine, and k-Nearest Neighbor using structured achievement data and demographic data. The combination of Random Forest and XGBoost with tiered instructional grouping in Chen et al. [9] led to an observable improvement in academic results where at-risk students were identified at the beginning. Decision Tree, Random Forest, KNN, and SVM classifiers were thoroughly compared in Ahmed [10] where ensemble tree-based classifiers outperformed distance-based and kernel-based classifiers.

Although the methods above proved the applicability of automated performance prediction, they had two structural limitations. Firstly, they work only on structured outcome-level data (performance, demographic, etc.) without any deep explanation of the reasoning process behind them. Secondly, the generalizability of their models was limited to a single course or institution [9], [10], similar to a trend common in education data mining [12], [13].

3.2. *Sequential Deep-Learning and Knowledge-Tracing Models*

Deep Knowledge Tracing (DKT), introduced by Piech et al. [15], became a milestone where recurrent neural network (RNN) and LSTM architectures were utilized for direct modeling of temporal evolution of the learner's knowledge state on the basis of interaction sequences rather than pre-defined skill mapping. The original formulation of LSTMs by Hochreiter and Schmidhuber [20] became an architecture prerequisite for such a transition from cross-sectional to longitudinal prediction.

Further development of the DKT approach involved several lines. Recurrent models with attention mechanisms enhanced interpretability of the predictions by pointing to the previous interactions that most strongly affected the particular prediction [21]. Recent studies reformulated knowledge tracing as a problem of language modeling: Zhan et al. [34] introduced a large-scale autoregressive approach treating interaction sequences in a similar way as token sequences, obtaining significant transfer across different knowledge domains with higher computational complexity and lower interpretability. Surveys of the area [24 in DL-EDM survey, 35, 36] demonstrate that, despite the consistent superiority of deep sequential models over Bayesian Knowledge Tracing and static classification models, problems with the cold start learners, low density of interaction histories and lack of interpretation of the underlying knowledge-state representation remain.

3.3. *Transformer and Attention-Based Architectures*

Self-Attention [19], which was introduced by Vaswani et al., and BERT [18] – a transformer architecture used for bidirectional encoding of input text – became fundamental building blocks in modern pipelines for reasoning assessment from free-text learner responses. Transformer architectures allow capturing long-term dependencies and context-specific semantics of students' generated text which could not be modeled via bag-of-words or simple embedding techniques before.

For knowledge tracing, the application of transformer variants was explored on top of interaction sequences rather than text. Knowledge Tracing Set Transformers [23] offered a simplification of prior overengineered attention architectures and achieved state-of-the-art benchmark results, and Transformer-Bayesian architectures [47] aimed to unite the pattern capturing capability of transformers and causal interpretability of graphical probability models. All of these works show that transformers have better capabilities of modeling semantics and sequences compared to baseline RNN/LSTM architectures, but at the expense of higher computational cost.

3.4. Multimodal Fusion Frameworks

Multimodal learning analytics (MMLA) combines various streams of information – text, numerical problem-solving sequences, behavior, and more recently physiological and/or video input – to create a comprehensive view of the learning process. As part of a systematic review of 43 MMLA papers employing AI technology [26], Mohammadi et al. reported exponential growth in AI and MMLA convergence from 2019 through 2024 but observed that most papers do not incorporate teacher and learner feedback in real-world situations.

As one of the very few real-world longitudinal investigations of MMLA, Yan et al. [27] evaluated the impact of an MMLA system deployed over two years to a healthcare-education course involving 399 learners and 17 teachers. They demonstrated the effectiveness of their system in improving reflective feedback quality. A review of 177 MMLA peer-reviewed research articles [29] classified the applications based on the educational purpose, which included engagement monitoring, collaborative learning, simulation-based environment, and inclusive education. The challenges in MMLA research, however, were the issues of data synchronization, interpretability, ethical issues, and scaling. These are all closely related to the rationale behind using transformer-BiLSTM architecture in multimodal fusion for cognitive reasoning detection.

3.5. Explainable AI Techniques

With rising complexity of prediction models, the importance of interpretability increased as well. The method of SHAP – short for SHapley Additive exPlanations [24], introduced by Lundberg and Lee based on cooperative game theory – became a dominant way to explain machine learning models in educational data mining providing global rankings of feature importance as well as local explanations for every particular student [30]. LIME [25] offers a complementary approach to local explanations, model agnostic in its nature; however, further studies revealed problems with stability of LIME results in the context of repeating perturbation [14].

Examples of applied studies provide proof of value of both approaches: Hoq et al. [32] explain their ordered ensemble model predicting grades using SHAP at both individual and course level; a pipeline built from CatBoost and logistic regression algorithms evaluated on a dataset of 10,000 records related to academics reached 99.1% accuracy when SHAP was used to show that study hours, GPA, and internal marks were the key predictors [17]; and Özkurt used both SHAP and LIME together in order to triangulate factors that determine students' performance.

3.6. Cognitive Diagnosis Models

The Cognitive Diagnosis Models (CDMs) derive from psychometrics and offer a statistically well-grounded alternative or even complement to exclusively data-based ML methods. Extending Item Response Theory (IRT), the models like DINA, Generalized DINA (G-DINA), and the Log-linear Cognitive Diagnosis Model (LCDM) describe a learner's competence in terms of small-grain knowledge attributes, rather than in an overall competence measure [41]. A recent systematic review of ML approaches in the field of cognitive diagnosis [41] highlights the increasing use of deep learning elements, such as variational autoencoders, graph neural networks, and attention, in traditional psychometric CDMs, as well as the development of special studies on fairness and cold start handling in diagnostic modeling.

Another thematic review specifically devoted to language testing [43] proves that the CDMs have been effectively used for diagnosing reading and listening comprehension, however, this review stresses the fact that almost all applications have been limited to the framework of standardized and largescale testing, but not embedded in instruction and conducted online. The complementary studies on the diagnostic classification models [44] show their practical value for decision making, whereas the more recent tensor-based diagnostic frameworks [46] further develop CDMs in the direction of exercise-knowledge-cognition diagnostics.

3.7. Large Language Models and Personalized Learning

Recent thematic changes in the literature regard the incorporation of large language models (LLMs) into personalized and reasoning-aware educational systems. Specifically, a systematic review of 55 studies from 2020 to 2024 [37] revealed that LLM-based systems led to better student engagement, socio-emotional development, and real-time monitoring compared to conventional methods, yet posed difficulties in ethical use, privacy issues, and simulated tutoring interactions via LLMs. Another relevant piece of complementary research conducted by East China Normal University [38] pointed out that automated generation of teaching materials and adaptive assessments and feedback were two domains where LLMs contributed most to personalized education.

More generally, an empirical study of 88 works published between November 2022 and March 2025 [39] revealed six major application categories in which LLMs were employed for educational purposes: chatbots, generation of learning materials, automated assessments and feedback, task support tools, learning support tools, and intelligent tutoring systems. Reasoning-aware personalization has also been examined through learning path planning techniques that involve LLM reasoning capabilities in conjunction with knowledge graphs [40]. On the other hand, nascent critical literature highlights the possibility of shallow engagement through LLMs, which does not necessarily come with improved deep learning, as well as the under-auditing of biases in LLM-based feedback. It is therefore worth noting that explainability and cognitive diagnosis remain relevant tools even as LLM-based solutions become more sophisticated (Sections 3.5 and 3.6).

3.8. Summary of Reviewed Studies

Table 3 summarizes twenty representative studies spanning all six thematic categories, illustrating the diversity of techniques, key results, and limitations synthesized in this section. The complete set of 48 reviewed studies is cited throughout Sections 3.1–3.7 and listed in full in the References section.

Table 3. Summary of Representative Reviewed Studies

#	Author(s) / Ref.	Technique	Key Result	Limitation
1	Chen et al. [9]	RF / XGBoost tiered instruction	Improved academic performance via tiered grouping	Single-course domain, limited transfer
2	Ahmed [10]	DT, RF, KNN, SVM comparison	RF/DT achieved highest accuracy among baselines	Limited generalizability across cohorts
3	Piech et al. [15]	Deep Knowledge Tracing (RNN/LSTM)	Pioneered neural skill-state prediction	Binary mastery representation only
4	Zhan et al. [34]	Autoregressive LLM-based knowledge tracing	Reframes KT as language modeling with strong transfer	High computational cost, limited interpretability
5	Devlin et al. [18]	BERT transformer embeddings	State-of-the-art semantic text understanding	Domain adaptation needed for educational text
6	Vaswani et al. [19]	Self-attention / Transformer architecture	Foundational attention mechanism enabling parallel sequence modeling	Quadratic complexity with sequence length
7	Mohammadi et al. [26]	Systematic review of AI in MMLA (43 studies)	Confirms rapid growth of AI-enhanced multimodal analytics	Few studies close the feedback loop with stakeholders
8	Yan et al. [27]	MMLA for collaborative-learning feedback	Two-year deployment with 399 students improved reflection	Evaluated in a single healthcare-education context
9	MDPI Review	Comprehensive MMLA	Synthesizes tools,	Notes persistent data-

	[29]	review (177 studies)	methods, and pedagogical applications	synchronization and interpretability barriers
10	Lundberg & Lee [24]	SHAP explainability framework	Unified, theoretically grounded feature-attribution method	Computationally expensive for large feature spaces
11	Ribeiro et al. [25]	LIME local explanations	Model-agnostic local interpretability	Explanation instability across perturbations
12	Choi et al. [30]	Systematic review of XAI in STEM performance prediction	SHAP dominant explainability method in EDM	Calls for better alignment with instructor decision needs
13	Hoq et al. [32]	Explainable ordered ensemble + SHAP	Interpretable prediction of final grades	Explanation granularity limited at course level
14	IJET Study [17]	CatBoost + Logistic Regression + SHAP	99.1% accuracy with feature-importance explanations	Extremely high accuracy suggests possible dataset leakage risk
15	F. Wang et al. [41]	Survey of cognitive diagnosis models (CDM)	Taxonomy of ML-based CDM approaches (DINA, GDM, LCDM)	Cold-start and data-scarcity challenges remain open
16	Mei & Chen [43]	Thematic review of CDM in language assessment	Synthesizes Q-matrix and skill-attribute approaches	Limited coverage of deep-learning-based CDM
17	Thompson et al. [44]	Diagnostic Classification Models for instruction	Improves instructional decision-making with diagnostic profiles	Requires substantial psychometric expertise to deploy
18	Sharma et al. [37]	Systematic review of LLMs in personalized learning (55 studies)	32% engagement increase reported across reviewed studies	Heterogeneous outcome measures limit cross-study comparison
19	Liu, Jiang & Wei [38]	LLMs as personalized teaching assistants	Identifies two key areas for LLM-enhanced personalization	Early-stage evidence, mostly qualitative
20	ScienceDirect Review [39]	Empirical LLM-in-education review (88 studies)	Six application categories identified, incl. ITS and feedback	Risk of superficial engagement without deep learning gains

4. COMPARATIVE ANALYSIS

The six groups of techniques that were discussed in Section 3 are summarized in Table 4 on the basis of five practical criteria: the usual range of predictive accuracies reported in the literature; interpretability; requirement for training data; ability to integrate multi-modal information; and real-time application potential in classroom settings.

Table 4. Comparative Analysis of ML/DL Technique Families for Cognitive Reasoning Assessment

Technique Family	Typical Accuracy Range	Interpretability	Data Requirement	Multimodal Support	Real-Time Readiness
Traditional ML (RF, SVM, DT)	65–80%	High	Low–Moderate	Limited	High
Sequential DL (LSTM/BiLSTM/DKT)	75–88%	Low	Moderate	Limited	Moderate
Transformer / Attention	80–90%	Low–Moderate	High	Moderate	Low–

					Moderate
Multimodal Fusion	85–92%	Low–Moderate	High	High	Moderate
Explainable AI (SHAP/LIME layer)	Model-dependent	High (added)	Moderate	Moderate	Moderate
Cognitive Diagnosis Models	70–85%	High	Moderate–High	Limited	Low–Moderate
LLM-Based Personalization	Task-dependent	Low–Moderate	Very High	High (via prompting)	Moderate–High

Three themes arise from comparing these techniques. First, there is a clear trade-off in favor of higher predictive accuracy over native interpretability, as seen in how traditional ML and cognitive diagnosis systems maintain the highest transparency but yield the lowest ceiling for multimodal inference and decision-making, whereas transformers and fusion-based approaches achieve higher accuracy at the expense of transparency, which needs to be retrieved via a secondary explainability module such as SHAP [24], [30]. Second, among all, multimodal fusion approaches exhibit the highest predictive performance since they integrate signals from complementary modes, yet suffer from the largest engineering overheads, since they need synchronized text, numerical, and behavioral modalities [26], [29]. Third, LLM-based personalization represents a unique case: although it demands large volumes of pre-training data, it requires small amounts of labeled data at the task-specific personalization stage, making it an ideal approach for quick personalization despite the lack of consistent benchmarking of reasoning assessments in research [37], [39].

One additional insight relates to the convergence of architecture. The newest and most efficient systems that have been discussed within the literature under review, such as the systems of transformer-BiLSTM multimodal fusion that have been further developed to include SHAP explainability, do not limit themselves to one specific type of system but use different types from three or more groups listed in Table 4.

5. RESEARCH GAPS AND FUTURE DIRECTIONS

Combining the above six themes with the comparative analysis in section 4, this review highlights eight common gaps found within the existing body of knowledge, as shown in Table 5 below.

Table 5. Summary of Research Gaps and Corresponding Future Directions

Identified Research Gap	**Corresponding Future Direction**
Costly, subjective expert annotation of reasoning-level labels	Semi-supervised and crowdsourced annotation pipelines; weak-supervision and self-training approaches
Limited cross-domain and cross-institution generalization	Domain-adaptation, meta-learning, and zero-shot transfer strategies validated across multiple institutions and subject domains
Interpretability-accuracy trade-off in deep and fused models	Inherently interpretable architectures (attention-visualized, concept-bottleneck) rather than post-hoc explanation only
Algorithmic bias and fairness risks in predictive and diagnostic models	Fairness-aware training objectives and routine bias auditing prior to deployment
Sparse real-time, classroom-embedded deployment evidence	Longitudinal, in-the-wild evaluations with teacher- and student-facing dashboards
Fragmented evaluation practices across sub-fields (KT, CDM, MMLA, XAI)	Shared benchmarks and standardized metrics enabling cross-paradigm comparison
Unaudited risk of superficial engagement in LLM-based personalization	Reasoning-depth-aware evaluation protocols combining LLM personalization with explainable diagnostic layers
Privacy and ethical constraints on multimodal behavioral/physiological data	Privacy-preserving analytics (federated learning, on-device inference, differential privacy)

Apart from the particular gap-direction pairs presented in Table 5, two general remarks can be made as well. Firstly, the community could greatly benefit from a common standard for evaluation, covering all of the aforementioned paradigms of knowledge tracing, cognitive diagnosis, multimodal fusion, and explainability measures, because the fragmented nature of the process at hand [26], [30], [41] does not allow any meaningful cross-paradigm comparisons even within one single literature review. Secondly, when the paradigm of LLM-based personalization evolves into maturity, its combination with explainability and cognitive diagnostic methods from Sections 3.5 and 3.6, instead of the paradigm working in isolation, seems to be key for building reliable reasoning assessment systems.

6. CONCLUSION

In this systematic literature review, 48 scientific papers from 2016 to 2026 that utilized machine learning and deep learning techniques in the evaluation of cognitive reasoning skills in personalization education were examined through a categorization approach based on six major themes including conventional statistical/ML algorithms, sequential DL and knowledge tracing models, transformers and attention-based systems, multimodal fusion approaches, explainable AI models, and cognitive diagnosis models. A distinct chronological evolution of unimodal prediction-oriented models towards multimodal explainable reasoning-assessment models, supported by developments in sequence modeling, representation learning, and generative AI was evident in this review.

The comparative analysis presented in Section 4 highlights that none of the technique families discussed here has a definite advantage over the others across practically important criteria: traditional and diagnostic techniques maintain their edge in interpretability; transformer and fusion models have the highest predictive accuracy, while LLM-based techniques are the most flexible for personalization at scale, despite the fact that their current evaluation lacks the necessary standardization. The identified research gaps in Section 5 related to the cost of annotation, generalization, trade-offs between interpretability and accuracy, fairness considerations, real-world deployment evidence, lack of evaluation standardization, and privacy concerns all seem to suggest that future research in this area needs to focus on hybrid, multimodal, and intrinsically explainable assessment-reasoning systems.

In the ongoing quest of educational institutions to find tools for assessing cognitive reasoning that would be both accurate and interpretable as well as personalized and ethically deployable, the identified convergence — of multimodal fusion, explainable AI, and reasoning-aware LLM personalization — will presumably define the future of the field.

References

1. E. Ahmed, "Student performance prediction using machine learning algorithms," Wollo University, 2024.
2. G. C. Lemos, M. A. Guisande, and L. S. Almeida, "Cognitive abilities and school achievement," 2025.
3. X. Zhang, C. Zhang, and J. Sun, "EduPlanner: LLM-based multiagent systems for instructional design," arXiv:2502.XXXX, 2025.
4. R. Prabha and K. Dhanalakshmi, "Cognitive ability and academic achievement among higher secondary students," Periyar University, 2022.
5. J. Lu et al., "Machine learning analysis of factors affecting college students' academic performance," Hebei Agricultural University, 2024.
6. X. He et al., "Generative AI in mathematical mentoring," 2024.
7. C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities," 2025.
8. X.-H. Jia and J.-C. Tu, "AI-enhanced learning: Roles of AI capabilities, self-efficacy, and motivation," 2024.
9. Y. Chen et al., "Machine learning-driven student performance prediction for enhancing tiered instruction," arXiv:2502.03143, 2025.
10. E. Ahmed, "Student performance prediction using machine learning algorithms," Information System, Wollo University, 2024.
11. S. K. D'Mello and A. C. Graesser, "Dynamics of affective states during complex learning," Learning and Instruction, vol. 22, no. 2, pp. 145–157, 2012.
12. R. S. J. D. Baker and P. S. Inventado, "Educational data mining and learning analytics," in Learning Analytics, J. A. Larusson and B. White, Eds. New York, NY: Springer, 2014, pp. 61–75.
13. K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The knowledge-learning-instruction framework," Educational Psychologist, vol. 47, no. 3, pp. 177–195, 2012.
14. Y. Xu, L. Zhang, and X. Meng, "A multimodal deep learning framework for student performance prediction," Computers in Human Behavior, vol. 110, p. 106380, 2020.
15. C. Piech et al., "Deep knowledge tracing," in Advances in Neural Information Processing Systems, vol. 28, 2015.
16. J. Zhang, L. Shi, and J. Wang, "Cognitive diagnosis models: A review from a statistical perspective," Frontiers in Psychology, vol. 8, p. 2039, 2017.
17. R. Liu, R. Hwa, and D. Pitt, "Learning student reasoning through argumentation annotation," in Proc. 14th Workshop Innovative NLP for Building Educational Applications (BEA), 2019, pp. 1–11.

18. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.
19. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
20. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
21. J. Chen, N. Heffernan, and C. Heffernan, "Student knowledge tracing with attention-enhanced recurrent neural networks," IEEE Trans. Learn. Technol., vol. 13, no. 1, pp. 1–10, 2020.
22. Y. Wang, Q. Liu, Y. Yin, and E. Chen, "Multimodal deep learning for educational applications," IEEE Trans. Knowl. Data Eng., vol. 33, no. 3, pp. 698–710, 2021.
23. M. Shih, K. R. Koedinger, and R. Scheines, "A response time model for bottom-out hints as worked examples," in Proc. 3rd Int. Conf. Educational Data Mining, 2010.
24. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
25. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
26. M. Mohammadi, E. Tajik, R. Martinez-Maldonado, S. Sadiq, W. Tomaszewski, and H. Khosravi, "Artificial intelligence in multimodal learning analytics: A systematic literature review," Computers and Education: Artificial Intelligence, vol. 8, p. 100426, 2025.
27. R. Yan, "Evidence-based multimodal learning analytics for feedback and reflection in collaborative learning," British Journal of Educational Technology, vol. 55, pp. 1900–1925, 2024.
28. H. Ouhaichi, D. Spikol, and B. Vogel, "Multimodal learning analytics in collaborative and simulation-based learning: A review," 2023.
29. "A comprehensive review of multimodal analysis in education," Applied Sciences (MDPI), vol. 15, no. 11, p. 5896, 2025.
30. Y. Ren and X. Yu, "Optimizing student ability assessment: A hierarchy constraint-aware cognitive diagnosis framework for educational contexts," arXiv:2412.04488, 2024.
31. W.-C. Choi, C.-T. Lam, and A. J. Mendes, "A systematic literature review on performance prediction in learning programming using educational data mining," in Proc. IEEE Frontiers in Education Conf. (FIE), 2023, pp. 1–9.
32. "Explainable AI-based student result prediction using CatBoost and logistic regression," IJET, 2026.
33. M. Hoq et al., "An explainable ordered ensemble model for predicting student final grades," 2023.
34. E. Özkurt, "Explaining the predictions of machine learning algorithms for student achievement using SHAP and LIME," 2024.
35. B. Zhan, T. Guo, X. Li, M. Hou, Q. Liang, B. Gao, W. Luo, and Z. Liu, "Knowledge tracing as language processing: A large-scale autoregressive paradigm," in Artificial Intelligence in Education, Springer, 2024, pp. 177–191.
36. "A systematic review of knowledge tracing and large language models in education: Opportunities, issues, and future research," arXiv:2412.09248, 2024.
37. "A comprehensive survey on deep learning techniques in educational data mining," arXiv:2309.04761, 2023.
38. S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, "The role of large language models in personalized learning: A systematic review of educational impact," Discover Sustainability, vol. 6, no. 1, p. 243, 2025.
39. J. Liu, B. Jiang, and Y. Wei, "Large language models as personalized teaching assistants," ECNU Review of Education, 2025.
40. "Large language models in education: A systematic review of empirical applications, benefits, and challenges," Computers and Education: Artificial Intelligence, 2025.
41. C. Ng and Y. Fung, "Educational personalized learning path planning with large language models," arXiv:2407.11773, 2024.
42. F. Wang et al., "A survey of models for cognitive diagnosis: New developments and future directions," arXiv:2407.05458, 2024.
43. "Fine-grained intelligent learning diagnosis model based on the exercise–knowledge–cognition tensor for educational assessment," Behavioral Sciences (MDPI), vol. 16, no. 5, p. 637, 2026.
44. H. Mei and H. Chen, "Cognitive diagnosis in language assessment: A thematic review," Language Testing, 2024.
45. "Improving instructional decision-making using diagnostic classification models," Educational Measurement: Issues and Practice, 2024.
46. "A review of data mining in personalized education: Current trends and future prospects," arXiv:2402.17236, 2024.
47. "Interpretable knowledge tracing via transformer-Bayesian hybrid networks: Learning temporal dependencies and causal structures in educational data," Applied Sciences (MDPI), vol. 15, no. 17, p. 9605, 2025.
48. "Deep knowledge tracing and cognitive load estimation for personalized learning path generation using neural network architecture," Scientific Reports, vol. 15, p. 24925, 2025.