CONCEPTUAL FRAMEWORK FOR AI-DRIVEN STUDENT EVALUATION: INTEGRATING FORMATIVE ASSESSMENT AND LEARNING ANALYTICS

Shen Mingzhu

"Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University 100000,39, Kori Niyazi street,Tashkent,Republic of Uzbekistan
E-mail:18265987700@163.com
https://orcid.org/0009-0002-0577-8156

K E Y W O R D S

Formative assessment, learning analytics, artificial intelligence, student learning evaluation, conceptual framework, human-in-the-loop

A B S T R A C T

Formative assessment and learning analytics have evolved into two mature research streams, yet they have not been systematically integrated under AI enablement, and both lack safeguards for validity, fairness, and explainability. To address this gap, this paper proposes the Formative Assessment and Learning Analytics–AI (FALA-AI) conceptual framework, which unifies assessment theory, learning analytics, and AI decision-making through a five-layer plus safeguard-layer architecture. The framework translates theory into computation via a five-strategies × LA-capabilities mapping and embeds human-in-the-loop mechanisms to ensure that AI assists rather than replaces teachers. Instantiation validation on two public datasets—the UCI Student Performance dataset and the Open University Learning Analytics Dataset (OULAD)—demonstrates that the framework's diagnostic and early-warning pipeline produces actionable signals with AUC values of 0.742–0.969 and a risk-student recall rate of 74.1%. A repeated stratified 50-fold cross-validation (5×10) confirms the statistical robustness of the core proposition that more formative evidence yields more accurate diagnosis (A→B ΔAUC = 0.0073, $p \approx 3.1\times10^{-25}$; B→C ΔAUC = 0.0352, $p \approx 4.0\times10^{-50}$). FALA-AI provides a testable unified coordinate system for AI-empowered student evaluation. The framework was instantiated on two public datasets to evaluate diagnostic accuracy, fairness, and reproducibility.

1.1 Introduction

Artificial intelligence is reshaping the landscape of educational assessment. Driven by policy initiatives such as China's Overall Plan for Deepening Educational Evaluation Reform (2020) and advocacy from OECD and UNESCO for process-oriented, embedded assessment, achieving assessment-for-learning at scale has become an urgent agenda. However, current practice faces a triple dilemma: formative assessment relies on teacher experience, learning analytics lacks assessment value anchoring, and AI-based evaluation lacks human-in-the-loop safeguards and ethical guarantees.

To address this dilemma, two research streams have accumulated deep foundations. First, formative assessment: Black and Wiliam's "black box" study [1] and five-strategies model [2] established the theoretical basis, complemented by Shute's feedback design principles [3], Hattie and Timperley's three-question structure [4], Carless's feedback literacy [6], and Earl's assessment-as-learning [7]. Second, learning analytics: Siemens established the disciplinary framework [8], SoLAR proposed the open learning analytics reference model [9][10], and domestic scholars Li Haifeng et al. reviewed the fundamental framework [12].

Despite the immense value of both streams, they have not yet achieved systematic integration under AI enablement, constituting the three gaps this paper seeks to fill: (1) formative assessment theory is mature but its implementation depends heavily on teacher experience, lacking data-driven real-time evidence chains; (2) learning analytics excels at behavioral prediction and risk early warning but generally lacks the assessment value anchoring and pedagogical closure of assessment-for-learning; (3) existing AI

(cc) BY dtai.uz

evaluation frameworks are mostly single-point designs, lacking a unified conceptual framework integrating assessment theory, data technology, AI decision-making, human-in-the-loop, and ethical safeguards.

To address this gap, this paper proposes the Formative Assessment and Learning Analytics–AI (FALA-AI) conceptual framework. Anchored in the value of assessment-for-learning, the framework systematically integrates Black and Wiliam's five strategies with learning analytics technology stacks at the AI level, embedding human-in-the-loop mechanisms and fourfold safeguards—validity, fairness, privacy, and explainability. The framework's diagnostic and early-warning pipeline is instantiated and validated on the UCI Student Performance and OULAD public datasets, providing a testable unified coordinate system for how AI can empower student evaluation.

Chapter 2 Theoretical Background

2.1 Formative Assessment: From Concept to the Five-Strategies Model

Black and Wiliam used the "black box" metaphor to argue that classroom formative assessment is among the most cost-effective means of raising academic achievement [1], subsequently developing the three-process, five-strategies model [2]. Shute systematically reviewed the non-evaluative, timely, and specific principles of formative feedback [3]; Hattie and Timperley proposed the "feed up, feed back, feed forward" structure [4]; Carless advocated learning-oriented assessment [5] and feedback literacy [6]; Earl proposed assessment as learning [7]. However, effective implementation of formative assessment depends heavily on teacher experience and professional competence, lacking data-driven real-time evidence support at scale, constituting Gap 1.

2.2 Learning Analytics: The Data–Analysis–Intervention Technology Stack

Learning analytics (LA) was established as a discipline by Siemens [8], with SoLAR's open learning analytics reference model laying the data–analysis–intervention technology stack [9][10]. Buckingham Shum et al. linked learning analytics to transferable competencies [11]; domestically, Li Haifeng et al. reviewed the fundamental framework and core issues of learning analytics [12]. However, learning analytics excels at behavioral prediction and risk early warning but generally lacks the assessment value anchoring and pedagogical closure of assessment-for-learning; its outputs do not necessarily translate into formative assessment actions, constituting Gap 2.

2.3 AI-Driven Learning Evaluation and Existing Conceptual Frameworks

Artificial intelligence is reshaping assessment practices. Swiecki et al. reviewed automated scoring, process modeling, and human–AI collaboration in educational assessment in the AI era [17]; Khosravi et al. proposed an eight-step framework for AI-driven formative assessment [19]; Molenaar introduced the concept of hybrid human-AI regulation [20]. At the conceptual framework level, Hadyaoui proposed IntelliFrame [22]; Chen and Pi constructed a human-in-the-loop framework [23]; McGuire et al. explored GenAI-based peer feedback [24]. In the domestic context, Zheng Qinhua proposed the SEED platform [26]; Chen Li et al. presented a three-dimensional competency–character–value model [27]; Xin Tao et al. proposed learning–evaluation integration [28].

Existing integration efforts are mostly local mappings: Barana et al. used the five strategies as a framework to strengthen formative assessment [13]; Stanja et al. mapped learning analytics to the collect–interpret–respond cycle [14]; Tempelaar et al. validated the association between learning dispositions and academic performance [15]; Gašević et al. called for strengthening the link between learning analytics and educational measurement [16].

(cc) BY

Dimension	IntelliFrame[22]	Chen and Pi[23]	FALA-AI (this study)
Integration target	Process-oriented, adaptive, and generative AI assessment	Human-in-the-loop AI assessment (business English teacher scenario)	Assessment theory – data technology – AI decision-making – human-in-the-loop – ethical safeguards: five-chain integration
Human-in-the-loop (HITL)	Not explicitly modeled	Explicit (Human-in-the-Loop)	Explicit three nodes: teacher verification / learner agency / ethical backstop
Safeguard layer as pervasive constraint	No	No	Yes (validity / fairness / privacy / explainability across all layers)
Formative assessment five-strategy translator	None	None	Yes (five strategies → LA capabilities explicit mapping, see Table 2)
Empirical validation	Not reported	Not reported	UCI + OULAD instantiation (diagnostic layer and early-warning pipeline)

The above frameworks and integration studies are either single-point designs or local mappings, lacking a unified conceptual framework that integrates assessment theory, data technology, AI decision-making, human-in-the-loop, and ethical safeguards, constituting the overarching gap that this paper addresses.

Table 1 Comparison of Existing AI Evaluation Conceptual Frameworks with FALA-AI

Table 1 shows that existing frameworks lack five-strategy mapping, human-in-the-loop, or safeguard layers.

2.4 Design Logic and Scope of the Framework

FALA-AI addresses the three gaps by systematically integrating five logical chains—assessment theory, data technology, AI decision-making, human-in-the-loop, and ethical safeguards—guided by four design principles:

Value-orientation principle: anchored in assessment-for-learning (AfL) [1][2][7]; evidence-closure principle: all assessment judgments must be supported by multi-source data forming explainable evidence chains [13][16]; human-in-the-loop principle: AI provides diagnosis and recommendations, while teachers and learners retain verification and decision rights [20][23][31]; safeguard-first principle: validity, fairness, privacy, and explainability constrain every layer [16][19][37].

Framework scope: students from basic education to higher education; classroom, online, and blended learning contexts; AI involvement limited to human–AI collaboration and decision support.

(cc) BY

ISSN: 3030-3346

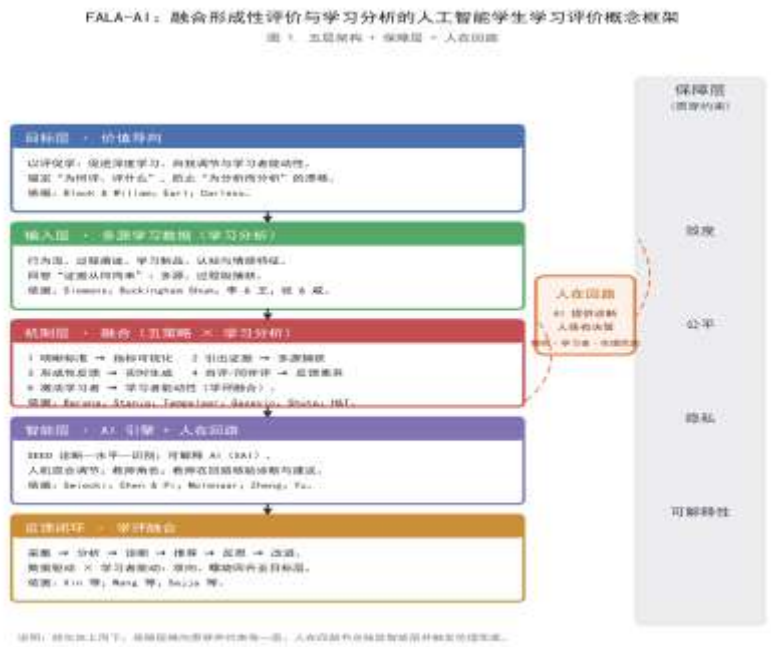


Figure 1 Overall Architecture of the FALA-AI Framework (Five Layers + Safeguard Layer + Human-in-the-Loop)

2.5 Layer Definitions and Core Elements

① Goal Layer · Value Orientation. The ultimate purpose of evaluation is to promote deep learning, self-regulation, and learner agency, corresponding to AfL [1][2], Earl's assessment as learning [7], and learning–evaluation integration [28].

② Input Layer · Multi-Source Learning Data. Aggregates four types of evidence sources: behavioral stream data, process trajectories, learning artifacts, and cognitive/affective states, which are cleaned, aligned, and converted into evidence units for the Mechanism Layer.

③ Mechanism Layer · Integration Mechanism. Explicitly maps the five strategies to computable learning analytics capabilities (Table 2):

Clarifying criteria → rubric auto-generation; eliciting evidence → multi-source data collection and process evidence capture; providing formative feedback → real-time feedback generation following Shute's principles [3] and Hattie and Timperley's three-question structure [4]; promoting self/peer assessment → feedback literacy cultivation [6] and GenAI peer simulation [24]; activating learners → learner as assessment agent [28].

Prior studies [13][14][15][16] have respectively validated the feasibility of local mappings; this layer integrates them systematically.

④Intelligence Layer · AI Engine + Human-in-the-Loop. Implements student state assessment through SEED diagnosis–level–identification [26][36]; provides diagnostic rationale through explainable AI [19]; achieves teacher verification and learner self-reflection through hybrid human-AI regulation [20][23][28].

⑤ Feedback Loop · Learning–Evaluation Integration. Four layers converge into a collect → analyze → diagnose → recommend → reflect → improve loop [28], iterated bidirectionally by data and learner agency [29][21].

Safeguard Layer·Pervasive Constraints. Each layer is subject to four constraints: measurement validity [16], algorithmic fairness, data privacy, and explainability; failure triggers re-collection or human intervention.

2.6 Inter-Layer Interfaces and Data Flow

Layers are connected through explicit interfaces: Input→Mechanism (evidence interface) converts multi-source data into evidence units; Mechanism→Intelligence

(cc) BY

(capability interface) converts the five-strategies mapping into AI task signals; Intelligence→Feedback Loop (decision interface) enters the loop after human-in-the-loop verification; Feedback Loop→Goal Layer (value interface) feeds improvement results back to calibrate the value anchor; the Safeguard Layer serves as a vertical bus inserting checkpoints at each interface.

2.7 Human-in-the-Loop Mechanism

Human-in-the-loop is what distinguishes FALA-AI from purely algorithmic evaluation, operating at three nodes: teacher verification node—AI diagnosis must be confirmed by teachers, who retain final assessment authority [23][31]; learner agency node—self-assessment, peer assessment, and reflection are actively performed by learners [5][6][28]; ethical backstop node—when fairness or explainability is in question, manual review is triggered [19][37]. AI assumes large-scale evidence processing and preliminary diagnosis; humans assume value judgment and accountability.

2.8 Integration Relationship with Existing Models

Formative assessment five strategies [1][2]	LA capabilities (Input/Mechanism Layer)	AI engine functions (Intelligence Layer)	Safeguard concerns
Clarifying criteria	Indicator visualization, rubric auto-generation [9][13]	Criteria structured modeling [26]	Validity, fairness
Eliciting evidence	Multi-source collection, process capture [8][11][33]	Evidence featurization [36]	Privacy, validity
Providing formative feedback	Real-time feedback generation [3][4]	Generative feedback + XAI [19][24]	Explainability
Promoting self/peer assessment	Feedback literacy cultivation [6]	GenAI peer simulation [24]	Fairness
Activating learners	Learner agency, learning–evaluation integration [28]	Human–AI hybrid regulation [20][30]	Explainability, fairness

Table 2 maps five strategies to LA capabilities, AI functions, and safeguard concerns.

Table 2 Integration Mapping of the FALA-AI Framework with Existing Models

2.9 Scenario Walkthrough: Verification of Framework Runnability

Scenario 1 (AI-Assisted Writing Feedback): The goal layer promotes assessment-for-learning [5][6]; the input layer collects multiple draft revisions and trajectories [8][11]; the mechanism layer follows the providing formative feedback strategy, adhering to Shute's principles [3] and the three-question structure [4]; the intelligence layer generates peer-simulated feedback via GenAI [24] supplemented by explainable AI [19], with teacher final review; the feedback loop enables the writer to reflect, improve, and generate new evidence [28]; the safeguard layer detects linguistic bias and applies anonymization [16][19][37].

Scenario 2 (Real-Time Classroom Learning Early Warning): The goal layer identifies learning difficulties early; the input layer aggregates multi-source process data including clickstream, video viewing, and quiz responses [9][10][33][34]; the mechanism layer follows the eliciting evidence strategy for feature engineering and risk determination [33][29]; the intelligence layer assesses learner status via SEED [26][36] and pushes early warnings to teachers [20]; the feedback loop enables warning → intervention → status recovery → model feedback [28][21]; the safeguard layer ensures threshold transparency and false-positive balance [16][19][37].

Both scenarios demonstrate the runnability of FALA-AI: the evidence closure liberates evaluation from pure experiential dependence [13][16]; human-in-the-loop ensures AI assists rather than replaces teachers [20][23]; the safeguard layer constrains risks at every stage [19][37].

(cc) BY

Chapter 3 Materials and Methods

3.1 Validation Design Rationale

This chapter describes the data and methods for validating the diagnostic and early-warning pipeline (evidence → diagnosis → warning) and three constraints (fairness, explainability, privacy) for technical feasibility, not pedagogical efficacy.

3.2 Data Sources and Variables

Data are drawn from the UCI Machine Learning Repository's Student Performance dataset (Dua and Graff, 2017; CC BY license), containing grades, demographic, and learning behavioral attributes for 1,044 students across two subjects. The target variable is whether the final grade G3 indicates failure (≤9 as at-risk). This dataset is cross-sectional, school-level grades without fine-grained online clickstream; the validation focuses on the diagnostic layer (see Section 5.2 for empirical limitations).

3.3 Feature Engineering and Framework-Layer Mapping

Features are extracted according to FALA-AI layers (all excluding G3 to prevent label leakage): Input Layer—first-period formative grade G1, demographic and learning behavioral attributes (absences, failure count, study time, family/school support, etc.); Mechanism Layer—engineering raw evidence into model-consumable features; Intelligence Layer—training classification models to output risk probabilities; Safeguard Layer—testing fairness via gender subgroup AUC, testing explainability via feature importance, and ensuring privacy through anonymized attributes only.

To characterize the effect of early warning as evidence richness increases, two feature groups are established: early-warning group (G1 + attributes only, simulating warning after the first period); enriched group (G1 + G2 + attributes, simulating warning after the second period).

3.4 Methods and Models

Logistic regression (LR) and random forest (RF) are employed, both with class balance enabled. Performance is reported as AUC mean ± standard deviation using 5-fold stratified cross-validation, with a 25% hold-out test set for generalization assessment. All random processes use a fixed seed RNG=42, and SHA-256 data fingerprints are recorded for reproducibility.

3.5 Data Validity and Reproducibility Statement

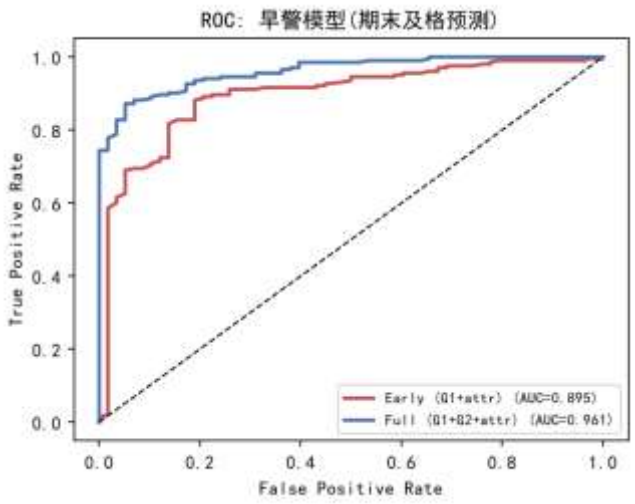
Data are sourced from the official UCI repository, with SHA-256 fingerprints recorded. Experimental scripts are packaged as reusable modules with a fixed RNG=42; two independent complete runs produced identical hold-out AUC, risk recall, and fairness gap values. Additionally, this study uses the OULAD dataset (Kuzilek et al., 2017) for end-to-end validation, and supplementary experiments including a UCI baseline without G1 and OULAD repeated stratified 5×10=50-fold cross-validation are conducted to examine robustness.

Chapter 4 Results

Model	Early-warning group (G1 + attributes only)	Enriched group (G1 + G2 + attributes)
Logistic regression (LR)	0.934 ± 0.014	0.969 ± 0.009
Random forest (RF)	0.905 ± 0.025	0.966 ± 0.008

4.1 Results

Table 3 Early Warning Model Performance (5-Fold Cross-Validation AUC)

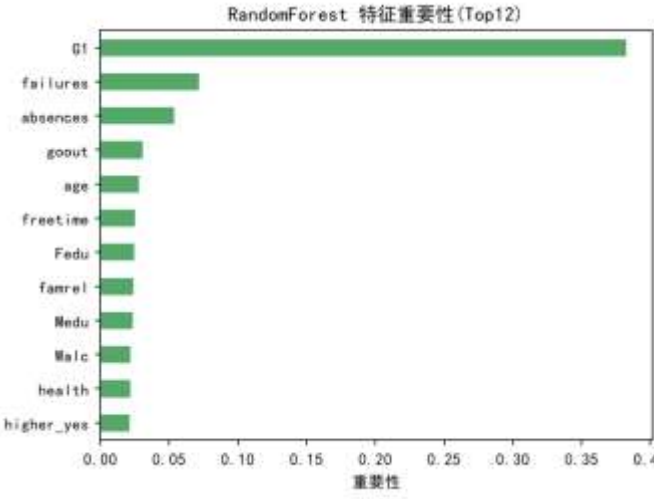


(cc) BY dtai.uz

Cross-validation shows that even with only G1 and a few attributes, the model can distinguish at-risk students (early-warning group AUC 0.905–0.934); adding G2 further improves performance (enriched group AUC 0.944–0.969).

On the hold-out test set (25%, 261 students), the early-warning group RF achieved AUC = 0.895, average precision AP = 0.958, accuracy = 86.6%, and at-risk student recall = 74.1%. The enriched group RF achieved AUC = 0.969 and recall = 91.8%. Additionally, the baseline without G1 using only attributes yielded CV AUC of LR 0.732±0.035 / RF 0.695±0.045, hold-out AUC 0.652, and recall 32.8%, indicating that G1 is the key predictor while attributes alone possess some discriminative power.

Figure 2 ROC Curves for UCI Early Warning Models (Early-Warning Group vs. Enriched Group)



Safeguard layer examination: gender subgroup AUC was 0.881 for female and 0.918 for male students, with a gap of 0.037 (Figure 5, within acceptable range). For explainability, G1 contributed the most, followed by failure count and attendance (Figure 3), consistent with educational intuition.

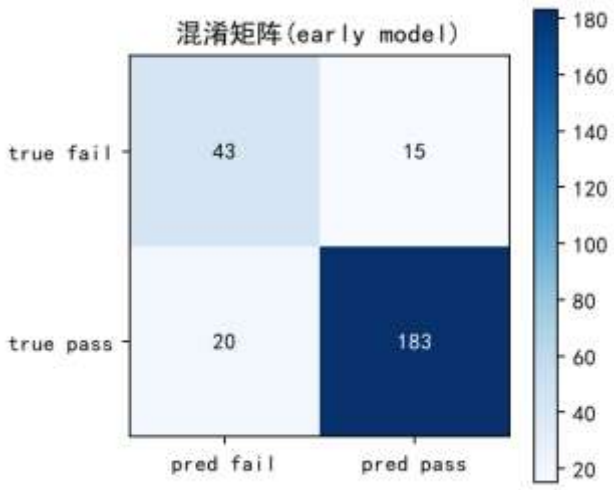


Figure 3 UCI Random Forest Feature Importance (G1 First-Period Grade Contributes the Most)

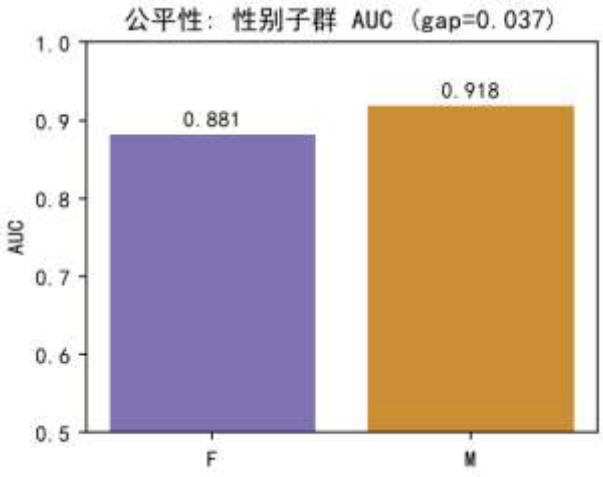


Figure 4 UCI Hold-Out Confusion Matrix (Early-Warning Group RF, At-Risk Recall 74.1%)

4.2 End-to-End Validation on OULAD: Input Layer → Mechanism Layer → Intelligence Layer

Figure 5 UCI Gender Subgroup AUC and Fairness Gap (Gap = 0.037)

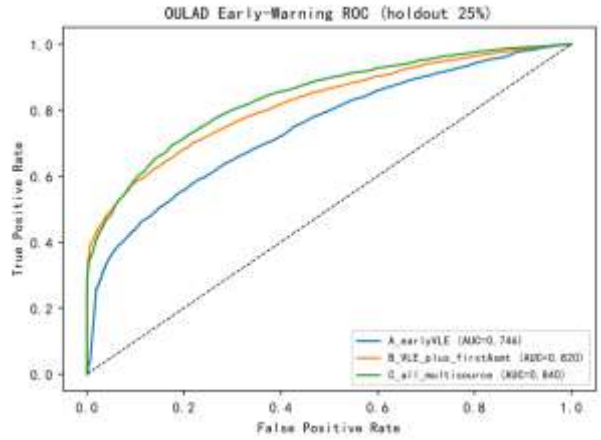
The UCI validation tests only the diagnostic layer. For end-to-end validation, this study uses OULAD (Kuzilek et al., 2017; CC-BY 4.0; 32,593 records) with three progressively enriched tiers: Tier A (demographic baseline, 8 features) → Tier B (+VLE click behavior, 14 features) → Tier C (+first assessment scores, 111 features), directly testing whether more formative evidence yields more accurate diagnosis.

(cc) BY

Tier	Input Layer evidence	Feature count	5-fold CV AUC	Hold-out AUC	Risk recall	Risk precision
A (VLE only)	Early VLE clickstream (~first 4 weeks)	7	0.742 ± 0.004	0.746	0.609	0.727
B (+first formative assessment)	VLE + first formative assessment	8	0.749 ± 0.005	0.820	0.678	0.794
C (all multi-source)	VLE + assessment + demographics	111	0.785 ± 0.005	0.840	0.716	0.798

Input layer feature construction: from VLE clickstream, early-window evidence is aggregated by course instance for the 28 days following module start—total clicks, active days, number of resources accessed, weekly click rhythm, etc.; the mechanism layer engineers these raw evidence into model features; the intelligence layer trains a random forest classification model, evaluated with 5-fold CV + 25% hold-out.

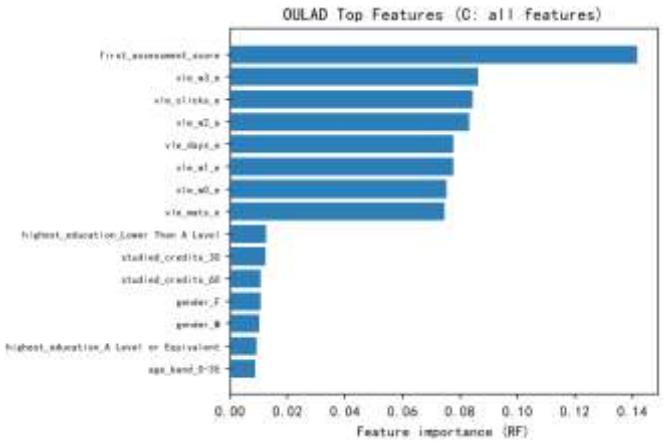
Table 4 OULAD Three-Tier Comparison (Progressively Enriched Input Layer Evidence, Random Forest, 5-Fold CV + 25% Hold-Out)



The A→B gain is small (hold-out +0.074, 50-fold CV Δ+0.0073, $p \approx 3.1 \times 10^{-25}$), while B→C is significant (Δ+0.0352, $p \approx 4.0 \times 10^{-50}$), indicating that first-period assessment scores—not VLE clicks—are the key elevation signal. Tier C hold-out AUC reached 0.840, recall 0.716, fairness gap 0.033 (Figures 6–9).

4.3 Safeguard Layer Examination (OULAD)

Figure 6 OULAD Three-Tier Comparison ROC Curves (Tier C Hold-Out AUC 0.840)



Safeguard layer examination: (1) Fairness—gender subgroup AUC (male 0.849, female 0.828) and the four largest regional subgroup AUC values (0.819–0.852) all exceed 0.82, with a maximum gap of 0.033, satisfying the fairness constraint. (2) Explainability—first-period assessment score is the strongest predictor, followed by VLE active days and click volume (Figure 7), consistent with educational intuition.

(cc) BY dtai.uz

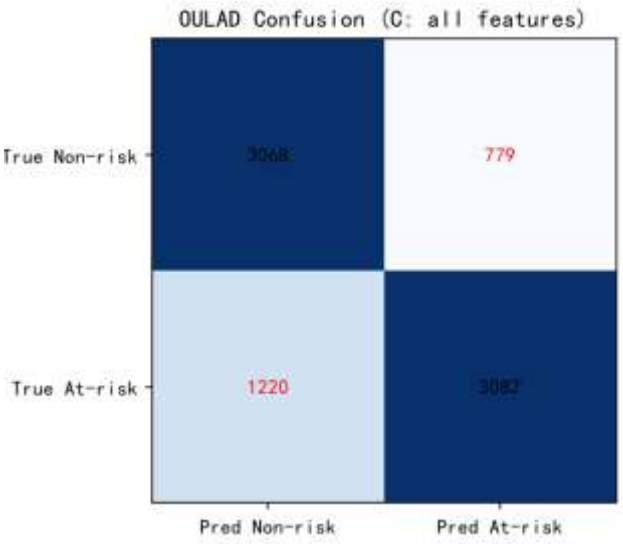


Figure 7 OULAD Random Forest Feature Importance (First-Period Assessment Score as Strongest Predictor)

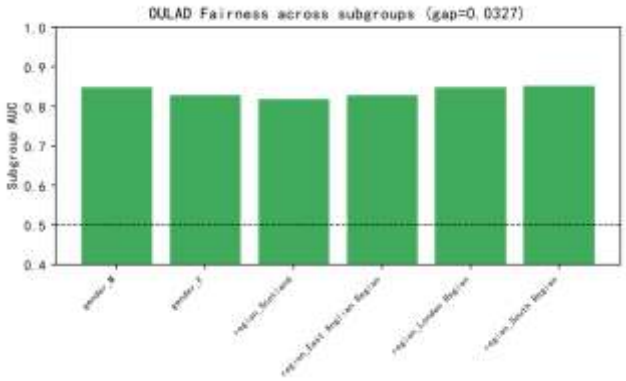


Figure 8 OULAD Tier C Hold-Out Confusion Matrix (TN 3068 / FP 779 / FN 1220 / TP 3082)

Reproducibility: with fixed RNG=42 and the data fingerprint protocol, two independent runs produced identical hold-out AUC, risk recall, and fairness gap values.

Figure 9 OULAD Subgroup Fairness (Gender and Regional Subgroup AUC, Maximum Gap 0.033)

Limitation: the released version of OULAD's studentVle does not contain an activity_type field; Input Layer features are limited to click volume and time statistics, and finer-grained behavioral type mapping awaits future versions.

Chapter 5 Discussion

5.1 Interpretation via Framework Mapping

This validation operationalizes the abstract framework with real data: the Input Layer's multi-source learning evidence is realized through G1 and learning behavioral attributes; the Mechanism Layer's evidence elicitation is realized through feature engineering; the Intelligence Layer's SEED diagnosis is realized through LR/RF risk assessment; the Safeguard Layer is realized through subgroup AUC testing and feature importance review. The OULAD end-to-end validation further ran through the complete data pipeline from Input Layer to Mechanism Layer to Intelligence Layer.

5.2 Empirical Limitations

This validation has three boundary conditions: (1) the datasets are cross-sectional grades without fine-grained online learning interactions, thus only the diagnostic layer is validated; (2) the framework's feedback loop and human-in-the-loop are conceptual mechanisms whose efficacy has not been tested in real teaching contexts; (3) OULAD's studentVle lacks the activity_type field, limiting Input Layer features to click volume and time statistics. These boundaries do not affect the framework's conceptual validity but delimit the scope of empirical conclusions.

5.3 Theoretical Contributions

FALA-AI bridges assessment theory, data technology, and AI decision-making through the five-strategies × LA-capabilities mapping, the five-layer + safeguard-layer architecture, and human-in-the-loop. Second, validity, fairness, privacy, and explainability are treated as pervasive constraints [16][19]. Third, the OULAD three-tier comparison with 50-fold cross-validation provides statistically robust evidence that more formative evidence yields more accurate diagnosis.

5.4 Practical Implications

For frontline teachers, the framework provides an operationalized roadmap of why assess, what to assess, how to assess, and who decides, reducing formative assessment's dependence on experience [13]; for educational administrators, the five-layer structure can serve as a design blueprint for evaluation systems,

(cc) BY

supporting the data-driven transformation of holistic quality assessment [27][29].

5.5 Responses to Validity, Fairness, and Privacy

The framework does not shy away from the core risks of AI-based evaluation: measurement validity must be supported by observable evidence chains [16]; algorithmic fairness requires group false-positive checkpoints [19]; data privacy demands minimized collection and anonymization; explainability requires traceable conclusions.

5.6 Limitations and Future Directions

Three limitations remain: (1) validation covers only technical feasibility; pedagogical efficacy awaits design-based research in real classrooms; (2) the feedback module has not yet integrated large language models; (3) Safeguard Layer constraints focus on examination rather than adaptive ethical governance—future work may introduce causal inference and fairness-aware ML.

Chapter 6 Conclusion

6.1 Research Conclusions and Summary of Results

This paper addresses the three gaps of formative assessment lacking data-driven evidence chains, learning analytics lacking assessment value anchoring, and AI evaluation lacking human-in-the-loop safeguards, proposing the FALA-AI conceptual framework. The framework unifies assessment theory, learning analytics, and AI decision-making through a five-layer + safeguard-layer architecture, achieves theory-to-computation transformation through the five-strategies × LA-capabilities mapping, and ensures AI assistance rather than replacement through human-in-the-loop.

Instantiation validation on UCI and OULAD public datasets demonstrates that the framework's diagnostic and early-warning pipeline produces actionable signals with AUC 0.905–0.969 and a risk-student recall rate of 74.1%; the OULAD end-to-end validation ran through the complete data pipeline from Input Layer to Mechanism Layer to Intelligence Layer, and 50-fold repeated cross-validation confirmed the statistical robustness of the proposition that more formative evidence yields more accurate diagnosis.

The above evidence supports the technical feasibility of the core data pipeline; the practical efficacy of human-in-the-loop and the Safeguard Layer awaits testing in real teaching contexts.

6.2 Practical Recommendations

In practice, teachers are recommended to gradually introduce the framework following the "why assess—what to assess—how to assess—who decides" roadmap, and educational administrators may use the five-layer structure as a design blueprint for evaluation systems.

6.3 Concluding Remarks

As artificial intelligence deeply penetrates educational evaluation, what determines its value is not the algorithm itself, but whether it genuinely serves assessment-for-learning. The FALA-AI framework offers one answer: let data construct evidence, let theory guide direction, and let humans master decisions.

References

1. Black P., Wiliam D. Inside the Black Box: Raising Standards Through Classroom Assessment. Phi Delta Kappan, 1998, No.80(2), pp.139–148.
2. Black P., Wiliam D. Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 2009, No.21(1), pp.5–31. DOI: 10.1007/s11092-008-9068-5.
3. Shute V.J. Focus on Formative Feedback. Review of Educational Research, 2008, No.78(1), pp.153–189. DOI: 10.3102/0034654307313795.
4. Hattie J., Timperley H. The power of feedback. Review of Educational Research, 2007, No.77(1), pp.81–112. DOI: 10.3102/003465430298487.
5. Carless D. Exploring learning-oriented assessment processes. Higher Education, 2015, No.69(6), pp.963–976. DOI: 10.1007/s10734-014-9816-z.

(cc) BY

6. Carless D., Boud D. The development of student feedback literacy: enabling uptake of feedback. Assessment & Evaluation in Higher Education, 2018, No.43(8), pp.1315–1325. DOI: 10.1080/02602938.2018.1463354.
7. Earl L. Assessment as Learning: Using Classroom Assessment to Maximize Student Learning. Thousand Oaks: Corwin Press, 2003. 176 p.
8. Siemens G. Learning analytics: The emergence of a discipline. American Behavioral Scientist, 2013, No.57(10), pp.1380–1400. DOI: 10.1177/0002764213498851.
9. Siemens G., Gasevic D., Haythornthwaite C., Dawson S., Shum S.B., Ferguson R., Duval E., Verbert K. Open Learning Analytics: An Integrated & Modularized Platform. Society for Learning Analytics Research (SoLAR), 2011. 119 p.
10. Lang C., Siemens G., Wise A.F., Gasevic D., Merceron A. (Eds.). The Handbook of Learning Analytics (2nd ed.). Vancouver: SoLAR, 2022. 280 p.
11. Buckingham Shum S., Deakin Crick R. Learning dispositions and transferable competencies: pedagogy, modelling and learning analytics. In: Proc. 2nd Int. Conf. on Learning Analytics and Knowledge (LAK '12). New York: ACM, 2012, pp.92–101. DOI: 10.1145/2330601.2330629.
12. Li Haifeng, Wang Wei. Research on learning analytics: Basic framework and core issues. Open Education Research, 2022, No.28(6), pp.60–71. DOI: 10.13966/j.cnki.kfjyyj.2022.06.007.
13. Barana A., Conte A., Fissore C., Marchisio M., Rabellino S. Learning analytics to improve formative assessment strategies. Journal of E-Learning and Knowledge Society, 2019, No.15(3), pp.75–88. DOI: 10.20368/1971-8829/1135057.
14. Stanja J., Gritz W., Krugel J., Hoppe A., Dannemann S. Formative assessment strategies for students' conceptions—The potential of learning analytics. British Journal of Educational Technology, 2023, No.54(1), pp.58–75. DOI: 10.1111/bjet.13288.
15. Tempelaar D., Rienties B., Giesbers B. Dispositional learning analytics and formative assessment: an inseparable twinship. International Journal of Educational Technology in Higher Education, 2024, No.21(1), 57. DOI: 10.1186/s41239-024-00489-8.
16. Gasevic D., Greiff S., Shaffer D.W. Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond. Computers in Human Behavior, 2022, No.134, 107304. DOI: 10.1016/j.chb.2022.107304.
17. Swiecki Z., Khosravi H., Chen G., Martinez-Maldonado R., Lodge J.M., Milligan S., Selwyn N., Gasevic D. Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence, 2022, No.3, 100075. DOI: 10.1016/j.caeai.2022.100075.
18. Gasevic D., Siemens G., Sadiq S. Empowering learners for the age of artificial intelligence (editorial). Computers and Education: Artificial Intelligence, 2023, No.4, 100130. DOI: 10.1016/j.caeai.2023.100130.
19. Khosravi H., Buckingham Shum S., Chen G., Conati C., Tsai Y.-S., Kay J., Knight S., Martinez-Maldonado R., Sadiq S., Gasevic D. Explainable Artificial Intelligence in education. Computers and Education: Artificial Intelligence, 2022, No.3, 100074. DOI: 10.1016/j.caeai.2022.100074.
20. Molenaar I. The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. Computers and Education: Artificial Intelligence, 2022, No.3, 100070. DOI: 10.1016/j.caeai.2022.100070.
21. Sajja R., Sermet Y., Cwiertny D., Demir I. Integrating AI and Learning Analytics for Data-Driven Pedagogical Decisions and Personalized Interventions in Education. arXiv, 2023, arXiv:2312.09548 [cs.CY]. URL: https://arxiv.org/abs/2312.09548 (accessed: 17.08.2026).
22. Hadyaoui A., Cheniti-Belcadhi L. IntelliFrame: A Framework for AI-Driven, Adaptive, and Process-Oriented Student Assessments. In: Proc. 20th Int. Conf. on Web Information Systems and Technologies (WEBIST), 2024, pp.441–448. DOI: 10.5220/0013070800003825.

(cc) BY

23. Chen X., Pi X. Human-in-the-Loop: A Conceptual Framework for Business English Teachers' AI-Empowered Assessment. In: Lecture Notes in Computer Science (Proc. ICWL 2024). Springer, 2025, pp.204–213. DOI: 10.1007/978-981-96-4407-0_17.
24. McGuire A., Qureshi W., Saad M. A Constructivist Model for Leveraging GenAI Tools for Individualized, Peer-Simulated Feedback on Student Writing. International Journal of Technology in Education, 2024, No.7(2), pp.326–352. DOI: 10.46328/ijte.639.
25. Guo H., Johnson M.S., Ercikan K., Saldivia L., Worthington M. Large-Scale Assessments for Learning: A Human-Centred AI Approach to Contextualizing Test Performance. Journal of Learning Analytics, 2024, No.11(2), pp.229–245. DOI: 10.18608/jla.2024.8007.
26. Zheng Qinhua. AI-empowered educational evaluation reform and innovation. Journal of Wuhan University (Natural Science Edition), 2025, No.71(5), pp.589–590. DOI: 10.14188/j.1671-8836.2025.0189.
27. Chen Li, Lin Wenjing, Zheng Qinhua. Innovative exploration of AI-empowered comprehensive quality assessment for primary and secondary school students. China Examinations, 2025, No.2, pp.41–49. DOI: 10.19360/j.cnki.11-3303/g4.2025.02.005.
28. Xin Tao, Nie Zhuming, Shi Wenjing, Zhang Sheng. Learning–evaluation integration: Conceptual shift and implementation of educational evaluation. Open Education Research, 2025, No.31(5), pp.30–38. DOI: 10.13966/j.cnki.kfjyyj.2025.05.004.
29. Wang Huaibo, Chai Huanyou, Guo Liming, Liu Zijing, Zheng Qinhua. Intelligent technology-empowered student comprehensive competency assessment: Framework design and implementation path. China Educational Technology, 2022, No.8, pp.16–23.
30. Yu Shengquan. The future role of AI teachers. Open Education Research, 2018, No.24(1), pp.16–28. DOI: 10.13966/j.cnki.kfjyyj.2018.01.003.
31. Peng Hongchao, Zhu Zhiting. Human–machine collaborative data wisdom mechanism: Data value alchemy in smart education. Open Education Research, 2018, No.24(2), pp.41–50.
32. Liu Sanya, Hao Xiaohan, Li Qing. A new paradigm of educational research: AI-driven educational science research. Educational Research, 2024, No.45(3), pp.147–159.
33. Zhang Zhi, Qi Yeguo. Construction of a multi-source multi-dimensional comprehensive quality assessment model based on big data. China Educational Technology, 2017, No.9, pp.69–77.
34. Tian Aili. Comprehensive quality assessment: Reform and implementation of learning assessment in the intelligent era. China Educational Technology, 2020, No.1, pp.109–113.
35. Luo Haifeng, Luo Yang, Liu Jian. Educational evaluation reform in the era of artificial intelligence. China Examinations, 2024, No.3, pp.8–17. DOI: 10.19360/j.cnki.11-3303/g4.2024.03.002.
36. Wu Yao, Li Luyue, Zheng Qinhua. Theoretical reconstruction and practical pathways of AI-empowered student comprehensive quality assessment. China Educational Information Technology, 2025, No.31(10), pp.111–118.
37. Huang Ronghuai, et al. Artificial Intelligence and Future Educational Development. Beijing: Science Press, 2023. 276 p.
38. Dua D., Graff C. UCI Machine Learning Repository [Dataset]. Irvine, CA: University of California, School of Information and Computer Sciences, 2017. DOI: 10.24432/C5SG7M. URL: https://archive.ics.uci.edu/dataset/320/student+performance (accessed: 17.08.2026).
39. Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset. Scientific Data, 2017, No.4, 170171. DOI: 10.1038/sdata.2017.171.

(cc) BY dtai.uz