Research on the Reconstruction of Intelligent Evaluation System for Higher Education Classroom Teaching Quality from the Perspective of Digital Intelligence Governance

Baorong Zhan[1], Lin Yuan[2], Feng Wang[3]

[1,2,3] *School of Computer Science and Artificial Intelligence, Guangzhou Xinhua University, Guangzhou,510520 China*
[3]*Corresponding author.*

ABSTRACT
University classroom evaluation is often assembled from questionnaires, periodic inspections, and peer observation. These sources describe overall perceptions but leave much of the within-classroom process unobserved. We formulate a multimodal evaluation scheme from the standpoint of digital-intelligence governance. The scheme separates acquisition, modality-specific processing, evidence integration, and feedback. Camera streams, classroom audio, LMS events, and environmental records are not converted directly into quality scores; they are used to derive observable descriptors such as participation events, speaking turns, response latency, instructional pacing, and room conditions. Evidence is then interpreted together with course context and human review. The design also specifies data-retention rules, role-based access, uncertainty reporting, and a route for contesting consequential decisions. Particular attention is given to temporal misalignment, missing modalities, construct validity, and discipline-dependent meaning of behavioral indicators. The resulting architecture is intended as a reference design rather than a validated scoring instrument. Its measurement properties and transferability must be examined through course-level and cross-institutional studies.

Keywords: *Digital intelligence governance, Higher education, Classroom teaching quality, Intelligent evaluation, Multi-modal data fusion.*

1. INTRODUCTION

Current university practice commonly combines student questionnaires, inspection records, and peer observation [1-3]. These instruments answer useful questions, such as how a course is perceived or whether basic teaching requirements are met, but they sample only selected moments of instruction. They seldom preserve the temporal sequence of questioning, student response, pacing, and participation. Sensors and learning platforms can fill part of that gap by producing time-stamped traces. The methodological issue is not data volume. It is whether a trace has a defensible interpretation and whether that interpretation can be audited when it enters a teaching-quality decision [4].

In this paper, digital-intelligence governance refers to the arrangement by which data collection, algorithmic processing, and human authority are linked [5]. This definition sets a boundary on automation. A model may summarize evidence or flag an interval for review, but it does not by itself determine teaching quality. The decision path must remain recoverable: what was collected, how it was processed, which uncertainty was reported, who reviewed it, and how the result was used. Such traceability is especially important because the same observable event can have different meanings in a lecture, laboratory, discussion class, or project session.

A smart classroom may provide video, audio, LMS events, response-system records, and environmental measurements. These streams are heterogeneous not only in format but also in sampling rate, missing-data pattern, and privacy risk. A low level of LMS activity, for instance, can

be expected in a paper-based exercise, while head orientation in video indicates where a student is facing rather than what the student understands. We therefore retain modality-specific limitations during analysis. Fusion is used to assemble complementary evidence, not to turn behavioral proxies into direct measures of attention, emotion, or learning.

This study develops a reference framework with three contributions. First, it specifies a four-layer architecture connecting data acquisition, feature extraction, multimodal analysis, and feedback. Second, it organizes candidate indicators into engagement, interaction, instructional delivery, affective/contextual conditions, and links them to feasible analytical methods. Third, it defines governance requirements for data minimization, access control, model interpretation, human review, and appeal. The framework is conceptual and design-oriented; it does not report a multi-institutional validation experiment. Performance claims are therefore treated as hypotheses for future testing rather than as established system outcomes.

2. CONCEPTUAL BASIS AND DESIGN PRINCIPLES

2.1 From Episodic Judgment to Process Evidence

Retrospective ratings remain useful because they capture perceptions and professional judgments [6-7], but their temporal resolution is coarse. Process data can add repeated observations, for example, speaking turns, response latency, participation events, and changes in teaching pace. The additional observations do not automatically increase validity. Each indicator needs an explicit interpretation tied to the decision for which it will be used; the consequences of that use are also part of the validity argument. A high-frequency measure that has an uncertain construct meaning is not necessarily better evidence than a less frequent but well-understood measure.

AI also changes the distribution of evidential authority in an evaluation system [8]. A model can generate summaries at a speed that human observers cannot match, but the availability of a summary does not settle its meaning. We therefore separate computation from judgment. Algorithms identify or quantify observable events; teachers and evaluators interpret those events against course objectives, instructional design, and local conditions. When a result is challenged, the system should expose the source record, processing step, and model uncertainty rather than present the output as a neutral fact. This separation makes responsibility easier to assign when an error or bias is discovered.

2.2 Socio-Technical Validity and Context Dependence

Contextual validity is important because an indicator that is meaningful in one classroom may be misleading in another [9]. Speaking frequency may be informative in a seminar but less informative in a laboratory demonstration; gaze direction may indicate orientation toward a screen rather than engagement; low platform activity may be normal in a paper-based class. The framework therefore treats indicators as evidence conditional on course type, teaching task, and data quality. Institution-level evaluation should compare similar contexts and should not infer teaching or learning quality from a single behavioral proxy.

Learning-science theories help identify potentially relevant observations, but they do not authorize a direct mapping from sensor signal to psychological state. Embodied cognition, cognitive-load theory, and control-value theory, for example, can motivate attention to posture, speech, task difficulty, and self-reported experience [10]. Operational interpretation still requires context. Where an inference concerns affect or cognitive load, self-report or teacher annotation can provide evidence that a camera or wearable cannot supply. Physiological and facial signals deserve a particularly high threshold for use because they are sensitive and their construct meaning can be unstable across individuals and settings. For this reason, the design favors triangulation with the minimum set of necessary sources.

2.3 Human-AI Division of Labor in Teaching Evaluation

Within this governance model, AI has a supporting rather than substitutive role in teaching evaluation [11]. Its most defensible use is to reduce the search space for human review. A system may, for example, mark an interval in which speaking balance changed sharply or response latency increased. The teacher can inspect that interval together with the lesson plan and classroom context before deciding whether the event has instructional significance. Analytical automation therefore concerns detection and summarization; pedagogical interpretation remains a professional task.

The same distinction allows standardization without assuming that every course should be judged by the same behavioral pattern. Data definitions, audit rules, and processing procedures can be common across the institution, while the indicators used for interpretation can vary. A seminar may make use of turn-taking measures; a laboratory class may place more weight on task progression or equipment interaction. Institution-wide indicators should therefore be restricted to constructs that remain comparable across settings, with course-specific indicators retained for local review. This arrangement avoids forcing disciplinary differences into a single rubric.

3. REFERENCE ARCHITECTURE OF THE INTELLIGENT EVALUATION SYSTEM

3.1 Data and System Architecture

Based on the functions required in previous work on digital governance and classroom analytics [12-13], the proposed architecture is divided into four operational layers. The collection layer acquires only the video, audio, platform events, or environmental measurements needed for a defined indicator. Modality-specific processing then converts raw records into time-stamped features. A fusion layer aligns those features or combines model outputs, and the application layer exposes the resulting evidence to authorized users together with context and uncertainty. The four layers describe responsibilities rather than a fixed implementation; individual institutions may use different sensors or model families.

Governance parameters are attached to the data flow at the point of collection. Each source is assigned a purpose, owner, retention period, and access scope. Shared timestamps and data dictionaries support alignment across systems; access controls, encryption, and audit logs record who can retrieve or modify the data. Retaining complete classroom video by default is unnecessary when an aggregate feature or short review segment can support the stated purpose. Limiting retention in this way reduces exposure of sensitive data and also makes later audits more informative.

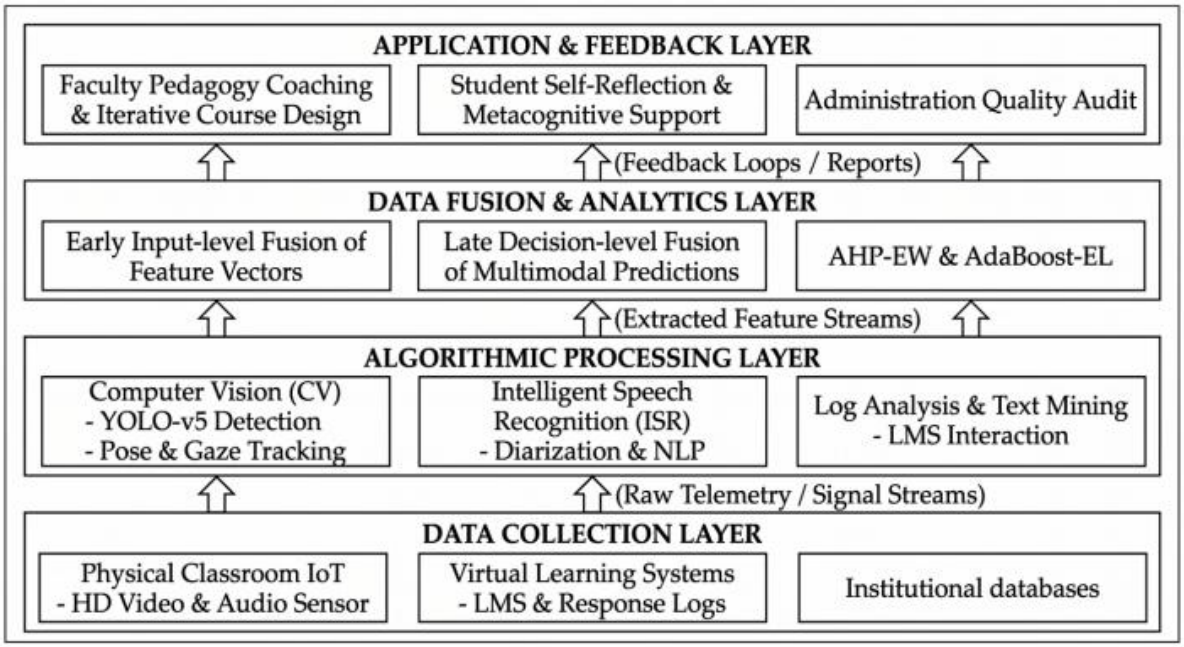


Figure 1 Systemic and multi-layered conceptual framework for classroom teaching quality evaluation.

"Figure 1" shows this sequence at system level. Classroom and platform records first remain separate while modality-specific features are extracted. After temporal alignment, the fusion stage produces indicator-level evidence; it does not collapse the lesson into a single irreversible score. For each indicator, the feedback layer can retain the relevant source interval, model confidence, and rubric entry. An authorized teacher or evaluator can therefore move from a reported value back to the evidence on which it was based. Depending on the permitted use, the same record may support lesson reflection, teaching consultation, or formal quality review.

3.2 Feature Extraction and Analytical Models

The processing layer converts raw video, audio, and textual records into features defined by the indicator scheme [14]. For video, redundant frames can be reduced before object or pose analysis. One simple option is correlation-based keyframe selection: a new frame is retained when the correlation coefficient R between adjacent frames falls below a threshold, indicating sufficient visual change:

$$R = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-\bar{A})(B_{ij}-\bar{B})}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij}-\bar{A})^2 \sum_{i=1}^{m}\sum_{j=1}^{n}(B_{ij}-\bar{B})^2}} \quad (1)$$

Here, m and n denote successive frame matrices, while $\bar{A}$ and $\bar{B}$ denote their mean pixel intensities. After keyframe selection, a detector or pose-estimation model can be applied to observable events such as hand raising, standing, or board interaction. For audio, automatic speech recognition and speaker diarization can estimate speaking time, turn-taking, question frequency, and response latency. These outputs are treated as observable descriptors; they are not, by themselves, direct measurements of comprehension or motivation.

Analytical methods should be selected by indicator rather than assigned once for the entire system. For classification or regression problems involving mixed behavioral features, AdaBoost, Random Forest, and XGBoost are reasonable candidates for comparison [15]; their suitability must be established on the target data. Weighting problems are different. When a teacher-evaluation construct contains an expert-defined hierarchy, AHP can represent that structure, while entropy weighting can reflect dispersion in the observed data. Combining the two is an option, not a default. Selection should be based on out-of-sample error, calibration, missing-data robustness, and consistency across course types.

3.3 Multimodal Fusion and Feedback

Fusion becomes difficult when modalities are sampled at different rates or disappear for part of a lesson. Early fusion places synchronized features in one representation before prediction. Late fusion keeps separate modality models and combines their outputs. We use the latter as a practical baseline when temporal alignment is weak or a source is intermittently unavailable, because an absent modality can be handled without reconstructing the entire feature vector. When reliable synchronization exists, early fusion can also be tested. No fusion rule is assumed to be generally superior; the choice is an empirical question for each indicator.

The output of the fusion stage is treated as a reviewable record. At minimum, that record should identify the indicator, the time interval, the contributing source, the model confidence, and the rule or model that generated the value. This information allows a teacher to return to a particular episode rather than interpret a score in isolation. In higher-stakes use, an evaluator can compare the same episode with the lesson objective, observation notes, or course materials. Keeping this evidential trail is what makes an automated result open to correction.

4. IMPLEMENTATION, INDICATOR DESIGN, AND GOVERNANCE

4.1 Phased Implementation and Validation

Deployment is organized around validation rather than immediate campus-wide use. The initial step is to define the decision being supported and to restrict data collection to sources required for that decision. A pilot can then be run in a small set of courses to estimate missing-data rates, model error, review workload, and user acceptance. The next step is comparison with existing evidence, such as peer-observation records or course artifacts, followed by revision of thresholds and rubrics where disagreement is systematic. Expansion to additional courses is justified only after these checks show that the indicators are sufficiently stable for their intended use. Work on AI-supported assessment likewise indicates that institutional policy and user capability must develop alongside the technical system [16].

Operational evidence should also feed back into governance. Rather than relying only on accuracy, institutions can track sensor failures, missing segments, disagreement with human review, appeals, and variation across course types. These records reveal whether the system fails technically or is being used outside its intended scope. Training for faculty should focus on interpretation: what each indicator represents, where uncertainty enters, and when an automated result should be ignored. Periodic review by teachers, students, technical staff, and administrators can then be used to revise rules that cannot be evaluated from model metrics alone.

4.2 Assessment Redesign in AI-Supported Learning

Generative AI alters what can be inferred from a finished student product. A final essay or program may show that a task was completed without showing how reasoning and authorship developed. Where process is part of the learning outcome, supplementary records such as drafts, oral explanation, revision history, peer contribution, and project milestones can provide relevant evidence [17]. These records are not intended as a general-purpose AI-detection mechanism. Their use should be specified in the assessment design, together with the forms of AI assistance that are permitted.

These process records can also be incorporated into the intelligent evaluation framework when they are pedagogically relevant. For example, oral-defense performance can be combined with project revisions to examine consistency between a submitted artifact and the student's explanation; peer-interaction records can document contribution patterns in group work. Such evidence should be interpreted at task level and should not be converted automatically into a general judgment about a student's integrity or ability. Human review remains necessary when automated indicators trigger consequential decisions.

4.3 Multimodal Indicator Matrix and Interpretive Boundaries

The proposed indicator matrix in "Table 1" organizes classroom evidence into five dimensions: engagement, cognitive interaction, instructional delivery, affective state, and environmental context. The table links each dimension to candidate data sources and analytical methods. Its purpose is to make the measurement chain explicit, from sensor or log to observable feature, analytical method, and intended interpretation, so that unsupported inference can be identified during design. Prior proof-of-concept research shows that machine-learning methods can classify selected teaching micro-skills, but such results do not establish the validity of a general classroom-quality score [17].

Table 1. Multi-modal indicator matrix for classroom teaching quality evaluation

Dimension	Source	Input	Feature	Method	Interpretive limit
Engagement	Video / events	Classroom video; response events	Head orientation; hand raising; activity episodes	Detection; pose / event counting	Behavioral participation only; not direct evidence of attention or learning
Interaction	Audio / text / LMS	Audio; transcripts; LMS logs	Speaking time; turns; questions; platform events	ASR; diarization; text / event analysis	Describes interaction structure; meaning depends on lesson task
Instruction	Video / text	Video; transcript; teaching materials	Question type; scaffolding; media use; pace	NLP; rule coding; AHP-EW if weighting is needed	Supports process review; requires a discipline-specific rubric
Affective state	Self-report; optional physiology	Brief survey; approved wearable signals	Self-reported load; optional HRV / conductance	Statistics or supervised models	Supplementary only; sensitive signals should not drive individual judgments alone
Environment	Room / IoT	Noise; CO2; light; temperature	dB; CO2; lux; temperature	Threshold / trend analysis	Context variable; not a teacher-performance measure by itself

"Table 1" is intended to keep interpretation close to the observable data. Speaking time, turn-taking, and response latency characterize interaction structure, but they do not establish whether the dialogue was productive. Head orientation and hand raising have the same limitation: they record behavior, not attention or learning. Reports should therefore show the indicator alongside its source interval, confidence information, and classroom context. A teacher may use that package to locate episodes for reflection; a formal evaluator should combine it with observation notes, course design, and student work rather than treat it as standalone evidence.

4.4 Ethical Governance, Reliability, and Limitations

When automated evidence enters a consequential decision, the burden of governance increases. Work on AI-based learning-outcome assessment emphasizes human oversight, transparency, technical robustness, and the possibility of contesting a result [18]. In the proposed design, a consequential output must be traceable to its source data and model information and must be reviewable by a qualified person. An appeal or correction route is therefore part of the evaluation process. Tools such as SHAP or LIME can assist interpretation for some predictive models, but the presence of an explanation method does not by itself make a decision understandable; explanation quality has to be assessed for the user and task.

Technical reliability also varies with local conditions. Dialects, classroom acoustics, occlusion, lighting, and network interruptions can all change model error rates. Institutions should validate speech and vision components on local data before operational use and should define minimum data-quality thresholds below which an indicator is withheld. Sensitive data, especially physiological or identifiable video data, require explicit necessity assessment, access restrictions, and limited retention. Where a less intrusive source can support the same evaluation purpose, it should be preferred.

Potential benefits should be evaluated rather than asserted. Automated preprocessing may reduce the time required to review long classroom recordings, but the amount of saved work depends on model accuracy, interface design, and the extent of required human verification. Likewise, automated feedback may encourage reflection for some users while creating surveillance concerns for others. Pilot studies should therefore measure workload, acceptance, error, behavioral change, and unintended effects alongside predictive performance. These implementation outcomes are part of system validity, not secondary operational details.

Further work is needed before the architecture can support routine institutional use. Validation should examine whether indicator performance transfers across courses, how missing modalities affect results, and whether reported confidence is calibrated. Subgroup analysis is also necessary because behavioral proxies may vary with discipline, culture, or accessibility needs. These tests should be repeated when models or data sources change. A deployed model should have explicit criteria for revalidation, restriction, or withdrawal if its performance or acceptability deteriorates [19-20].

5. CONCLUSION

We present a reference design for classroom teaching-quality evaluation in which multimodal analytics is embedded in an auditable governance process. The architecture links data collection, modality-specific processing, fusion, and feedback, but it does not give algorithms final evaluative authority. Purpose and retention rules are attached to the data source, indicator values remain traceable to supporting records, and consequential interpretations are reviewed by people with responsibility for the decision.

This design is not evidence that multimodal analytics yields a more objective or more accurate evaluation. That claim would require empirical comparison. Future validation should test the reliability of individual indicators, the behavior of fusion methods across course types, and the effect of feedback on instructional practice. Multi-course and multi-institution studies should also examine privacy, fairness, workload, and user acceptance. Until such evidence is available, the architecture is best understood as a specification for piloting and validation rather than as a ready-made scoring system.

ACKNOWLEDGMENTS

This research was supported by the Guangzhou Xinhua University Teaching Quality and Teaching Reform Projects for the 2025 Academic Year (Nos.2025J055, 2025J054, 2025J004).

REFERENCES

[1] HU Q T, WU W Y, FENG G, PAN T F, CHEN Z, QIU K X. Key technologies and practice of higher education teaching evaluation in the era of artificial intelligence[J]. Open Education Research, 2021, 27(5): 15-23.

[2] ZHU N X, GUO L J, YI J. Dilemmas and breakthroughs of university classroom teaching quality evaluation in the era of artificial intelligence[J]. China Agricultural Education, 2023, 24(4): 49-56.

[3] QIAN M X, ZHAO L L. Big data enabling undergraduate teaching quality evaluation: Value implication, practical dilemmas and path selection[J]. Chongqing Higher Education Research, 2023, 11(5): 78-89.

[4] LIU B Q, WANG Y X. Digital transformation of regional education governance: Challenges, logical framework and practical strategies[J]. China Educational Technology, 2023, (10): 89-97.

[5] LUO S Q, CHEN Z. Value reconstruction and logical rationale of educational evaluation in the era of big data[J]. Journal of Guizhou Normal University (Social Science Edition), 2023, (4): 116-128.

[6] NIE X L, ZHANG J X, HUANG Q Q, LIU B Q. Intelligent technology enabling digitalized governance of education: Connotation, framework and path[J]. China Education Informatization, 2024, 30(4): 21-31.

[7] TENG C L, DENG R P. From digital attachment to high-quality assessment: Paradigm transformation of higher education quality assessment[J]. Jiangsu Higher Education, 2023, (1): 17-24.

[8] WANG Z J, LI Y N. Transformation and restructuring of China's higher education evaluation system in the digital intelligence era[J]. University Education Science, 2024, (2): 106-117.

[9] LU X N, GAO X W. Artificial intelligence empowering educational evaluation reform: Development trends, risk inspection and countermeasures[J]. Chinese Journal of Education, 2023, (2): 48-54.

[10] GU R J, ZHAO L L. Teaching quality monitoring in vocational colleges in the intelligent era: Logic and dilemma breakthrough[J]. Chinese Vocational and Technical Education, 2021, (35): 11-18.

[11] ZHU J J. Design of intelligent classroom teaching monitoring and evaluation model in universities based on deep learning[J]. China Computer & Communication, 2024, 36(6): 95-97.

[12] GUO Y L, LIN Z X, XIAO X M. Research on the implementation of higher education quality evaluation driven by digital governance[J]. Straits Science, 2021, (5): 54-57.

[13] GUO J, BAI L, YU Z, ZHAO Z, WAN B. An AI-Application-Oriented In-Class Teaching Evaluation Model by Using Statistical Modeling and Ensemble Learning[J]. Sensors, 2021, 21(1): 241.

[14] TIAN K, HE Y, XU X, FU J. Toward a data analytics-based approach for computing and predicting context-sensitive teaching effectiveness[J]. Data Science and Informetrics, 2021, 1(4): 26-46.

[15] GRASSINI S. Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings[J]. Education Sciences, 2023, 13(7): 692.

[16] KOLADE O, OWOSENI A, EGBETOKUN A. Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework[J]. Heliyon, 2024, 10(4): e25953.

[17] DANN C, O'NEILL S, GETENET S, CHAKRABORTY S, SALEH K, YU K. Improving Teaching and Learning in Higher Education through Machine Learning: Proof of Concept of AI's Ability to Assess the Use of Key Microskills[J]. Education Sciences, 2024, 14(8): 886.

[18] MANGANELLO F, NICO A, BOCCUZZI G. Theoretical Foundations for Governing AI-Based Learning Outcome Assessment in High-Risk Educational Contexts[J]. Information, 2025, 16(9): 814.

[19] CROMPTON H, BURKE D. Artificial intelligence in higher education: The state of the field[J]. International Journal of Educational Technology in Higher Education, 2023, 20(1): 22.

[20] TANG Z, WEN T. Research on the Development Trend of Science and Education Evaluation in the Context of New Technology[J]. Data Science and Informetrics, 2022, 2(4): 74-89.