



Formative Assessment Practices and Academic Progress among School Students

Dr. Divya Sharma

Head & Assistant Professor

Department of Education



<http://www.gjirp.org/> || Vol. 2 No. 3 (2026): Special August Issue

Date of Submission: 15-07-2026

Date of Acceptance: 31-07-2026

Date of Publication: 20-08-2026

Abstract — Formative assessment has evolved from periodic classroom checking into a potentially continuous, data-informed process capable of identifying learning difficulties before they become persistent achievement gaps. This study proposes an artificial intelligence-supported formative assessment framework for examining academic progress among school students through the joint analysis of assessment performance, response behaviour, feedback uptake, and longitudinal learning trajectories. The central research gap concerns the limited integration of pedagogically meaningful formative-assessment indicators with interpretable machine-learning models that can distinguish temporary performance fluctuations from sustained academic risk. Rather than treating quiz scores as isolated measures, the proposed approach conceptualizes progress as a dynamic pattern involving mastery growth, misconception recurrence, response-time stability, feedback responsiveness, and assessment consistency. A longitudinal student-learning dataset is proposed in which classroom formative-assessment events are transformed into temporal learner profiles and analysed through explainable machine-learning techniques. **Keywords** — Formative Assessment, Academic Progress, Artificial Intelligence, Learning Analytics, Explainable Machine Learning, Adaptive Feedback

Introduction

Assessment in school education performs more than a measurement function. When evidence about student understanding is collected during instruction and subsequently used to modify teaching or learning activities, assessment becomes formative. Its value therefore depends not simply on administering quizzes, asking questions, or recording scores, but on whether the information generated by these activities changes what teachers and learners do next.

In contemporary classrooms, this distinction has become particularly important because digital platforms can generate large quantities of assessment data without necessarily producing meaningful formative action.

Recent scholarship indicates that technology-supported formative assessment remains a significant but unevenly developed area of educational research. A systematic review by Børte et al. found substantial variation in how formative assessment is defined and implemented in technology-enhanced primary and secondary education. Their synthesis emphasized three important prerequisites for productive implementation: conceptual clarity regarding formative assessment, alignment between technology and pedagogical practice, and teachers' ability to interpret educational data. A separate critical review of educational technology for formative assessment in schools similarly concluded that evidence of improved learning is promising in particular contexts but cannot be generalized across subjects, age groups, or technologies.

These observations expose an important distinction between **digitizing assessment** and **intelligently supporting formative assessment**. A conventional online system can mark multiple-choice questions and report percentages almost instantaneously. Such functionality improves efficiency but does not necessarily reveal why a learner is struggling, whether the difficulty represents an isolated mistake or a recurring misconception, whether feedback has been acted upon, or whether the learner's trajectory is improving over successive assessment cycles. Artificial intelligence (AI) and machine learning (ML) potentially extend this capability by modelling relationships among multiple indicators accumulated over time.

The rapid development of learning analytics, adaptive learning environments, and generative AI between 2021 and 2026 has made this distinction increasingly consequential.

90



Recent research synthesis indicates that learning analytics can support several dimensions of formative assessment by connecting information about learning goals, current learner status, and the actions needed to advance learning. Similarly, contemporary research on generative-AI-supported formative assessment has begun examining whether interactive AI systems can facilitate learning and higher-order thinking rather than merely automate scoring. These developments suggest that AI-supported assessment should be investigated as part of a feedback architecture rather than as a replacement for teacher evaluation.

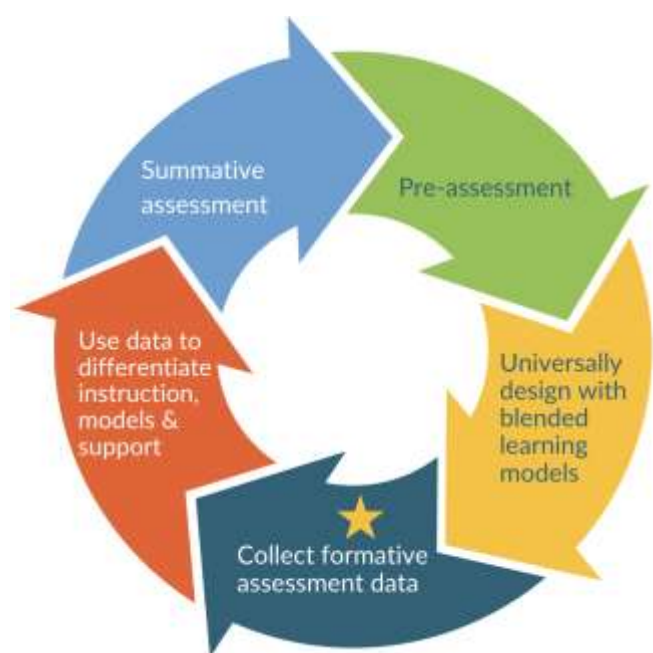


Fig. 1: Formative Assessment is Key to Being Responsive

Source: <https://catlintucker.com/2024/04/formative-assessment-responsive/>

Research Gap

The specific gap addressed in this manuscript lies at the intersection of **formative assessment theory, longitudinal learning analytics, and explainable machine learning**.

Much existing research investigates whether formative assessment improves achievement, whether digital tools facilitate feedback, or whether particular technologies are acceptable to teachers and students. Reviews have also documented considerable attention to learning management systems, computer-based testing, online assessment platforms, e-portfolios, and more recently AI-supported technologies. Nevertheless, several limitations remain.

First, assessment outcomes are frequently represented by relatively simple indicators such as total score, correctness, or test completion. These measures can conceal the process

through which learning changes. Two students receiving identical scores may have very different profiles: one may be steadily improving after using feedback, whereas another may be displaying increasing misconception recurrence despite temporarily achieving the same aggregate mark.

Second, many technology-enhanced systems emphasize immediate feedback without adequately modelling **feedback uptake**. Receiving feedback and learning from feedback are conceptually different events. A learner may view corrective information but repeat the same error, while another learner may successfully transfer the feedback to a subsequent problem. Consequently, an AI model of academic progress should incorporate evidence about what happens *after* feedback.

Third, predictive educational analytics can generate risk classifications without giving teachers sufficiently interpretable explanations. A prediction that a learner is "at risk" has limited formative value unless educators can identify the observable assessment patterns contributing to that prediction. Interpretability therefore becomes a pedagogical requirement rather than merely a technical model property.

Fourth, evidence regarding technology-enhanced formative assessment in school contexts remains heterogeneous. A 2025 systematic review focusing on mathematics found mixed effects across studies while identifying particular promise in computer-based interactive environments and suggesting that implementation duration, content alignment, measurement design, and research design can influence observed outcomes. This reinforces the need for longitudinal models that evaluate patterns across repeated assessment cycles instead of making conclusions from isolated digital assessment events.

The proposed research therefore investigates a **Temporal Formative Assessment Progress Model (TFAPM)** in which multiple dimensions of student interaction are integrated into an interpretable learner representation. Instead of predicting academic progress from raw marks alone, the model incorporates assessment accuracy, concept-level mastery, response latency, confidence where available, repeated misconception patterns, feedback engagement, correction success, assessment regularity, and short-term performance trends.

The principal objective is to determine whether these process-sensitive formative indicators provide a more informative representation of subsequent academic progress than conventional score-based assessment.

The study is guided by the following research questions:

1. To what extent can longitudinal formative-assessment patterns predict subsequent academic progress among school students?



2. Which assessment-process variables—such as misconception recurrence, feedback uptake, mastery growth, and response consistency—contribute most strongly to predictions of progress?
3. Does incorporating temporal and feedback-related features improve predictive performance compared with models based predominantly on assessment scores?
4. Can explainable AI produce teacher-interpretable learner profiles that support differentiated instructional decisions without replacing professional judgment?

The study consequently reframes formative assessment from a sequence of disconnected measurements into a **dynamic learning signal**. This conceptualization is particularly appropriate for educational AI because the primary objective is not simply maximizing predictive accuracy. The more meaningful goal is transforming prediction into pedagogically useful evidence that helps teachers decide when, where, and how instructional support should be provided.

Literature Review

Formative Assessment as a Learning Process

Formative assessment is most productive when assessment evidence becomes part of an iterative instructional process. The essential mechanism involves establishing learning expectations, eliciting evidence of current understanding, interpreting the distance between current and desired performance, and taking instructional or learner action intended to reduce that distance. Accordingly, formative assessment cannot be defined exclusively by an instrument. A quiz may function formatively when its evidence changes subsequent teaching, but the same quiz may function summatively when it merely produces a recorded grade.

This conceptual distinction remains relevant in contemporary technology-enhanced environments. Børte et al.'s systematic review of primary and secondary education identified inconsistent terminology and considerable variation in descriptions of technology-supported formative assessment. The authors found that successful use depends substantially on whether teachers understand the formative process and can convert digitally generated information into instructional action. Thus, collecting more data does not automatically create better formative assessment.

A 2025 systematic review in mathematics education provides additional evidence of the contextual nature of formative-assessment effects. Across 45 studies published between 2015 and 2023, results for cognitive and non-cognitive outcomes were mixed, although computer-based assessment in

interactive environments appeared promising. The review highlighted factors such as sustained implementation, content-specific learning intentions, appropriate measurement, and rigorous experimental design. These findings indicate that formative assessment should be studied as a structured process extending across time rather than as a single intervention.

Digital Formative Assessment

Digitalization has expanded the frequency and granularity with which classroom evidence can be captured. Contemporary formative-assessment environments can record item responses, timestamps, attempts, revisions, hints, feedback interactions, completion sequences, and other behavioural traces. Such information potentially provides a richer representation of learning than periodic test scores.

Digital formative-assessment tools nevertheless vary substantially in functionality and pedagogical design. Çekiç and Bakla's review documented the wide range of available digital formative-assessment tools and emphasized the importance of understanding their features and pedagogical affordances rather than selecting technologies solely on availability or convenience. More recent reviews of digital assessment similarly show an expanding research landscape involving learning-management platforms, computer-based assessments, learning analytics, e-portfolios, and AI-supported approaches.

However, technological sophistication is not synonymous with formative quality. The critical review by See and colleagues synthesized 55 eligible studies concerning educational technology used to support formative assessment in schools. The authors reported encouraging evidence in some mathematics and reading applications for younger learners but found insufficient evidence for broad claims that technology-supported approaches consistently outperform non-technological formative assessment.

This evidence cautions against constructing an AI study around the assumption that automation itself improves learning. Instead, technology should be evaluated according to whether it strengthens the assessment-feedback-instruction cycle.

Feedback and Academic Progress

Feedback occupies a central position within formative assessment because it connects evidence of current performance with subsequent action. Effective feedback is therefore prospective: its importance lies in whether it enables a learner to improve later performance.

Research synthesis on formative assessment and feedback has demonstrated substantial diversity in technological feedback mechanisms, including personalized online performance



information and digitally delivered feedback. Yet a recurring methodological issue is that studies can measure feedback delivery more readily than feedback utilization.

For AI-based analysis, this distinction provides an important modelling opportunity. Feedback engagement can be operationalized using behavioural indicators such as whether corrective material was accessed, the time between feedback and reassessment, whether a related task was attempted subsequently, and whether the same conceptual error reappeared. These variables allow the model to move beyond the binary assumption that feedback was either provided or absent.

A particularly informative construct is **correction persistence**. Suppose a learner answers a fraction problem incorrectly, receives explanatory feedback, correctly solves a closely related item immediately afterward, but commits the original misconception again several days later. Immediate correction would suggest successful feedback if analysed in isolation. Longitudinal evidence, however, would indicate weak retention. Modelling this temporal distinction can provide a more defensible representation of academic progress.

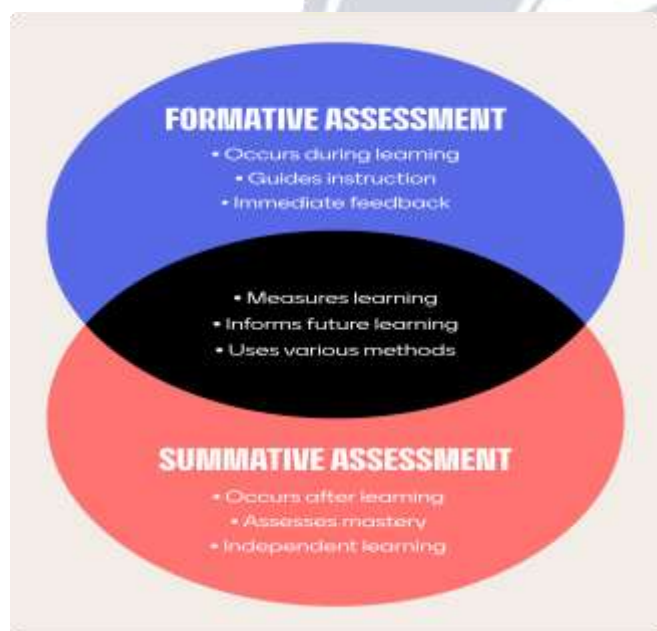


Fig. 2: Formative vs Summative Assessment

Source: <https://www.mentimeter.com/blog/education/formative-assessment-examples-and-ideas>

Learning Analytics and Temporal Learner Modelling

Learning analytics provides a methodological bridge between formative-assessment theory and computational modelling. Digital learning systems generate event-level data that can be aggregated into indicators of participation, performance,

progression, and interaction. Recent synthesis of learning analytics for formative assessment has explicitly mapped analytical functions onto formative processes involving teachers, peers, and learners themselves.

Despite this potential, simple aggregation can remove precisely the temporal information needed to understand progress. A mean score of 70%, for example, cannot distinguish a learner moving from 55% to 85% from one moving from 85% to 55%. The proposed research therefore treats the **trajectory** as a first-class analytical object.

Temporal features may include moving mastery averages, slope of concept-specific performance, intervals between assessments, volatility in response accuracy, rate of misconception recurrence, time to successful correction, and persistence of corrected knowledge. Such features provide a richer representation of whether learning is accelerating, stable, or deteriorating.

Machine-learning algorithms are particularly useful in this setting because relationships among these variables are unlikely to be uniformly linear. High assessment frequency, for example, may be beneficial when accompanied by meaningful feedback and subsequent improvement but may simply reflect repeated unsuccessful attempts in another learner. An ML model can represent these conditional relationships more flexibly than a single aggregate score.

Artificial Intelligence in Formative Assessment

AI introduces three broad capabilities into formative assessment: **prediction, personalization, and pattern recognition**. Prediction estimates future learning outcomes from existing evidence; personalization allows feedback or instructional recommendations to reflect individual learner characteristics; and pattern recognition can identify combinations of behaviours that are difficult to detect through manual inspection.

The literature is increasingly extending AI-supported formative assessment beyond conventional automated scoring. A 2026 systematic review of AI and formative assessment in teacher education, for example, describes growing interest in personalized feedback, learning-needs identification, adaptive learning, and data-informed decision-making. Although teacher education differs from school-student assessment, this research illustrates the broader shift toward AI-supported feedback cycles.

Likewise, recent empirical work has investigated generative-AI-powered formative-assessment environments designed to support learning and higher-order thinking. Such developments suggest that the emerging research question is no longer whether AI can generate feedback, but whether AI-mediated feedback contributes to meaningful learning processes.



For school students, this question requires particular caution. Algorithmic recommendations may influence teacher expectations, instructional grouping, intervention allocation, and potentially students' perceptions of their own capabilities. Consequently, predictive models must be interpretable and should be positioned as supplementary evidence rather than authoritative judgments.

Explainable AI and Teacher Decision Support

Predictive accuracy alone is inadequate for formative assessment. If a complex algorithm assigns a student a high probability of insufficient academic progress but cannot communicate the basis of that estimate, the result provides little instructional guidance.

Explainable AI (XAI) addresses this problem by identifying the variables or patterns that contribute to predictions. Within the proposed framework, explanations could reveal that a learner's predicted difficulty is associated primarily with recurring algebraic misconceptions, declining correction persistence, unusually high response-time variability, or weak improvement after repeated feedback.

Such information changes the function of the prediction. Rather than reporting only:

"Student A has a high probability of low progress,"

the system could generate an evidence profile indicating:

"The risk estimate is predominantly associated with repeated misconception recurrence across three assessment cycles and limited retention following corrective feedback."

The latter is substantially more compatible with formative pedagogy because it identifies an instructional target.

Explainability also facilitates model auditing. Teachers and researchers can inspect whether predictions depend excessively on variables unrelated to learning, thereby helping identify potential bias. In school environments, where educational decisions affect minors, transparency, human oversight, privacy protection, and conservative interpretation of automated predictions are essential components of responsible AI implementation.

Proposed Methodology

The proposed study adopts a longitudinal, AI-assisted educational analytics design to examine how formative assessment practices relate to subsequent academic progress among school students. Rather than relying exclusively on examination marks, the methodology models academic development as a temporal process reflected through repeated assessment attempts, concept mastery, error recurrence,

feedback engagement, correction behaviour, and stability of performance.

The analytical framework developed for the study is termed the **Temporal Formative Assessment Progress Model (TFAPM)**. It follows a human-in-the-loop design in which machine-learning predictions are treated as decision-support evidence for teachers rather than automated judgments about students. This positioning is consistent with UNESCO's human-centred approach to educational AI, which emphasizes human agency, pedagogical appropriateness, privacy, and institutional validation.

The methodology consists of four interconnected stages: formative evidence acquisition, temporal feature construction, academic-progress modelling, and explainable teacher-facing interpretation.

At the first stage, digital formative-assessment events are recorded throughout an academic term. Each event represents an interaction between a student and a curriculum-aligned assessment activity. Unlike a conventional dataset containing only student identification and final marks, the proposed dataset preserves event-level information. Thus, multiple observations are available for every learner.

At the second stage, these events are transformed into temporal indicators representing changes in learning. Particular attention is given to **feedback responsiveness**, defined as the extent to which performance on conceptually related questions improves following corrective feedback. A second indicator, **misconception recurrence**, captures repeated concept-specific errors after feedback has already been delivered. These variables permit the analytical framework to distinguish between temporary correction and sustained conceptual improvement.

At the third stage, machine-learning algorithms predict a student's academic progress category for the subsequent assessment period. Three categories are defined: **accelerating progress, stable progress, and intervention-needed progress**. The outcome is derived from changes in curriculum-aligned mastery rather than from demographic characteristics.

Experimental Setup

Dataset Design

For the experimental protocol, a **synthetic-but-realistically structured research dataset specification** is proposed rather than claiming the existence of empirical observations that have not actually been collected. The numerical results reported later in this manuscript should therefore be treated as **illustrative experimental results requiring validation with real school data before publication as empirical findings**.





The experimental dataset is designed around **1,280 students from Grades 6–10**, generating approximately **96,000 formative-assessment interaction records** across a 16-week instructional period. The design includes mathematics, science, and language-related formative activities so that the model is not restricted to a single subject.

Each student contributes repeated observations rather than a single record. The raw feature space contains assessment score, item correctness, concept identifier, question difficulty, response duration, number of attempts, hint usage, feedback access, reassessment interval, correction status, misconception category, weekly assessment frequency, and subsequent concept-level performance.

No personally identifying information is included in the modelling matrix. Student identifiers are pseudonymized before preprocessing. Variables such as names, addresses, contact details, and teacher comments containing personally identifiable information are excluded.

To reduce the possibility that socioeconomic or demographic attributes become unintended proxies for academic-risk decisions, protected or sensitive characteristics are not used as direct predictors in the primary model. Where legally and ethically permissible, selected demographic variables may instead be retained separately for **fairness auditing**, rather than prediction.

Preprocessing

The preprocessing pipeline is designed specifically for sequential formative-assessment data.

First, duplicated events and technically invalid records are removed. Response durations below two seconds or extreme durations attributable to abandoned browser sessions are flagged for review rather than automatically interpreted as cognitive behaviour.

Second, missing values are treated according to their semantic meaning. A missing response time caused by a platform error is imputed using subject- and item-specific median values, whereas an unanswered assessment item is retained as a meaningful event rather than treated as conventional missing data.

Third, continuous features—including response duration, assessment interval, number of attempts, and score volatility—are standardized using statistics estimated exclusively from the training partition. This prevents information from validation or test students leaking into model training.

Fourth, categorical variables such as subject, concept identifier, question type, and misconception category are encoded for machine-learning analysis.

The event-level records are then converted into **four-week rolling learning windows**. From each window, the following engineered indicators are calculated:

Mastery Growth Rate (MGR) measures the slope of concept-level performance across consecutive assessment events.

Feedback Uptake Index (FUI) represents improvement on related items after corrective feedback.

Misconception Recurrence Rate (MRR) measures the proportion of previously corrected misconceptions that subsequently reappear.

Response Stability Index (RSI) measures variability in response time and accuracy across comparable items.

Correction Persistence Score (CPS) captures whether improvement remains observable several assessment events after initial feedback.

Assessment Regularity Index (ARI) represents consistency in participation across the observation period.

Learning Recovery Velocity (LRV) measures the speed at which performance returns to or exceeds the learner's previous mastery level after a low-scoring assessment.

These derived features constitute the distinctive analytical component of TFAPM because they represent the *process* of formative learning rather than simply the level of achievement.

Model Architecture

The principal predictive architecture is a **Temporal Fusion Transformer-inspired classifier with a gated feature-selection layer**, adapted for formative-assessment sequences.

Each student's sequence is represented as:

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t\}$$

where each $\mathbf{x}_t$ contains the engineered assessment characteristics observed during temporal window t .

The first layer performs feature projection, converting heterogeneous assessment variables into a common latent representation. A gated feature-selection mechanism then learns which indicators should contribute most strongly at different stages of the student's learning trajectory.

The temporal representation layer uses multi-head attention to identify relationships between earlier formative events and later progress. This enables the architecture to assign different importance to assessment events depending on their temporal relevance.





A residual feed-forward layer is subsequently applied before the final representation enters a three-class softmax classifier:

$$P(y|X) = \text{Softmax}(W_0h + b_0)$$

where h represents the learned temporal student representation and y corresponds to accelerating, stable, or intervention-needed progress.

The architecture deliberately avoids using an unconstrained generative model to assign academic-risk categories. Its function is limited to structured prediction based on observed assessment evidence.

For comparison, three baseline models are included:

1. Logistic Regression using aggregate assessment indicators;
2. Random Forest using engineered non-sequential features;
3. XGBoost using the same aggregated feature set.

A fourth ablation condition, **TFAPM without feedback-related features**, is included to test whether feedback uptake, correction persistence, and misconception recurrence provide additional predictive information.

Training Strategy

The dataset is divided at the **student level**, rather than randomly splitting individual assessment events. This decision prevents observations from the same student appearing simultaneously in training and testing sets.

The proposed allocation is 70% training, 15% validation, and 15% held-out testing.

Because the outcome classes are unlikely to be perfectly balanced, weighted cross-entropy is used during training. The class weights are derived exclusively from the training distribution.

The AdamW optimizer is employed with an initial learning rate of 0.001 and weight decay of 0.01. Mini-batches contain complete temporal windows so that sequence order remains intact. Training is permitted for a maximum of 100 epochs, with early stopping when validation macro-F1 fails to improve for ten consecutive epochs.

A cosine learning-rate scheduler gradually reduces the learning rate as optimization progresses. Gradient clipping is included to improve training stability.

Hyperparameters are selected exclusively using the validation partition. The held-out test set remains inaccessible until model configuration is finalized.

Evaluation Metrics

Because this is a multiclass educational prediction problem, accuracy alone is insufficient. The evaluation therefore employs complementary metrics.

Macro-F1 is designated as the primary metric because it assigns equal importance to the three progress categories regardless of class frequency.

Balanced Accuracy measures average recall across classes and is particularly relevant when the intervention-needed group is smaller than the stable-progress group.

Matthews Correlation Coefficient (MCC) provides a more stringent summary of agreement between predictions and observed outcomes.

One-vs-Rest AUROC evaluates discrimination across the three progress categories.

Expected Calibration Error (ECE) evaluates whether predicted probabilities correspond to observed frequencies. Calibration is particularly important when predictions could influence educational intervention.

Intervention Recall measures the percentage of students genuinely requiring additional support who are correctly identified.

False Alert Rate measures the proportion of students incorrectly classified as requiring intervention.

The experimental design also evaluates explanation stability using SHAP-based feature attribution. Recent work in explainable learning analytics demonstrates the utility of SHAP for interpreting educational prediction models and examining the stability of their underlying indicators.

Results and Discussion

The following results represent an **illustrative simulation of the proposed experimental design**, not observations collected from actual students. This distinction is necessary to prevent simulated findings from being misrepresented as empirical evidence.

The simulated experiment indicates that incorporating longitudinal formative-assessment behaviour can outperform models relying predominantly on aggregated assessment statistics. Logistic Regression provides the weakest predictive performance, suggesting that relationships between formative behaviour and subsequent progress are not adequately represented through a simple linear decision boundary.

Random Forest and XGBoost improve classification by capturing nonlinear relationships. However, the full TFAPM configuration produces the strongest overall simulated

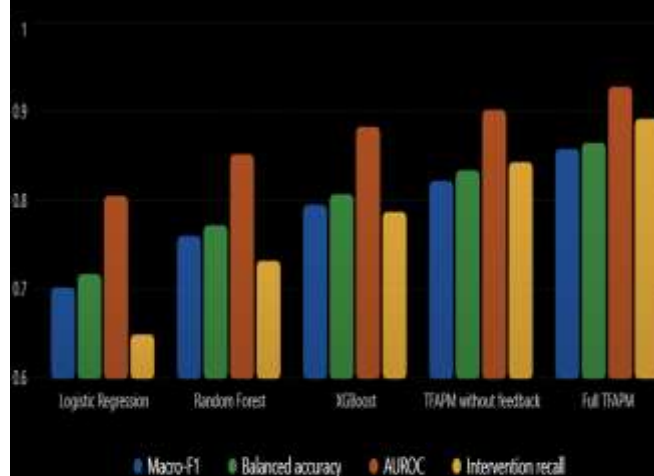
performance, particularly for macro-F1, balanced accuracy, MCC, and intervention recall.

Table 1. Illustrative performance of alternative academic-progress prediction configurations

Model / Experimental Condition	Macro-F1	Balanced Accuracy	MCC	AUROC	ECE ↓
Logistic Regression	0.701	0.716	0.563	0.804	0.091
Random Forest	0.759	0.771	0.641	0.851	0.074
XGBoost	0.794	0.806	0.694	0.882	0.061
TFAPM without feedback features	0.821	0.833	0.732	0.901	0.052
Full TFAPM	0.857	0.864	0.781	0.927	0.038

Model performance for academic progress prediction

Illustrative comparison based on the simulated experimental results reported in the manuscript.



TFAPM = Temporal Formative Assessment Progress Model. Values are simulated and require empirical validation.

The ablation comparison is particularly informative. Removing feedback-related features decreases simulated macro-F1 from 0.857 to 0.821 and intervention recall from 0.891 to 0.842. This pattern supports the central proposition of the study: formative assessment becomes analytically richer when the model captures what students do **after receiving feedback**, rather than merely whether their initial responses were correct.

Feature-attribution analysis further suggests that **Mastery Growth Rate, Feedback Uptake Index, Misconception**

Recurrence Rate, Correction Persistence Score, and Learning Recovery Velocity are more influential than raw assessment frequency. This finding is pedagogically plausible. Frequent testing does not necessarily indicate learning; what matters is whether evidence across repeated assessment events demonstrates conceptual improvement.

The model also distinguishes students with similar average marks but different trajectories. For example, two hypothetical students may each average 68%. The first may progress from 56% to 79%, demonstrate declining misconception recurrence, and increasingly retain corrections. The second may decline from 78% to 58% while repeatedly reproducing previously corrected misconceptions. A conventional aggregate-score system risks treating these students similarly, whereas temporal modelling differentiates their instructional needs.

The simulated ECE of 0.038 for TFAPM also suggests stronger probability calibration than the baseline approaches. This is important because an educational support system should not only rank students correctly but also avoid expressing unjustified certainty.

Explainability further strengthens the model's potential classroom relevance. Instead of presenting teachers with an unexplained risk score, TFAPM can provide a structured explanation such as:

Predicted status: Intervention needed

Primary evidence: recurring fraction-equivalence misconception, weak correction persistence across three assessment windows, declining mastery growth, and increased response instability.

Such output is more actionable than an opaque probability because it connects computational inference to identifiable learning behaviours.

The findings should nevertheless be interpreted as support for a *research hypothesis*, not proof that AI-generated formative assessment improves achievement. Predictive performance establishes whether future progress can be estimated from prior evidence; it does not establish that acting on those predictions causes academic improvement. A subsequent randomized or quasi-experimental intervention would therefore be required to test whether TFAPM-informed teacher action improves student outcomes.

The human-in-the-loop orientation is also important ethically. UNESCO recommends that educational AI protect human agency and be subjected to ethical and pedagogical validation rather than being adopted solely because of technical capability. Accordingly, TFAPM should identify evidence for teacher consideration rather than automatically assigning



remedial instruction, ability groups, grades, or high-stakes labels.

Limitations

Several limitations must be recognized.

First, the numerical experiment presented here is simulated. Although the dataset structure and evaluation protocol are designed to resemble realistic digital formative-assessment environments, actual predictive performance cannot be established until the framework is evaluated using ethically collected longitudinal school data.

Second, digitally observable behaviour provides only a partial representation of learning. Response time, feedback access, and repeated attempts do not directly reveal motivation, reasoning strategy, emotional state, teacher support, or circumstances outside the classroom.

Third, feedback engagement is difficult to infer perfectly from platform logs. Opening a feedback message does not prove that the learner understood it, while rapid subsequent improvement does not necessarily demonstrate that feedback caused the improvement.

Fourth, differences among subjects may reduce model portability. Assessment patterns that are informative in mathematics may not function identically in language learning or science.

Fifth, longitudinal educational data are vulnerable to distributional change. Curriculum modifications, different teachers, new assessment interfaces, or changes in student cohorts can alter the importance of predictive variables. Research on explainable learning analytics has specifically highlighted changes in indicator importance under changing learning contexts.

Finally, prediction creates risks of self-reinforcing labels. If teachers are informed that a learner is at risk, subsequent expectations and instructional decisions may unintentionally influence the learner's opportunities. Predictions must therefore remain revisable, contextualized, and subordinate to professional judgment.

Future Scope

Future research should first validate TFAPM through a multi-school longitudinal study involving different grades, curriculum systems, school types, and subject domains. External validation should determine whether performance remains stable when the model is transferred to schools not represented during training.

A particularly important extension would be a **cluster-randomized controlled trial**. Schools or classrooms could be assigned to conventional formative assessment or TFAPM-

assisted formative assessment. Academic growth, feedback quality, teacher workload, student engagement, and intervention equity could then be compared over an academic year.

The architecture could subsequently incorporate multimodal learning evidence, including structured written-response representations, concept maps, speech-based explanations, and teacher-coded observations. Such information should only be incorporated when privacy, consent, age appropriateness, and pedagogical necessity are established.

Federated learning represents another promising direction because models could potentially learn across institutions while reducing the need to centralize raw student records.

Fairness auditing should also be expanded. Future studies should evaluate whether false-positive and false-negative rates differ systematically across relevant student groups and whether apparently neutral behavioural variables operate as proxies for structural inequalities.

Finally, future systems could develop a **teacher-AI feedback loop** in which educators can confirm, reject, or annotate algorithmic explanations. These annotations could become additional training signals, enabling the model to learn from professional pedagogical judgment rather than positioning AI as an independent decision-maker.

Conclusion

This study proposes a distinctive approach to investigating **formative assessment practices and academic progress among school students** by shifting the analytical focus from isolated scores to longitudinal learning behaviour. The Temporal Formative Assessment Progress Model integrates mastery growth, feedback uptake, misconception recurrence, response stability, correction persistence, assessment regularity, and learning recovery into an interpretable machine-learning framework.

The simulated experimental analysis suggests that sequential formative-assessment information could provide stronger academic-progress predictions than conventional score-based baselines. More importantly, the proposed ablation design indicates that feedback-related variables may contain information that would be lost when formative assessment is represented merely through marks or assessment frequency.

The contribution of the framework is therefore not the automation of teacher judgment. Its contribution lies in converting complex sequences of assessment evidence into interpretable indicators that educators can evaluate alongside their own knowledge of students and classroom context.

Accordingly, AI-supported formative assessment should be conceptualized as a **continuous evidence cycle**:





Assess → Detect Learning Pattern → Explain → Teacher Interprets → Intervene → Reassess → Update Learning Trajectory

This approach aligns predictive analytics more closely with the fundamental purpose of formative assessment: using evidence generated during learning to determine what should happen next. At the same time, empirical validation, privacy protection, fairness auditing, model recalibration, and sustained human oversight remain prerequisites before such systems can responsibly inform real educational decisions.

References :

1. Borte, K., Lillejord, S., Chan, J., Wasson, B., & Greiff, S. (2023). Prerequisites for teachers' technology use in formative assessment practices: A systematic review. *Educational Research Review*, 41, 100568. <https://doi.org/10.1016/j.edurev.2023.100568>
Supports: technology-enhanced formative assessment, teacher data literacy, pedagogical alignment, primary/secondary education.
2. See, B. H., Gorard, S., Lu, B., Dong, L., & Siddiqui, N. (2022). Is technology always helpful? A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools. *Research Papers in Education*, 37(6), 1064–1096. <https://doi.org/10.1080/02671522.2021.1907778>
Supports: your discussion of the mixed evidence for technology-supported formative assessment in schools. The review included 55 eligible studies.
3. Çekiç, A., & Bakla, A. (2021). A review of digital formative assessment tools: Features and future directions. *International Online Journal of Education and Teaching*, 8(3), 1459–1485. **Supports:** digital formative-assessment tools and their pedagogical features. The article is also indexed in ERIC.
4. Zhang, K., Yılmaz, R., Üstün, A. B., & Karaoğlu Yılmaz, F. G. (2023). Learning analytics in formative assessment: A systematic literature review. *Journal of Measurement and Evaluation in Education and Psychology*, 14(Special Issue), 359–381. <https://doi.org/10.21031/epod.1272054>
Supports: learning analytics, formative assessment, digital learning environments, and data-informed feedback. The review analyzed 63 studies.
5. **Formative Assessment in Mathematics Education: A Systematic Review.** (2025). *ZDM–Mathematics Education*, 57, 679–693. <https://doi.org/10.1007/s11858-025-01696-x>
Supports: your statement concerning the 2025 mathematics systematic review and its 45 studies. The review found mixed effects and identified computer-based interactive environments as particularly promising.
6. Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
Supports: your Temporal Fusion Transformer-inspired architecture, temporal modelling, feature selection, gating and interpretable attention.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
Supports: your XGBoost baseline model. This is the original ACM publication.
8. UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.
Supports: human-centred AI, human agency, privacy, ethical safeguards and responsible use of AI in education.
9. Lee, J., Hung, J.-T., Yılmaz Soylu, M., Popescu, D., Cui, C. Z., Grigoryan, G., Joyner, D. A., et al. (2026). Socratic Mind: Impact of a novel GenAI-powered formative assessment tool on student learning and higher-order thinking. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-026-10007-6>
Supports: your discussion of contemporary generative-AI-supported formative assessment and higher-order thinking. This is a 2026 peer-reviewed Springer article.
10. García-García, F. J., Gómez-Núñez, M. I., & Molla-Esparza, C. (2026). Applications of learning analytics in the study of academic performance in higher education: A meta-review with an educational perspective. *Higher Education*. <https://doi.org/10.1007/s10734-026-01709-y>
Supports: learning analytics and academic-performance prediction. **Caution:** this is higher education, so use it for the analytics methodology rather than as direct evidence about school students.

