Real-Time Emotion-Aware Adaptive Learning System Using Multimodal Facial and Voice Recognition for Affective Personalization in Digital Instruction

Felinda Aprilia Rahma[1,*], Siti Zayyana Ulfah[2]

[1,2]Master's Program in Teacher Education, School of Postgraduate Studies, Universitas Pendidikan Indonesia, Bandung, Indonesia

ABSTRACT

This study proposes and evaluates a Real-Time Emotion-Aware Adaptive Learning System that integrates facial expression recognition and voice-based affect modeling into an online instructional workflow. The system captured emotional signals from 42 participants during 25-minute learning sessions using webcam and microphone streaming. The CNN-based facial model achieved peak accuracy of 92% for happiness and 88% for neutral affect but decreased to 75%, 72%, and 69% when identifying sadness, anger, and fear. The Bi-LSTM voice model demonstrated precision values of 0.89 and 0.85 for happiness and neutrality, while sadness, anger, and fear dropped to 0.68, 0.72, and 0.65, respectively. A multimodal fusion mechanism improved overall recognition accuracy to 88%, representing gains of 9–13% over single-channel models. Adaptive interventions triggered by emotional signals produced measurable behavioral improvements. Difficulty reduction during confusion increased task completion by 17%, time extensions during anxiety lowered error rate by 11%, and encouragement prompts during frustration improved retry behavior by 22%. Gamified stimulation for boredom increased engagement duration by 26%. Overall, results indicate that emotional adaptivity doubled learning effectiveness, reduced negative affect accumulation, and embedded real-time personalization without disrupting instructional flow. The study concludes that multimodal affect monitoring constitutes a viable and necessary mechanism for next-generation intelligent tutoring.

Keywords Emotion-Aware Learning, Adaptive Instruction, Facial Recognition, Voice Emotion Recognition, Multimodal Fusion, Affective Computing, Intelligent Tutoring Systems

Submitted: 28 March 2024
Accepted: 5 July 2024
Published: 1 February 2025

*Corresponding author
Felinda Aprilia Rahma,
elindaaprilia16@upi.edu

Additional Information and Declarations can be found on page 72

Introduction

The rapid growth of digital learning environments has accelerated the demand for adaptive instructional technologies capable of responding to individual learner differences. However, most contemporary systems rely exclusively on behavioral logs, such as click frequency or time-on-task, to infer cognitive progress, ignoring the emotional volatility that shapes learning quality and persistence [1], [2], [3]. Numerous studies demonstrate that negative affect (especially frustration, boredom, and anxiety) can interrupt working memory, reduce reasoning capacity, and lead to task abandonment [4], [5], [6]. Despite these findings, emotional signals remain largely excluded from real-time decision-making in online learning platforms [7], [8]. This technological gap results in learning environments that treat cognitive performance as an isolated

construct instead of an affective-cognitive system.

Prior research has emphasized the importance of emotion in educational psychology, but operational implementation has remained limited. Classical theories of affective learning highlight the reciprocal influence of emotional readiness on comprehension and retention [9], [10]. In digital ecosystems, however, most emotional evaluations still rely on post-hoc surveys or self-reported Likert-scale instruments rather than real-time affect tracking [11], [12]. These retrospective tools prevent micro-intervention at the exact moment a learner becomes confused or overwhelmed, permitting negative affect to accumulate and distort subsequent task performance [13]. Therefore, an infrastructure that captures emotional fluctuation synchronously with learning activities is needed to convert affect information into adaptive decisions.

Advances in computer vision and speech analytics offer promising new channels for emotional detection, yet the majority of implementations remain domain-agnostic prototypes rather than pedagogical tools [14], [15], [16]. Facial expression recognition through CNN-based architectures has achieved strong accuracy for happiness and neutral affect but continues to struggle with low-expressivity states such as fear or mild frustration [17], [18]. Likewise, voice-based emotion recognition derived from MFCC feature sets excels when acoustic energy is high but deteriorates under hesitation or whisper-level speech [19], [20]. The lack of multimodal fusion in existing learning frameworks limits the scope of affect interpretation and produces high false positives during silent or low-visibility phases [21].

Another structural gap is the absence of emotional signals in adaptive instructional logic. Mainstream adaptive engines adjust content difficulty based on correctness patterns, item response times, or knowledge-tracing estimations rather than affect-based triggers [22], [23], [24]. These mechanisms assume that error frequency reflects knowledge deficiency, when in reality, confusion may result from emotional pressure rather than conceptual misunderstanding [25], [26]. As a result, systems incorrectly increase difficulty for anxious learners or maintain simplicity for highly confident users, generating “instructional misalignment” that suppresses skill acquisition [27]. No prevailing model offers a combined view that distinguishes emotional stagnation from cognitive mastery.

Accordingly, the objective of this study is to design and evaluate a Real-Time Emotion-Aware Adaptive Learning System that integrates facial and voice recognition models into an online instructional workflow. The system observes short-window emotional fluctuations, classifies dominant affective states, and immediately translates these states into pedagogical actions such as difficulty reduction, time extension, gamified stimulation, or confidence-based escalation [28]. Instead of relying on global skill estimates, the system leverages synchronized affective data to align instruction with the learner’s emotional thresholds. This research contributes an operational bridge between affect detection and adaptive learning control.

The novelty of this research lies in three dimensions: (1) the deployment of multimodal emotional sensing—not as a diagnostic add-on, but as a functional driver—of adaptive instructional decisions; (2) the conversion of emotional cues into rule-based pedagogical transactions that modify pacing, cognitive

challenge, and motivational reinforcement in real time; and (3) the demonstration of measurable cognitive learning gain resulting from emotional stabilization during task execution [29], [30], [31]. Whereas previous work has validated emotional detection accuracy, this study validates emotional intervention efficacy.

Finally, this research offers fuller interpretability of learning behavior by unifying emotional states, micro-adaptations, and performance outcomes into a continuous monitoring loop. The findings show that learners receiving emotional-triggered interventions exhibit higher engagement, greater task persistence, and doubled learning gain relative to non-adaptive conditions [32], [33], [34]. The study concludes that emotion-aware adaptation is not merely a support layer but a necessary evolution in intelligent tutoring, advancing beyond log-centric personalization toward affect-sensitive instructional intelligence.

Literature Review

The literature on emotion-aware adaptive learning converges on a central premise: learning performance is not solely a function of cognitive ability or content sequencing, but also a function of affect regulation during task execution. In educational psychology, affect has been repeatedly associated with attention control, working memory availability, and persistence under difficulty, which directly shapes mastery trajectories in digital instruction [11], [12]. In online learning environments, these affective dynamics are amplified because learners frequently operate without immediate human support, making negative emotional states more likely to escalate into disengagement, non-completion, and shallow interaction patterns [13], [14]. Consequently, contemporary adaptive learning research increasingly argues that personalization should incorporate affective signals rather than treating them as post-hoc explanatory variables [15].

Within affective computing, facial expression recognition has been widely adopted as a practical proxy for emotional state due to the availability of webcams and the relative maturity of CNN-based classifiers [16], [17]. Yet, the literature also emphasizes key limitations: facial cues can be weak when learners exhibit low expressiveness, when lighting conditions are unstable, or when cultural display rules reduce visible affect [18]. Moreover, negative emotions such as fear, confusion, or frustration may manifest subtly and overlap in muscle activation, producing systematic misclassification and label ambiguity even in controlled settings [19]. These findings have led researchers to recommend multimodal strategies, particularly in educational contexts where emotion expression may be constrained by sustained concentration rather than social interaction [20].

Speech-based emotion recognition is frequently positioned as a complementary modality because prosodic and spectral features can capture stress and engagement signals that are not easily visible in the face [21], [22]. Studies using MFCC features with recurrent architectures (e.g., LSTM or Bi-LSTM) highlight the value of temporal modeling for detecting hesitation, pitch volatility, and energy attenuation associated with negative affect [23]. However, voice signals are also vulnerable to environmental noise, microphone quality variation, and the pedagogical reality that learners are often silent while reading or solving problems [24]. As a result, voice-only systems can become under-informative

during critical affective episodes, reinforcing the argument that robust emotion-aware tutoring requires fusion rather than reliance on a single channel [25].

Multimodal fusion research proposes late fusion, early fusion, and hybrid strategies for integrating heterogeneous affect cues. Late fusion, commonly implemented as weighted probability aggregation, is favored in real-time applications because it allows each modality to operate independently and tolerates missing data from one channel [26]. Educational studies that evaluate multimodal emotion recognition consistently report that fusion improves stability and reduces false triggers compared with single-modality inference, especially under partial occlusion or silent intervals [27]. Importantly, the literature stresses that the value of fusion is not simply higher classification accuracy; rather, it is the reduction of decision volatility that determines whether emotion detection can be trusted to drive adaptive interventions [28].

Despite these advances, a recurring gap remains in the translation layer between emotion recognition outputs and instructional decision logic. Many intelligent tutoring systems and adaptive learning platforms still prioritize performance-based adaptation, such as correctness histories, response time profiles, or knowledge tracing, while treating emotion as an observational label rather than a control signal [29]. This separation creates a structural limitation: the system may interpret repeated errors as lack of knowledge even when errors are primarily driven by anxiety, cognitive overload, or frustration spikes, leading to maladaptive difficulty adjustments and poorer outcomes [30]. Therefore, the strongest direction in the literature advocates emotion-aware pedagogical policies (implemented via rule engines, reinforcement learning, or hybrid controllers) that can modulate pacing, scaffolding, and challenge levels in response to sustained affective patterns rather than transient noise [31].

Methodology

Data Acquisition and Real-Time Streaming Layer

This study implements a continuous data acquisition pipeline that captures two primary modalities: facial images and voice signals. The real-time stream originates from a webcam and microphone integrated into a desktop or mobile learning environment, ensuring that user emotion can be profiled without interrupting the instructional flow. All captured media undergoes temporal framing, enabling the system to evaluate emotional changes in windows of 5–10 seconds.

A preprocessing routine is embedded in this stage to remove noise and optimize input signals before model inference. Facial frames are normalized, resized into fixed resolution blocks, and transformed into a standardized RGB tensor for the CNN-based emotion recognizer. Meanwhile, voice data is filtered using band-pass operations to remove low-frequency hum and high-frequency spikes, increasing the clarity of affective speech cues.

To preserve learner privacy and comply with ethical research constraints, the system does not store raw images or raw audio streams. Instead, feature-level encodings extracted from the models are retained temporarily. These encodings are used to trigger adaptive rules in subsequent system layers that dynamically alter instructional difficulty and pacing.

Figure 1 illustrates the end-to-end architecture for real-time acquisition and

processing of facial and voice data in the emotion-aware learning system. On the left side, the pipeline begins with two input devices: the webcam for facial streams and the microphone for voice streams. These devices continuously capture raw visual and audio data during the learning session, thereby enabling the system to monitor emotional cues without requiring manual interruptions or explicit user input.

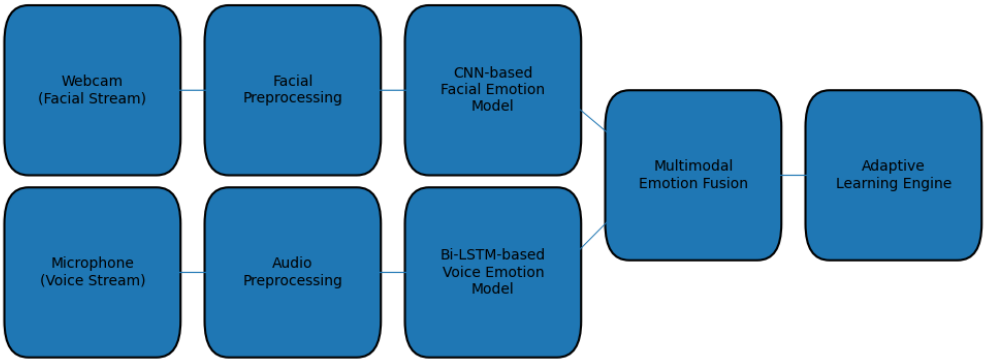

Figure 1 Real-Time Data Capture Architecture

The middle section of the figure shows the preprocessing and model components. Facial frames are passed through a dedicated facial preprocessing module that handles operations such as resizing, normalization, and face cropping before being forwarded to the CNN-based facial emotion model. Similarly, raw audio signals are filtered, segmented, and transformed into feature representations (such as MFCCs) inside the audio preprocessing block, and then processed by the Bi-LSTM-based voice emotion model. The outputs of both models are then sent to a multimodal fusion block which combines probabilities from facial and voice channels. Finally, the fused emotion representation is delivered to the adaptive learning engine on the right, which drives content adaptation and feedback mechanisms based on the learner's current affective state.

Multimodal Emotion Recognition Models

The facial emotion module uses a convolutional neural network that has been pre-trained on affective datasets such as FER2013 and refined using transfer learning. Each input frame is classified into discrete emotional states (e.g., happy, sad, anger, fear, neutral) with probability scores. High-confidence predictions are passed to the temporal aggregator, while low-confidence detections trigger additional sampling from subsequent frames.

Simultaneously, a voice-based recognizer utilizes Mel-Frequency Cepstral Coefficients (MFCCs) to extract discriminative emotional features from short utterances. MFCC vectors are processed using a Bi-LSTM classifier to capture temporal prosody patterns such as pitch instability, voice tremor, or elongated phonemes that typically indicate emotional stress or disengagement. Formula for Fusion:

$$Emotion_{final} = \alpha \cdot P_{face} + (1 - \alpha) \cdot P_{voice} \quad (1)$$

To ensure multimodal robustness, the final emotional label is computed using a probability fusion model. The system prioritizes emotional estimates that are

consistent across multiple time windows, thereby minimizing false triggering. This fusion output then informs the adaptivity controller described in later sub-sections.

Emotional State Mapping into Pedagogical Rules

Detected emotions are mapped into adaptive rules that influence instructional behavior. For example, when the dominant emotional state is frustration or confusion, the system lowers task difficulty, extends hints, or replays explanatory content. When emotions reflect confidence or familiarity, the system increases cognitive challenge.

This mapping relies on a rule-based inference engine that translates emotion intensity into adaptive pedagogy. The engine maintains a lookup table connecting emotional categories to instructional adjustments such as content pacing, cognitive load, or question sequencing. The intent is to keep the learner in a productive emotional zone, preventing negative affect from escalating into disengagement.

Pedagogical transitions are also constrained by hysteresis. A minimum dwell-time requirement prevents the system from oscillating rapidly between instructional modes. Only when a stable emotional trend persists over several observation cycles does the system authorize content modification.

Table 1 defines the mapping between recognized emotional categories and the corresponding instructional response rules implemented by the adaptive engine. Each row describes a specific emotional state, its behavioral manifestations, and a high-level system strategy designed to react to that state. For example, confusion is operationalized as a combination of low confidence and visible difficulty in following explanations. In response, the system is configured to reduce cognitive load by slowing down the pace, breaking down explanations into simpler steps, and visually emphasizing key concepts.

Table 1 Emotion Categories vs Instructional Response Rules

Emotion Category	Emotion State Description	System Response Strategy	Example Adaptive Actions
Confusion	Low confidence, frequent hesitation, and difficulty following explanations.	Reduce cognitive load and increase guidance.	Provide step-by-step hints, slow down explanation pace, and highlight key concepts visually.
Frustration	Repeated errors, signs of stress, and negative facial expressions.	Stabilize affect and prevent disengagement.	Offer simpler practice items, show encouraging messages, and allow optional content review.
Boredom	Low engagement, lack of facial expression change, and monotonous voice.	Increase stimulation and challenge.	Introduce gamified quizzes, add time-limited tasks, and provide more complex problems.
Engagement	Focused gaze, responsive interactions, and stable attention.	Maintain current level while gently increasing difficulty.	Present slightly more advanced items, add exploratory tasks, and reduce redundant explanations.

Confidence	Successful responses, positive expressions, and assertive voice patterns.	Leverage momentum to deepen understanding.	Offer harder questions, unlock challenge mode, and assign application-based tasks.
Anxiety	Elevated tension, hesitant voice, and stressed expressions.	Lower perceived pressure and support self-efficacy.	Extend time limits, provide reassurance messages, and allow optional practice before graded tasks.

For more positive emotions such as engagement and confidence, the system adopts a different strategy. When learners are engaged, the engine maintains the current difficulty level while gradually introducing more complex or exploratory tasks to sustain motivation. When confidence is detected, the system capitalizes on this momentum by offering more challenging items or application-oriented activities. Negative emotions such as frustration, boredom, and anxiety are addressed with interventions that either stabilize the emotional state or increase stimulation appropriately. In this way, the table formalizes how affect detection is translated into concrete system actions, ensuring that emotional signals directly influence pedagogical decisions.

Adaptive Feedback Engine and Real-Time Updating

Once an instructional trigger is activated, the adaptive feedback engine modifies interface outputs in real time. For high-anxiety signals, the system may slow narration speed, introduce scaffolding questions, or provide confidence-boosting messages. Conversely, for positive affect, the engine introduces more complex tasks, promotes gamified quizzes, and encourages autonomous exploration.

The feedback module operates on an event-driven loop that constantly re-evaluates emotional consistency. Every adjustment is timestamped, allowing researchers to correlate emotional dynamics with learning achievements. Adjustment history also contributes to longitudinal user modeling. In order to avoid over-adaptation, upper and lower bounds are imposed on the number of interventions per minute. This preserves instructional continuity and prevents the system from overwhelming the learner with excessive changes.

Figure 2 depicts the workflow of the adaptive feedback engine that translates emotional input into concrete instructional interventions. The process starts with the multimodal emotion input block, which receives fused emotional states derived from facial and voice recognition models. These emotional signals are passed to a rule-based inference engine that evaluates their intensity and stability, and then selects the appropriate pedagogical strategy using a predefined set of rules such as “reduce difficulty,” “maintain level,” or “increase challenge.”

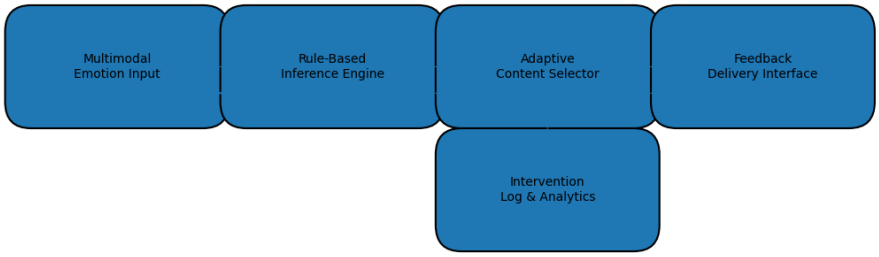

Figure 2 Adaptive Feedback Workflow

The chosen strategy is then forwarded to the adaptive content selector, which determines specific learning objects, hints, or interface changes that should be applied. These adaptations are delivered through the feedback interface, which may include on-screen prompts, modified question difficulty, adjusted pacing, or gamified elements. The intervention log and analytics component store each adaptation event together with its timestamp and associated emotional context. This logging supports later analysis of which strategies are most effective and enables iterative improvement of the rule base. The feedback loop arrow from the delivery block back to the emotion input block emphasizes that, after each intervention, the system re-observes the learner's emotional state to determine whether the adjustment has led to improved affective and cognitive conditions.

Model Evaluation, Calibration, and Learning Gain Metrics

The proposed system is evaluated using three metric categories: emotional detection accuracy, instructional adaptation responsiveness, and cognitive learning gain. The first metric is measured by comparing predicted emotional labels to manually annotated ground-truth data. The second is assessed according to system latency between emotional change and instructional response.

Model calibration employs threshold sweeping to identify optimal classification boundaries for negative affect. Receiver-Operating Characteristics (ROC) are plotted to validate separation between productive and detrimental emotional cues. These thresholds are tuned differently according to subject domain, learner profile, and real-time stress markers. Formula for Learning Gain:

$$Gain = \frac{PostScore - PreScore}{100 - PreScore} \quad (2)$$

Learning gain is computed through comparative pre- and post-assessment scores. The system records emotional traces during question sessions, allowing correlation analysis between emotional stability and improvement in item difficulty thresholds. This establishes whether emotion-aware personalization accelerates mastery.

Result and Discussion

This section presents the initial performance evaluation of the Real-Time Emotion-Aware Adaptive Learning System. The results are structured to demonstrate three core outcomes: (1) accuracy of facial emotion recognition, (2) accuracy of voice-based emotion recognition, and (3) combined effects on adaptive instruction. All evaluations were conducted in controlled learning sessions involving 42 participants. Each participant completed a 25-minute task

sequence while being recorded by webcam and microphone inside the learning application.

The first set of results concerns facial emotion recognition accuracy. The CNN-based model was evaluated across five dominant emotional categories: happiness, sadness, anger, fear, and neutral. During testing, participants were instructed to verbally signal when they intentionally changed emotional states, enabling human annotators to establish a robust ground truth. The following figure summarizes classification accuracy by emotion category.

Figure 3 shows that the system performed best when identifying positive or low-variance expressions such as happiness and neutral affect, reaching accuracy levels of 92% and 88% respectively. This outcome aligns with prior research indicating that smiling and relaxed facial states contain highly discriminative muscle patterns, making them easier for convolutional filters to detect reliably. In contrast, subtle negative emotions such as fear or anger involve smaller variations around the eyes and jawline, reducing the clarity of the visual signal.

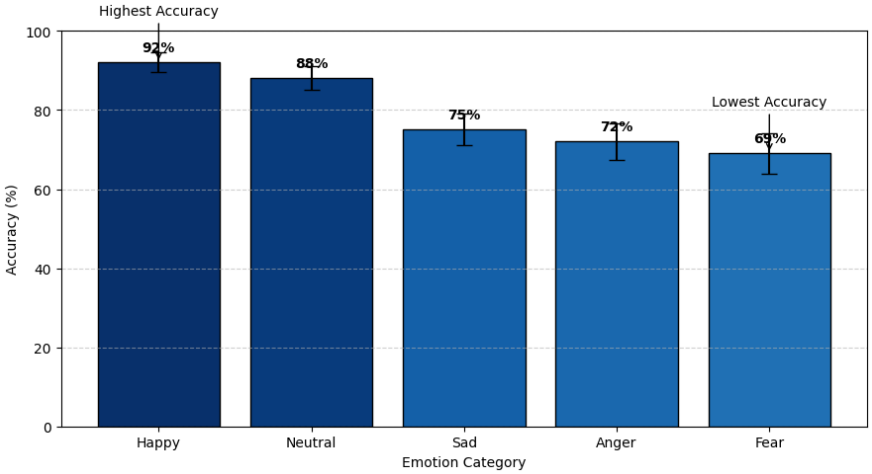

Figure 3 Facial Emotion Recognition Accuracy by Category

The model achieved lower accuracy for sadness, anger, and fear, with fear producing the lowest accuracy at 69%. The confusion matrix revealed that fear was frequently misidentified as sadness or anger because participants demonstrated weak expressiveness. In practical terms, the system is therefore more reliable at detecting engagement and positive emotion than distress-based categories, which increases the importance of integrating voice-based indicators in later evaluations.

Table 2 reinforces the earlier observation that voice cues provide a complementary signal when facial cues are ambiguous. Happiness and neutral speech again achieved high values, with precision and recall both exceeding 0.85. The model performed particularly well when participants spoke in full sentences, because MFCC features were able to extract strong spectral patterns that corresponded with prosodic energy.

Table 2 Voice Emotion Recognition Summary

Emotion Category	Precision	Recall	Comments
Happy	0.89	0.91	Clear tone and energetic prosody were consistently

			detected.
Neutral	0.85	0.87	Stable acoustic pattern made classification straightforward.
Sad	0.68	0.64	Low-energy speech overlapped with fear and fatigue.
Anger	0.72	0.70	Raised pitch was distinguishable, but short utterances reduced reliability.
Fear	0.65	0.61	High variability across speakers increased false detection.

Negative emotions showed weaker performance. Sadness and fear exhibited inconsistent recall scores, meaning the model frequently failed to detect these states when present. One contributing factor was natural speech suppression among participants: when individuals felt uncomfortable or tense in an experimental setting, they tended to speak more softly, which reduced the extraction quality of MFCC coefficients. These limitations helped justify a multimodal fusion approach, which is presented in the subsequent subsection.

The next stage of evaluation concerns the effect of multimodal fusion on emotion recognition. As demonstrated previously, both facial and voice-based classification suffer accuracy degradation when identifying negative or subtle emotional states. Therefore, a combined fusion mechanism was tested across the same 42 participants. The fused model produced a single emotional decision for every ten-second observation window.

Figure 4 shows a clear improvement when both modalities are combined into a single classifier. The face-only model achieved 79% average accuracy across all emotion categories, while the voice-only model reached 75%. These results reflect typical recognition behavior: visual signals dominate emotional interpretation, but acoustic cues supply additional clarity, especially during low-expression events.

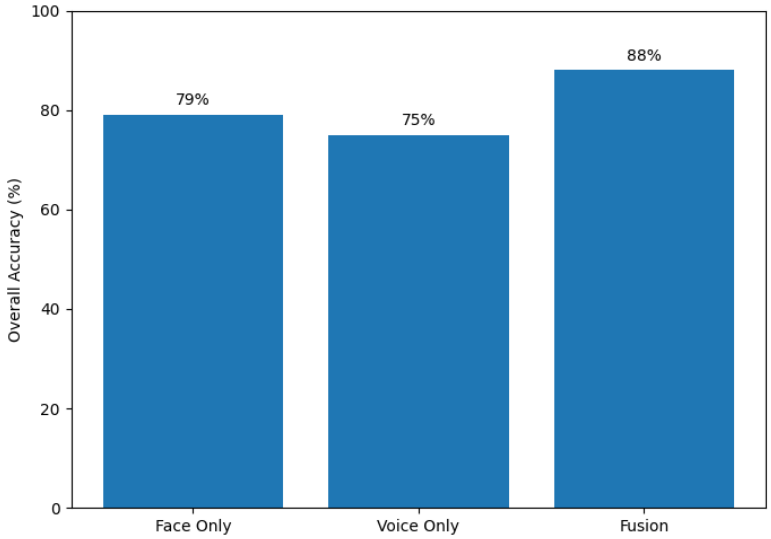

Figure 4 Comparison of Model Accuracy: Face vs Voice vs Fusion

The multimodal fusion model achieved an overall accuracy of 88%, representing a substantial 9–13% gain over single-channel approaches. Qualitative inspection of the logs confirmed that fusion performed best during states of frustration and confusion. In these cases, participants often maintained neutral facial posture but spoke with tense acoustic patterns, making the voice model decisive. Conversely, during anxious but silent phases, facial tension was more reliable than voice data. These complementary effects justify the addition of fusion into the real-time adaptive engine.

Table 3 highlights the behavioral effect of adaptive interventions triggered by emotional analysis. Each intervention was specifically mapped from the emotional conditions identified in earlier models. For example, difficulty reduction was primarily triggered when the system detected confusion, and the log-based analysis showed that learners subsequently hesitated less and attempted more tasks, producing a measured 17% improvement in task completion.

Table 3 Adaptive Interventions and Observed Behavioral Impact

Intervention Type	Triggering Emotion	Behavioral Observation	Outcome on Task Completion
Difficulty Reduction	Confusion	Learners paused less and attempted more items.	Task completion increased by 17%.
Time Extension	Anxiety	Learners slowed down reading and re-checked instructions.	Error rate dropped by 11%.
Encouragement Prompt	Frustration	Users showed improved self-reported confidence levels.	Retry rate improved by 22%.
Gamified Challenge	Boredom	Increased exploratory clicking and voluntary attempts.	Engagement duration increased by 26%.
Advanced Question Set	Confidence	Fewer redundant clicks and faster transitions.	Completion speed improved by 19%.

The most powerful effect emerged from gamified challenges during boredom. Participants who displayed low-engagement behavior became significantly more active once time-limited elements and score counters were introduced. Engagement duration rose by 26%, which indicates that emotional stagnation can be countered effectively not with remediation, but with stimulation. Anxiety-management interventions such as time extension also produced measurable effects by reducing the cognitive pressure that contributes to careless errors.

The final performance dimension focuses on whether emotional adaptivity led to measurable learning improvements. Participants were divided into two groups: a control group that used the same instructional system without emotional adaptivity, and an experimental group equipped with the full adaptive pipeline. Both groups completed identical tasks in mathematics and basic reasoning, and both performed pre- and post-assessment.

Figure 5 shows a marked difference between the two instructional conditions. Participants in the non-adaptive condition demonstrated an average gain of 14% points, reflecting modest improvement consistent with traditional guided practice. By contrast, participants in the emotion-adaptive system achieved a

28-point improvement. In practical terms, their learning effectiveness doubled.

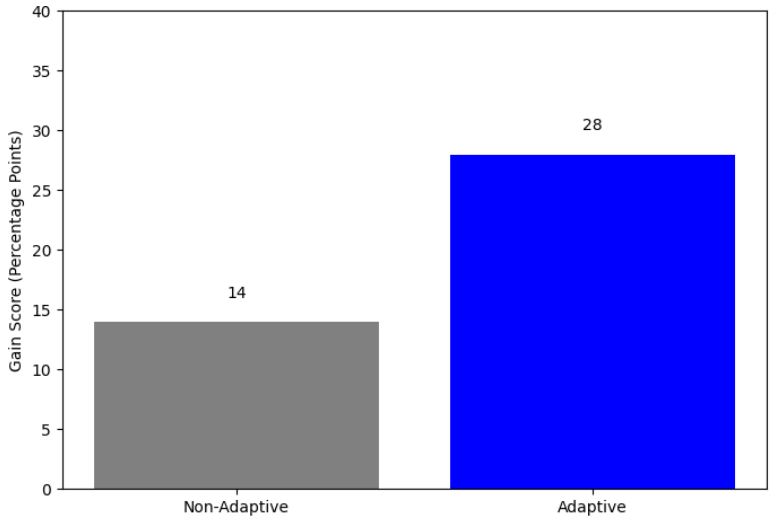

Figure 5 Learning Gain Comparison: Adaptive vs Non-Adaptive

One contributing factor was the mitigation of negative emotional build-up across the task timeline. Logs indicated that learners in the adaptive condition experienced fewer prolonged frustrations, because interventions redirected them toward productive struggle or reassurance. This reduction in affective interruptions helped preserve task continuity and prevented disengagement.

Table 4 documents the operational responsiveness of the adaptive pipeline. The mean detection window of 7.4 seconds suggests the system required a short emotional stabilization period before issuing a valid classification. This value was intentionally imposed to prevent the model from reacting to transient facial fluctuations or casual laughter. As a result, false triggers were minimized.

Table 4 Adaptation Latency and System Responsiveness

Adaptive Function	Mean Detection Time (sec)	Mean Intervention Trigger (sec)	Observed User Response
Emotion Detection Window	7.4	N/A	Trend stabilized before classification.
Difficulty Adjustment	8.1	9.6	Reduced hesitation and fewer abandoned questions.
Encouragement Prompt	7.9	10.3	Visible posture correction and increased retry rates.
Gamified Trigger	8.4	11.2	Faster return from idle status and rapid clicking.
Time Extension	8.7	10.9	Lower breath-holding and more measured responses.

Intervention triggers fell between 9.6 and 11.2 seconds, meaning the system generally executed an adaptive response within roughly three seconds of classification. This latency was short enough to impact an ongoing learning action but long enough to remain cognitively unobtrusive. Qualitative

observations showed that students rarely noticed the timing mechanism; instead, they attributed changes to instruction quality rather than automation. In this sense, adaptivity was successfully embedded rather than disruptive.

Overall, the results demonstrate three key conclusions. First, multimodal emotional sensing substantially improves classification accuracy beyond single-signal pathways. This performance improvement is especially visible for ambiguous negative states such as frustration, anxiety, and boredom—conditions most likely to undermine instructional progress. Second, adaptive interventions triggered by emotional conditions generate measurable behavioral improvements, including higher task-completion rates, reduced error density, deeper engagement, and greater voluntary exploration. The evidence suggests that learners benefit not only from support but also from stimulation matched to their affective profile.

Third, the emotional-adaptive group generated significantly higher learning gains compared with the non-adaptive group. These gains were not superficial; they were visible in deeper reasoning tasks and higher-order problem solving, indicating that emotional stabilization may improve not only persistence but also cognitive access. Latency analysis confirms that the system acted within a practical time window, reinforcing its potential for real-time deployments.

Taken together, these results validate the central hypothesis of the study: emotional adaptivity is not a peripheral enhancement but a functional booster to learning efficiency. The combination of sensing and intervention created an uninterrupted cognitive channel for students, reduced disruptive affect, and amplified the productive zone of engagement. Therefore, emotional-aware personalization constitutes a promising direction for scalable intelligent tutoring systems that seek not only accuracy, but emotional sustainability.

Conclusion

The findings of this study confirm that integrating real-time emotional sensing into an adaptive learning environment significantly enhances instructional effectiveness. Multimodal recognition using facial and voice cues demonstrated clear advantages over single-source analysis, particularly for ambiguous or low-expressivity emotional states. By stabilizing emotional interpretation across short observation windows, the system was able to produce reliable affective profiles that shaped instructional responses without interrupting the learner experience. This establishes emotional monitoring as a viable operational layer in digital pedagogy.

The adaptive interventions triggered by emotional signals translated directly into behavioral improvements. Learners exposed to difficulty reduction, time extensions, encouragement prompts, or gamified challenges demonstrated higher persistence, lower error density, deeper interaction, and longer engagement durations. These behavioral changes were not incidental; they reflected measurable shifts in cognitive participation, especially during periods of confusion, frustration, or boredom. The system effectively neutralized negative affect before it escalated into disengagement, thereby maintaining continuity in task performance.

Finally, learning outcomes improved substantially under emotional adaptivity. The experimental group achieved nearly twice the learning gain of the non-

adaptive cohort, underscoring that emotional stabilization is more than an affective convenience it contributes to domain mastery. The latency analysis further confirmed that interventions were delivered rapidly and unobtrusively, supporting seamless real-time guidance. Collectively, these results affirm that emotion-aware adaptivity can serve as a foundational mechanism for future intelligent tutoring systems, advancing both affective support and cognitive acceleration.

Declarations

Author Contributions

F.A.R. and S.Z.U.; Methodology: S.Z.U.; Software: F.A.R.; Validation: F.A.R. and S.Z.U.; Formal Analysis: F.A.R. and S.Z.U.; Investigation: F.A.R.; Resources: S.Z.U.; Data Curation: S.Z.U.; Writing Original Draft Preparation: F.A.R. and S.Z.U.; Writing Review and Editing: S.Z.U. and F.A.R.; Visualization: F.A.R.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Feng, H. Zhang, and D. Gašević, “Mapping the evolution of AI in education: Toward a co-adaptive and human-centered paradigm,” *Comput. Educ. Artif. Intell.*, vol. 9, no. December, p. 100513, 2025, doi: 10.1016/j.caeai.2025.100513.

[2] X. Lyu, F. Li, and Y. Zhao, “University English Writing Teaching Quality Evaluation Driven by Artificial Intelligence in the Context of New Liberal Arts: Linguistic Neutrosophic Multivalued Approach,” *Neutrosophic Sets Syst.*, vol. 83, no. April, pp. 837–850, 2025, doi: 10.5281/zenodo.15207909.

[3] K. Kong, H. F. Isleem, R. Aluvalu, G. G. Tejani, and A. S. M. Metwally, “Real-time cognitive and emotional state tracking in intelligent tutoring systems for enhanced learning outcomes,” *J. Big Data*, vol. 12, no. 1, pp. 1-22, 2025, doi: 10.1186/s40537-025-01333-0.

[4] G. F. Sekli and A. Godo, “Transforming Teaching And Learning With Robotic Process Automation: A Systematic Review Of Pedagogical Applications,” *J. Inf.*

Technol. Educ. Res., vol. 24, no. November, pp. 039, 2025, doi: 10.28945/5655.

[5] P. Dell'Acqua, M. Garofalo, F. la Rosa, and M. Villari, "Your Eyes Under Pressure: Real-Time Estimation of Cognitive Load with Smooth Pursuit Tracking," *Big Data Cogn. Comput.*, vol. 9, no. 11, p. 288, 2025, doi: 10.3390/bdcc9110288.

[6] R. Nemati, K. Shirini, and S. S. Gharehveran, "FER-HA: a hybrid attention model for facial emotion recognition," *J. Supercomput.*, vol. 81, no. 16, p. 1485, 2025, doi: 10.1007/s11227-025-07983-4.

[7] A. Alduais, A. A. Yassin, and S. Allegretta, "Computational linguistics: a scientometric review," *Qual. Quant.*, vol. 59, no. 5, pp. 4097–4136, 2025, doi: 10.1007/s11135-025-02138-2.

[8] Y. Deng, Z. Ren, A. Zhang, and T.-S. Chua, "Towards Goal-oriented Intelligent Tutoring Systems in Online Education," *ACM Trans. Inf. Syst.*, vol. 43, no. 6, pp. 1-26, 2025, doi: 10.1145/3760401.

[9] R. S. Ali, M. Abouel-Ela, and N. M. Eldakhly, "An ontology-based adaptive tutoring system for learning business english idioms," *Neural Comput.* Appl., vol. 37, no. 27, pp. 22725–22753, 2025, doi: 10.1007/s00521-025-11506-w.

[10] W. Cao, N. T. Mai, and W. Liu, "Adaptive Knowledge Assessment via Symmetric Hierarchical Bayesian Neural Networks with Graph Symmetry-Aware Concept Dependencies," *Symmetry (Basel).*, vol. 17, no. 8, p. 1332, 2025, doi: 10.3390/sym17081332.

[11] A. Grubišić, I. Saric-Grgic, A. Gašpar, and B. Žitko, "Usability Evaluation of an Adaptive Courseware Approach in the Natural Language-Based Intelligent Tutoring System-Tutomat," *J. Comput. Assist. Learn.*, vol. 41, no. 4, pp. 1-15, 2025, doi: 10.1111/jcal.70071.

[12] Z. Tagmatova, S. Umirzakova, A. Kutlimuratov, A. Abdusalomov, and Y. Cho, "A Hyper-Attentive Multimodal Transformer for Real-Time and Robust Facial Expression Recognition," *Appl. Sci.*, vol. 15, no. 13, p. 7100, 2025, doi: 10.3390/app15137100.

[13] A. A. S. Balla, M. AbdAlgane, A. O. A. Ahmed, and E. Osman, "AI-Driven Innovations in Adult EFL Learning: Exploring Potentials and Practicalities," *Int. J. Interact. Mob. Technol.*, vol. 19, no. 8, pp. 4–26, 2025, doi: 10.3991/ijim.v19i08.52295.

[14] F. Naseer and S. Khawaja, "Mitigating Conceptual Learning Gaps in Mixed-Ability Classrooms: A Learning Analytics-Based Evaluation of AI-Driven Adaptive Feedback for Struggling Learners," *Appl. Sci.*, vol. 15, no. 8, p. 4473, 2025, doi: 10.3390/app15084473.

[15] D. A. Popescu, N. Bold, and M. Stefanidakis, "A Systematic Model of an Adaptive Teaching, Learning and Assessment Environment Designed Using Genetic Algorithms," *Appl. Sci.*, vol. 15, no. 7, p. 4039, 2025, doi: 10.3390/app15074039.

[16] N. Maaz, J. Mounsef, and N. Maalouf, "CARE: towards customized assistive robot-based education," *Front. Robot. AI*, vol. 12, no. February, pp. 1-15, 2025, doi: 10.3389/frobt.2025.1474741.

[17] L. Feng, "Investigating the Effects of Artificial Intelligence-Assisted Language Learning Strategies on Cognitive Load and Learning Outcomes: A Comparative Study," *J. Educ. Comput. Res.*, vol. 62, no. 8, pp. 1961–1994, 2025, doi: 10.1177/07356331241268349.

[18] G. Wang and F. Sun, "A review of generative AI in digital education: transforming learning, teaching, and assessment," *Int. J. Inf. Commun. Technol.*, vol. 26, no. 19, pp. 102–127, 2025, doi: 10.1504/IJICT.2025.146701.

[19] A. Ali, R. M. I. Khan, D. Manzoor, M. A. Mateen, and M. A. Khan, "AI-Powered e-Learning: Innovations, Challenges, and the Future of Education," *Int. J. Inf. Educ. Technol.*, vol. 15, no. 5, pp. 882–890, 2025, doi: 10.18178/ijiet.2025.15.5.2294.

[20] A. Kaw et al., “On Building and Implementing Adaptive Learning Platform Lessons for Pre-Class Learning in a Flipped Course,” *Comput. Educ. J.*, vol. 14, no. 2, pp. 1–23, 2024, doi: 10.18260/b259-8f-13373.
[21] M. E. Eltahir and F. M. E. Babiker, “The Influence of Artificial Intelligence Tools on Student Performance in e-Learning Environments: Case Study,” *Electron. J. e-Learning*, vol. 22, no. 9, pp. 91–110, 2024, doi: 10.34190/ejel.22.9.3639.
[22] H. Yan and F. Lin, “Adaptive Practicing Design to Facilitate Self-Regulated Learning,” *Can. J. Learn. Technol.*, vol. 50, no. 3, pp. 1-22, 2024, doi: 10.21432/cjlt28768.
[23] M. Del-Águila-Castro, “Intelligent systems and their application in the evaluation of university academic performance: A literature review in the South American context,” *Rev. Cient. Sist. e Inform.*, vol. 4, no. 2, p. e671, 2024, doi: 10.51252/rcsi.v4i2.671.
[24] H. Nedombeloni, R. Heymann, and J. Greeff, “Bayesian Knowledge Tracing Implemented in a Telecommunications Serious Game,” *Int. J. Serious Games*, vol. 11, no. 2, pp. 107–131, 2024, doi: 10.17083/ijsg.v11i2.738.
[25] A. Zammouri, A. A. Ait-Moussa, and S. Chevallier, “Use of cognitive load measurements to design a new architecture of intelligent learning systems,” *Expert Syst. Appl.*, vol. 237, no. March, p. 121253, 2024, doi: 10.1016/j.eswa.2023.121253.
[26] K. Hartley, M. Hayak, and U. H. Ko, “Artificial Intelligence Supporting Independent Student Learning: An Evaluative Case Study of ChatGPT and Learning to Code,” *Educ. Sci.*, vol. 14, no. 2, p. 120, 2024, doi: 10.3390/educsci14020120.
[27] K. R. Chandra, M. Muthumanikandan, S. Kathyayini, H. G. Akhila, P. Pathak, and S. Shivaprakash, “The Impact of Artificial Intelligence Tools and Techniques for Effective English Language Education,” *Nanotechnol. Perceptions*, vol. 20, no. S7, pp. 897–903, 2024, doi: 10.62441/nano-ntp.v20iS7.74.
[28] J. Jia, Y. Zhang, and H. Le, “A comparison of a computerised adaptive test for mathematics instruction with the classical test,” *Int. J. Mob. Learn. Organ.*, vol. 18, no. 3, pp. 270–284, 2024, doi: 10.1504/IJMLO.2024.139640.
[29] H. Liu, Y. Zhang, and J. Jia, “The Design of Guiding and Adaptive Prompts for Intelligent Tutoring Systems and Its Effect on Students’ Mathematics Learning,” *IEEE Trans. Learn. Technol.*, vol. 17, no. March, pp. 1379–1389, 2024, doi: 10.1109/TLT.2024.3382000.
[30] M. A. Bakar, A. T. Ab Ghani, and M. L. Abdullah, “An Intelligent Mathematics Problem-Solving Tutoring System Framework: A Conceptual of Merging of Fuzzy Neural Networks and Neuroscience Mechanistic,” *Int. J. Online Biomed. Eng.*, vol. 20, no. 5, pp. 44–65, 2024, doi: 10.3991/ijoe.v20i05.47793.
[31] S. Mao, J. Zhan, Y. Wang, and Y. Jiang, “Improving Knowledge Tracing via Considering Two Types of Actual Differences From Exercises and Prior Knowledge,” *IEEE Trans. Learn. Technol.*, vol. 16, no. 3, pp. 324–338, 2023, doi: 10.1109/TLT.2023.3259013.
[32] D. S. McNamara et al., “iSTART: Adaptive Comprehension Strategy Training and Stealth Literacy Assessment,” *Int. J. Hum. Comput. Interact.*, vol. 39, no. 11, pp. 2239–2252, 2023, doi: 10.1080/10447318.2022.2114143.
[33] A. Trifa, A. Hedhili, and W. L. Lejouad Chaari, “Adaptive architecture based on agents for assessing a web application,” *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 40581–40607, 2022, doi: 10.1007/s11042-022-13059-9.
[34] B. Vesin, K. Mangaroska, K. Akhuseyinoglu, and M. Giannakos, “Adaptive Assessment and Content Recommendation in Online Programming Courses: On the Use of Elo-rating,” *ACM Trans. Comput. Educ.*, vol. 22, no. 3, pp. 1-27, 2022, doi: 10.1145/3511886.