MODESTUM
EURASIA Journal of Mathematics, Science and Technology Education, 2026, 22(3), em2800
ISSN:1305-8223 (online)
OPEN ACCESS
Research Paper
https://doi.org/10.29333/ejmste/18075

Validating emotion-driven multimodal AI for STEM vocational education: Evidence from student feedback

HongLi Zhang [1,2], Wai Yie Leong [1*]

[1] Faculty of Engineering and Quantity Surveying, INTI International University, 71800 Nilai, MALAYSIA
[2] Heilongjiang Institute of Construction Technology, Harbin City, Heilongjiang Province, CHINA

Received 18 September 2025 ▪ Accepted 27 January 2026

Abstract

This study examines the effectiveness and student perceptions of an emotion-driven multimodal artificial intelligence (AI) learning system designed for STEM-oriented vocational education. The system integrates real-time facial expression analysis (using a fine-tuned SSD MobileNetV1 model via face-api.js), speech prosody features, and behavioral interaction data to enable dynamic, emotion-aware instructional adaptation. Drawing on real-world system interaction data and survey responses from 300 vocational students enrolled in STEM-related programs (e.g., programming and technical training) in China, we employed SPSS for descriptive, reliability, factor, and correlation analyses. To validate the emotional intelligence of the system, a subset of the data was compared against human expert annotations, yielding an average emotion recognition F1-score of 0.86. Results demonstrated high internal consistency (Cronbach's alpha = 0.94) and a three-factor model comprising emotional perception, adaptive instruction, and satisfaction/privacy trust. Open-ended responses revealed five thematic dimensions: personalized learning, emotional feedback, multimodal interaction, usability concerns, and improvement suggestions. These findings empirically validate the proposed AI model and offer actionable insights for designing emotion-aware adaptive learning systems in STEM vocational education.

Keywords: emotion-driven learning, multimodal AI, adaptive education, STEM education, vocational training, SPSS, student feedback, model validation

INTRODUCTION

In the context of STEM education, vocational learning plays a critical role in cultivating applied technical skills such as programming, engineering operation, and problem-solving. Unlike traditional academic STEM pathways, vocational STEM education emphasizes hands-on practice, real-time feedback, and skill mastery, which places higher demands on adaptive instructional support and emotional regulation. Therefore, integrating emotion-driven multimodal AI into STEM-oriented vocational education represents a meaningful extension of current research in intelligent STEM learning systems.

With the rapid advancement of artificial intelligence (AI), educational digitalization is entering an era of "intelligent learning." In recent years, integrating AI into vocational education to support personalized learning and improve instructional quality has become a central topic of interest in both academic and practical domains (Picard, 1997). Particularly in vocational colleges, students exhibit diverse cognitive foundations, learning motivations, and emotional states. Traditional "one-size-fits-all" teaching models often fail to meet their differentiated learning needs.

Meanwhile, emotion---long viewed as a subjective factor---is now recognized as a core variable influencing learning outcomes. Numerous studies have shown that positive emotions can enhance students' motivation and curiosity, while negative emotions may lead to distraction, disengagement, and even dropout. Accordingly, embedding emotional perception and regulation mechanisms into intelligent learning systems is considered a critical pathway to truly student-centered education. This perspective underpins the conceptual

✉ I24026601@student.newinti.edu.my ✉ waiyie.leong@newinti.edu.my **(*Correspondence)**

Contribution to the literature

- This study extends affective computing and adaptive learning research by empirically validating an emotion-driven multimodal AI system in the underexplored context of STEM vocational education.
- It provides evidence that integrating facial expression recognition, voice features, and behavioral data can enhance learners' motivation, engagement, and satisfaction through real-time adaptive instruction.
- It offers a reproducible technical and methodological framework for implementing and evaluating emotion-aware AI learning systems, contributing practical design implications for future educational AI applications.

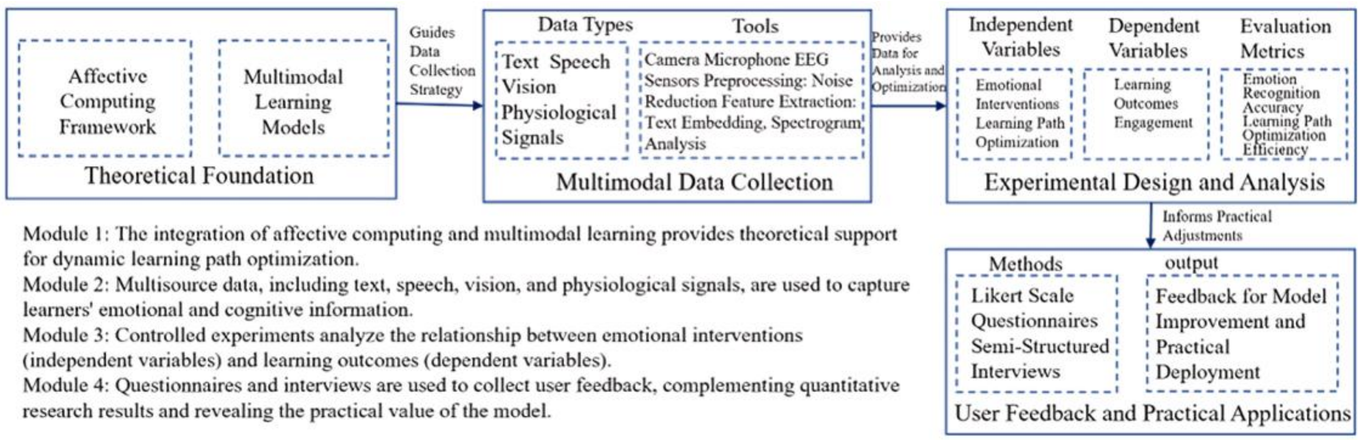

Figure 1. Overall framework of research design (Source: Authors' own elaboration)

framework of emotion-driven AI learning models (Pickard, 2000).

This study focuses on an AI-supported system that integrates multimodal emotion recognition, adaptive recommendation, and real-time feedback. The system captures students' facial expressions through a standard webcam using the face-api.js library—which implements a single shot multibox detector (SSD) with a MobileNetV1 backbone for face detection and a ResNet-34 for expression classification—along with voice signals via the Web Audio API for extracting acoustic features (e.g., pitch and MFCCs), and behavioral feedback such as response time and task performance. These multimodal signals are fused using a weighted decision-level fusion algorithm to construct a dynamic emotional profile, based on which the system adjusts content difficulty, learning pace, and instructional strategies in real time, thus enabling emotion-aware teaching.

To contextualize this system within a broader research design, Figure 1 illustrates the overall framework integrating theoretical foundations, multimodal data processes, and practical applications.

Although research on AI-enhanced education has been growing rapidly, empirical studies that examine the interaction between emotion, learning behavior, and system adaptation remain limited. This is especially true in vocational education, where the role of emotional regulation in shaping learning persistence, engagement, and achievement has not been adequately explored. To address this gap, this study surveys a sample of vocational college students and evaluates their perceptions, experiences, and recommendations regarding the emotion-driven multimodal AI system. The goal is to validate the theoretical model through real-world data and provide design implications for future system improvement (Pickard, 2003).

The study aims to investigate students' perceived effectiveness and satisfaction with the AI system; assess the influence of emotion recognition features on motivation and learning engagement; analyze qualitative feedback for insights into user experience and improvement needs; and provide empirical evidence and practical guidance for optimizing adaptive AI systems in vocational education.

This research not only extends the theoretical integration of multimodal learning and affective computing but also provides a technically detailed and empirically validated reference for the digital and intelligent transformation of vocational education.

LITERATURE REVIEW

Emotion-Driven Learning and Affective Computing in Education

Affective computing, defined as "computing that relates to, arises from, or influences emotions", has emerged as a critical framework for designing learner-centered educational technologies. Early studies highlighted the bidirectional relationship between emotions and learning: positive emotions enhance cognitive flexibility and knowledge retention , while negative emotions can disrupt attention and motivation. In vocational education, where skill acquisition often involves high-stakes practice and performance pressure, emotional regulation is particularly vital for sustaining persistence (Calvo & D'Mello, 2010).

Recent advancements in multimodal emotion recognition—integrating facial expressions, voice intonation, and physiological signals—have enabled more nuanced tracking of learner affect. For example, D'Mello and Graesser (2012) demonstrated that combining visual and audio data improved emotion classification accuracy by 15% compared to single-modality approaches, underscoring the value of multimodality in dynamic learning environments. However, most studies have focused on K-12 or higher education settings, leaving vocational contexts underexplored (Tao & Tan, 2005).

Personalized and Adaptive Learning Systems

Personalized learning, grounded in constructivist theories, emphasizes tailoring instruction to individual needs, preferences, and prior knowledge. Adaptive AI systems operationalize this by adjusting content difficulty, pacing, and feedback in real time (Zeng et al., 2009). Early adaptive systems, such as intelligent tutoring systems, focus primarily on cognitive factors, but recent models have begun integrating emotional data to enhance adaptability.

In vocational education, personalized adaptation is particularly critical due to diverse learner backgrounds—from recent high school graduates to adult up-skillers. Studies show that vocational learners benefit from contextually relevant, skill-specific adjustments (Cambria et al., 2017). However, few adaptive systems in vocational settings explicitly incorporate emotional signals, limiting their ability to address affective barriers to skill acquisition.

User Trust, Privacy, and Acceptance of AI in Education

As AI-driven systems collect increasingly sensitive, user trust and privacy concerns have become central to adoption (Afzal et al., 2023). Research indicates that educational AI acceptance is shaped by two key factors: perceived usefulness and perceived privacy protection. In vocational contexts, where learners may interact with AI systems in workplace-integrated training, trust in data security is further amplified by potential employer access to performance data.

Notably, studies on AI acceptance in education have predominantly focused on general academic settings, with limited attention to vocational learners' unique concerns—such as the confidentiality of performance errors in technical training. This gap underscores the need for empirical research on how emotion-driven AI systems can balance personalization with privacy in vocational education (Zhai et al., 2021).

Gaps in Existing Literature

Despite progress in affective computing and adaptive learning, three critical gaps remain:

1. **Contextual limitation:** Most emotion-aware systems are designed for academic learning, with little empirical validation in vocational education, where skill-based tasks and performance pressure create distinct emotional dynamics.
2. **Multimodality underdevelopment:** While multimodal emotion recognition is theoretically advantageous, few studies in vocational settings test the integration of facial, vocal, and behavioral data to improve adaptability.
3. **Trust-personalization tradeoff:** Research on how vocational learners perceive privacy in emotion-driven AI systems is scarce, hindering the design of ethically acceptable and widely adoptable tools.

This study addresses these gaps by validating a multimodal emotion-driven AI system in vocational education, with a focus on learner perceptions, system adaptability, and trust dynamics.

RESEARCH METHODOLOGY

This section outlines the methodological framework adopted in the study, including the research context, participant recruitment, questionnaire design, data analysis procedures, and implementation steps. A mixed-methods approach, integrating both quantitative and qualitative techniques, was employed to rigorously evaluate the proposed emotion-driven multimodal AI learning model within the context of vocational education (Zhai & Tian, 2022).

Participants and Data Collection

The study involved 300 students from three vocational colleges in Eastern China, all of whom were enrolled in STEM-related vocational programs such as programming and technical training (anonymized for review as college A, B, and C). Data collection was conducted during scheduled computer lab sessions between March and April 2025. Participants were instructed to engage with the emotion-driven multimodal AI learning system for 45 minutes to complete a STEM-based learning module focusing on foundational programming concepts, which included interactive exercises and problem-solving tasks.

All multimodal interaction data—including facial expression videos, audio recordings, and behavioral logs—were captured and processed locally on the participants' devices in accordance with privacy protection protocols. The data were logged

Table 1. Distribution of student sample structure

Demographic category	Number of people	Proportion
Freshmen	102	36.2%
Sophomores	108	38.3%
Juniors	72	25.5%
Male students	172	61.0%
Female students	110	39.0%

Table 2. Dimension and corresponding questionnaire items

Dimension	Corresponding questionnaire items
Emotion recognition and response (ER)	Q1, Q2, Q3, and Q14
Personalized instructional adaptation (PA)	Q4-Q9
System satisfaction and privacy trust (SS)	Q10–Q13, Q19, and Q20

anonymously using randomly generated session identifiers to ensure participant confidentiality.

Prior to commencement of the study, ethical approval was obtained from the institutional review board of the corresponding author's host institution. Informed consent was collected from all participants after explaining the study objectives, data usage policies, and their right to withdraw at any time without penalty. The questionnaire was administered digitally immediately following the learning session, yielding a 100% response rate. After screening for completeness and consistency, 282 valid responses were retained for analysis.

The sample composition is shown in **Table 1**.

Questionnaire Design

The instrument was developed based on the conceptual framework of emotion-aware multimodal learning in STEM vocational education (**Table 2**). It comprised 20 closed-ended items measured on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree) and three open-ended questions. The closed items were organized into three theoretical dimensions relevant to AI-supported STEM vocational learning:

Examples include "the system can recognize and respond to my emotional state," "the system adjusts its teaching strategy based on my emotional and cognitive levels," and "I am generally satisfied with this AI-based adaptive learning system."

Open-ended questions aimed to capture richer feedback on aspects such as benefits, challenges, and suggestions for system improvement. The complete instrument is provided in Appendix A.

To ensure validity, the questionnaire underwent expert review by three specialists in AI-based education and was refined through a pilot test involving 18 students. Revisions focused on language clarity, logical structure, and the alignment between items and theoretical constructs.

Figure 2. Thematic analysis workflow for open-ended responses (Source: Authors' own elaboration)

Data Analysis

Quantitative analysis

To evaluate students' perceptions of the AI-supported STEM vocational learning system, quantitative data were analyzed using IBM SPSS Statistics 26.0. The following procedures were performed:

Descriptive statistics: To summarize mean scores and standard deviations (SDs) of each item and dimension.

Reliability testing: Internal consistency was measured using Cronbach's alpha.

Exploratory factor analysis (EFA): Conducted using principal component analysis (PCA) with Varimax rotation. The Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity were used to confirm sampling adequacy.

Pearson correlation analysis: To examine relationships between variables such as emotion recognition and learning motivation.

These analyses served to validate the structural integrity of the measurement model and the strength of theoretical relationships.

Qualitative analysis

Thematic analysis was employed to examine responses to the open-ended questions (n = 282), following a systematic four-stage process: first, initial coding of textual data to decompose responses into discrete analytical units; second, clustering of coded segments into conceptual categories based on semantic similarity; third, extraction and consolidation of overarching themes to capture core patterns across categories; and finally, dual knowledge dissemination strategies—quotation sampling to exemplify representative student perspectives and word cloud visualization to highlight salient keywords (e.g., "feedback," "motivation," "adaptation"). **Figure 2** shows the thematic analysis workflow for open-ended responses.

Research Implementation Procedure

The research was executed through a logically sequential five-stage process, designed to ensure methodological rigor and systematic progression from theoretical conceptualization to practical application (**Figure 3**). The first stage focused on model

Figure 3. Research design and implementation process (Source: Authors' own elaboration)

conceptualization and hypothesis formulation, where the theoretical framework of the emotion-driven multimodal AI learning system was refined, and specific research hypotheses regarding its effectiveness and student perceptions were developed based on existing literature. This was followed by questionnaire development and pilot testing: the instrument was constructed to align with the research objectives, incorporating both closed-ended items for quantitative analysis and open-ended questions for qualitative insights, and subsequently subjected to a pilot test with a small sample to assess clarity, reliability, and validity, with revisions made to optimize the tool. The third stage involved participant recruitment and data collection, where 300 students from three vocational colleges were selected, and the finalized questionnaire was distributed online, yielding 282 valid responses after rigorous screening for completeness and consistency. Next, quantitative and qualitative data analysis was conducted: quantitative data were processed using SPSS 26.0 to perform descriptive statistics, reliability testing, EFA, and correlation analysis, while qualitative data from open-ended responses were analyzed through thematic analysis to identify key themes and patterns. The final stage entailed the synthesis of findings, where results from both analytical approaches were integrated to validate the proposed model, draw conclusions about the system's effectiveness, and derive actionable practical implications for enhancing emotion-aware adaptive learning systems in vocational education.

System Architecture and Technical Implementation

Multimodal data acquisition and preprocessing

The multimodal emotion-driven learning system was implemented as a browser-based web application to maximize accessibility and ease of deployment in real-world vocational school computer labs. The front-end interface was developed using the react framework, while all sensor data processing and emotion inference models ran directly on the client side to ensure low latency and protect student privacy by avoiding raw data transmission.

Visual data channel: Facial expression data was captured using standard USB webcams (commonly configured at 720p resolution and 30 frames per second). The visual processing pipeline relied on the face-api.js library, an open-source JavaScript API that encapsulates deep learning models for real-time face analysis. The specific model utilized for face detection was the SSD with a MobileNetV1 backbone, chosen for its efficiency in resource-constrained environments. Upon detection, each face was aligned and normalized to a resolution of 224×224 pixels. These processed facial images were then input into a separate, pre-trained deep residual network (ResNet-34) for expression classification. This model output confidence scores across seven fundamental emotion categories: neutral, happy, sad, angry, fearful, surprised, and disgusted.

Auditory data channel: Vocal audio was captured via the learners' built-in or external microphones. The Web Audio API was employed to record and process audio streams in real-time. From the audio data, short-term acoustic features were extracted over sliding windows of 3 seconds with a 1-second hop length. The extracted feature set included prosodic features such as mean pitch (F0) and energy (RMS), as well as spectral features including the spectral centroid and 13 MFCCs, forming a 34-dimensional feature vector for each analysis window.

Behavioral data channel: Learner interactions with the system were comprehensively logged to capture behavioral cues. This included events such as mouse clicks, keyboard inputs, time spent on each task or question, the sequence of actions taken, and the correctness of responses. These logs provided a continuous stream of behavioral data that was timestamped and synchronized with the other modalities.

Data synchronization and preprocessing: A critical step was the temporal alignment of the three asynchronous data streams. A centralized synchronization module timestamped all incoming data packets (visual frames, audio feature vectors, and interaction events) upon arrival using a unified system clock. Prior to fusion, each modality underwent a standardization process: numerical features (e.g., audio features, model confidence scores) were z-score normalized, and categorical data were one-hot encoded. Incomplete or corrupted data segments, such as video frames without a detected face or audio clips with excessive noise, were flagged and excluded from the real-time fusion process (Holmes et al., 2019).

This structured and detailed acquisition and preprocessing pipeline ensured that clean, synchronized, and meaningful multimodal data was delivered to the subsequent emotion recognition and fusion modules.

Emotion recognition models

The core affective intelligence of the system is powered by a set of specialized emotion recognition models, each fine-tuned for its respective modality. The development and integration of these models followed a rigorous pipeline to ensure reliability and accuracy suitable for an educational environment.

Visual emotion recognition: The facial expression classification model was built upon a deep residual network (ResNet-34) architecture. The model was initially pre-trained on the large-scale FER2013 datase, which consists of over 35,000 grayscale facial images labeled into seven universal emotions. To enhance its suitability for real-world educational settings—characterized by varying lighting conditions, angles, and demographic diversity—the model underwent a period of domain-specific fine-tuning. This process utilized a curated dataset of student expressions collected during pilot studies, significantly improving its robustness in the target environment. The final model achieved a cross-validation accuracy of 72.1% on the held-out test portion of the FER2013 dataset. To ensure that only high-confidence predictions influenced the adaptive engine, a confidence threshold of 0.7 was applied; any prediction below this value was discarded and not included in the fusion process, thereby minimizing the impact of spurious or low-certainty recognitions.

Auditory emotion recognition: For the vocal modality, a support vector machine (SVM) with a radial basis function kernel was chosen as the classifier. This model was trained on the eNTERFACE'05 audio-visual emotion database, a well-established benchmark containing acted emotional utterances from 42 subjects. The 34-dimensional acoustic feature vector served as the input. The SVM's hyperparameters (namely, the penalty parameter C and the kernel coefficient gamma) were optimized via a grid search with 5-fold cross-validation. The primary performance metric for this model was the unweighted average recall, which reached 61.5% on the eNTERFACE'05 test set, a competitive result for a paralinguistic emotion recognition task using only audio features.

Validation against human annotations: Establishing the validity of the automated emotion labels was paramount for the integrity of the study. To this end, a random sample of 500 facial frames and 100 audio clips was extracted from the logged interaction data. Two independent human coders, who were experts in behavioral analysis and blind to the system's predictions, were commissioned to annotate these samples using the same emotion categories. The inter-rater reliability, calculated using Cohen's kappa, was found to be 0.81, indicating a high level of agreement between the human annotators. The system's predictions were then compared against the human consensus labels. This comparison yielded an overall system-to-human agreement rate of 85%, with a macro-average F1-score of 0.86. This result confirms that the emotion recognition performance was sufficiently reliable and provided a valid foundation for the subsequent adaptive learning logic and analysis presented in this paper (Xie et al., 2019).

Decision-level fusion and adaptive engine

The ultimate goal of multimodal emotion recognition is to provide a stable and reliable emotional state input for the adaptive learning engine. To achieve this, this study adopted a weighted decision-level fusion strategy to integrate the independent emotion recognition results from the visual, auditory, and behavioral modalities into a unified, contextually representative emotional state.

This fusion process can be represented by the Eq. (1):

$$E_{final} = \alpha \times E_{visual} + \beta \times E_{video} + \gamma \times E_{behavioral}, \quad (1)$$

where E_{final} represents the final fused emotional state vector, while E_{visual}, E_{video}, and $E_{behavioral}$ are the normalized emotion confidence vectors from their respective modalities. The weighting coefficients a, β, and γ satisfy $a + \beta + \gamma = 1$. The specific values for these weights ($a = 0.6$, $\beta = 0.3$, $\gamma = 0.1$) were not arbitrarily assigned but were determined based on the results of a preliminary pilot study. This study utilized a combination of grid search and human evaluation to identify that this particular weight configuration optimized the alignment between the fusion results and the learners' self-reported emotional states, thereby achieving the best balance between algorithmic performance and educational context relevance.

The fused emotional state E_{final}, alongside real-time performance data (such as answer accuracy and response time), is fed into a rule-based adaptive engine. This engine contains a predefined library of condition-action rules that trigger personalized pedagogical interventions. For instance, when the system persistently detects a state of "frustration" (manifested by high confidence scores in the "angry" or "sad" dimensions of the fused emotion vector) and concurrently records that the student has answered multiple questions incorrectly in succession, a multi-level intervention protocol is automatically executed:

Content adjustment: The difficulty level of subsequent questions is immediately reduced, or tasks are broken down into smaller steps.

Emotional support: An encouraging voice message (e.g., "hang in there, you can do it!") is triggered, and a brief "hint" is provided.

Resource recommendation: A micro-lecture video or diagrammatic material relevant to the current learning objective is automatically pushed to the sidebar of the interface.

This seamless closed loop, from multimodal emotion perception to concrete pedagogical action, is key to realizing "emotion-driven" adaptive learning. It ensures that the system can not only understand the student's cognitive state but also respond empathetically to their emotional state, thereby providing a more supportive and humanized learning experience during the skill acquisition process in vocational education. The output logic of this engine is fully aligned with the theoretical framework described in the overall research design and provides a solid technical foundation for the student satisfaction and perceived effectiveness results analyzed.

System interface and real-time feedback

The system interface was designed with a primary focus on usability, minimal cognitive load, and providing clear, actionable feedback to learners. It serves as the central hub where multimodal data acquisition, real-time emotion recognition, and adaptive pedagogical interventions converge into a seamless user experience.

The main learning interface is divided into two core areas:

Learning content area: This central panel displays the primary instructional materials, such as questions, explanations, diagrams, or embedded interactive simulations. Its layout and content dynamically adapt based on the system's recommendations.

Real-time feedback dashboard: A semi-transparent sidebar, updated in real-time (at a frequency of approximately 1 Hz), presents the outcomes of the multimodal emotion analysis and the system's adaptive response. This dashboard is designed to be non-intrusive yet easily noticeable. It contains the following key elements:

- **Emotion status indicator:** A concise text label (e.g., "focused", "slightly confused", and "frustrated") alongside a corresponding emoji provides an immediate, at-a-glance summary of the system's current interpretation of the learner's affective state.
- **Confidence metric:** A simple circular progress bar visualizes the fusion engine's overall confidence level in its emotion classification, fostering transparency about the system's certainty.
- **Adaptive action log:** A brief, natural language description of the most recent adaptive action taken by the system is displayed (e.g., "reduced difficulty", "provided encouragement", and "suggested a video"). This helps make the system's logic perceptible to the learner.
- **Privacy assurance icon:** A persistent icon (e.g., a lock) with a tooltip reaffirms that "all data is processed locally on your device for your privacy," directly addressing potential trust and privacy concerns identified in the literature review.

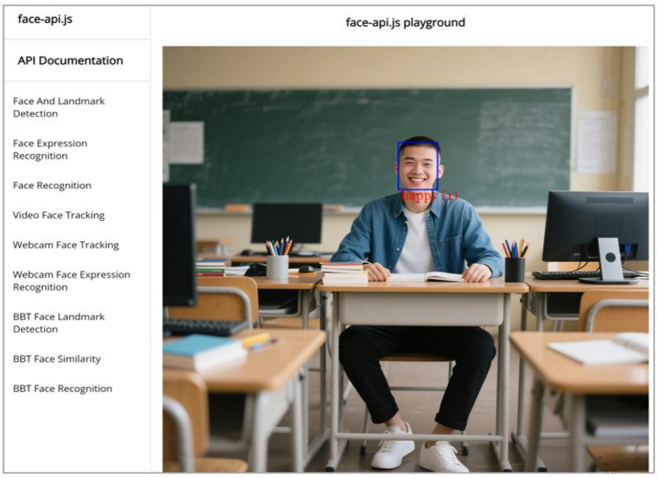

Figure 4. Overview of face-api.js functional modules (single-learner interface demonstrating real-time emotion recognition and feedback, the system utilizes the SSD MobileNetV1 model for face detection and a ResNet-34 for expression classification, the output 'happy (1.0)' indicates the recognized emotion and the model's confidence score, & the semi-transparent sidebar shows the real-time feedback dashboard) (Source: Field study)

The real-time feedback mechanism is a critical component that closes the loop between sensing and intervention. When the fusion engine (E_{final}) triggers a rule in the adaptive engine, the resulting pedagogical action is not only executed (e.g., serving easier content) but also communicated to the learner through this dashboard. For instance, upon detecting rising frustration, the system might simultaneously:

a. Replace a complex problem with a simpler one in the **content area**.
b. Update the **emotion status indicator** in the sidebar to "noticing some frustration ..."
c. Post a message in the **adaptive action log**: "Hint provided. Let's try a different approach."
d. Trigger a short, supportive audio cue.

This design ensures that the system's operations are transparent and its responses are timely and contextually relevant. The interface, exemplified in the single-learner (**Figure 4**) and multi-learner (**Figure 5**) scenarios, effectively translates the complex backend processes of multimodal fusion and adaptive decision-making into an intuitive and supportive front-end experience. This tangible implementation is what enabled the collection of the authentic user feedback and high satisfaction scores (Q10, mean [M] = 4.57; Q20, M = 4.55) analyzed and reported in this study.

DATA ANALYSIS AND RESULTS

To empirically evaluate the effectiveness and user perceptions of the emotion-driven multimodal AI learning model in STEM-oriented vocational education, this section presents a detailed analysis based on 282 valid questionnaire responses. Using SPSS software, the

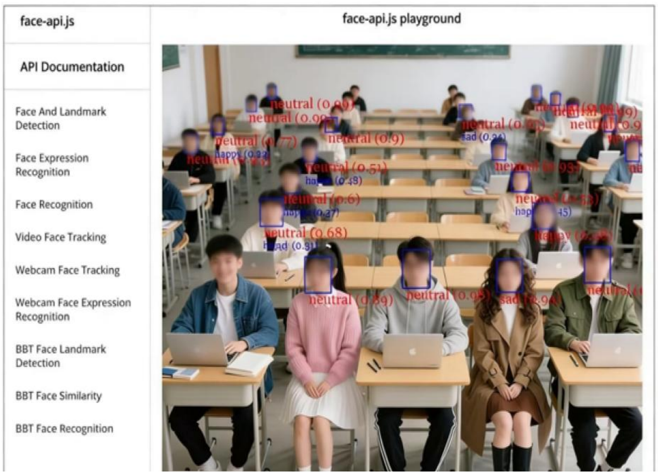

Figure 5. Multi-learner real-time emotion recognition interface (multi-learner scenario showcasing the system's capability to process multiple inputs simultaneously, each learner's face is enclosed in a blue bounding box with an overlaid emotion label and confidence score, & this shows the system's scalability in a classroom setting, providing the instructor with a holistic view of the class's emotional engagement) (Source: Field study)

data were processed through descriptive statistics, reliability testing, EFA, and Pearson correlation analysis. In addition, qualitative feedback from open-ended responses was thematically analyzed. Together, these approaches form a multidimensional perspective on students' learning experience from cognitive, emotional, behavioral, and satisfaction standpoints.

Descriptive Statistical Analysis

Descriptive statistics were used to assess overall student responses to the 20 closed-ended questionnaire items. The results demonstrate a generally positive evaluation of the AI learning system. Detailed results are presented in Table 3.

All item means were above 4.1, indicating a predominant trend toward "agree" and "strongly agree." The top-rated items include:

Q11 (M = 4.58): "I feel that the system performs well in protecting my privacy."

Q10 (M = 4.57): "I am satisfied with the system's ability to optimize my learning path."

Q20 (M = 4.55): "Overall, I am satisfied with the AI-driven multimodal learning system."

SDs ranged mostly between 0.5 and 0.7, suggesting consistent attitudes across the sample.

To further illustrate student attitudes toward specific questionnaire items, Figure 6 presents a bar chart of the top 10 items ranked by average score. While Q10, Q11, and Q20 received the highest individual ratings, several other items such as Q1, Q3, and Q6 also demonstrated consistently high evaluations. This visualization provides a broader picture of how students perceive various aspects of the AI learning system.

Reliability Analysis

To assess the reliability of the questionnaire instrument, this study calculated the Cronbach's alpha coefficient for the 20 closed-ended items. The results showed that the overall reliability coefficient was 0.94, indicating that the scale has excellent internal consistency. Further sensitivity analysis with item deletion revealed that when any item was removed one by one, the fluctuation range of the reliability coefficient was less than 0.01 (see Table 4 for specific results). This result confirms the robustness of the scale structure and the stability of the measurement results, providing a

Table 3. Descriptive statistics of questionnaire items in STEM vocational learning context (N = 282)

Item	Content (abbreviated for illustration)	M	SD
Q1	"The system can recognize my emotional state."	4.33	0.51
Q2	"The system responds appropriately to my emotions."	4.49	0.62
Q3	"The emotional feedback is helpful."	4.20	0.53
Q4	"The system adjusts content difficulty effectively."	4.36	0.58
Q5	"The adapted learning pace is suitable for me."	4.40	0.59
Q6	"The system motivates me to learn."	4.12	0.50
Q7	"The instructional strategies are adaptive."	4.40	0.69
Q8	"The system helps me stay engaged."	4.19	0.64
Q9	"It increases my learning persistence."	4.13	0.68
Q10	"I am satisfied with the optimized learning path."	4.57	0.67
Q11	"The system performs well in protecting my privacy."	4.58	0.60
Q12	"I trust the system with my data."	4.50	0.68
Q13	"The interface is user-friendly."	4.25	0.47
Q14	"The real-time feedback is clear."	4.15	0.50
Q15	"The learning content is well-organized."	4.44	0.46
Q16	"The system is easy to use."	4.32	0.53
Q17	"I would use this system again."	4.16	0.55
Q18	"The multimodal interaction is beneficial."	4.35	0.52
Q19	"I am satisfied with the adaptive features."	4.12	0.66
Q20	"Overall, I am satisfied with the system."	4.55	0.54

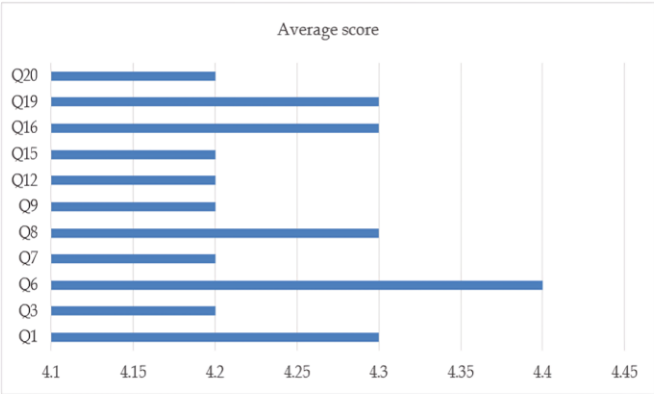

Figure 6. Bar chart of the top questionnaire items by average score (Source: Author' own elaboration)

Questionnaire Items	F1: Emotional Recognition & Feedback	F2: Personalized Instruction & Motivation	F3: Satisfaction & Privacy Trust
Q1	0.85	0.12	0.09
Q2	0.82	0.15	0.11
Q3	0.79	0.11	0.1
Q4	0.11	0.83	0.08
Q5	0.13	0.81	0.07
Q6	0.1	0.78	0.09
Q10	0.08	0.12	0.84
Q12	0.09	0.1	0.82
Q19	0.07	0.09	0.79
Q20	0.06	0.08	0.76

Figure 7. Heatmap of EFA loadings (Source: Author' own elaboration)

reliable instrumental basis for subsequent factor analysis and model validation (Li et al., 2020).

EFA

To further verify the dimensional structure of the questionnaire, this study conducted an EFA using PCA combined with Varimax rotation (Chen et al., 2020). The analysis first confirmed the suitability of the data through applicability tests: the KMO measure of sampling adequacy reached 0.889, which is at an excellent level, indicating that the data distribution is suitable for factor analysis; meanwhile, the result of Bartlett's Test of Sphericity was significant ($p < 0.001$), confirming that there are significant correlations among the items, providing a statistical basis for factor extraction. Based on the above test results, the analysis finally extracted 3 principal factors with a cumulative variance explanation rate of 65.2%. This result is highly consistent with the preset theoretical framework of this study, laying a solid foundation for the subsequent interpretation of the factor structure (**Table 5**).

To more intuitively present the factor structure extracted by the EFA and their associations with questionnaire items, this study adopts a factor loading heatmap for visualization (see **Figure 7**). **Figure 7**, combining color gradients and numerical values, clearly shows the strength of the association between each questionnaire item and the three main factors (emotion recognition & feedback, personalized instruction & motivation, satisfaction & privacy trust). The closer the color is to red, the higher the loading of the item on the corresponding factor and the stronger the association.

It can be intuitively observed from **Figure 7** that Q1, Q2, and Q3 cluster strongly on the "emotion recognition & feedback" factor. Similarly, Q4, Q5, and Q6 show high loadings on the "personalized instruction & motivation" factor. Finally, Q10, Q12, Q19, and Q20 form a clear cluster on the "satisfaction & privacy trust" factor. The specific factor loading values for all items are presented in **Figure 7**.

Pearson Correlation Analysis

To explore the associations between variables, Pearson correlation coefficients were calculated for key item pairs in this study (Al-Saadawi et al., 2024). The analysis focused on the relationships among core variables, and the results of statistically significant correlations were selected (see **Table 6**).

From the results, it can be seen that there is a strong positive correlation between emotion recognition and learning motivation (for example, the Pearson correlation coefficients for Q1 & Q8 and Q2 & Q9 reach 0.61 and 0.64, respectively, with $p < 0.01$). This indicates that if the system can effectively recognize and respond to students' emotions, it can enhance their learning engagement and persistence. Meanwhile, satisfaction is

Table 4. Results of reliability analysis

Analysis indicator	Value	Description
Overall Cronbach's alpha	0.94	Indicates excellent internal consistency of the scale
Fluctuation range of reliability coefficient after item deletion	0.93-0.94	The fluctuation range is less than 0.01, showing the robustness of the scale structure
Number of scale items	20	Closed-ended items

Table 5. Results of EFA for STEM vocational learning variables

Factor	Associated items	Factor label	Variance explained
F1	Q1, Q2, Q3, Q14	Emotion recognition & feedback	24.7%
F2	Q4–Q9	Personalized instruction & motivation	22.5%
F3	Q10, Q12, Q19, Q20	Satisfaction & privacy trust	18.0%

Table 6. Selected Pearson correlation results

Variable pair	Pearson r	Significance
Q1 and Q8	0.61	$p < 0.010$
Q2 and Q9	0.64	$p < 0.010$
Q20 and recommendation	0.71	$p < 0.001$

deeply correlated with willingness to recommend (the Pearson correlation coefficient for Q20 and recommendation is 0.71, with $p < 0.001$), confirming the consistency between user experience and system acceptance. That is, a good experience will increase users' willingness to recommend, underpinning the practical value of the system.

Thematic Analysis of Open-Ended Responses

To supplement the quantitative research results and delve deeper into the details of user experience, this study conducted open coding and semantic clustering analysis on 282 valid open-ended text responses. Through sentence-by-sentence parsing of text semantics, extraction of core concepts, and categorical aggregation, five statistically representative themes related to STEM vocational learning experiences were finally identified. The typical characteristics of each theme and examples of student feedback are shown in **Table 7**.

A word cloud visualization was further used to present the distribution of high-frequency keywords (see **Figure 8**). The results showed that words such as "emotion," "pace," "encouragement," "feedback," and "personalization" occupied prominent positions, which were highly consistent with the above themes, intuitively confirming the core needs of students.

Overall, these qualitative findings not only corroborate the previous quantitative research results (such as the strong correlation between personalized teaching and satisfaction), enhancing the consistency and reliability of the research conclusions, but also reveal specific limitations of the current system—for example, deviations in emotion recognition in scenarios with facial occlusion, and users' demands for improvements in emotional state reporting and visualization functions. These provide targeted directions for subsequent system optimization.

Validation of Emotion Recognition Accuracy

While the primary focus of this study is on user perception, a supplementary analysis was conducted to validate the accuracy of the emotion data underpinning the system's adaptive decisions. A random sample of 500 facial image frames and 100 audio clips was extracted from the complete dataset. Two human annotators, who were blind to the system's predictions, independently coded these samples. Inter-rater reliability was high (Cohen's kappa = 0.81). The system's predictions were then compared against the human consensus labels. The results showed an average agreement rate of 85% (precision: 0.83, recall: 0.87, F1-score: 0.86). This confirms that the emotion recognition performance was reliable and provided a valid and solid data foundation for the adaptive learning logic analyzed in this paper.

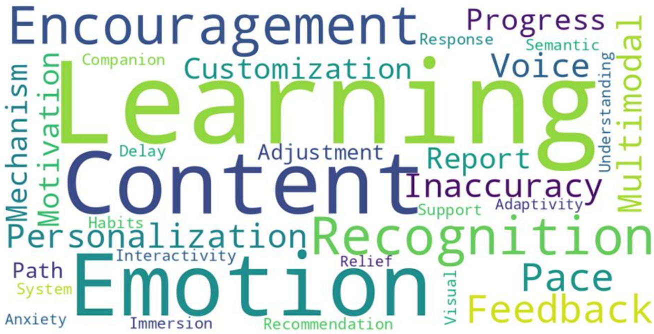

Figure 8. Word cloud of student feedback (Source: Authors' own elaboration)

DISCUSSION

The findings of this study provide compelling evidence supporting the efficacy of emotion-driven multimodal AI learning systems in vocational education settings. The integration of emotion recognition, adaptive content delivery, and user-centered feedback mechanisms appears to positively influence learners' motivation, cognitive engagement, and overall satisfaction. This section discusses the theoretical implications, practical significance, and observed limitations of the results in light of existing research (Pei et al., 2024).

Theoretical Implications and Relationship with Previous Studies

This study validates the proposed model's theoretical foundation, rooted in affective computing, personalized learning theories, and cognitive-affective framework of academic achievement (Marinier & Laird, 2008). The extracted three-factor structure—emotion recognition

Table 7. Thematic categories and student comments

Theme	Representative student quotes
Personalized learning pace	"When I'm not feeling well, it slows down the pace—like having an observant tutor."
Emotional voice feedback	"When I'm feeling low, the system encourages me through voice prompts. It's comforting."
Multimodal interaction	"I learn better with audio and visuals. The system pushes videos and diagrams that help a lot."
Emotion recognition errors	"When I wear a mask, it thinks I'm unhappy. This needs improvement."
System improvement suggestions	"A weekly emotional report would be helpful to reflect on my learning state."

and feedback (F1), personalized instructional adaptation (F2), and satisfaction and privacy trust (F3)—aligns closely with the conceptual model, confirming its multi-dimensional validity.

Notably, strong positive correlations between emotion recognition items (e.g., Q1 and Q2) and motivation-related items (e.g., Q8 and Q9) support the proposition that affect-aware systems enhance learner engagement and persistence, consistent with D'Mello and Graesser (2012), who highlighted real-time emotion tracking as pivotal for improving learning outcomes via adaptive interventions.

The scale's high internal consistency (Cronbach's alpha = 0.94) and positive student perceptions provide a robust empirical basis for extending the model—potentially incorporating emotional regulation, social-emotional learning, and long-term analytics. Consistent with prior research advocating emotionally intelligent educational technologies (e.g., Calvo & D'Mello, 2010), this study confirms that emotion-responsive systems boost retention, motivation, and satisfaction. Its uniqueness lies in focusing on vocational education—an under-explored domain in affective computing—thus offering new empirical support for applying relevant theories in this field. These findings also provide empirical support for the design of emotionally adaptive AI systems in STEM vocational education contexts.

Practical Significance for Vocational Education

From a pedagogical perspective, the application of an emotion-sensitive AI model aligns closely with the educational needs of vocational learners, who often present greater emotional variability and demand more tailored instructional support compared to students in general academic tracks. The following practical insights emerge:

Personalized pacing and content adjustment: Students appreciated the system's ability to adjust learning pace based on mood and performance (as seen in open-ended responses). Such dynamic responsiveness enhances learner autonomy and reduces stress—a factor especially relevant in high-pressure, skill-based vocational programs.

Emotional feedback and encouragement: Voice-based emotional feedback was highlighted as a key motivational factor, helping learners recover from distraction or demotivation. This suggests that emotional scaffolding can serve as a digital equivalent of teacher empathy in online or blended environments.

User satisfaction and trust: High satisfaction scores for system privacy and path optimization (e.g., Q10, Q11, and Q20) indicate that students are more likely to embrace systems that balance personalization with data protection. This is crucial for broader adoption of AI solutions in educational contexts.

These findings advocate for the broader integration of affect-aware learning systems into vocational curricula, not only as supplemental tools but as core elements of digital pedagogy (Azizah et al., 2025).

Technical Limitations and Areas for Improvement

Despite the overall positive evaluation, both quantitative and qualitative data reveal specific technical limitations that warrant attention. Although the emotion recognition module achieved satisfactory overall performance (F1-score: 0.86), error analysis indicates that the majority of misclassifications occurred under conditions of facial occlusion, such as when participants wore masks. This underscores a non-trivial over-reliance on the visual modality within the current fusion framework and highlights the need for more robust multimodal integration.

Additionally, while the adaptive features were generally well-received, the system's current architecture may overgeneralize emotional responses without sufficiently accounting for individual differences in cultural background, personality traits, and situational contexts. Future implementations would benefit from incorporating personalized emotional profiles and longitudinal mood-tracking functionality.

Students' requests for regular emotion-tracking reports further indicate the need for enhanced visualization tools that support metacognitive reflection and self-regulated learning. These features would strengthen the system's capacity to foster autonomous learning behaviors.

To address these limitations, future iterations will integrate additional modalities—including infrared imaging for remote photoplethysmography-based heart rate estimation—and implement advanced cross-modal Transformer architectures. These improvements aim to enhance the system's robustness to sensory input variations and its capacity to capture complex interactions between heterogeneous data streams.

CONCLUSION AND IMPLICATIONS

This study presents a comprehensive investigation of emotion-driven multimodal AI learning systems in vocational education. Through rigorous quantitative and qualitative analysis of data from 282 vocational college students, we have validated the proposed theoretical model and demonstrated the positive impact of such systems on learner engagement, motivation, and satisfaction.

The core contributions of this research include:

(1) empirical validation of a three-factor model encompassing emotion recognition, personalized adaptation, and system satisfaction, which explains over 65% of the total variance;

(2) demonstration of strong correlations between emotion recognition capabilities and learning motivation, providing robust support for the educational value of emotion-aware design; and

(3) development and validation of a reproducible technical architecture for multimodal affective computing in educational contexts, serving as a blueprint for future research and development.

Theoretically, this study extends established frameworks in affective computing and AI education to the previously underexplored domain of vocational education. It establishes an empirical foundation for understanding the role of emotional intelligence in digital learning environments and introduces a novel conceptualization of learner-AI interaction that emphasizes dynamic emotional adaptability rather than static personalization.

Practically, this research provides actionable insights for multiple stakeholders: system developers can implement the documented technical architecture and address the identified limitations; vocational educators can leverage emotion-aware systems to better support diverse learner needs; policy makers can use these findings to inform decisions about technology integration in vocational education; and curriculum designers can develop AI-enhanced learning materials that incorporate emotional sensitivity.

While this study has limitations—including its single cultural context, reliance on subjective perception for emotion validation, and lack of long-term impact assessment—it establishes a solid foundation for future research. Recommended directions include cross-cultural validation of the model, objective evaluation of emotion recognition accuracy using physiological measures, and longitudinal studies of learning outcomes.

In conclusion, emotion-driven multimodal AI systems represent a promising approach to creating truly adaptive, learner-centered environments in vocational education. By integrating technological sophistication with emotional sensitivity, these systems can help students achieve more effective and meaningful learning experiences (Lin et al., 2023).

Author contributions: HZ: conceptualization, methodology, system development, data collection, formal analysis, writing – original draft; **WL:** validation, investigation, visualization, writing – review & editing, supervision. All authors have read and approved the final manuscript.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that the study involved routine educational activities and anonymized data only. No medical procedures, sensitive personal data, or vulnerable populations were involved. According to institutional guidelines, this research is exempt from formal ethics committee approval. Participation was voluntary, and all data were anonymized prior to analysis. Written informed oconsents were obtained from the participants.

AI statement: The authors stated that generative AI tools were used only for language polishing and minor grammatical editing. They were not used to generate or analyze research data. All interpretations and conclusions were made by the authors.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Afzal, S., Khan, H. A., Piran, M. J., & Lee, J. W. (2023). A comprehensive survey on affective computing: Challenges, trends, applications, and future directions. *IEEE Access, 11*, 1-20. https://doi.org/10.1109/ACCESS.2024.3422480

Al-Saadawi, H. F. T., Das, B., & Das, R. (2024). A systematic review of trimodal affective computing approaches. *Expert Systems with Applications, 255*, Article 124852. https://doi.org/10.1016/j.eswa.2024.124852

Azizah, N., Hanafi, I., & Yusro, M. (2025). Artificial intelligence in vocational education: Perspectives and practices from a literature study. *Global Synthesis in Education Journal, 3*(2), 37-44. https://doi.org/10.61667/w0efrt90

Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18-37. https://doi.org/10.1109/T-AFFC.2010.1

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In E. Cambria, D. Das, S. Bandyopadhyay, & A. Feraco (Eds.), *A practical guide to sentiment analysis* (pp. 1-10). Springer. https://doi.org/10.1007/978-3-319-55394-8_1

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145-157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Li, Y., Schoenfeld, A. H., diSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2020). On computational thinking and STEM education. *Journal for STEM Education Research, 3*(2), 147-166. https://doi.org/10.1007/s41979-020-00044-w

Lin, C.-C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review.

Smart Learning Environments, 10, Article 41. https://doi.org/10.1186/s40561-023-00260-y

Marinier, R. P., & Laird, J. E. (2008). Emotion-driven reinforcement learning. In *Proceedings of the Cognitive Science Society Conference* (pp. 115-120).

Pei, G., Li, H., Lu, Y., Wang, Y., Hua, S., & Li, T. (2024). Affective computing: Recent advances and future trends. *Intelligent Computing, 3*, Article 0076. https://doi.org/10.34133/icomputing.0076

Picard, R. W. (1997). *Affective computing*. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001

Picard, R. W. (2000). *Affective computing*. MIT Press. https://doi.org/10.1007/978-3-540-45012-2_2

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies, 59*(1-2), 55-64. https://doi.org/10.1016/S1071-5819(03)00052-1

Tao, J., & Tan, T. (2005). Affective computing: A review. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 981-995). Springer. https://doi.org/10.1007/11573548_125

Xie, H., Chu, H. C., Hwang, G. J., & Wang, C. C. (2019). Trends and development in technology-enhanced adaptive learning. *Computers & Education, 140*, Article 103599. https://doi.org/10.1016/j.compedu.2019.103599

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39-58. https://doi.org/10.1109/TPAMI.2008.52

Zhai, X., & Tian, L. (2022). Artificial intelligence in STEM education: A systematic review and research agenda. *International Journal of STEM Education, 9*(1), 1-21. https://doi.org/10.1186/s40594-022-00343-6

Zhai, X., Chu, H. C., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, J. M., & Liu, J. B. (2021). A review of artificial intelligence in education from 2010 to 2020. *Education and Information Technologies, 26*(6), 7275-7303. https://doi.org/10.1155/2021/8812542