AI-Based Analysis of Student Frustration: Speech and Facial Expression Recognition

Akniyet Tokhtarov[1], Nurgul Toxanbayeva[2], Meirambala Seisekenova[3], Talgat Baidildinov[4], Dametken Baigozhanova[5] and Asylbek Abden[6]

[1]Department of Pedagogy and Psychology, I. Zhansugurov Zhetysu University, Taldykorgan, Kazakhstan
[2]Department Of General and Implied Psychology, Al-Farabi Kazakh National University, Almaty, Kazakhstan
[3]Department Of Economics and Service, I. Zhansugurov Zhetysu University, Taldykorgan, Kazakhstan
[4]Department of Special Pedagogy, Abay Kazakh National Pedagogical University, Almaty, Kazakhstan
[5]Higher School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan
[6]Department of Physical Culture and Primary Military Training, I. Zhansugurov Zhetysu University, Taldykorgan, Kazakhstan

akniyettokhtarov@outlook.com
toxanbayevanurgul@outlook.com
meirambalaseisekenova@gmail.com (corresponding author)
baidildinovt@outlook.com
dbaigozhanova@gmail.com
asylbekabden.edu@outlook.com

https://doi.org/10.34190/ejel.23.2.4043

Abstract: Frustration is a key affective state that affects student engagement and learning outcomes. While mild frustration can promote persistence in problem-solving, prolonged frustration often leads to disengagement and reduced academic performance. In traditional learning environments, instructors rely on facial expressions, vocal cues, and behavioral indicators to identify frustration and provide timely support. Such monitoring becomes impractical in large or digital classrooms. Artificial intelligence (AI)-based emotion recognition offers a scalable solution by automatically detecting frustration through facial and speech analysis, enabling adaptive interventions in real time. This study proposes a multimodal AI system that integrates facial expression recognition using a Convolutional Neural Network (CNN) and speech emotion recognition with a Transformer-based model. The system uses attention-based feature fusion to improve accuracy by weighting the more informative modalities. The model was trained on benchmark datasets, including DAiSEE, IEMOCAP, and RAVDESS, and evaluated in a real-world study involving 160 Kazakhstani university students in online and in-person learning sessions. AI-generated predictions were compared with instructor assessments to validate the system's performance. Results indicate that the multimodal system outperforms unimodal approaches, achieving 85% accuracy, 83% precision, and 86% recall on benchmark data, with 84% accuracy and precision in real-world conditions. Comparative analysis reveals that speech-based cues are more informative than facial expressions, particularly when frustration is masked or internalized. The system is less effective at detecting subtle frustration, highlighting the need for greater contextual sensitivity. Although limitations remain, the results demonstrate the system's potential for scalable implementation in classrooms and online platforms. These findings support the integration of AI-driven frustration detection into adaptive learning platforms to help educators identify students at risk of disengagement. By enabling timely intervention and support, such tools can contribute to more responsive and inclusive educational environments. Future research should explore cultural variation in emotional expression and long-term effects on learning outcomes.

Keywords: Frustration detection, Emotion recognition, Multimodal learning, Facial analysis, Speech emotion recognition, AI in education

1. Introduction

Frustration in learning arises when students face challenges that hinder knowledge acquisition. It can result from unresolved confusion, complex tasks, or inadequate instructional support, influencing academic engagement and performance (Baker et al., 2025). Minor frustration may promote persistence, but prolonged frustration often leads to disengagement and stress (Rahman et al., 2024). A survey involving 22,983 Chinese college students found that 59.9% experienced academic burnout, which can be associated with frustration, particularly in high-pressure environments (Liu et al., 2023). Therefore, the recognition and mitigation of frustration are

Cite this article: Tokhtarov, A. et al. 2025. "AI-Based Analysis of Student Frustration: Speech and Facial Expression Recognition", *Electronic Journal of e-Learning*, 23(2), pp 143-157, https://doi.org/10.34190/ejel.23.2.4043

essential to support student motivation and performance. Additionally, understanding student frustration can help educators foster a psychologically supportive learning environment, allowing early interventions to reduce emotional strain and prevent dropout.

Traditional classroom instructors can detect frustration through facial expressions, voice tone, and behavioral cues, which enables timely intervention. In digital learning environments, frustration is harder to identify and address. Studies show that approximately 59% of students' frustration with e-texts is linked to extraneous cognitive load, 19% stems from technological difficulties, and 28% from curriculum-related issues (Novak, McDaniel and Li, 2023). If left unresolved, these frustrations can adversely affect motivation and learning outcomes, underscoring the need for more effective digital support systems.

In online learning environments and large classroom settings, personalized support is limited. Artificial intelligence (AI)-based automatic emotion recognition can bridge this gap by detecting frustration in real time from students' affective cues (Henderson et al., 2021; Corza-Vargas et al., 2024). Beyond academic interventions, automated frustration recognition can serve as an early-warning system for educators to identify students who may need additional psychological consultation. By tracking emotional trends, teachers and counselors can proactively provide emotional support or recommend mental health resources as needed.

Emotion recognition technology typically utilizes facial expressions and voice signals. Computer vision is employed to track facial movements, while speech analysis examines vocal features such as pitch and intensity (Malekshahi, Kheyridoost and Fatemi, 2024). However, single-modal approaches have notable limitations. Facial expressions may be ambiguous, and speech analysis can be unreliable in noisy conditions (Agung, Rifai and Wijayanto, 2024). Frustration can also be masked in one modality while being evident in another. Moreover, models trained on controlled datasets may not generalize well to real-world educational settings. A multimodal framework that integrates facial and speech cues is therefore required to improve accuracy and robustness (Henderson et al., 2021). Most prior studies in emotion recognition have concentrated on engagement and boredom, often relying on single-modal data such as facial expressions or interaction logs, which do not adequately capture the complexity of frustration (Moon et al., 2022). Speech-based emotion recognition remains underexplored in this domain (Qian and Han, 2022). This study seeks to address these gaps by developing a multimodal approach for more accurate frustration detection.

The research involves the development and validation of an AI-based system for detecting student frustration through facial expression and speech emotion analysis. The research consists of two phases: (1) developing a hybrid model that integrates a CNN for facial analysis and a recurrent neural networks (RNN) or a transformer-based model for speech recognition, and (2) empirical validation through controlled learning experiments.

This study addresses the following research questions (RQ):

RQ1: Can a multimodal AI model significantly improve frustration detection compared to single-modal approaches?

RQ2: How well do automated predictions align with human (instructor) assessments of frustration in real learning scenarios?

RQ3: What are the practical benefits and challenges of implementing such technology in educational settings?

RQ4: How can machine-driven frustration detection assist educators in identifying students who may need additional psychological support or consultation?

The findings of this study contribute to the growing field of affective computing by presenting a novel multimodal approach for frustration detection in education. Scientifically, the research evaluates the effectiveness of integrating facial and speech cues for emotion recognition and identifies sources of classification errors. Comparisons between different model variations (e.g., face-only vs. voice-only vs. multimodal) provide insights into the added value of each modality. Practically, the study offers a prototype system that could be incorporated into e-learning platforms or intelligent tutoring systems to enhance student support. Additionally, the discussion on ethical implications, such as obtaining student consent, avoiding biases, and protecting privacy, provides clear guidance for responsible use of AI in education. Moreover, intelligent frustration tracking can help educational institutions improve their psychological climate by identifying patterns of emotional distress among students. By integrating frustration detection with psychological counseling services, schools can provide targeted support, ensuring that students receive the help they need before frustration negatively impacts their well-being and academic performance.

By bridging the gap between theoretical advancements in artificial intelligence and real-world applications, this study lays the foundation for intelligent frustration detection systems that foster a psychologically supportive learning environment, improve student engagement, and enhance educational outcomes.

2. Literature Review

2.1 AI in Education: Advances in Multimodal Emotion Recognition

The integration of artificial intelligence in education has led to numerous emotion-aware systems capable of detecting and analyzing students' affective states, including frustration (Bustos-López et al., 2022). AI-driven emotion recognition methodologies predominantly leverage computer vision and speech analysis, utilizing CNNs for facial expression classification and RNNs or transformers for speech-based affect detection (Abbaschian, Sierra-Sosa and Elmaghraby, 2021; Wang, 2022). Despite advances, real-world use faces challenges like expression variability, cultural differences, and ethical concerns pertaining to data privacy and surveillance (Banzon, Beever and Taub, 2024).

While machine-driven affect recognition offers considerable advantages over conventional self-reporting mechanisms or interaction log analyses, its practical efficacy is constrained by dataset limitations and generalization challenges (Moon et al., 2022). Many widely utilized datasets, such as DAiSEE (Gupta et al., 2022), provide valuable benchmark resources but lack ecological validity due to controlled conditions (Aguilera, Mellado and Rojas, 2023). Similarly, speech-based models often struggle with spontaneous discourse, regional accent variations, and ambient noise present in authentic classroom interactions (Song et al., 2021). Addressing these constraints necessitates the development of more robust, adaptable models trained on heterogeneous datasets reflective of real-world learning environments.

2.2 Frustration in Learning: Cognitive and Behavioral Correlates

Frustration in education involves cognitive load, emotional distress, and behavioral disengagement when students face academic obstacles (Pekrun and Marsh, 2022). It manifests as an emotional response to perceived obstacles in learning (Baker et al., 2025). Frustration arises from diverse sources, including unclear instructional guidance, excessive task complexity, and delayed instructor feedback, all of which influence learning outcomes (Henderson et al., 2021). Moderate frustration levels may foster problem-solving skills, sustained frustration is correlated with increased stress, academic disengagement, and attrition (Graesser and D'Mello, 2012).

Empirical research has demonstrated that frustration can be conveyed through a combination of facial, vocal, and behavioral indicators, including tense expressions, strained vocal tone, and task disengagement (Moon et al., 2022; Shou et al., 2024). However, AI-based frustration detection models frequently exhibit classification errors due to emotional similarity with states such as confusion and boredom. Confusion, for instance, is a precursor to frustration but does not inherently signal emotional distress, thereby complicating automated classification (Rahman et al., 2024). Moreover, frustration expression is context-dependent, influenced by task complexity, individual learning history, and cultural norms, necessitating intelligent recognition systems capable of integrating contextual variables alongside multimodal affective cues (Henderson et al., 2021).

2.3 Comparative Evaluation of AI-Based Frustration Detection Models

AI-based frustration detection methodologies typically follow unimodal or multimodal analytical frameworks. Unimodal models, such as facial expression (Solanki and Mandal, 2022) or speech-based systems (Song et al., 2021), often perform poorly due to limited input. CNN-based facial models, though accurate in benchmarks, are sensitive to lighting, occlusion, and expression variability (Pordoy et al., 2024; Pham et al., 2023). Similarly, speech emotion recognition models, while effective in controlled environments, exhibit performance degradation in real-time applications due to background noise and spontaneous linguistic variations (Villegas-Ch et al., 2023).

Multimodal fusion models have demonstrated superior performance by integrating facial and vocal features, resulting in higher frustration classification accuracy (Moon et al., 2022). Such models often use feature-level fusion, combining visual and vocal embeddings to improve robustness. Despite their advantages, multimodal approaches face challenges related to computational cost, real-time deployment, and dataset bias. Models trained on narrow datasets often generalize poorly across student demographics, highlighting the need for broader data sources (Bustos-López et al., 2022). Inconsistent annotation practices make it harder to reach consensus on what qualifies as frustration in different educational settings. Table 1 summarizes key models and their methodological strengths and limitations.

Table 1: Comparative Analysis of AI-Based Frustration Detection Approaches

Study	Modalities	Model Approach	Key Findings	Limitations
Solanki and Mandal (2022)	Facial video	Custom CNN + ANN on DAiSEE	86.6% accuracy for frustration detection	Limited to visual cues; lacks contextual integration
Moon et al. (2022)	Facial video + interaction logs	Supervised multimodal fusion on custom dataset	10% performance improvement over unimodal models	Small sample (31 students); tested in controlled settings
Song et al. (2021)	Speech audio	Wide ResNet on spectrograms from game-play corpus	Enhanced classification accuracy over baseline CNN	Absence of visual cues; non-educational domain focus
Rahman et al. (2024)	Facial video + speech	Deep learning fusion model on EmoDetect	Improved robustness in online learning	Ethical concerns regarding student privacy

These findings confirm the advantages of multimodal fusion, despite challenges in dataset diversity, real-world use, and ethical concerns (Mamieva et al., 2023). Multimodal methods consistently outperform single-modality approaches, supporting the use of combining facial and contextual data. However, small samples and narrow models scopes still limit generalizability. For example, Solanki and Mandal (2022) reported high accuracy using detailed facial features, though retraining is likely required for other contexts.

Recent studies demonstrate the growing relevance of multimodal emotion recognition (MER) for educational contexts where accurate detection of affective states such as frustration is critical. Transformer-based architectures offer state-of-the-art performance by capturing complex dependencies between modalities (Lian et al., 2023). Dual-attention mechanisms enhance cross-modal alignment, particularly in speech and facial inputs (Zaidi, Latif and Qadir, 2024). Models using tensor product fusion and transformer backbones have surpassed 93 percent accuracy in recognizing student emotions during learning tasks (Xiang et al., 2024). Body gesture data has also proven valuable, with trimodal systems achieving high accuracy by integrating facial expressions, speech, and posture (Yan et al., 2024). Graph-based reasoning networks like Emotion-LLaMA support fine-grained emotion interpretation and contextual reasoning (Cheng et al., 2024). Systematic reviews emphasize the need for broader dataset diversity and ethical deployment in classrooms (Ahmed, Al Aghbari and Girija, 2023; Khare et al., 2024). Overall, these advancements confirm the potential of MER technologies to support emotionally responsive learning environments when implemented with consideration for practical, cultural, and ethical constraints.

2.4 Ethical Considerations in AI-Based Emotion Recognition

The deployment of automated emotion recognition in education requires close examination of ethical implications, particularly concerning privacy and algorithmic bias. These systems rely on sensitive biometric data, such as facial images and voice recordings, raising concerns about data security and informed consent (Mattioli and Cabitza, 2024). If unregulated, such tools may create a surveillance-oriented environment, where students modify their behavior due to constant monitoring, potentially undermining pedagogical efficacy (Rhue, 2018). Ethical technology deployment mandates transparency, student autonomy in data sharing, and localized data processing to mitigate privacy risks (Mattioli and Cabitza, 2024). Recent studies also emphasize the need for transparent and secure educational technology infrastructures (Sakhipov et al., 2022).

Algorithmic bias is another key concern. Emotion classifiers often vary in accuracy across demographic groups, leading to differential classification outcomes (Rhue, 2018). Cultural differences in emotional expression further complicate generalizability, calling for fairness-aware design and diversified training datasets (Corza-Vargas et al., 2024). Interdisciplinary research is essential to ensure ethical alignment with pedagogical and legal standards.

Future research should focus on expanding dataset diversity, refining fusion models, and enabling adaptive real-time learning environments. Longitudinal studies are also needed to assess the long-term impact of frustration detection on motivation and academic resilience (Baker et al., 2025). Additionally, incorporating context, such as task difficulty and prior performance, may enhance classification and intervention accuracy (Moon et al., 2022). With stronger technical design and ethical safeguards, emotion-aware AI can become a transformative tool for supporting student learning while preserving privacy and autonomy.

Overall, the literature affirms the promise of multimodal emotion detection in education. Yet, issues remain around generalizability, data inclusivity, and ethical implementation. Comparative analyses reveal that while

multimodal models outperform unimodal ones, their success depends on diverse data, adaptable design, and bias mitigation. This study addresses these challenges by proposing a robust frustration detection framework, validating it in authentic settings, and offering ethical guidelines for fair adoption.

3. Materials and Methods

3.1 Data Sources and Preprocessing

This study investigates AI-based frustration detection by developing and evaluating a multimodal recognition system. The research comprises two primary phases: (1) the development of an AI model integrating facial expression and speech emotion recognition and (2) empirical validation through controlled learning experiments.

The model was trained using benchmark datasets, including DAiSEE (frustration-labeled video data), IEMOCAP (emotionally expressive speech), and RAVDESS (acted multimodal emotional expressions). To assess real-world applicability, an experimental study was conducted with 160 university students in structured learning scenarios designed to elicit frustration through technical disruptions, complex problem-solving tasks, and delayed instructor feedback. AI-based predictions were systematically compared with instructor evaluations to assess classification accuracy, practical feasibility, and ethical considerations related to privacy and bias. A detailed breakdown of the datasets used is presented in Table 2.

Table 2: Summary of Datasets Used for Model Training

Dataset	Samples	Participants	Emotion Labels	Modality
DAiSEE (Gupta et al., 2022)	9,068 videos	112 users	Boredom, Confusion, Engagement, Frustration (4 intensity levels)	Video
IEMOCAP (Busso et al., 2008)	302 dialogues	10 speakers (5 pairs)	Angry, Excited, Fear, Sad, Surprised, Frustrated, Happy, Disappointed, Neutral	Audio-Video
RAVDESS (Livingstone and Russo, 2018)	7,356 audio-visual files	24 actors (12 female, 12 male)	Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust (speech) + Song emotions	Audio-Video, Audio-only, Video-only

3.2 AI Model Architecture

The multimodal AI architecture, illustrated in Figure 1, integrates two primary branches: a convolutional neural network for facial expression analysis and a transformer-based model for speech emotion recognition.

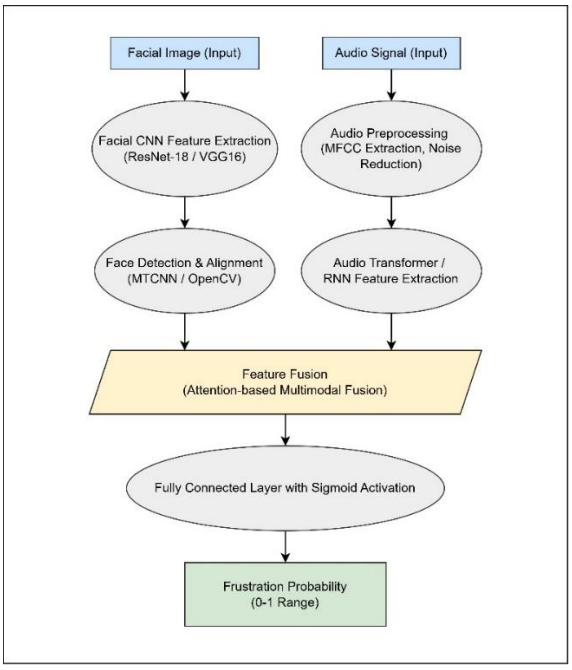

Figure 1: Multimodal AI Architecture for Frustration Detection, Illustrating the Facial CNN, Audio Transformer, and Fusion Layers

The CNN, based on ResNet-18 or VGG16, extracts both low- and high-level features relevant to frustration detection, such as textures, furrowed brows, and narrowed eyes. A feature pyramid module enhances the model's ability to capture fine-grained facial expressions (Mamieva et al., 2023). Instead of direct classification, the CNN generates an embedding vector, which is smoothed over time to reduce sensitivity to brief fluctuations. For speech processing, the model analyzes Mel-frequency cepstral coefficients (MFCCs) to extract frustration-related vocal cues. Both LSTM and Transformer architectures were considered, with the Transformer outperforming LSTM in capturing early vocal indicators of frustration.

Feature fusion was implemented using an attention-weighted strategy, dynamically prioritizing facial or vocal cues depending on their informativeness (Wang et al., 2023; Zaidi, Latif and Qadir, 2023). The final classification layer applies a sigmoid activation function, using binary cross-entropy loss for optimization. Dropout and L2 regularization were included to prevent overfitting.

The CNN and speech models were pre-trained separately before being combined for joint fine-tuning. The fusion mechanism adapts dynamically, giving priority to facial expressions when vocal signals are unclear, and vice versa. This approach significantly improves accuracy over unimodal models (Moon et al., 2022). To minimize false positives, frustration is detected only when both modalities indicate it, or when at least one parameter exceeds a critical level. The architecture is depicted in Figure 1, illustrating the integration of facial and vocal modalities within the frustration detection system.

3.3 Experimental Setup

The proposed model was evaluated through two main experimental phases: (1) offline evaluation on curated datasets, using train/validation/test splits to measure baseline performance, and (2) real-time experimental testing with student volunteers in real learning scenarios to assess real-world effectiveness and compare AI-based frustration detection with human observations. The overall structure of these experimental phases is illustrated in Figure 2.

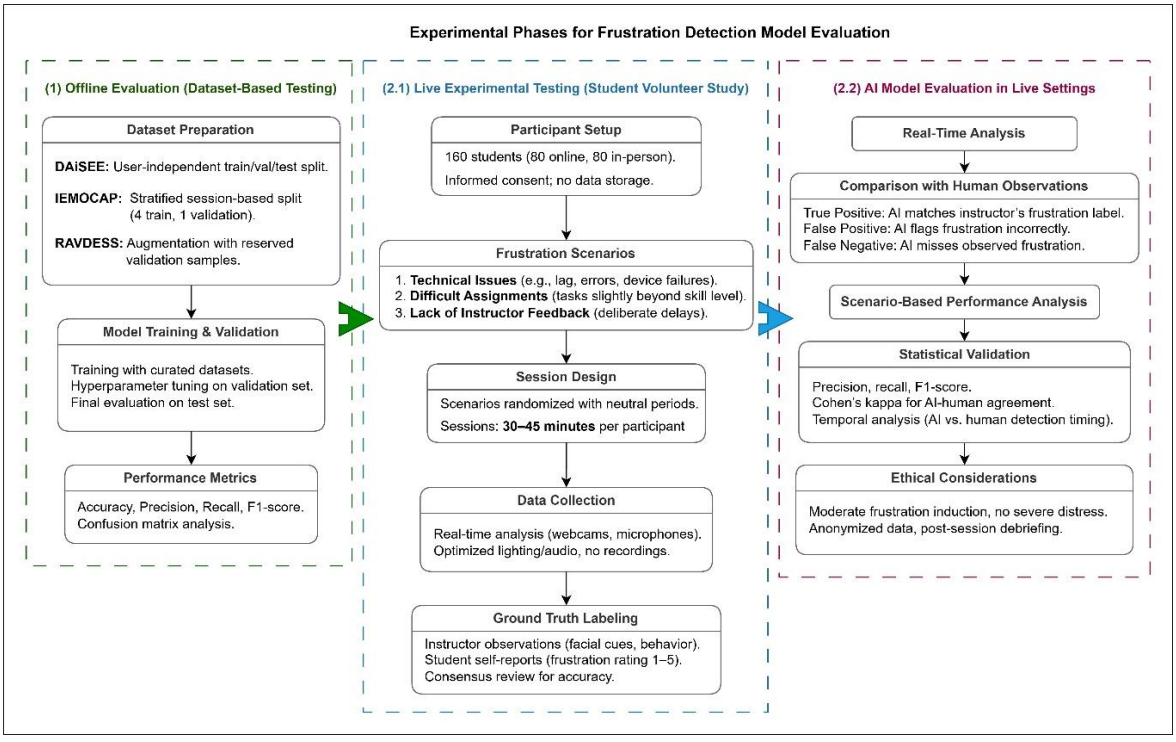

Figure 2: Experimental Phases for Frustration Detection Model Evaluation

This research followed the official DAiSEE partitions: the training set for model training, the validation set for hyperparameter tuning, and the test set for final evaluation. DAiSEE ensures a user-independent split (no overlap between train/val/test), which helps reduce overfitting. For the IEMOCAP audio, following the approach of Busso et al. (2008), a stratified split was conducted based on the five dialogue sessions. Four sessions were used for training and one for validation, rotating for cross-validation due to the small speaker pool. RAVDESS,

following the framework outlined by Livingstone and Russo (2018), primarily served to augment training. A subset was used for mini-validation to monitor overfitting.

The evaluation employed standard classification metrics: Accuracy, Precision (*true positives / [true positives + false positives]*), Recall (*true positives / [true positives + false negatives]*), and F1-score (harmonic mean of precision and recall). These metrics were calculated on both held-out dataset partitions and subsequently validated with volunteer testing data. Confusion matrices helped analyze frequent misclassifications (e.g., confusion vs. frustration).

A total of 160 university students joined controlled learning sessions. Participants provided informed consent, and no media were stored. Half participated online via personal computers and webcams, and half attended in-person classes with cameras and microphones, addressing different practical challenges related to equipment setup.

Frustration scenarios were based on common triggers from prior research: (1) technical issues (e.g., slow loading, errors, or hardware malfunctions); (2) difficult assignments, slightly beyond skill levels to induce productive struggle; and (3) lack of instructor feedback, where instructor responses were intentionally delayed, creating feelings of being unsupported. Each participant faced randomized frustration scenarios alongside baseline tasks. Sessions lasted approximately 30–45 minutes.

Student facial and vocal data were analyzed in real time via webcams (online) or classroom cameras and microphones (in-person), without storing recordings. Lighting and audio conditions were adjusted for realistic settings. Frustration instances were identified using scenario timestamps, instructor observations, and student self-reports. Instructors or trained researchers monitored facial expressions, vocal reactions, and behavioral cues (e.g., sighs, frowns, disengagement). Participants rated their frustration on a 1–5 scale, refining observer assessments. Discrepancies were resolved through second rater consensus to ensure labeling accuracy.

After obtaining informed consent, the model processed facial and speech data in real time using a sliding-window approach, generating frustration probabilities flagged when exceeding optimized thresholds. Detected instances, logged with timestamps, were compared to human observations to assess true positives, false positives, and false negatives. Performance metrics such as precision, recall, accuracy, and F1-score were used to evaluate scenario-specific strengths and weaknesses (Table 3). Ethical guidelines ensured minimal distress, with real-time processing conducted anonymously without data retention. Post-session debriefings confirmed participant well-being. Statistical analyses, including Cohen's kappa, measured alignment between AI predictions and human labels, validating the model's effectiveness and identifying areas for improvement.

3.4 Ethical Considerations

Ethical principles guided all stages of this study, with particular focus on privacy, informed consent, and bias mitigation. To protect participants, no video, photo, or audio recordings were stored during testing; only real-time outputs were analyzed. Results were anonymized, and all visuals used in analysis were either blurred or abstracted. The system is designed to function without storing raw data, allowing real-time, local processing on user devices. These measures ensure compliance with the General Data Protection Regulation (GDPR), including principles of data minimization and informed processing.

All participants provided informed consent after being clearly briefed on the study's goals, data use, and their right to withdraw. It was explicitly stated that the AI was a research tool and not a diagnostic system, reducing any psychological pressure. Bias mitigation efforts included the use of diverse datasets (e.g., DAiSEE, IEMOCAP, RAVDESS) and a participant pool representing multiple genders and ethnicities. A multimodal fusion approach helped reduce bias from any single input source.

The system is intended to support learning, not monitor or penalize students. Its outputs are meant to prompt supportive interventions, not judgments. Ethical safeguards were applied throughout to prioritize student well-being and ensure the system remains a responsible and learner-centered tool.

4. Results and Discussion

4.1 Performance Analysis of AI-Based Frustration Detection

4.1.1 Phase 1: Results of offline evaluation (dataset-based testing)

As shown in Table 3, the multimodal model outperformed the single-modality baselines. Precision (83%) and recall (86%) show the model detects frustration accurately without excessive false positives. This balance

suggests the model is both effective at identifying frustration and avoiding excessive false positives – a desirable trait for practical use. The model's success is attributed to the complementary strengths of facial and voice signals: when one modality failed, the other often compensated. For instance, a sample with a neutral facial expression but frustrated vocal tone was correctly flagged by the audio model. Conversely, another case with a calm voice but a scowling face was accurately classified by the vision model. The attention-based fusion mechanism dynamically prioritized the more informative input, improving robustness. Benchmarks from prior studies suggest that this F1-score (0.85) is competitive. While direct comparisons are difficult due to dataset differences, the performance is higher than some prior multimodal models, particularly in binary frustration classification.

Table 3: Performance of Unimodal vs. Multimodal Models on Test Data

Model Variant	Accuracy	Frustration		
		Precision	Recall	F1-score
Facial CNN (vision only)	75%	0.78	0.70	0.74
Audio Transformer (audio only)	78%	0.80	0.75	0.77
Multimodal CNN+Audio (fusion)	**85%**	**0.83**	**0.86**	**0.85**

As summarized in Table 3, the multimodal fusion achieved a notable F1-score improvement (+11%) over unimodal models, demonstrating its practical value in capturing frustration that might be missed when relying solely on facial or vocal cues.

4.1.2 Phase 2: Results of real-time experimental testing

The real-world evaluation tested the model on 160 students across 216 frustration-inducing scenarios. Instructor observations confirmed 207 actual frustration instances, while the AI flagged 186 instances (Table 4), of which 178 were true positives, 8 were false positives, and 30 were missed detections. Additionally, the model correctly identified 32 cases where students were not experiencing frustration (true negatives), distinguishing them from similar emotions like confusion or concentration. The model achieved 84% precision and 86% recall, closely aligning with its offline test performance. This suggests it generalizes well, though some cases of internalized frustration were missed, and confusion was occasionally misclassified as frustration.

To further analyze the model's classification performance, Figure 3 presents the confusion matrix for frustration detection. The matrix illustrates true and false classifications in the test dataset. The true label represents the actual frustration state as determined by human observers, while the predicted label indicates the AI's classification output.

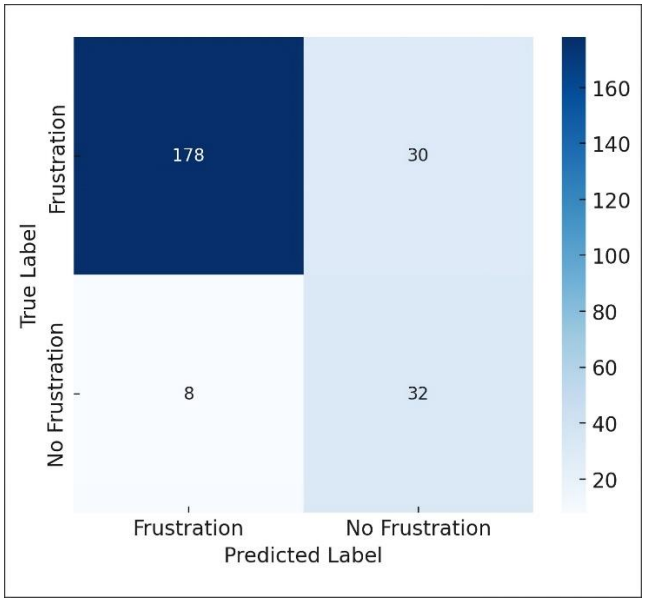

Figure 3: Confusion Matrix for Frustration Detection

Table 4 shows that frustration induced by delayed feedback was the most difficult for students, and also yielded the highest detection rate (87%). However, confusion during challenging tasks led to more false positives, suggesting a need for better differentiation.

Table 4: Frustration Detection Results in Different Learning Scenarios (N=160 students)

Frustration Scenario	Instances (students) with frustration (per instructor)	Detected Cases of Frustration	Detection Rate (Recall)	Noted False Alarms
Technical Issues (online)	40/80 (online group only, N=80 online students, 40 showed frustration)	36 out of 40 actual cases	90%	1 (in 2 cases model thought frustration during minor lag that student shrugged off)
Difficult Assignment	56/160 (some students enjoyed challenge)	46 out of 56 cases	~82.14%	3 (flagged students as frustrated, but they reported only mild confusion)
Lack of Feedback	120/160 (most students found being ignored frustrating)	104 out of 120 cases	~86.7%	4 (brief frustration was flagged in cases where students were only slightly annoyed)

(Note: Each scenario was conducted for each student. "Instances with frustration" is how many students actually felt frustrated in that scenario according to observation/self-report. "Detected Cases" is how many of those instances the system successfully flagged. False alarms indicate cases where frustration was flagged, but observers did not confirm its presence.)

Tables 3 and 5 show the model's high precision (83% on test data, 84% in experiments), meaning most flagged frustration instances were accurate. However, optimizing the detection threshold is crucial because lowering it boosts recall but risks more false alerts, which may be impractical in a classroom. Qualitative observations revealed a reliance on overt signals. Vocal expressions such as sighs triggered instant detection, while silent frustration took longer to register due to temporal smoothing. Masked emotions, such as polite smiles covering irritation, often went undetected, reflecting a limitation shared by human observers. Interestingly, participants who quickly shifted from frustration to focus were sometimes initially misclassified as frustrated, but the probability decreased as their demeanor stabilized. This suggests the system captured momentary states rather than persistent frustration. Overall, the model effectively detected frustration when clear cues were present but struggled with internalized frustration and confusion misclassification. Further improvements may include context-aware features (e.g., task difficulty) and adaptive sensitivity based on individual patterns.

Figure 4 shows that AI-based frustration detection closely matches instructor evaluations, particularly for technical issues (90%) and lack of feedback (87%). While the model slightly underperforms in recognizing frustration from difficult assignments (82% vs. 95%), overall results indicate its reliability. This suggests the system operates effectively without constant instructor oversight and is scalable for large or online classes.

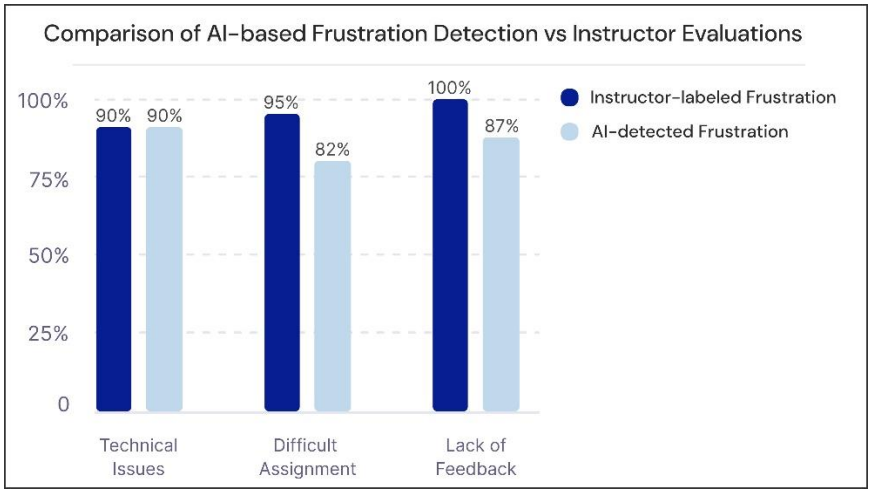

Figure 4: AI vs. Instructor Frustration Detection

4.2 Comparative Analysis of Model Variations

To evaluate the effectiveness of the multimodal approach, experiments were conducted on the same group of students, with each modality activated sequentially before combining them. Frustration detection was first assessed using facial expressions, then vocal cues, and finally both together in a multimodal fusion model. This sequential approach enabled a direct comparison, confirming the advantages of multimodal fusion. As shown in Table 5, the fusion model outperformed unimodal models by ~10% in accuracy, demonstrating that frustration is best detected through a combination of facial and vocal signals rather than a single modality alone.

Among single-modality models, the audio model (78%) slightly outperformed the facial model (75%), suggesting that vocal cues were more reliable for frustration detection. This could be due to facial masking in formal settings, while vocal tone is harder to control. Additionally, DAiSEE's frustration labels may have included milder cases that were visually ambiguous, whereas IEMOCAP's frustration-labeled utterances were clearer. The attention-based fusion further improved performance (F1 = 0.85) by dynamically adjusting modality weights, favoring voice when facial cues were neutral and vice versa. Multimodal fusion improved detection confidence and reliability. Notably, hidden frustration misses were lowest with the fusion model, indicating better capture of nuanced emotional signals.

Table 5: Analysis of Frustration Detection by Parameter During the Experimental Phase

	Performance				Errors	Confidence	
	Accuracy	Precision	Recall	F1-score	Hidden Frustration Misses	Average Confidence (Conf. Avg)	Max Confidence (Conf. Max)
Facial CNN (Vision)	77%	0.79	0.70	0.74	7	0.71	0.92
Audio Transformer	76%	0.79	0.76	0.77	6	0.74	0.95
Multimodal Fusion	84%	0.84	0.86	0.85	10	0.81	0.98
+% from CNN	**+9.1%**	**+6.3%**	**+22.9%**	**+14.9%**	**+42.9%**	**+14.1%**	**+6.5%**
+% from Audio	**+10.5%**	**+6.3%**	**+13.2%**	**+10.4%**	**+66.7%**	**+9.5%**	**+3.2%**

Error analysis identified key challenges in unimodal models. The facial model often misclassified concentration as frustration, while the audio model confused high-arousal emotions like excitement with frustration. The fusion model reduced these errors by leveraging cross-modal context, though misclassifications persisted when both modalities misaligned, such as strained vocal tones paired with a furrowed brow. Transformer-based audio processing outperformed LSTM (~2-3% higher F1-score), likely due to its ability to capture key vocal cues early. The model's decision window (2-5s) balanced reactivity and stability, avoiding the noise of shorter windows while preventing missed transient frustration signals in longer ones.

This analysis directly addresses the research questions (RQ).

RQ1: Can a multimodal AI model significantly improve frustration detection compared to single-modal approaches?

The findings demonstrate that the multimodal AI model (Accuracy = 85%, F1-score = 0.85) significantly outperforms unimodal approaches, including facial expression analysis alone (75%) and speech emotion recognition alone (78%). These results highlight the advantage of integrating both modalities, as they provide complementary information that enhances the reliability of frustration detection.

RQ2: How well do automated predictions align with human (instructor) assessments of frustration in real learning scenarios?

The study demonstrates a strong alignment between AI-based frustration detection and instructor evaluations, with a precision of 84% and recall of 86%. This suggests that the AI system is highly effective in identifying frustration when clear affective cues are present. However, some limitations persist, particularly in cases where frustration is internalized or subtly expressed, leading to occasional false negatives.

RQ3: What are the practical benefits and challenges of implementing such technology in educational settings?

The system's ability to detect frustration in real-time suggests its potential integration into adaptive learning platforms. However, effective implementation requires careful calibration of detection thresholds to balance sensitivity and specificity, minimizing false alarms while ensuring meaningful intervention opportunities. Additionally, ethical considerations related to privacy, consent, and potential bias must be addressed to facilitate responsible deployment in educational environments.

RQ4: How can machine-driven frustration detection assist educators in identifying students who may need additional psychological support or consultation?

The system's capacity to detect early indicators of frustration enables educators to identify students at risk of disengagement or heightened emotional distress. This proactive approach can support timely psychological interventions, fostering a more supportive learning environment. Future refinements, particularly in context-aware modeling and personalized sensitivity adjustments, will further enhance the system's capacity to assist educators in addressing students' emotional and academic needs.

4.3 Interpretation of Errors and Model Limitations

While the model effectively detected frustration, certain limitations led to errors. Key issues include: (1) *Confusion vs. Frustration Misclassification:* The model often mistook confusion for frustration, especially in students concentrating intensely. Expressions like furrowed brows and squinting appeared in both states, making differentiation difficult. The binary classification approach did not explicitly account for confusion, leading to misclassifications. A multi-label model or task performance data (e.g., tracking incorrect attempts) could help distinguish these emotions. (2) *Subtle or Internalized Frustration Misses:* Some students exhibited frustration in subtle ways, such as posture changes, which the AI struggled to detect. Unlike a human instructor who could infer frustration from context, the model lacked situational awareness. Temporal smoothing, while reducing false positives, sometimes ignored brief frustration moments. A more context-aware approach could improve detection of mild frustration. (3) Short-lived or Contextual Frustration: The model flagged temporary frustration that quickly resolved, such as brief reactions to technical issues. Since it did not track frustration duration, alerts were sometimes unnecessary. Future refinements could incorporate frustration persistence tracking (e.g., only triggering alerts if frustration lasts over 30 seconds). (4) Sensor/Input Limitations: Real-world conditions, such as face occlusion, poor lighting, or background noise, impacted detection. In group settings, isolating a single student's voice was challenging, leading to misattributions (e.g., detecting another student's sigh as frustration). The audio model also struggled with soft-spoken frustration, as it prioritized vocal intensity. Integrating speech-to-text analysis could help by identifying explicit frustration-related words rather than relying solely on tone.

These limitations are consistent with the Cognitive Disequilibrium Theory (Graesser and D'Mello, 2012), which suggests that frustration and confusion lie on a continuum, where unresolved confusion may develop into frustration. This may partly explain the model's difficulty in distinguishing between the two. A lack of contextual awareness also led to occasional false positives unrelated to coursework. Incorporating task-related metrics, such as incorrect attempts or delayed responses, and applying adaptive thresholds based on individual expressiveness could improve detection accuracy. Although these refinements were not implemented due to sample size constraints, the system still provides valuable early indicators to support instructors.

4.4 Ethical Considerations and Bias in AI-Based Emotion Detection

The study involved university students from Kazakhstan, representing both European and Central Asian ethnicities. This reflects local diversity and helped calibrate the model to common facial and vocal expressions in the region. While performance was consistent across tested groups, generalizability to other populations remains uncertain due to cultural and linguistic differences in emotional expression.

The study identified no major biases across gender or ethnicity, though the limited sample size prevents definitive conclusions. The model may still underperform for individuals expressing frustration atypically, including culturally diverse or neurodivergent students. This echoes concerns from AI proctoring systems about fairness and unintended impacts (Sakhipov, Omirzak and Fedenko, 2025). Both false positives and missed detections could affect student support, reinforcing the need for larger, more varied datasets and ongoing evaluation.

Participants sometimes altered their behavior due to awareness of being monitored, highlighting the importance of minimizing observer effects. Emotion detection tools should assist, not surveil, with final decisions left to

educators. Ethical deployment requires full transparency, informed consent, and compliance with GDPR and UNESCO principles to ensure fairness, privacy, and student trust.

4.5 Implications and Applications

The promising performance of the frustration detection model opens up several applications in adaptive learning systems. One immediate use is real-time alerts for instructors. In in-person or online classroom settings, a dashboard could highlight students displaying frustration, such as by placing a red border around their video feed, allowing teachers to check in before a student disengages. This early warning system would help guide targeted interventions, enabling more proactive classroom management. An early version of the frustration analysis interface is shown in Figure 5; future versions may introduce design improvements and expand visualization capabilities.

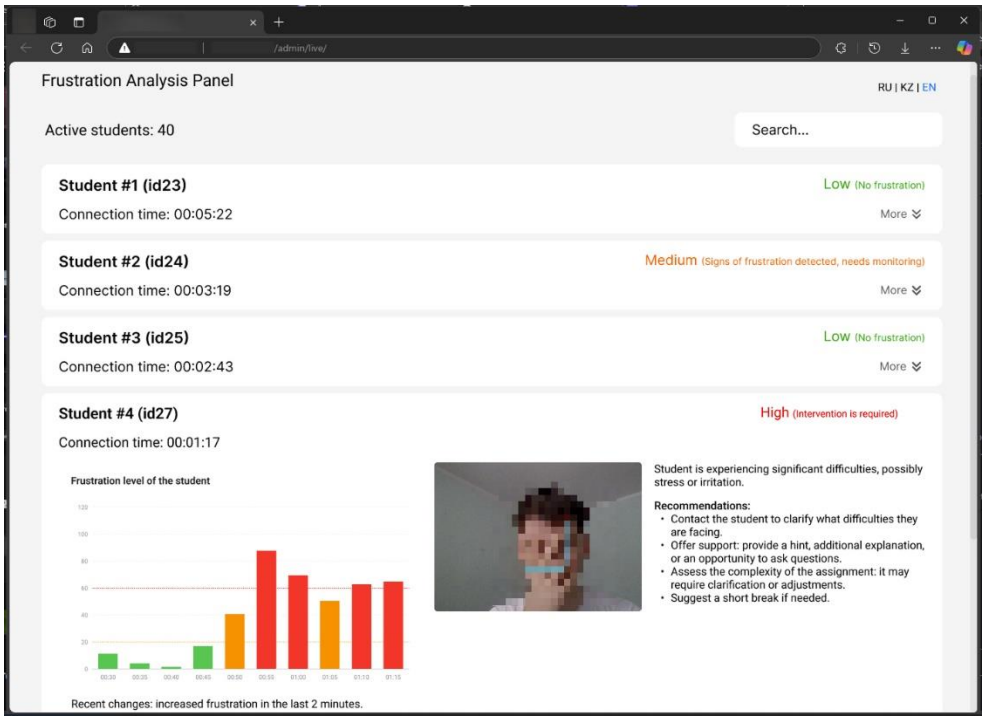

Figure 5: Real-Time Frustration Analysis Panel: Student Frustration Detection with Intervention Recommendations

The AI model can enhance adaptive tutoring by adjusting content based on frustration detection. When a student struggles, it can offer hints, encouragement, or modify the difficulty level to maintain learning flow. Self-awareness feedback could help students regulate emotions by suggesting breaks or review sessions. Aggregated frustration data informs curriculum improvements, highlighting lessons that need refinement. Tracking frustration trends enables personalized support, while emotion data can optimize group learning by balancing teams. Expanding detection to boredom or confusion could refine engagement strategies. Real-time frustration indicators and adaptive feedback foster student engagement, emotional regulation, and curriculum optimization.

The findings highlight the practical viability of multimodal AI systems for real-time frustration detection in educational environments. Technologically, the proposed model can be integrated into adaptive learning platforms, enabling timely and personalized interventions. Pedagogically, it supports instructors in identifying at-risk students, facilitating early emotional and academic support. At the institutional level, aggregated frustration trends may inform curriculum design and psychological service allocation. The findings provide a foundation for implementing frustration-aware feedback in intelligent tutoring systems and for guiding institutional strategies that enhance student support and well-being.

5. Conclusion

This study demonstrates the potential of AI-based emotion recognition to support education through a multimodal system that detects student frustration using facial and vocal cues. The model outperformed

unimodal approaches and showed high accuracy in both controlled and real-world conditions, offering practical benefits for adaptive learning. By providing real-time emotional feedback, the system can help instructors identify at-risk students and respond with timely support, especially in digital or large-scale classrooms where individual monitoring is limited. In addition to individual interventions, such tools may also contribute to a more emotionally supportive classroom climate by making students' affective states more visible and actionable for educators.

The findings underscore the importance of integrating multiple modalities to capture both overt and subtle expressions of frustration. While the system shows promise, it was tested in structured scenarios and uses binary classification, which may oversimplify emotional dynamics. Its broader effectiveness and long-term impact on learning remain to be explored. Future work will also focus on refining multi-label classification, expanding cross-cultural validation, and embedding emotion-aware systems into diverse educational technologies. Educators and edtech developers may consider integrating such systems as part of responsive instructional design, guided by clear protocols for transparency, consent, and student support.

6. Limitations and Future Work

6.1 Delimitations of the Study

This study intentionally focused on the detection of frustration as a target emotional state, excluding other related emotions such as confusion, boredom, or anxiety. The participant pool was limited to university students in Kazakhstan to ensure contextual consistency and manageability, and findings may not generalize to other demographics or educational settings. Only facial and vocal modalities were used for emotion recognition; behavioral logs, physiological data, and contextual cues were deliberately excluded. The system was tested in short, individual learning sessions, and long-term emotional trends or academic outcomes were not examined. Additionally, considerations such as data storage, privacy, and large-scale deployment in real classroom environments were beyond the scope of this initial study, which was primarily focused on developing and validating the core AI model. These aspects are recognized as important directions for future research and practical implementation.

6.2 Research Limitations

The sample consisted of 160 university students from Kazakhstan, reflecting local ethnic diversity. While this ensures cultural relevance, it limits generalizability to other populations where frustration may be expressed differently. The model was also trained on benchmark datasets containing acted emotional responses, which may not fully reflect how frustration occurs in real educational settings. As a result, the model's behavior in natural classroom environments remains untested.

The system was evaluated using binary classification, which simplifies emotional states and does not capture transitions between related emotions such as confusion or disengagement. In addition, the study did not examine whether using frustration detection leads to improved academic performance, motivation, or persistence. This remains an open question for future applied research. Although behavioral context such as task progress was observed during annotation, the model itself does not use these signals as input. Incorporating them could improve detection of subtle or internalized frustration.

6.3 Future Directions

Future research should explore the impact of AI-based frustration detection on academic outcomes, including learning performance, task completion, and knowledge retention. Experimental studies could assess whether real-time emotional feedback enhances student engagement, motivation, or instructional responsiveness. It is also important to examine effects on metacognitive regulation, help-seeking behavior, and persistence in cognitively demanding contexts. Embedding the system into existing educational platforms would allow for usability testing and evaluation of how instructors and learners interpret and apply emotional insights. Further work should address generalizability by testing across age groups, languages, and educational systems, as well as through longitudinal designs that capture emotional dynamics over time. Improving detection sensitivity may require integrating behavioral indicators such as gaze, response timing, or interaction patterns. Models that account for emotional transitions could support more adaptive and context-aware feedback. Finally, the pedagogical use of emotion data must be guided by principles of transparency, privacy, and student autonomy, requiring close collaboration between AI developers, educators, and institutional stakeholders.

AI Statement: AI tools were not used for writing or generating the content of this research paper. Artificial intelligence was only employed during model development and performance evaluation for frustration

detection. All research design, data analysis, and manuscript preparation were conducted manually by the authors.

Ethics Statement: This study followed strict privacy and confidentiality standards. No personally identifiable information was collected, stored, or accessed by the researchers at any point. Facial and vocal inputs were processed locally and automatically by the AI model in real time, with no human access, recording, or transmission. No raw biometric or behavioral data were retained or available during or after the study. All analysis was based on fully anonymized outputs. Participants provided informed consent prior to participation, and the study was designed to ensure minimal intrusion and no harm. The research complies with the GDPR, ensuring lawful, fair, and transparent processing aligned with privacy-by-design principles for AI applications in education.

References

Abbaschian, B.J., Sierra-Sosa, D. and Elmaghraby, A., 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4), p.1249. https://doi.org/10.3390/s21041249.

Aguilera, A., Mellado, D. and Rojas, F., 2023. An assessment of in-the-wild datasets for multimodal emotion recognition. *Sensors*, 23(11), p.5184. https://doi.org/10.3390/s23115184

Agung, E.S., Rifai, A.P. and Wijayanto, T., 2024. Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Scientific Reports*, 14, p.14429. https://doi.org/10.1038/s41598-024-65276-x

Ahmed, N., Al Aghbari, Z. and Girija, S., 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17, p.200171. https://doi.org/10.1016/j.iswa.2022.200171

Baker, R.S., Cloude, E., Andres, J.M.A.L. and Wei, Z., 2025. The confrustion constellation: A new way of looking at confusion and frustration. *Cognitive Science*, 49(1), e70035. https://doi.org/10.1111/cogs.70035.

Banzon, A.M., Beever, J. and Taub, M., 2024. Facial expression recognition in classrooms: Ethical considerations and proposed guidelines for affect detection in educational settings. *IEEE Transactions on Affective Computing*, 15(1), pp.93–104. https://doi.org/10.1109/TAFFC.2023.3275624.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), pp.335–359. https://doi.org/10.1007/s10579-008-9076-6

Bustos-López, M., Cruz-Ramírez, N., Guerra-Hernández, A., Sánchez-Morales, L.N., Cruz-Ramos, N.A. and Alor-Hernández, G., 2022. Wearables for engagement detection in learning environments: A review. *Biosensors*, 12(7), p.509. https://doi.org/10.3390/bios12070509.

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X. and Hauptmann, A., 2024. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint*. Available at: https://arxiv.org/abs/2406.11161 [Accessed 16 June 2025].

Corza-Vargas, V.M., Martinez-Maldonado, R., Escalante-Ramirez, B. and Olveres-Montiel, J., 2024. Students' ethical, privacy, design, and cultural perspectives on visualizing cognitive-affective states in online learning. *Journal of Learning Analytics*, 11(3), pp.24–40. https://doi.org/10.18608/jla.2024.8483.

Graesser, A.C. and D'Mello, S., 2012. Chapter five - Emotions during the learning of difficult material. In: B.H. Ross, ed. *Psychology of Learning and Motivation*. Vol. 57. Academic Press, pp.183-225. https://doi.org/10.1016/B978-0-12-394293-7.00005-4.

Gupta, A., D'Cunha, A., Awasthi, K. and Balasubramanian, V., 2022. DAiSEE: Towards user engagement recognition in the wild. *arXiv preprint*. Available at: https://arxiv.org/abs/1609.01885v7 [Accessed 10 March 2025].

Henderson, N., Min, W., Rowe, J. and Lester, J., 2021. Enhancing multimodal affect recognition with multi-task affective dynamics modeling. *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Nara, Japan, pp.1–8. https://doi.org/10.1109/ACII52823.2021.9597432.

Khare, S.K., Blanes-Vidal, V., Nadimi, E.S. and Acharya, U.R., 2024. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102, p.102019. https://doi.org/10.1016/j.inffus.2023.102019

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C. and Zong, Y., 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10), p.1440. https://doi.org/10.3390/e25101440

Liu, Z., Xie, Y., Sun, Z., Liu, D., Yin, H., and Shi, L., 2023. Factors associated with academic burnout and its prevalence among university students: a cross-sectional study. *BMC Med Educ,* 23, 317. https://doi.org/10.1186/s12909-023-04316-y.

Livingstone, S.R. and Russo, F.A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391.

Malekshahi, S., Kheyridoost, J.M. and Fatemi, O., 2024. A general model for detecting learner engagement: Implementation and evaluation. *arXiv preprint*. Available at: https://doi.org/10.48550/arXiv.2405.04251 [Accessed 10 March 2025].

Mamieva, D., Abdusalomov, A.B., Kutlimuratov, A., Muminov, B. and Whangbo, T.K., 2023. Multimodal emotion detection via attention-based fusion of extracted facial and speech features. *Sensors (Basel, Switzerland)*, 23(12), p.5475. https://doi.org/10.3390/s23125475.

Mattioli, M. and Cabitza, F., 2024. Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. *Machine Learning and Knowledge Extraction*, 6(4), pp.2201-2231. https://doi.org/10.3390/make6040109.

Moon, J., Ke, F., Sokolikj, Z. and Dahlstrom-Hakki, I., 2022. Multimodal data fusion to track students' distress during educational gameplay. *Journal of Learning Analytics*, 9(3), pp.75-87. https://doi.org/10.18608/jla.2022.7631.

Novak, E., McDaniel, K. and Li, J., 2023. Factors that impact student frustration in digital learning environments. *Computers and Education Open*, 5, p.100153. https://doi.org/10.1016/j.caeo.2023.100153.

Pekrun, R. and Marsh, H.W., 2022. Research on situated motivation and emotion: Progress and open problems. *Learning and Instruction*, 81, p.101664. https://doi.org/10.1016/j.learninstruc.2022.101664.

Pham, T.-D., Duong, M.-T., Ho, Q.-T., Lee, S. and Hong, M.-C., 2023. CNN-based facial expression recognition with simultaneous consideration of inter-class and intra-class variations. *Sensors*, 23(24), p.9658. https://doi.org/10.3390/s23249658

Pordoy, J., Farman, H., Dicheva, N.K., Anwar, A., Nasralla, M.M., Khilji, N. and Rehman, I.U., 2024. Multi-frame transfer learning framework for facial emotion recognition in e-learning contexts. *IEEE Access*, 12, pp.151360–151381. https://doi.org/10.1109/ACCESS.2024.3478072.

Qian, F. and Han, J., 2022. Contrastive regularization for multimodal emotion recognition using audio and text. *arXiv preprint* arXiv:2211.10885. Available at: https://arxiv.org/abs/2211.10885 [Accessed 16 June 2025].

Rahman, M.M., Munir, M.U., Rahman, M.M. and Badiuzzaman, M., 2024. EmoDetect: A learning-centred affective database for detecting student frustration in online learning. *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*. Lahore, Pakistan, pp.1-6. https://doi.org/10.1109/ICACS60934.2024.10473235.

Rhue, L., 2018. Racial influence on automated perceptions of emotions. *SSRN*. Available at: https://ssrn.com/abstract=3281765 [Accessed 4 March 2025].

Sakhipov, A., Yermaganbetova, M., Latypov, R. and Ualiyev, N., 2022. Application of blockchain technology in higher education institutions. *Journal of Theoretical and Applied Information Technology*, 100(4), pp.1138–1147.

Sakhipov, A., Omirzak, I. and Fedenko, A., 2025. Beyond face recognition: A multi-layered approach to academic integrity in online exams, *Electronic Journal of e-Learning*, 23(1), pp.81-95. https://doi.org/10.34190/ejel.23.1.3896.

Shou, Z., Huang, Y., Li, D., Feng, C., Zhang, H., Lin, Y. and Wu, G., 2024. A student facial expression recognition model based on multi-scale and deep fine-grained feature attention enhancement. *Sensors*, 24(20), p.6748. https://doi.org/10.3390/s24206748.

Solanki, N. and Mandal, S., 2022. Engagement analysis using DAiSEE dataset. *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, pp.223-228. https://doi.org/10.1109/ICARCV57592.2022.10004250.

Song, M., Parada-Cabaleiro, E., Liu, S., Milling, M., Baird, A., Yang, Z. and Schuller, B.W., 2021. Supervised contrastive learning for game-play frustration detection from speech. *In: M. Antona and C. Stephanidis, eds. Universal Access in Human-Computer Interaction. Design Methods and User Experience*. Cham: Springer International Publishing, pp.617–629. https://doi.org/10.1007/978-3-030-78092-0_43.

Villegas-Ch, W.E., García-Ortiz, J. and Sánchez-Viteri, S., 2023. Identification of emotions from facial gestures in a teaching environment with the use of machine learning techniques. *IEEE Access*, 11, pp.38010–38022. https://doi.org/10.1109/ACCESS.2023.3267007.

Wang, C., 2022. Emotion recognition of college students' online learning engagement based on deep learning. *International Journal of Emerging Technologies in Learning (iJET)*, 17(06), pp.110–122. https://doi.org/10.3991/ijet.v17i06.30019.

Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., Li, C. and Quan, D., 2023. Multimodal transformer augmented fusion for speech emotion recognition. *Frontiers in Neurorobotics*, 17, 1181598. https://doi.org/10.3389/fnbot.2023.1181598.

Xiang, A., Qi, Z., Wang, H., Yang, Q. and Ma, D., 2024. A multimodal fusion network for student emotion recognition based on transformer and tensor product. *arXiv preprint*, arXiv:2403.08511. Available at: https://arxiv.org/abs/2403.08511 [Accessed 16 June 2025].

Yan, J., Li, P., Du, C., Zhu, K., Zhou, X., Liu, Y. and Wei, J., 2024. Multimodal emotion recognition based on facial expressions, speech, and body gestures. *Electronics*, 13(18), p.3756. https://doi.org/10.3390/electronics13183756

Zaidi, S.A.M., Latif, S. and Qadir, J., 2023. Cross-language speech emotion recognition using multimodal dual attention transformers. *arXiv preprint*. Available at: https://doi.org/10.48550/arXiv.2306.13804 [Accessed 10 March 2025].

Zaidi, S.A.M., Latif, S. and Qadir, J., 2024. Enhancing cross-language multimodal emotion recognition with dual attention transformers. *IEEE Open Journal of the Computer Society*, 5, pp.684–693. https://doi.org/10.1109/OJCS.2024.3486904