Do Emotion Recognition Models Generalize to Classrooms? Robustness and Fairness Analysis

Ashwin T S Srigowri Mayasandra Prasanna Joyce Horn Fonteles
Gautam Biswas
Vanderbilt University
Nashville, TN 37212
ashwindixit9@gmail.com, srigowri.mayasandra.prasanna@vanderbilt.edu,
joyce.horn.fonteles@vanderbilt.edu, gautam.biswas@vanderbilt.edu

Abstract

Emotion recognition models are widely used in affective computing and learning analytics. However, most models are trained on web-scale or controlled datasets that differ substantially from real classroom environments. Classrooms introduce variations in camera angle, lighting, image resolution, and interaction structure, which may affect both prediction reliability and fairness across student populations. This paper evaluates emotion recognition models in classroom learning environments. We analyze three state-of-the-art models (EmoNet, HSEmotion, and iMotions) using data collected from 69 students across three learning environments: game-based, embodied, and collaborative activities. The evaluation examines model sensitivity to visual perturbations and demographic variation across Fitzpatrick skin tone groups. Results show that prediction errors increase under changes in camera angle, lighting, and resolution, and vary systematically across skin tone groups. We further observe that embodied learning environments exhibit higher sensitivity to visual perturbations compared to game-based and collaborative settings. Lightweight mitigation strategies using data augmentation and partial fine-tuning reduce both prediction error and fairness gaps. These findings highlight challenges in deploying emotion recognition models in classroom learning environments and motivate evaluation practices that account for real-world classroom conditions.

1. Introduction

Learning can be examined both as an outcome and as a process. While learning outcomes describe what students ultimately achieve such as grades, assessment scores, or demonstrated knowledge; understanding the learning process requires examining how students regulate cognition, motivation, and emotions during learning activities [9, 39]. Research in the learning sciences often analyzes these processes through constructs such as self-regulated learning (SRL), socially shared regulation of learning (SSRL), and student engagement. Emotions play a central role in these processes, influencing how learners interpret challenges, persist through difficulties, and collaborate with peers during complex tasks. Consequently, understanding students' emotional states provides important insight into how learning unfolds in real classroom environments [12, 46, 50].

Advances in affective computing have enabled automated emotion recognition using computer vision models that analyze facial expressions [1, 13, 27]. State-of-the-art models such as EmoNet, HSEmotion, and commercial systems like iMotions[1] estimate continuous emotional states such as valence and arousal from facial images [45, 47]. These approaches provide scalable alternatives to traditional methods such as self-reports or observational coding [8, 10, 33]. However, their reliability in classroom environments remains uncertain.

One challenge arises from the datasets used to train these models. Most emotion recognition models are trained on datasets containing adult facial expressions collected in controlled or semi-controlled settings. Although some datasets include images captured "in the wild," they rarely reflect the visual conditions present in classrooms [41]. In addition, publicly available datasets containing facial expressions of K–12 students are limited. Existing educational datasets are often collected within specific contexts or demographic populations, which restricts their ability to represent the diversity of classroom environments [7, 26]. As a result, models trained on these datasets may not generalize well when applied to different educational contexts or student populations.

Classroom environments introduce additional complexi-

[1] https://imotions.com/

ties for emotion recognition systems. Students interact with peers, change their body orientation, and move within the learning space. These conditions introduce variations in camera angle, lighting, and image resolution, which can affect the visibility of facial features used by emotion recognition models [24]. Prior work has also highlighted potential algorithmic biases in facial analysis systems, suggesting that prediction errors may vary across demographic groups and environmental conditions [5, 36].

Despite increasing use of automated emotion recognition in educational research, systematic evaluations of model behavior under classroom conditions remain limited. In particular, it is unclear how visual perturbations commonly observed in classrooms influence emotion predictions, whether prediction errors vary across skin tone groups, and whether different learning environments introduce additional sources of bias. Addressing these questions is essential before deploying emotion recognition systems in educational settings.

This study evaluates the behavior of emotion recognition models in classroom learning environments. We analyze three widely used models: EmoNet, HSEmotion, and iMotions, using data collected from students participating in three learning environments: game-based, embodied, and collaborative learning activities. The analysis examines prediction robustness under variations in camera angle, lighting, and image resolution, as well as fairness across Fitzpatrick skin tone groups [11, 22]. We further investigate whether lightweight mitigation strategies improve robustness and reduce fairness gaps.

The study addresses the following research questions:

- *RQ1:* How sensitive are emotion recognition models to visual perturbations commonly observed in classroom environments?
- *RQ2:* Do prediction errors vary systematically across Fitzpatrick skin tone groups?
- *RQ3:* Does the learning environment influence the magnitude of prediction bias?
- *RQ4:* Can lightweight mitigation strategies reduce both prediction error and fairness gaps?

To address these questions, this paper makes the following contributions:

- Evaluation of emotion recognition models in K–12 classroom settings, a domain that is underrepresented in existing affective computing benchmarks.
- Application of a paired-instance perturbation framework to isolate the effect of camera angle, lighting variation, image resolution, and skin tone on valence–arousal predictions in classroom data.
- Joint analysis of robustness and fairness under real classroom conditions, revealing that prediction instability and demographic disparities are amplified in this setting.
- Empirical evidence that learning environments, particularly embodied settings, introduce additional sources of variation that impact model reliability.
- Assessment of lightweight mitigation strategies, showing that while improvements are possible, significant robustness and fairness gaps remain.

2. Related Work

Emotion Recognition Models for Valence–Arousal Estimation: In educational settings, affect is often expressed through states such as *confusion*, *frustration*, and *boredom*, which do not map cleanly to basic categorical emotions [12, 19]. For this reason, dimensional affect representations based on *valence* and *arousal* are more appropriate for classroom emotion analysis than basic emotion classifiers alone, since they allow these learning-centered states to be interpreted in a continuous affective space [2, 3, 23, 43].

A large number of FER models have been proposed for affect estimation, including ResNet-based affect predictors, Face-SSD, AffWildNet, VGG-FACE-based models, EmotionGCN, and recent challenge systems built on ensembles, visual transformers, or multimodal fusion [41, 49]. However, many of these methods are either optimized for a single benchmark, designed primarily for categorical emotion recognition, computationally heavy, or not released with convenient pretrained weights for direct reuse in downstream applied settings. By contrast, *EmoNet* [47] and *HSEmotion* [44] remain strong choices for this study because both provide publicly usable pretrained models and have been evaluated across multiple affect benchmarks rather than a single dataset.

EmoNet was designed specifically for joint facial landmark detection, categorical emotion recognition, and continuous valence–arousal estimation in a single architecture. It was evaluated on *AffectNet*, *AFEW-VA*, and *SEWA*, and reported improved performance over baselines such as *ResNet-18*, *Face-SSD*, and *VGG-FACE+2M images* for valence and arousal [47]. *HSEmotion* provides EfficientNet-based pretrained models for facial emotion recognition, including a variant that predicts valence and arousal, and was evaluated across several benchmarks including *AffectNet*, *AFEW*, *VGAF*, *LSD*, and *ABAW4 MTL*. In addition to benchmark performance, HSEmotion was explicitly designed as a high-speed, reusable FER library with public pretrained weights, making it practical for empirical evaluation in new application domains [44, 45]. Thus, although newer FER systems exist, *EmoNet* and *HSEmotion* remain appropriate choices here because they combine continuous affect prediction, multi-dataset evaluation, and public availability. In addition to open-source models, we included *iMotions* as a commercial baseline. iMotions is widely used in applied behavioral and educational research and provides an integrated pipeline for facial expression analysis and affect inference [46]. This makes it a relevant comparison

point for evaluating how research-oriented FER models and commercial systems behave in classroom environments.

Although several FER datasets support affect modeling, such as *AffectNet* [37], *AFEW-VA* [30], and *SEWA* [31], these datasets primarily contain facial imagery of adult subjects collected in laboratory, media, or general in-the-wild contexts. Some datasets related to learning environments have also been introduced, including *DAiSEE* [26] and the *EmotiW engagement challenge datasets* [17]. However, these datasets predominantly involve adult learners, such as undergraduate or graduate students, rather than school-aged children. As a result, existing FER models are typically trained and evaluated on adult facial expressions and may not generalize well to K–12 classroom environments, where facial morphology, expressions, and interaction contexts differ substantially. This gap motivates the need to evaluate how current emotion recognition models perform on data collected from school-aged students in real classroom learning environments.

Bias, Fairness, and Robustness in Facial Emotion Recognition: Bias in computer vision systems has been widely documented and is often attributed to dataset imbalance, demographic underrepresentation, and domain shifts between training and deployment environments [20, 36, 40]. In facial emotion recognition (FER), disparities in prediction performance have been observed across demographic attributes such as skin tone, age, and gender, particularly when models are trained on datasets with uneven representation. Prior work has explored mitigation strategies including data rebalancing, adversarial debiasing, domain adaptation, and data augmentation techniques to improve fairness and robustness in FER models. In addition to demographic bias, studies have shown that FER performance is sensitive to image-level perturbations such as variations in camera angle, lighting, occlusion, and image resolution, which can degrade facial feature extraction and affect valence–arousal estimation [5, 21, 48, 51].

While these issues have been studied extensively in computer vision benchmarks, their impact in educational environments remains relatively underexplored. Most FER evaluations rely on benchmark datasets collected in laboratory or general "in-the-wild" settings, and it remains unclear how these models behave when applied to real K–12 classroom environments with diverse learning activities. Moreover, classroom settings introduce recurring visual conditions—such as non-frontal camera viewpoints, fluctuating lighting, varying webcam resolution, and diverse student populations—that differ substantially from benchmark datasets. In this work, we focus on these commonly observed classroom perturbations to evaluate how state-of-the-art FER models behave under realistic learning conditions.

Emotion Recognition in Educational and Classroom Settings: Modern learning environments include a wide range of platforms such as intelligent tutoring systems, simulation-based learning environments, open-ended learning environments, game-based learning environments, collaborative learning settings, and embodied or immersive learning environments [4, 16, 25, 32]. These modalities differ substantially in interaction structure and physical dynamics, which directly influence the visual characteristics of recorded facial data.

In particular, collaborative and embodied learning environments introduce interaction patterns that differ from individual computer-based learning. Collaborative activities and socially shared regulation of learning can produce a broad spectrum of affective expressions that are not easily captured by basic emotion categories but can instead be represented within a continuous valence–arousal space [18, 28]. Embodied activities additionally involve physical movement and spatial interaction with peers and artifacts [16, 35]. These conditions introduce variations in camera viewpoint, head pose, facial visibility, occlusion, and image resolution as students move, interact with peers, or engage with shared artifacts. Such variations create conditions under which facial emotion recognition models may produce unstable predictions [29, 34]. Evaluating FER models across representative classroom learning environments therefore provides insight into how prediction errors or potential biases emerge under realistic classroom conditions.

3. Learning Environment and Dataset

To evaluate emotion recognition models in real classroom data, we analyze video data collected from three learning environments: (1) a narrative-centered learning environment, (2) an embodied learning environment, and (3) a collaborative computational modeling environment. These environments differ in interaction structure and student activity patterns, providing varied conditions for evaluating facial emotion recognition models.

Data were collected from three independent classroom studies involving middle- and high-school students in the United States. Participants across the three environments were distinct, meaning that no student appears in more than one dataset. Recordings were captured using a combination of webcam and multi-camera classroom setups. Table 1 summarizes the environments, participants, and recording configurations.

In total, the dataset includes recordings from 69 students aged 11–18 years (51% female, 49% male). All data collection procedures received Institutional Review Board (IRB) approval, and informed consent was obtained from students and their guardians prior to participation.

Narrative-centered learning environment. The first dataset was collected from EcoJourneys [38], a narrative-centered, game-based collaborative learning environment designed for middle school science learning. Students work

Table 1. Overview of learning environments and recording setups used in the study.

Environment	Students	Group Size	Camera Setup	Session Length
Narrative-centered (EcoJourneys)	25	3–4	Laptop webcam	45 min × 2 sessions
Embodied learning (GEM-STEP)	20	4–6	4 classroom cameras	25 min × 6 sessions
Collaborative modeling (C2STEM)	24	2	Laptop webcam	90 min × 4 sessions

in groups of three to four to investigate the cause of illness affecting tilapia fish at a local aquaculture farm. The activity is organized into sequential quests requiring students to gather data, analyze environmental conditions, and construct explanatory models. Facial video was captured using laptop webcams during gameplay (Figure 1). Students worked on individual laptops while collaborating with peers, occasionally turning toward group members or away from the screen, producing variations in head pose and facial orientation.

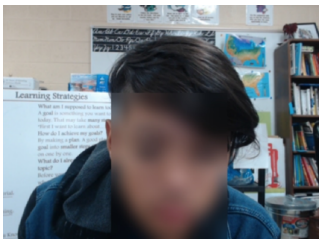

Figure 1. Example frame from the narrative-centered learning environment (EcoJourneys).

Embodied learning environment. The second dataset was collected from GEM-STEP [24], an embodied learning environment in which students enact scientific processes such as molecular motion or photosynthesis. Students move within a classroom space while interacting with projected simulations. Four cameras positioned around the classroom recorded student movement and facial activity, introducing variations in viewpoint and visibility (Figure 2). Because students move within the classroom space during the activity, the recordings contain substantial variations in camera viewpoint, facial visibility, and distance from the camera.

Collaborative computational modeling environment. The third dataset was collected from a collaborative computational modeling activity where high school students worked in pairs to construct simulations of physical motion using C2STEM [15]. Students built and evaluated models representing variables such as velocity and acceleration. Video recordings captured student interactions while they collaborated and interacted with shared displays (Figure 3). Students interacted primarily through a shared display while collaborating with peers, resulting in frequent head-pose changes and partial facial visibility when students turned toward each other or the shared screen.

Figure 2. Example frame from the embodied learning environment (GEM-STEP).

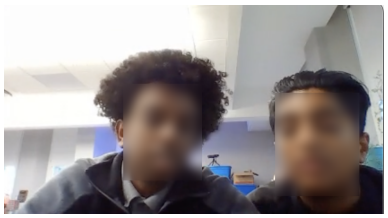

Figure 3. Example frame from the collaborative computational modeling environment (C2STEM).

4. Methodology

Emotion Recognition Models: Three facial emotion recognition systems were evaluated: EmoNet, HSEmotion, and iMotions (version 10.1). These systems were selected because they provide predictions in the valence–arousal affective space, enabling analysis of affective states beyond discrete emotion categories.

EmoNet is a deep neural network based on the EmoFAN architecture, which integrates a Face Alignment Network (FAN) with emotion prediction layers to estimate facial landmarks, categorical emotions, and continuous affect dimensions. In this study, the pretrained EmoNet model was used to obtain continuous valence–arousal predictions from facial images. For HSEmotion, we used the pretrained $enet\_b0\_8\_va\_mtl$ model, an EfficientNet-B0–based architecture designed for multi-task facial emotion recognition including valence and arousal prediction.

This configuration was selected because it is reported in the original implementation as a high-performing model for affect prediction. The commercial facial emotion recogni-

tion platform iMotions (version 10.1) was included as an additional baseline system and used in inference without any model modification.

Face Detection, Sampling, and Preprocessing: All classroom videos were first processed with MTCNN for face detection. This detector was selected because the dataset includes both webcam-based and multi-person classroom recordings, including embodied activities with substantial movement and partial occlusion. Detected facial regions were cropped and used as input to the emotion recognition models after resizing to the required input resolution.

For HSEmotion, facial crops were resized to 224 × 224 pixels, consistent with the EfficientNet-B0 input size used in the official implementation. For EmoNet, facial crops were resized to 256 × 256 pixels, following the preprocessing configuration used in the EmoNet implementation. No modifications were made to the default inference settings of the pretrained models.

To reduce redundant predictions from temporally adjacent frames, we applied temporally sparse frame sampling and extracted one frame per minute for analysis. This approach reflects the focus of the study on affective states in learning contexts, which typically evolve over longer intervals than instantaneous facial expressions.

Bias Mitigation Strategies: To examine whether simple model adaptations can reduce performance disparities in classroom data, we evaluated two mitigation strategies: targeted data augmentation and partial model fine-tuning.

Data augmentation: Augmentation was applied to detected facial crops in the training set to simulate visual conditions commonly observed in classroom recordings. Three transformations were used: small in-plane rotations and mild affine transformations to simulate camera-angle and head-pose variation; brightness and contrast jitter to represent classroom lighting changes; random downsampling followed by resizing to the model input resolution to simulate reduced image resolution. Augmentations were applied only to training images to maintain a clean evaluation of model robustness and fairness. Validation and test samples were kept unmodified.

Fine-tuning strategy: To adapt pretrained models to classroom data while preserving learned facial representations, we performed partial fine-tuning. The majority of the pretrained backbone was frozen, while the final prediction layers (and the last feature block when applicable) were updated. Training used the Adam optimizer for 10 epochs with early stopping based on validation error.

Fine-tuning used the training portion of the classroom dataset, pooling samples from the available learning environments within each fold. In student-independent cross-validation, students appearing in the training fold were excluded from validation and test sets [6]. For environment-held-out evaluation, fine-tuning was performed only on source-environment data, excluding the target environment from training.

Fairness evaluation: To measure demographic disparities in prediction error, we computed group-wise mean absolute error (MAE) for each Fitzpatrick skin tone group (Types I-VI). Fairness disparity was summarized using the Type VI minus Type I MAE gap, which captures the largest observed skin-tone-related error spread in the dataset.

Evaluation Metrics: To quantify sensitivity to visual perturbations, we compute *Bias Magnitude (BM)* as the absolute difference between predictions for paired instances:

$$BM = |\hat{y}_{\text{ref}} - \hat{y}_{\text{perturbed}}| \tag{1}$$

where $\hat{y}_{\text{ref}}$ and $\hat{y}_{\text{perturbed}}$ denote valence or arousal predictions for the reference and perturbed frames. Paired instances were constructed to preserve the underlying facial expression while varying a single visual factor. For camera angle, lighting, and resolution, paired samples were obtained from naturally occurring variations in classroom recordings. For apparent skin tone, paired samples were generated by applying a computer-vision-based facial appearance transformation to the same facial crop [14, 42]. This transformation was intended to alter skin-color appearance while preserving facial geometry and expression, thereby reducing variation in identity and expression across the pair.

In addition to robustness analysis, we report prediction error using mean absolute error (MAE) for valence and arousal. Ground-truth affect values were obtained from human-annotated classroom data, where two annotators independently assigned valence–arousal labels to sampled frames. Only frames with annotator agreement were retained for analysis.

We also report Expected Calibration Error (ECE) for valence predictions, computed by binning predicted values and measuring the deviation between predicted affect and observed annotation averages within each bin.

Human Validation of Expression Consistency: Across the three learning environments, the dataset contains approximately 55,000 sampled frames. From these, paired perturbation instances were constructed and independently reviewed by two annotators to ensure that the underlying facial expression remained consistent across perturbations. Only pairs with annotator agreement (Cohen's $\kappa = 0.81$) were retained, resulting in 1800 (angle), 1680 (lighting), 1550 (resolution), and 2400 (skin tone) validated pairs used in the analysis. This validation step was also used to confirm that skin-tone transformations preserved the underlying facial expression.

Table 2. Baseline emotion prediction performance across evaluated models.

Model	MAE (Valence)	MAE (Arousal)
EmoNet	0.21	0.23
HSEmotion	0.24	0.26
iMotions (v10.1)	0.28	0.30

5. Results

Baseline Emotion Prediction Performance: We first evaluate the baseline performance of three facial emotion recognition systems—EmoNet, HSEmotion, and iMotions (v10.1)—on classroom video data without any mitigation or model adaptation. All models produce continuous valence–arousal predictions, enabling direct comparison of prediction errors across the dataset.

Table 2 reports the overall mean absolute error (MAE) for valence and arousal predictions computed across the analyzed classroom frames. Among the evaluated systems, EmoNet achieves the lowest prediction error, with MAE values of 0.21 for valence and 0.23 for arousal. HSEmotion exhibits slightly higher errors (0.24 / 0.26), while the commercial system iMotions produces the largest prediction errors (0.28 / 0.30). These results indicate that models explicitly designed for dimensional affect prediction transfer more effectively to classroom data than systems trained using proprietary pipelines. However, even the best-performing model exhibits nontrivial prediction error, suggesting that emotion recognition models trained primarily on web-based or laboratory datasets do not fully generalize to the visual conditions present in real classroom recordings.

Robustness to Classroom Perturbations: To examine the sensitivity of emotion recognition models to visual variations commonly observed in classroom recordings, we evaluated prediction stability under four perturbation conditions: camera angle, lighting variation, resolution degradation, and skin tone differences. Robustness was measured using the Bias Magnitude (BM) metric, which quantifies the average prediction shift between paired reference and perturbed instances representing the same student expression.

Table 3 summarizes the BM values for valence and arousal predictions across the evaluated models. Across all perturbation conditions, prediction shifts are consistently observed for each system. For EmoNet, BM values range between 0.20–0.24 for valence and 0.23–0.27 for arousal depending on the perturbation type. HSEmotion exhibits larger shifts across most conditions, with BM values between 0.25–0.29 for valence and 0.26–0.31 for arousal. The commercial system iMotions shows the largest sensitivity to perturbations, with BM values exceeding 0.30 for both affect dimensions across several conditions.

Table 3. Robustness via Bias Magnitude (BM): average prediction shift under perturbations. Higher values indicate greater sensitivity. Models: EN (EmoNet), HS (HSEmotion), IM (iMotions).

M	P	BM_Va	BM_Ar	d	p
EN	Angle	0.24	0.23	1.00	1×10^{-12}
EN	Lighting	0.20	0.27	0.93	1×10^{-13}
EN	Resolution	0.21	0.23	0.66	5×10^{-15}
EN	Skin Tone	0.23	0.26	0.71	5×10^{-8}
HS	Angle	0.28	0.26	0.98	4×10^{-12}
HS	Lighting	0.29	0.28	1.08	1×10^{-14}
HS	Resolution	0.29	0.27	0.73	3×10^{-6}
HS	Skin Tone	0.25	0.31	0.92	2×10^{-15}
IM	Angle	0.31	0.30	1.15	2×10^{-14}
IM	Lighting	0.34	0.35	1.20	5×10^{-16}
IM	Resolution	0.30	0.32	0.88	7×10^{-9}
IM	Skin Tone	0.33	0.36	1.12	3×10^{-15}

Among the evaluated perturbations, lighting variation and camera angle changes produce the largest prediction shifts across models. Resolution degradation produces moderate shifts, while skin tone variation also introduces measurable changes in prediction outputs. These results indicate that facial emotion recognition systems trained on web-based datasets exhibit substantial sensitivity to visual conditions commonly present in classroom recordings. The magnitude and consistency of these shifts suggest that model predictions are influenced not only by the underlying facial expression but also by environmental factors such as camera angle and lighting condition. This observation motivates further analysis of fairness disparities and learning-environment effects presented in the following sections.

Fairness Across Skin Tone Groups: To evaluate demographic disparities in emotion prediction performance, we analyze model errors across Fitzpatrick skin tone groups (Types I–VI). Fitzpatrick labels were assigned by two annotators based on visual inspection of reference facial images using the Fitzpatrick scale, with disagreements resolved through discussion. For each group, we compute mean absolute error (MAE) for valence and arousal predictions. Fairness disparity is summarized using the MAE gap between Type VI and Type I, representing the largest observed error difference across skin tone groups.

Table 4 reports group-wise prediction errors for each evaluated model. Across all systems, prediction error increases progressively from lighter skin tones (Type I) to darker skin tones (Type VI). For EmoNet, valence MAE increases from 0.179 (Type I) to 0.230 (Type VI), producing a fairness gap of 0.051. A similar pattern is observed for HSEmotion, where MAE increases from 0.211 to 0.262, yielding an identical fairness gap of 0.051. The commercial

Table 4. Fairness across Fitzpatrick groups (I–VI). Metrics: MAE for valence (Va) and arousal (Ar), and ECE for valence; Fitzpatrick Type (FT).

Model	FT	MAE_Va	MAE_Ar	ECE_Va
EmoNet	I	0.179	0.198	0.051
EmoNet	II	0.192	0.210	0.058
EmoNet	III	0.198	0.221	0.071
EmoNet	IV	0.211	0.229	0.080
EmoNet	V	0.220	0.240	0.092
EmoNet	VI	0.230	0.251	0.100
EmoNet	Gap (VI–I)	0.051	–	–
HSEmotion	I	0.211	0.231	0.047
HSEmotion	II	0.219	0.239	0.059
HSEmotion	III	0.229	0.253	0.068
HSEmotion	IV	0.242	0.257	0.079
HSEmotion	V	0.250	0.271	0.091
HSEmotion	VI	0.262	0.279	0.099
HSEmotion	Gap (VI–I)	0.051	–	–
iMotions	I	0.240	0.260	0.082
iMotions	II	0.255	0.272	0.094
iMotions	III	0.268	0.289	0.108
iMotions	IV	0.281	0.300	0.120
iMotions	V	0.295	0.318	0.137
iMotions	VI	0.312	0.337	0.155
iMotions	Gap (VI–I)	0.072	–	–

system iMotions exhibits the largest disparity, with MAE increasing from 0.240 (Type I) to 0.312 (Type VI), corresponding to a fairness gap of 0.072.

In addition to prediction error, calibration differences are also observed across skin tone groups. Expected Calibration Error (ECE) increases monotonically across Fitzpatrick groups for all evaluated systems, indicating that prediction confidence becomes less reliable for darker skin tones.

These results show that demographic disparities persist across both research and commercial emotion recognition systems when applied to classroom data. The consistent increase in prediction error across skin tone groups suggests that models trained on web-based datasets may not generalize equally across diverse student populations.

Learning Environment Effects: To examine whether classroom activity structure influences prediction stability, we analyze Bias Magnitude (BM) across the three learning environments used in this study. Table 5 reports the BM values across perturbation types for each learning environment. Across all evaluated systems, embodied learning environments produce the largest prediction shifts. For example, EmoNet exhibits a BM of 0.24 for camera angle and 0.28 for skin tone in the embodied setting, compared to 0.21 and 0.21 respectively in the game-based environment. Similar trends are observed for HSEmotion and iMotions, where prediction shifts increase across most perturbation types in the embodied environment.

Table 5. Bias Magnitude (BM) across environments. Higher values indicate greater sensitivity. Models: EN (EmoNet), HS (HSEmotion), IM (iMotions). Environments: Gb (Game-based), Eb (Embodied), Co (Collaborative).

M	Env	BM_Ang	BM_Light	BM_Res	BM_ST
EN	Gb	0.21	0.22	0.25	0.21
EN	Eb	0.24	0.25	0.26	0.28
EN	Co	0.23	0.24	0.24	0.22
HS	Gb	0.30	0.25	0.24	0.25
HS	Eb	0.35	0.33	0.30	0.33
HS	Co	0.25	0.29	0.27	0.27
IM	Gb	0.34	0.36	0.32	0.31
IM	Eb	0.39	0.41	0.38	0.37
IM	Co	0.32	0.33	0.30	0.29

Table 6. Mixed-effects regression results estimating the influence of perturbation factors on prediction bias. Student identity is modeled as a random effect. O: Outcome; Va: Valence; Ar: Arousal.

O	Fixed Effect	Beta (β)	SE	t/z	p-value
Va	Angle	0.08	0.02	3.4	7×10^{-4}
Va	Lighting	0.06	0.03	5.6	2×10^{-8}
Va	Resolution	0.07	0.03	5.1	3×10^{-7}
Va	Skin Tone	0.09	0.02	2.4	1×10^{-2}
Va	Embodied	0.10	0.01	3.2	1×10^{-3}
Ar	Angle	0.07	0.03	5.0	6×10^{-7}
Ar	Lighting	0.09	0.02	4.3	2×10^{-5}
Ar	Resolution	0.06	0.02	2.1	3×10^{-2}
Ar	Skin Tone	0.08	0.01	2.4	1×10^{-2}
Ar	Embodied	0.07	0.03	4.3	2×10^{-5}

These results indicate that activity structures involving student movement and multi-camera viewpoints amplify perturbation sensitivity in emotion recognition models. In contrast, game-based environments with primarily frontal webcam recordings produce smaller prediction shifts, while collaborative computational modeling settings fall between these two conditions. To further quantify these effects while accounting for repeated observations from the same students, we fit a mixed-effects regression model with student identity as a random effect and perturbation factors as fixed effects. Table 6 reports the estimated coefficients for valence and arousal outcomes.

Across both affect dimensions, camera angle, lighting variation, resolution degradation, skin tone, and embodied learning modality emerge as statistically significant predic-

Table 7. Effect of mitigation strategies on prediction accuracy, robustness, and fairness. BM denotes average Bias Magnitude across perturbations. Fairness gap is computed as MAE(Type VI) - MAE(Type I). Models: EN (EmoNet), HS (HSEmotion), IM (iMotions). Va: Valence; Ar: Arousal; FG: Fairness Gap

Model	Setting	MAE (Va)	MAE (Ar)	Avg BM (Va)	Avg BM (Ar)	FG (Va)	FG (Ar)	ECE (Va)
EM	Baseline	0.21	0.23	0.22	0.25	0.06	0.07	0.09
EM	+ Augmentation	0.20	0.22	0.19	0.21	0.04	0.05	0.08
EM	+ Fine-tuning	0.19	0.21	0.17	0.18	0.03	0.04	0.07
HS	Baseline	0.24	0.26	0.28	0.28	0.06	0.07	0.09
HS	+ Augmentation	0.23	0.25	0.23	0.25	0.04	0.05	0.08
HS	+ Fine-tuning	0.22	0.24	0.20	0.22	0.03	0.04	0.07
iM	Baseline	0.28	0.30	0.32	0.33	0.08	0.09	0.14

tors of prediction bias. For example, camera angle perturbations increase valence prediction bias by $\beta = 0.08$, while embodied learning modality contributes an additional $\beta = 0.10$ shift. Similar effects are observed for arousal predictions. These results confirm that the observed prediction shifts are systematic and not attributable to random variation. Together, these findings show that classroom activity structure and recording conditions interact with visual perturbations to influence emotion prediction stability, highlighting the importance of evaluating emotion recognition systems under real classroom environments.

Mitigation Effects: To examine whether simple adaptation strategies can improve prediction stability and reduce demographic disparities, we evaluate two mitigation approaches: targeted data augmentation and partial model fine-tuning.

Table 7 summarizes the effect of these strategies on overall prediction error, perturbation sensitivity, and fairness disparity. For EmoNet, augmentation reduces the valence MAE from 0.21 to 0.20 and decreases the average Bias Magnitude (BM) from 0.22 to 0.19. Fine-tuning improves performance, reducing MAE to 0.19 and BM to 0.17. For HSEmotion, MAE decreases from 0.24 to 0.23 and then to 0.22, while Avg BM decreases from 0.28 to 0.23 and further to 0.20. Mitigation reduces fairness disparities across skin tone groups. For both EmoNet and HSEmotion, the fairness gap in valence predictions decreases from approximately 0.06 to 0.03 after fine-tuning. In addition, calibration improves, with Expected Calibration Error (ECE) decreasing from 0.09 to 0.07 for the evaluated models. The commercial system iMotions was evaluated only in the baseline configuration because model parameters are not accessible for adaptation. As a result, its performance and fairness gap remain unchanged. These results indicate that lightweight adaptation strategies can reduce both prediction error and fairness disparities when emotion recognition models are applied to classroom data, without requiring substantial architectural modification or large-scale retraining.

6. Conclusion

This paper presents a structured evaluation of emotion recognition models in classroom settings, focusing on robustness and fairness under realistic visual conditions. Using paired perturbation analysis, we show that camera angle, lighting variation, image resolution, and skin tone introduce systematic shifts in valence–arousal predictions, with additional amplification in embodied and collaborative learning environments. These results demonstrate that models validated on benchmark datasets do not reliably generalize to classroom data and may produce biased or unstable predictions in deployment scenarios. The goal of this work is not to optimize model performance but to identify failure modes in real-world settings; accordingly, mitigation strategies were intentionally limited to lightweight augmentation and partial fine-tuning. the evaluation is constrained by the need for manual validation of paired instances, which limits the number of analyzed frames but ensures reliability of the robustness analysis.

Future work should explore stronger domain adaptation approaches, multimodal integration, and the development of large-scale K–12 classroom datasets. Additional directions include scaling validation protocols, extending evaluation beyond paired perturbations, and designing models that explicitly account for interaction dynamics and visual variability in learning environments.

Acknowledgements

This research and development work was partially funded by the U.S. Army CCDC Soldier Center Award (#W912CG2220001) and supported by the National Science Foundation AI Institute Grant #DRL-2112635. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the Department of the Army, the Department of Defense, or the U.S. Government.

References

[1] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023. 1

[2] Celestine E Akpanoko, Gautam Biswas, et al. The interplay of affective states and cognitive processes in an open-ended learning environment: A case study. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 873-880*. International Society of the Learning Sciences, 2024. 2

[3] Celestine E. Akpanoko, Ashwin T. S., Grayson Cordell, and Gautam Biswas. Investigating the relations between studentsâ€™ affective states and the coherence in their activities in open-ended learning environments. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 511–517, Atlanta, Georgia, USA, 2024. International Educational Data Mining Society. 2

[4] Vincent Aleven, Ido Roll, Bruce McLaren, and Kenneth Koedinger. Help seeking and help design in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26:1–26, 2016. 3

[5] TS Ashwin and Gautam Biswas. Identifying and mitigating algorithmic bias in student emotional analysis. In *International Conference on Artificial Intelligence in Education*, pages 89–103. Springer, 2024. 2, 3

[6] TS Ashwin and Ram Mohana Reddy Guddeti. Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. *IEEE Access*, 7:150693–150709, 2019. 5

[7] TS Ashwin and Ram Mohana Reddy Guddeti. Affective database for e-learning and classroom environments using indian students' faces, hand gestures and body postures. *Future Generation Computer Systems*, 108:334–348, 2020. 1

[8] TS Ashwin, Caitlin Snyder, Srigowri Mayasandra Prasanna, Naveeduddin Mohammed, and Gautam Biswas. Using markov chain analysis to study the relations between emotions, cognitive actions, and performance in collaborative learning. In *18th International Conference on Computer-Supported Collaborative Learning (CSCL) 2025-CSCL Proceedings*, 2025. 1

[9] Roger Azevedo, Michelle Taub, and Nicholas V Mudrick. Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In *Handbook of self-regulation of learning and performance*, pages 254–270. Routledge, 2017. 1

[10] Ryan S Baker, Jaclyn L Ocumpaugh, and JMAL Andres. Bromp quantitative field observations: A review. *Learning Science: Theory, Research, and Practice. New York, NY: McGraw-Hill*, 2020. 1

[11] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018. 2

[12] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1): 18–37, 2010. 1, 2

[13] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582, 2022. 1

[14] Ying-Cong Chen, Xiaohui Shen, Zhe Lin, Xin Lu, I Pao, Jiaya Jia, et al. Semantic component decomposition for face attribute manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9859–9867, 2019. 5

[15] Clayton Cohn, Caitlin Snyder, Joyce Horn Fonteles, Ashwin TS, Justin Montenegro, and Gautam Biswas. A multimodal approach to support teacher, researcher and ai collaboration in stem+ c learning environments. *British Journal of Educational Technology*, 2024. 4

[16] Chris Dede. Immersive interfaces for engagement and learning. *Science*, 323(5910):66–69, 2009. 3

[17] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, page 784–789, New York, NY, USA, 2020. Association for Computing Machinery. 3

[18] Sidney D'Mello and Arthur Graesser. Autotutor and affective learning. *International Journal of Artificial Intelligence in Education*, 22:1–26, 2013. 3

[19] Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012. 2

[20] Wala Elsharif, Mahmood Alzubaidi, and Marco Agus. Cultural bias in text-to-image models: a systematic review of bias identification, evaluation and mitigation strategies. *IEEE Access*, 2025. 3

[21] Alex Fan, Xingshuo Xiao, and Peter Washington. Addressing racial bias in facial emotion recognition. *arXiv preprint arXiv:2308.04674*, 2023. 3

[22] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 2

[23] Joyce Fonteles, Eduardo Davalos, T. S. Ashwin, Yike Zhang, Mengxi Zhou, Efrat Ayalon, Alicia Lane, Selena Steinberg, Gabriella Anton, Joshua Danish, Noel Enyedy, and Gautam Biswas. A first step in using machine learning methods to enhance interaction analysis for embodied learning environments. In *Artificial Intelligence in Education*, pages 3–16, Cham, 2024. Springer Nature Switzerland. 2

[24] Joyce Horn Fonteles, Clayton Cohn, Efrat Ayalon, Mengxi Zhou, Ashwin TS, Eduardo Davalos, Zhijian Li, Surya Rayala, Divya Mereddy, Austin Coursey, Shruti Jain, Yike Zhang, Noel Enyedy, Joshua Danish, and Gautam Biswas. Analyzing Embodied Learning in Classroom Settings: A Human-in-the-loop AI Approach for Multimodal Learning Analytics. *Learning and Instruction*, 2026. 2, 4

[25] Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. Autotutor: A tutor with dialogue in natu-

ral language. *Behavior Research Methods*, 37(2):180–192, 2005. 3

[26] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild, 2022. 1, 3

[27] Maryam Imani and Gholam Ali Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019. 1

[28] Sanna Järvelä and Allyson F Hadwin. Socially shared regulation of learning: A review. *Educational Psychologist*, 48 (1):25–39, 2017. 3

[29] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the wild: Aff-wild2. *IEEE Transactions on Affective Computing*, 2021. 3

[30] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. *IEEE Transactions on Affective Computing*, 2021. 3

[31] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 3

[32] James C Lester, Hiller A Spires, John L Nietfeld, Wookhee Min, Bradford W Mott, and Erica V Lobene. Narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 24:1–21, 2014. 3

[33] Ryan Li, Tudur Sadashiva_Ashwin, Divya Mereddy, and Gautam Biswas. Do categorical emotions have intensity? modeling emotional intensity transitions by task, peer, and environment causes. *International Conference on Computers in Education*, 2025. 1

[34] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2019. 3

[35] Robb Lindgren and Mina Johnson-Glenberg. Embodied interaction for learning. *Educational Psychologist*, 48(3):1–15, 2013. 3

[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021. 2, 3

[37] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 3

[38] Bradford W Mott, Robert G Taylor, Seung Y Lee, Jonathan P Rowe, Asmalina Saleh, Krista D Glazewski, Cindy E Hmelo-Silver, and James C Lester. Designing and developing interactive narratives for collaborative problem-based learning. In *International Conference on Interactive Digital Storytelling*, pages 86–100. Springer, 2019. 3

[39] Ernesto Panadero and Sanna Järvelä. Socially shared regulation of learning: A review. *European Psychologist*, 2015. 1

[40] Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, 57(6):1–40, 2025. 3

[41] Amjad Rehman, Muhammad Mujahid, Alex Elyassih, Bayan AlGhofaily, and Ali Saeed. Comprehensive review and analysis on facial emotion recognition: Performance insights into deep and traditional learning with current updates and challenges. *Computers, Materials, & Continua*, 82(1):41, 2025. 1, 2

[42] Xingyu Ren, Jiankang Deng, Chao Ma, Yichao Yan, and Xiaokang Yang. Improving fairness in facial albedo estimation via visual-textual cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2023. 5

[43] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 2

[44] Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *International Conference on Machine Learning*, pages 30119–30129. PMLR, 2023. 2

[45] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022. 1, 2

[46] Michelle Taub, Nicholas V Mudrick, Ramkumar Rajendran, Yi Dong, Gautam Biswas, and Roger Azevedo. How are students' emotions associated with the accuracy of their note taking and summarizing during learning with itss? In *Intelligent Tutoring Systems: 14th International Conference, ITS 2018, Montreal, QC, Canada, June 11–15, 2018, Proceedings 14*, pages 233–242. Springer, 2018. 1, 2

[47] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021. 1, 2

[48] Udayini VEDANTHAM, Nithish CHOUTI, Avaneesh SUNDERARAJAN, Ashwin TUDUR SADASHIVA, Manjunath K VANAHALLI, Gautam Biswas, et al. Mapping bias: Visualizing valence-arousal distributions to reveal affective gaps in face datasets. In *International Conference on Computers in Education*, 2025. 3

[49] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022. 2

[50] Zi Yang Wong and Gregory Arief D Liem. Student engagement: Current state of the construct, conceptual refinement, and future research directions. *Educational Psychology Review*, 34(1):107–138, 2022. 1

[51] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 506–523. Springer, 2020. 3