Identifying and Mitigating Algorithmic Bias in Student Emotional Analysis

T. S. Ashwin(✉) and Gautam Biswas(✉)

Vanderbilt University, Nashville, TN, USA
ashwindixit9@gmail.com, gautam.biswas@vandderbilt.edu

Abstract. Algorithmic bias in educational environments has garnered increasing scrutiny, with numerous studies highlighting its significant impacts. This research contributes to the field by investigating algorithmic biases, i.e., selection, label, and data biases in the assessment of students' affective states through video analysis in two educational settings: (1) an open-ended science learning environment and (2) an embodied learning context, involving 41 and 12 students, respectively. Utilizing the advanced High-speed emotion recognition library (HSEmotion) and Multi-task Cascaded Convolutional Networks (MTCNN), and contrasting these with the commercially available iMotions platform, our study delves into biases in these systems. We incorporate real student data to better represent classroom demographics. Our findings not only corroborate the existence of algorithmic bias in detecting student emotions but also highlight successful bias mitigation strategies. The research advances the development of equitable educational technologies and supports the emotional well-being of students by demonstrating that targeted interventions can effectively diminish biases.

Keywords: Algorithmic Bias · Classification Bias · Open-ended Learning Environment · Embodied Learning · Bias Mitigation Techniques · Emotion Recognition Bias · Affective Computing · Facial Expressions

1 Introduction

In the digital era, where algorithms govern a spectrum of decisions from the mundane to the pivotal their neutrality and objectivity, especially in the machine learning and artificial intelligence domains, are being critically reevaluated [14]. Algorithmic bias, or the systematic skew that leads to unfair outcomes by favoring certain groups over others, has emerged as a significant concern. This concern is particularly pronounced in education applications, where such biases can amplify existing inequalities and compromise the integrity of educational fairness [3].

Recognizing the multifaceted nature of algorithmic bias is crucial. It manifests through various stages of the machine learning life cycle, encompassing historical, representational, measurement, aggregation, learning, evaluation, and

A. M. Olney et al. (Eds.): AIED 2024, LNAI 14830, pp. 89–103, 2024.
https://doi.org/10.1007/978-3-031-64299-9_7

deployment biases [18]. Each category, with its unique attributes, contributes to the complexities of ensuring equitable AI systems. Historical biases reflect entrenched societal inequities, while representation biases arise from data that fails to encapsulate the diversity of the target population. Measurement biases skew reality through oversimplified proxies, and aggregation biases ignore the nuances of diverse data subsets. Learning biases occur due to model choices that neglect fairness, evaluation biases emerge from unrepresentative benchmark datasets, and deployment biases arise when models are misapplied in practice. These biases present a formidable challenge in education, necessitating a nuanced understanding and a proactive stance to mitigate their multifarious impacts and steer towards a more equitable educational landscape [14].

Several research studies on algorithmic bias exist within the education domain. They have tried to understand the bias in applications such as dropout prediction and automated essay scoring, with a notable focus on racial, ethnic, and gender disparities, primarily within the United States [1]. While these studies have illuminated the differential effectiveness of algorithms across diverse student demographics, the exploration into vision data and affective state recognition, like student emotions and engagement, is relatively nascent. Noteworthy is the research indicating performance disparities in affect detection models that use the log data between rural and urban students, highlighting the need for contextually tailored approaches [15]. In the emotion recognition domain, efforts to tackle algorithmic bias are increasing, yet a comprehensive analysis of bias and the development of mitigation strategies are still in their infancy. Prevailing studies predominantly rely on preprocessed databases, characterized by predominantly frontal facial images captured in controlled environments, rather than the natural, varied settings of real life scenarios [11]. This approach tends to emphasize overtly expressive faces, usually associated with distinct actions or emotions, and often fails to capture the subtleties and complexities of spontaneous emotional expressions. Furthermore, these studies frequently utilize cloud-based algorithms or advanced techniques such as Generative Adversarial Networks (GANs) for data augmentation or bias mitigation, concentrating on specific demographic disparities or attribute biases [9]. Nevertheless, a significant gap remains: the educational domain, rich with context-specific nuances of facial expressions influenced by the learning environment, cultural background, and individual student experiences, is largely underrepresented in these datasets and methodologies. Furthermore, existing research often restricts its focus to primary emotions – happiness, sadness, anger, surprise, fear, and disgust-while neglecting academic emotions such as confusion, frustration, and boredom. This limitation represents a significant research gap in the study of bias within vision data-driven student emotion recognition, thereby hindering a comprehensive understanding of the interplay between cognition, emotion, and learning.

Addressing this gap, our study collects data from two distinct learning environments: an open-ended learning environment named Betty's Brain [4], and an embodied learning environment named GEM-STEP (Generalized Embodied Modeling - Science through Technology Enhanced Play) [6]. We broaden the

scope of bias analysis to encompass both basic and educational emotions, with a primary focus on *confusion*, thereby enriching our analysis of students' affective states. This paper emphasizes confusion as a pivotal educational emotion characterized by a state of disequilibrium, arising from cognitive conflict when learners encounter challenges in assimilating new information with prior knowledge [7]. Confusion, indicative of student engagement, presents an opportune moment for educators to provide targeted support [17]. Recognizing its critical role in learning, our investigation concentrates on understanding algorithmic biases in confusion in this paper.

To examine bias, a validated emotion recognition method is essential. AffectNet, a vast dataset of facial expressions, and the AFFDEX algorithm, integrated into the iMotions platform, are both trained on this extensive database, mitigating cultural and human labeling biases. Consequently, we selected iMotions for benchmarking because it is trained on multiple comprehensive datasets and offers detailed emotion and action unit analysis. In contrast, other methods that use the same dataset often lack such granularity, omitting action units or education-specific emotions. Furthermore, our study employs two novel techniques to address bias, contributing to the advancement of more equitable and precise emotion recognition within educational contexts.

The Key contributions of the paper are:

1. *Algorithmic Bias Analysis:* The study analyzes selection, label, and data biases in emotion recognition within educational settings.
2. *Emotions and Integration of Classroom Data:* Real student data from two different learning environments enrich the generalizability of our analyses. Additionally, it considers not only basic emotions but also learning-centered emotions, primarily confusion.
3. *Bias Mitigation Techniques:* The research introduces and applies state-of-the-art methods for identifying and mitigating biases, thereby enhancing the accuracy and fairness of emotion recognition in education.

2 Literature Review

Recent studies in higher education have leveraged log data to predict student dropout and course failure, revealing biases related to gender, ethnicity, race, and socioeconomic factors [5,12,21]. Some studies, such as [13] and [19], have focused on predicting academic achievements and evaluating speech, highlighting language-based and cross-cultural biases. A unique study by [15] explored geographical biases in detecting student affect using log data.

While existing research on algorithmic bias in education emphasizes factors like gender, ethnicity, and academic performance using log, text, or speech data, a notable gap exists regarding biases in vision data. Addressing this gap is essential for a more comprehensive understanding of algorithmic biases in educational contexts. Numerous studies have explored emotions, though not exclusively within the education domain. Many of these investigations have relied on publicly available data to address biases. In [11] study, the focus was on Cross-Database Emotion Classification bias. [20] extended the scope to include age, gender, and racial

bias, and considered Cross-Database Emotion Classification. On the other hand, [9] emphasized racial bias as a primary factor.

Several studies focus on basic emotions, often excluding learning-centered emotions. [20] expanded their dataset to include expressions like smiling. These studies primarily used image-based data and computer vision. [11] analyzed four child-focused datasets. [20] used two adult datasets: RAF-DB, with diverse subjects and conditions, and CelebA, featuring celebrities with annotated facial attributes. [9] employed the CAFE dataset, with diverse children aged 2–8, and AffectNet for adult facial emotion recognition. However, none of these datasets were from the education domain. Most datasets in the education domain include undergraduate or graduate populations [2,10]. This highlights a gap in emotion recognition within the education domain, where identifying and addressing bias remains under-explored. Current methods to address algorithmic emotion recognition from facial expressions are limited, with few focusing on mitigating overgeneralization and emphasizing important features [20].

This study aims to bridge these gaps by analyzing data from open-ended and embodied learning environments to understand basic and learning-centered emotions, and to investigate bias in computer vision data within educational contexts.

3 Learning Environment and Dataset

The data for this study is taken from two distinct learning environments. The first learning environment, Betty's Brain, is an Online Educational Learning Environment (OELE) designed for middle school students [4]. It utilizes a learning-by-teaching approach, and students actively engage in building causal models of scientific processes to teach a virtual character, the Teachable Agent (TA). This interactive platform provides students with resources such as a science book with hypermedia pages, enabling them to acquire knowledge by identifying relevant concepts and establishing causal relations between these concepts to model a scientific process (e.g., climate change, thermoregulation). Students also have opportunities to quiz the TA to check their learning progress, and the results, typically motivate the students to learn more and help improve their TAs performance. A Mentor agent (MA) observes the students' interactions with the TA and intervenes when it believes the student is having difficulties. In addition, students can also query the MA.

OELE Data: The data collection included 41 students, 12 males and 29 females, in the age range of 10–12. Students worked in the Betty's Brain environment for three days, approximately 40 min a day. This produced a total of around 5000 min of screen-recording videos. The video data, captured through iMotions using the laptop's webcam, maintained a resolution of 1092*614 at a frame rate of 30 frames per second. Emotion recognition was systematically performed on each visible face within the image frames to comprehensively analyze emotional states.

In GEM-STEP, the second learning environment, students learn by working in small groups to enact a scientific process (e.g., movement of molecules in solids,

liquids, and gas; or enacting the photosynthesis process). Students physically move around the classroom space, engaging in play-acting scenarios related to scientific concepts and processes, and witness their activities being projected into a computer simulation on a screen in front of the room. The simulated environment includes avatars and entities that represent key information, and direct students' attention to critical aspects of the science content. Overall, this represents a mixed-reality environment that captures students' physical activities and maps them onto the simulation of the scientific process.

Participant details: Data was collected over two days, with two groups each day. On Day 1, Group 1 had 3 girls and 2 boys, while Group 2 had 3 girls and 2 boys. On Day 2, Group 1 included 2 girls and 4 boys, and Group 2 had 2 girls and 3 boys. Each session lasted about 25 min. Videos, captured by four high-resolution cameras at the study area's corners, had a resolution of 1920*1080, processed at 30 frames per second (fps). All four cameras were utilized for emotion recognition, employing face detection and re-identification to track students' emotions as they played and moved in various directions. For both studies, the Institutional Review Board's approval was obtained, and all necessary participant consent procedures and formalities were diligently followed. Sample image screenshot of students from both the learning environment is shown in Fig. 1. The student population distribution was as follows: 60% White, 25% Black, 9% Asian, and 5% Hispanic.

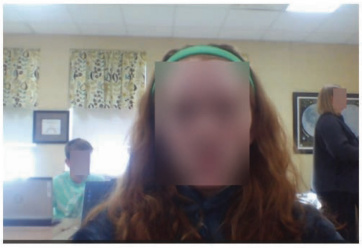
4 Methodology

In our methodology, inspired by Baker and Hawn's framework [3], we navigate the complex landscape of algorithmic bias in education through three key subsections. We begin by focusing on emotion recognition, a critical dimension of bias. Following Baker's progression, we identify unknown bias by scrutinizing misclassifications, moving from a lack of awareness to a comprehensive understanding of bias in specific contexts and among particular individuals.

The second phase involves meticulously examining misclassifications, aligning with Baker's concept of transitioning from *unknown bias to known bias*. This step allows us to pinpoint instances of bias, contributing valuable insights to the discourse on algorithmic fairness in education.

The third and final section addresses bias, where we actively work towards fairness in the algorithm. Drawing on the evolving understanding within the machine learning community, we aim to rectify identified biases, contributing to the broader goal of creating a more equitable learning environment.

In essence, our streamlined methodology systematically progresses from investigating emotion recognition biases and uncovering misclassifications to actively addressing and rectifying bias in pursuit of algorithmic fairness in education, aiming to foster equal opportunities (*fairness to equity*) for all learners.



(a) OELE Environment

(b) Embodied Learning Environment

Fig. 1. Sample image snapshot from learning environments

4.1 Emotion Recognition

Emotion recognition in the embodied learning environment differs from that in the OELE environment. In the case of OELE, we utilized data collected from iMotions, which included image frames captured at 30fps. Additionally, for each student, there was a corresponding CSV file containing information on basic emotions, confusion, and 25 selected action units. The CSV file facilitated a detailed analysis of emotional states. To ensure synchronization, the videos and CSV files were time-aligned using iMotions. As part of our methodology, we also applied HSEmotion to identify and recognize the same set of emotions. This dual approach allowed us to assess and identify potential misclassifications, HSEmotion (High-Speed Face Emotion Recognition) is used to predict the valence and arousal values. The architecture utilized the pre-trained model facial emotion recognition [16].

Confusion Annotation: As HSEmotion lacks native recognition for confusion emotion, we performed annotation and model retraining specifically for this emotion category. Using the Computer Vision Annotation Tool (CVAT), we meticulously annotated images independently by two different annotators. The inter-rater reliability, measured by Cohen's Kappa, was 0.71. For model training, we considered 500 instances of confusion emotion where both annotators completely agreed ($\kappa = 1$). To enrich the dataset, we applied data augmentation, increasing the sample size by 10 times. The resulting training accuracy achieved on HSEmotion was 94.3%.

Emotion Recognition in Embodied Learning Environment. Emotion recognition in an embodied learning environment poses unique challenges due to dynamic student movements and frequent occlusions. Our methodology addresses these challenges in three steps: face recognition using MTCNN, emotion recognition, and student re-identification. We employed MTCNN with fine-tuned thresholds (0.8 for P-Net, R-Net, and O-Net) to improve face detection accuracy across diverse skin tones, expressions, and lighting conditions. This setup successfully detected faces at a rate of 30 frames per second. Following face detection, emotional states were identified using the HSEmotion method

(mentioned in the previous subsection). The detected emotions and face coordinates were recorded in a CSV file for each frame.

Student Re-Identification: While MTCNN and HSEmotion are adept at facial detection and emotion recognition, respectively, they do not inherently perform re-identification-a crucial requirement in our study due to multiple students within the video frames. A robust re-identification strategy was imperative to track each student's emotions over time. Our approach involved the development of a sophisticated tracking algorithm that processed the CSV data obtained post-emotion recognition. This data comprised frame numbers, basic emotion, and face bounding box coordinates (x, y, width, height), providing comprehensive information for each detected face.

Acknowledging the principle of spatial continuity, our algorithm hypothesized that an individual's position changes incrementally between consecutive frames, considering the rapid frame rate of 30 frames per second (equivalent to 33.33 milliseconds per frame). Under this assumption, significant movement of a student within a single frame was deemed unlikely. The algorithm commenced by loading the data into a DataFrame and computing the center of each bounding box. This step was critical to establishing a reliable reference point for tracking individual movements across frames. To facilitate the tracking process, the algorithm introduced new columns in the DataFrame designated for unique Student IDs and the historical center positions ('Old_Center_X' and 'Old_Center_Y') of each identified face. Furthermore, it prepared columns for the predicted future positions ('Predicted_X' and 'Predicted_Y'), capitalizing on the inferred motion trajectory of each individual.

The core of the re-identification process relied on calculating the Euclidean distances between the centers of detected faces in consecutive frames. A distance threshold was established to determine whether a detected face in the current frame corresponded to any face from the previous frames. To account for potential changes in the pace or direction of the individuals, the algorithm incorporated a memory mechanism, considering a few past frames (denoted by memory_frames), and predicted future positions based on the velocities calculated in the X and Y coordinates. For each frame processed, if the center of a detected face was within the distance threshold of the predicted position of a previously identified individual, it was deemed the same individual. The corresponding Student ID was then assigned, and the historical center positions for that Student ID were updated. In cases where no match was found within the distance threshold, a new Student ID was assigned, indicating the detection of a new individual in the scene.

Upon the completion of the tracking process across all frames, the algorithm consolidated the valence and arousal scores for each Student ID, thereby crafting an emotional profile for each student throughout the video. The final output comprised Student IDs, bounding box coordinates, valence, arousal data, and frame numbers. This method facilitated the accurate re-identification of students with a success rate of 91%. For the remaining 9%, manual corrections were made, ensuring the integrity and continuity of the student tracking.

4.2 Identifying the Misclassification

In both studies, we employed the iMotions platform with the AFFDEX emotion recognition engine and HSEmotion, renowned for its efficacy on the AffectNet dataset. Data synchronized at 30 frames per second was processed using both models. Upon encountering misclassifications, selected frames were extracted for comparative analysis. This analysis entailed a manual review of each misclassified instance to ascertain whether it was a genuine error in emotion recognition or a result of extraneous factors such as occlusion. For the embodied learning dataset, we utilized Multi-task Cascaded Convolutional Networks (MTCNN) to crop facial images, which were then sequentially fed into iMotions for emotion classification.

4.3 Addressing the Bias

In this study, we investigate the implications of bias in facial expression recognition through a comparative analysis using three distinct methodologies, as detailed in [3]. Notably, the base architecture has been altered from $ResNet$ to $EfficientNet - B0$, and the labels are mapped to sensitivity, associating emotions considered in this study. The detailed implementation of these modifications is explained below. Initially, we establish a baseline using the $EfficientNet - B0$ model, a modification of the commonly used $ResNet$ architecture, particularly adapted for its efficiency and accuracy in various recognition tasks. This model is trained with a Cross-Entropy loss function to predict the facial expression label y_i for each input image (x_i), where the loss is calculated as $(L_{\exp}(x_i) = -sum_{k=1}^{K} 1_{[y_i=k]} log p_k)$. Here, (p_k) denotes the predicted probability of the input (x_i) belonging to class (k), and $(1[cdot])$ represents the indicator function. This baseline sets the stage for evaluating the performance enhancements and bias mitigation effects introduced by the subsequent methodologies.

To address bias, we introduce two sophisticated approaches: the Attribute-aware Approach and the Disentangled Approach. The Attribute-aware Approach, inspired by concepts from previous work, incorporates sensitive attribute information directly into the classification process [20]. In this method, an attribute vector (s_i) is transformed through a fully connected layer to match the feature vector $(\phi(x_i))$ obtained from the EfficientNet-B0 backbone, and the resultant combined feature is fed into the classification layer. This strategy is designed to assess how the inclusion of explicit attribute information influences the recognition of facial expressions and its potential to reduce bias. On the other hand, the Disentangled Approach aims to extract a feature representation $(\phi(x_i))$ that is devoid of any sensitive attribute information (s_i). This is achieved by splitting the network into branches: a primary branch for expression recognition and parallel branches that employ a confusion loss (L_{conf}) and an attribute predictive Cross-Entropy loss (L_s) to ensure the sensitive attributes cannot be predicted from $(\phi(x_i))$. These branches share the network layers until the final fully connected layer, where specific task-oriented branches are formed. The overall loss is a combination of the expression recognition loss, the attribute

predictive loss, and the confusion loss, weighted by a factor (α), ensuring that the final feature representation effectively encapsulates facial expression information while actively discarding sensitive attribute data. This balanced approach not only aids in mitigating bias but also enhances the robustness and fairness of the facial expression recognition task.

5 Results and Discussion

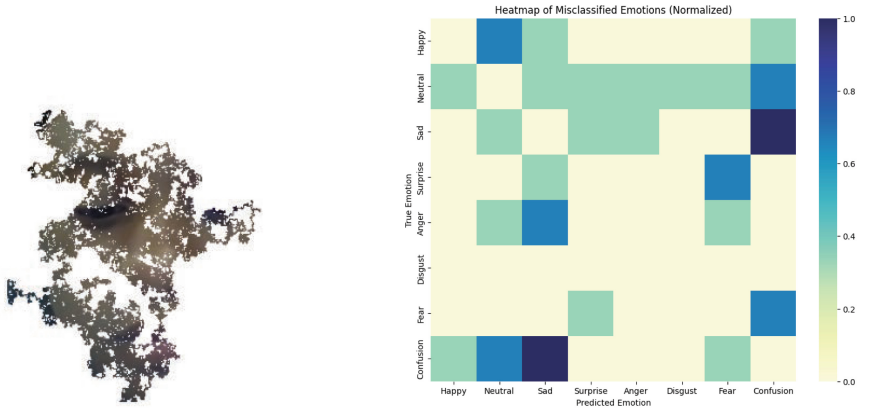
We utilized iMotions' emotion recognition engine (AFFDEX) and a retrained HSEmotion for emotion classification across two learning environments, identifying roughly 9.18 million class labels. Of these, 83.21% were consistent across both systems, while the remaining 16.79% differed, representing over 1,50,000 face image frames with misclassifications. Due to the vast number of misclassifications, manual verification was impractical. Consequently, we used iMotions results as a benchmark, investigating the misclassifications by HSEmotion and vice versa. This approach enabled us to quantify the total misclassifications for each emotion in our study. The distribution of emotions from both learning environments is as follows: Happy (23%), Neutral (20%), Sad (12%), Surprise (8%), Anger (7%), Disgust (7%), Fear (5%), Confusion (18%).

Given that the image frames include faces and the Facial Action Coding System (FACS) [8] provides a well-defined framework, our initial focus was on comprehending facial expressions using action units and assessing their influence on emotion classification. iMotions provides action units data, along with basic emotions in the CSV file. For our retrained HSEmotion model, we utilized Grad-CAM to visualize the selected face regions crucial for classification. This approach aids in understanding how specific facial features contribute to emotion predictions.

Emotions that are observed dominantly during learning are learning-centered emotions such as confusion, boredom, frustration, delight, and engagement. Instances of specific basic emotions like peak fear or peak anger are less frequent. Despite this, some images are detected as anger during student frustration or as disgust and sadness during boredom. Further, the embodied learning data had more than 40% instances of basic emotions where the students were surprised, sad, angry, happy/joy, and so on. Hence, There are ample instances classified under basic emotions.

In the misclassified basic emotions, common observations include surprise being misclassified with fear and vice versa, sadness being misclassified with confusion, anger being misclassified as fear and sadness, and fear also being misclassified as confusion, as evident in the confusion matrix heatmap 2 (b).

Other instances of misclassification occur when the face is not visible, either almost out of the frame or blurred due to the rapid movements of students in the embodied learning environment. Uncertain images, such as the transition from closed to opened eyes or the shift from one emotion to another, also contribute to misclassification in in-between frames. These nuanced scenarios highlight the challenges faced in accurately classifying emotions in dynamic learning environments.



(a) Action Unit Visualization (b) Confusion Matrix Heatmap

Fig. 2. Action unit and confusion matrices visualization

In each case, we analyzed the action unit values in the CSV file and noticed that not all action units were appropriately recognized for the intended emotion. For instance, Action Unit 4 was not identified, leading to a higher probability of sadness rather than confusion. Similarly, Action Units 20 and 7 were not recognized for fear, resulting in misclassification as surprise.

To comprehend the relationship between facial action units and their impact on emotion recognition, we referred to the iMotions documentation to understand the criteria for classification. Through examination of the AFFDEX 2.0 emotion recognition model within iMotions, we identified specific action units that play a significant role in determining certain emotions. To explore if there is any superset or subset relationship among these action units, the findings are presented in Fig. 3.

Similar to iMotions, HSEmotion also encountered misclassifications, primarily attributed to issues with properly recognizing action units. The use of Grad-CAM visualizations further highlighted these discrepancies. While misclassifications for confusion were relatively fewer due to prior training on educational data, the challenges persisted for other emotions, as depicted in Fig. 3.

In Fig. 2 (a), a slightly modified representation of Grad-CAM is presented, addressing concerns about full-face visibility to adhere to ethical considerations. It is evident from the visualization that none of the dominant action units are effectively considered (one eye slightly open and a part of the nose) for the classification process in this frontal face image frame.

Due to the misclassification at the feature level, we scrutinized the databases on which iMotions and HSEmotion systems were trained, specifically the Denver Intensity of Spontaneous Facial Action (DISFA), Affect in-the-wild (AFFwild 2), AffectNet, Audio/Visual Emotion Challenge (AVEC), Acted Facial Expressions in the Wild (AFEW), Video Level Group AFfect (VGAF), Video Game

Facial Animation (VGAF), and Engage Wild. The emergence of data, label, and selection biases is significant, largely attributable to the mismatch between the datasets' origin and their application within educational contexts. These biases manifest distinctly: Data bias arises from demographic discrepancies, as these databases, while comprehensive, do not cater specifically to the nuances of students' profiles, particularly those within the 10-14 age bracket or those with darker skin tones, resulting in a non-representative sample for the educational settings. Label bias is highlighted by the fact that expressions in these datasets, typically portrayed by adults, may not seamlessly align with the genuine emotional expressions of students in learning environments, underscoring the necessity for meticulous and context-aware annotation of action units. Selection bias is underscored by the fact that these datasets, not originally curated for educational purposes, stem from environments and contexts that starkly contrast with typical educational settings such as Betty's Brain and GEM-STEP, potentially compromising the systems' proficiency in accurately deciphering student emotions during educational activities.

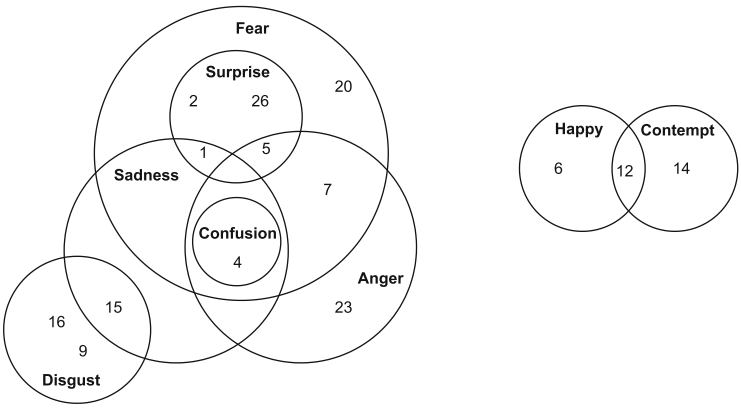


Fig. 3. The subset superset relation of facial action units with respect to emotions

Results on Addressing Bias. In this study, we explored methods to reduce bias in facial expression recognition using an EfficientNet-B0 model. The table shows the performance of different methods like Baseline, Attribute-aware (AW), Disentangled (DE), Data Augmentation Baseline (DA Baseline), Data Augmentation with Attribute-aware (DAWA), and Data Augmentation with Disentangled (DADE). The Baseline model sets a starting point with a 68.3% mAP. The AW approach improved recognition for emotions like Surprise, Fear, and Confusion, raising the mAP to 70.4%. The DE approach did well for Neutral but lowered the overall mAP to 64.6%, suggesting it might remove some useful features. Adding Data Augmentation (DA) raised the Baseline's mAP to 77.0%,

showing the value of a varied dataset. DAWA, combining DA and AW, reached the highest mAP of 78.1%, balancing data diversity and attribute sensitivity.

DADE also did well, with a 78.0% mAP, showing that removing sensitive attributes while augmenting data can effectively reduce bias. Performance details reveal significant improvements in recognizing Sad and Anger with DAWA and DADE. However, DE had some issues, possibly due to the complexity of removing sensitive attributes while keeping important features. While we've addressed bias with these methods, results could be further improved by using data more representative of this specific distribution and balancing the dataset (Table 1).

Table 1. Accuracy of model after addressing bias

in %	Baseline	AW	DE	DA Baseline	DAWA	DADE
mAP	68.3	70.4	64.6	77.0	78.1	78.0
Surprise	83.4	87.7	84.7	91.1	90.8	90
Fear	44.6	52.1	44.6	59.8	61.3	59.2
Disgust	45.2	45.2	36.4	56.8	59.2	59.5
Happy	96.8	96.8	96.6	97.1	97.3	95.1
Sad	69.5	70.6	60.9	80.4	88.7	85.5
Anger	73.2	69.1	59.1	81.2	81.8	89.1
Neutral	86.6	88.4	91.9	96.4	90.4	90.3
Confusion	47.1	53.6	43.1	53.4	55.7	55.5

5.1 Discussion

In the study of emotion recognition within the described educational environments, certain biases, and challenges that extend beyond the conventional scope of data, label, and selection bias were observed. These observations highlight the complexity and nuanced nature of implementing emotion recognition technology in diverse educational settings.

Algorithmic Bias and Misclassification Concerns: The emotion recognition technology demonstrated a higher rate of misclassification for students with darker skin tones, suggesting potential algorithmic biases in the system. This could be due to the model's training on datasets not representative of this demographic, leading to its inadequate performance. Specifically, the recognition of Action Unit 5 (AU5), associated with the upper facial movements, was notably low in the embodied learning data. This indicates that the model might not be adequately trained or tuned to capture the subtleties of these facial expressions, especially in the dynamic context of an embodied learning environment.

Contextual Bias and Technical Limitations: In the embodied learning environment, the detection of AUs was particularly challenging for students. This issue was compounded by inaccuracies in face tracking within the HSEmotion system, hinting at technical limitations or contextual biases within the technology. These inaccuracies could stem from various factors, such as lighting conditions, facial features, or the model's inability to generalize across educational contexts and demographics.

Performance Inconsistency Across Demographics: The technology's performance inconsistency, particularly regarding facial expression and emotion recognition for different skin tones, underscores the importance of ensuring diversity and representativeness in the training data. This is crucial to avoid perpetuating or exacerbating pre-existing biases.

Intrusiveness and Ethical Considerations: The high misclassification rate of confusion in iMotions might be indicative of selection bias, as most of the datasets used were composed of non-student facial expressions. This not only points to a potential selection bias but also raises ethical concerns regarding the appropriateness and relevance of the data used for training models intended for educational settings.

Unexplored Biases and Limitations in the Study: It's acknowledged that the study has not delved deeply into dissecting biases related to race, age, and gender. Given the relatively small sample size of 53 students, it's challenging to draw definitive conclusions or generalize findings across these dimensions. This limitation underscores the need for larger, more diverse datasets to understand and mitigate the various biases effectively.

6 Conclusion

This study researched algorithmic biases, focusing on selection, label, and data biases in emotion recognition within educational settings. Leveraging the retrained HSEmotion, this research meticulously evaluated and contrasted inherent biases in these novel methods against the biases in the commercially available iMotions platform, across emotions, including happy, sad, disgust, fear, anger, surprise, confusion, and neutral. The study offered a realistic and nuanced analysis of algorithmic biases by integrating real student data that accurately reflects classroom diversity. The study's thorough methodologies effectively pinpointed and addressed fundamental biases, demonstrating substantial potential in diminishing these biases through specific strategies. Notably, the results validated the presence of algorithmic bias in student emotion recognition systems while also showcasing viable paths to mitigate such biases significantly. This research enriches the dialogue on creating fairer educational technologies and underscores the importance of fostering the emotional well-being of students. Future research should diversify and enrich emotion recognition datasets, enhance real-time bias

monitoring and adjustment mechanisms, and closely examine the integration of these technologies within educational frameworks to ensure ethical usage and maximization of pedagogical benefits.

Acknowledgement. This research was supported by the National Science Foundation AI Institute Grant No. DRL-2112635. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Anderson, H., Boodhwani, A., Baker, R.S.: Assessing the fairness of graduation predictions. In: Proceedings of the 12th International Conference on Educational Data Mining, pp. 488–491 (2019)
2. Ashwin, T., Guddeti, R.M.R.: Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. Futur. Gener. Comput. Syst. **108**, 334–348 (2020)
3. Baker, R.S., Hawn, A.: Algorithmic bias in education. Int. J. Artif. Intell. Educ. 1–41 (2021)
4. Biswas, G., Segedy, J.R., Bunchongchit, K.: From design to implementation to practice a learning by teaching system: betty's brain. Int. J. Artif. Intell. Educ. **26**, 350–364 (2016)
5. Christie, S.T., Jarratt, D.C., Olson, L.A., Taijala, T.T.: Machine-learned school dropout early warning at scale. In: International Educational Data Mining Society (2019)
6. Danish, J.A., Enyedy, N., Saleh, A., Humburg, M.: Learning in embodied activity framework: a sociocultural framework for embodied cognition. Int. J. Comput.-Support. Collab. Learn. **15**, 49–87 (2020)
7. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learn. Instr. **22**(2), 145–157 (2012)
8. Ekman, P., Friesen, W.V.: Facial action coding system. Environ. Psychol. Nonverbal Behav. (1978)
9. Fan, A., Xiao, X., Washington, P.: Addressing racial bias in facial emotion recognition. arXiv preprint arXiv:2308.04674 (2023)
10. Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)
11. Howard, A., Zhang, C., Horvitz, E.: Addressing bias in machine learning algorithms: a pilot study on emotion recognition for intelligent systems. In: 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), pp. 1–7. IEEE (2017)
12. Lee, H., Kizilcec, R.F.: Evaluation of fairness trade-offs in predicting student success. arXiv preprint arXiv:2007.00088 (2020)
13. Li, X., Song, D., Han, M., Zhang, Y., Kizilcec, R.F.: On the limits of algorithmic prediction across the globe. arXiv preprint arXiv:2103.15212 (2021)
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
15. Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C.: Population validity for educational data mining models: a case study in affect detection. Br. J. Edu. Technol. **45**(3), 487–501 (2014)

16. Savchenko, A.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: International Conference on Machine Learning, pp. 30119–30129. PMLR (2023)
17. Sullins, J., Graesser, A.C.: The relationship between cognitive disequilibrium, emotions and individual differences on student question generation. Int. J. Learn. Technol. **9**(3), 221–247 (2014)
18. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–9 (2021)
19. Wang, Z., Zechner, K., Sun, Y.: Monitoring the performance of human and automated scores for spoken responses. Lang. Test. **35**(1), 101–120 (2018)
20. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating bias and fairness in facial expression recognition. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12540, pp. 506–523. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65414-6_35
21. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: evaluating different sources of student data. In: International Educational Data Mining Society (2020)