A Cross-Cultural Confusion Model for Detecting and Evaluating Students' Confusion In a Large Classroom

Yu Fang
Carnegie Mellon University
Pittsburgh, PA, USA
yufang@andrew.cmu.edu

Shihong Huang
Carnegie Mellon University
Pittsburgh, PA, USA
shihong@cmu.edu

Amy Ogan
Carnegie Mellon University
Pittsburgh, PA, USA
aeo@cs.cmu.edu

Abstract

In traditional lecture delivery setting, it is very challenging to identify which part of the lecture material that students are struggling with. One approach to identify difficult concepts is to capture students' confusion during class time. However, most existing confusion detectors focus on an individual student rather than a classroom, and only on a single ethnicity group which could propagate bias when developing pedagogical technologies. In this paper, we leverage two existing 'Confused' facial expression datasets (DAiSEE and DevEmo) with an East Asian 'Confused' facial expression dataset that we collected. Through model performance and explainableAI, we address potential cultural biases in detecting emotions, particularly in confusion, and identified culturally-specific features that align with prior research. As a proof-of-concept, we deployed this cross-cultural confusion machine learning model in a live semester-long class. This work to integrate cross-cultural facial features highlights the importance of fostering inclusivity in educational technologies.

Keywords

Cross-cultural models, Confusion, Affective computing, Retrieval-augmented generation

ACM Reference Format:
Yu Fang, Shihong Huang, and Amy Ogan. 2025. A Cross-Cultural Confusion Model for Detecting and Evaluating Students' Confusion In a Large Classroom. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706528

1 Introduction

Feeling confused is a very common state for students, especially in STEM education [15, 16]. Although confusion is not regarded as one of the seven universal emotions, it has been proven to be very important in learning science [9, 15, 16]. Usually, students express confusion when they are challenged [11]. While learners often believe that feeling confused negatively impacts their learning [29]; in fact, cognitive disequilibrium is often required for internalizing difficult topics [11, 15]. Broadly, confusion can be categorized into two contrasting types: 'productive' (beneficial) versus 'unproductive' (disruptive) confusion [9]. 'Productive' confusion is when a question is being resolved while 'unproductive' confusion is when a question cannot be resolved due to circumstances [28]. Although we cannot differentiate 'productive' from 'unproductive' confusion without collecting student feedback (whether they are still confused about the topic [9]), a good step forward is to develop a machine learning approach to identify the topics that the students feel confused about in a large classroom, where feedback is more difficult to obtain.

Previous work has successfully detected confusion in learners [12, 15, 18, 33], but the approach was deployed online at an individual level. However, most students learn through participation in class, where personalized interventions online are not appropriate. Additionally, self-reported confusion, when compared to objective measures, can be insensitive to contradictions in a simulated learning environment [12]. Other than forced-choice questions as an objective measure [12], new technologies have enabled us to potentially detect confusion through students' faces.

Additionally, most of these machine learning enabled confusion detection systems in place have not attempted at differentiating by ethnicity. In fact, the few available datasets used in these detection systems only focus on a single ethnicity [18, 33]. However, most classrooms in the United States are multicultural and it is known that different ethnicities express affect differently and they are culturally specific [6, 10, 22–24, 40]. Specifically, when tasked with decoding Eastern Asian learners' facial expressions, observers focus more on their eye areas while they focus more on the overall faces for the westerners [22]. Including these culturally-specific features can foster inclusivity in educational technology, specifically for application designs that aim to help students review course material [15, 19, 45].

In this paper, we built a cross-cultural dataset and deployed the confusion classification in a full-class. We were able to successfully detect confusion in Caucasian and Eastern Asian students. Although the South Asian accuracy was subpar, it was similar to previous work [18]. To our knowledge, no one has explored confusion modeling in Eastern Asian students. Additionally, confusion modeling has not been deployed in full-class, lecture, or discussion sessions. As a proof-of-concept, our work discovered where students struggled most in our lecture-based dataset and used retrieval-augmented responses to provide student review materials after class.

2 Related work

We first summarize how previous research has gauged students' understanding in class. Next, we highlight the importance of including

LAK 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706528

ethnic-group-specific data in using facial images to identify confusion. Finally, we discuss the importance of including contextual information to improve GenAI response validity.

2.1 Identifying difficult topics in a learning environment

Students often feel confused especially when they encountered difficult topics. To probe student's understanding of a lecture, professor can have their students fill out a 'One-Minute Paper' at the end of each class in which they answer a set of questions about what was challenging [19]. Not only did this practice prompt more student-instructor interactions in class, but also allowed instructors to be better prepared for their upcoming lectures. However, this method has several drawbacks. First, it depends on accurate recall from the students. Second, this 'One-Minute Paper' does not allow for more than one main point and question they have. Third, it requires additional time and effort from both instructors and students, which could lead to inconsistent implementation throughout the semester.

To address these concerns, Glassman et al. designed a method to do the aforementioned in class, as opposed to at the end of class. Their interactive dashboard allows students to mark and comment on a slide where terms/concepts are unclear [15]. This in-class exercise not only help students highlight more than one topic they have questions for at the moment, but the spatially displayed output also help aggregate students responses. Overall, this allows for a less time-consuming review for instructors. Certainly, this can only be done in an online learning environment; however, most classes are still held in traditional classroom settings.

Additionally, D'Mello et al. created a simulated learning environment where contains four conditions: both instructors and peer were correct (True-True), instructor was correct but peer was incorrect (True-False), instructor was incorrect but peer was correct (False-True), both instructor and peer were incorrect (False-False). The researchers then gauged the student's state of confusion through either self-report confusion or forced-choice questions. They found that self-reported confusion can only differentiate True-True from True-False but not False-True and False-False. In other words, student tend to report difference in confusion only when the authority (instructor) was correct, not the other way. This demonstrated that self-reported confusion can be confounded by student's perception. However, forced-choice questions revealed the learning outcomes of the students as expected [12]. These finding demonstrated that objective measures are potentially better at identifying where students struggles in class when comparing to self-reporting. To date, most systems that measure non-self-reported confusion data are online in front of a computer as part of an individualized system, rather than supporting full class engagement [12, 34].

2.2 Differences in facial expression in multi-cultural classrooms

Facial expression is a key indicator of understanding in learning environments. Importantly, however, prior research has found significant differences in facial expression amongst different ethnic groups. For instance, East Asians use their eyes to express emotion while Caucasians use their whole face [20–22]. Consequently, interpreting the emotions of East Asians learners is different when compared to those of Caucasian learners [22]. Additionally, people from more collectivist cultures (e.g., Indian, Chinese) are less facially expressive than more individualist ones (e.g., Western European); though even within collectivist cultures, there still appear to be differences in expressing emotions which could due to various philosophical foundations [10]. In fact, Karabuschenko et al. tested affect detection models of the seven universal emotions on people from various ethnicities and found that each model was best at recognizing people within the same ethnicity [24] .

Similar to other affective states, confusion expression is also culturally-specific [10, 22, 24, 40]. For instance, research has shown only 38 % of shared muscle movements (narrowing eyes, change in eyebrows, etc.) when expressing confusion among Caucasian, Indian, Chinese, Korean, and Japanese participants [10], although not demonstrated in a learning setting.

Confusion, while not one of the seven universal emotions, is a particularly important facial expression to understand in learning environments. Specifically, confusion is a proxy for cognitive disequilibrium [11]. While confusion is clearly an emotion with negative valence and medium arousal [5], there are inconsistencies with coincident universal emotions. In learning environments, Cloude et al. reported the confusion expressed by 6-graders in southeastern United States (multicultural) matched mainly with two universal emotions: 30 % disgust and 40 % surprise [9]; however, Manikowska et al. reported that college students' (all Caucasians) confusion coincided with mainly angry and disgust[33]. Although these discrepancies could be due to subjects' ages, it is not feasible to dissect confusion based on two or more universal emotions. It is crucial to understand these cultural nuances to accurately interpret and address student confusion in multicultural educational settings. Given that publicly available datasets we acquired only include Caucasian and South Asians [18, 33], and the knowledge gaps we discussed, in this work we collected a East Asian-specific learner dataset as a complement dataset to the existing two publicly available datasets .

2.3 Retrieval-Augmented Generation (RAG) for learning

Once we find moments of confusion, it is a reasonable next step to support instructors with identifying the specific parts of the lecture where difficult concepts that cause this confusion are located. One possible approach is to use ChatGPT to help find these moments; ChatGPT's responses often appear reliable, but they can sometimes be misleading due to large amounts of unverified training data, let alone new discoveries not included in the database [13]. The potential of ChatGPT and similar models in education has also raised questions about fairness and bias [25]. Alternatively, the concept of retrieval-augmented generation (RAG) combines retrieval-based and generative techniques to provide accurate and contextually relevant answers [3]. Thway et al. proposed a RAG solution that ingested all course materials (syllabus, course content, quiz answer, etc.) into the knowledge base [45]. The study found that the students used the chatbot more frequently when faced with challenging tasks, indicating their trust in the system.

In this work, as a proof-of-concept, we use Retrieval-Augmented Generation (RAG) to find confusing concepts based primarily on class transcripts.

3 Methods

3.1 Confusion meta-dataset curation

In this research, we used 3 datasets to explore confusion in a learning environment. The first dataset, DAiSEE, is an open-sourced collection that includes 4 emotions: engagement, boredom, confusion and frustration. The DAiSEE dataset is comprised of 113 subjects (9,068 videos) [18]. The median frame count per subject is 25272 frames. The frame count per subject ranges from 3960 frames to 35400 frames, excluding bottom and top 10 percentile. The median video count per subject is 85 10s videos. The video count per subject ranges from 29 videos to 118 videos, excluding bottom and top 10 percentile. In their original dataset, confusion is categorized into very low, low, high, and very high with 6024, 2191, 752 and 101 labels respectively. However, we excluded some of the frames that were not recognizable by our facial recognition extraction, yielding a total of 7752 (combined very low and low) and 819 (combined high and very high). For our research, we will simplify the confusion category by merging the first two and the last two to 'Not Confused' and 'Confused' for standardization purposes [16]. In DAiSEE, all videos have confusion annotations; however, the dataset contained much more 'Not Confused' annotation than 'Confused' (Fig. 1B). Additionally, each 10 second clip only contains 1 annotation. To address both issues, we opted to grab one frame (at 5 seconds) of the 'Not Confused' emotion while we selected one frame every 2 seconds for 'Confused' (Fig. 1A).

The second dataset, DevEmo, includes 217 video clips showing their facial expressions (anger, confusion, happiness, surprise and a neutral state) when they are working on programming questions [33]. For DevEmo, we selected 10 subjects that had confused emotion expression (2 subjects have 2 videos, while the rest have 1 video). The median frame count per subject is 539 frames. The frame count per subject ranges from 141 frames to 1248 frames, excluding bottom and top 10 percentile. We also applied a facial recognition extraction to minimize poor image quality frames. We used approximately 1:1 ratio of videos containing either confused or any of the following as not confused: anger, happy, surprised or neutral. Note that we dropped some videos due to encoding errors.

The third dataset was collected by the authors. It consists of college students, ages 20-23 years old. The participants were instructed to take a 1-min video of themselves watching a youtube video https://www.youtube.com/shorts/xJwkj25fFK4 explaining the Mendelbrot set. Most participants did not have prior knowledge. Our dataset included 17 subjects (10 females and 7 males). The median frame count per subject is 1815 frames. The frame count per subject ranges from 1526 frames to 1914 frames, excluding bottom and top 10 percentile. For all three datasets, the confusion labels were manually annotated using the same annotation tool "Boris" as in Manikowska et al. We also applied a facial recognition extraction to minimize poor image quality frames. [14, 33]. Similar data collection has also being used by other research papers in both natural and e-learning context [10, 27, 34, 39]. Some have used existing software for emotion recognition [27], while others used at least two experienced coders to annotate emotion [10, 34, 39].

3.2 CNN training and evaluation metrics

We trained a Convolutional Neural Network (CNN) by simplifying the existing DeepFace model [42–44] to three layers, preventing overfitting our limited dataset. In data preprocessing, we used MTCNN [49] to detect face bounds. Then, we formatted the colored images in grayscale and standardized the dimensions into 64 x 64 pixels. Finally, we one-hot encoded the 'Confused' facial expression annotations. The curated meta-dataset was split into training/test data (80%/20%). Our proposed CNN architecture involved the use of a Conv2D layer with MaxPooling2D for the extraction of prominent features, followed by an AveragePooling2D for noise reduction and a Dense layer for giving output. The Adam optimizer was applied to minimize the categorical cross-entropy loss function, enhancing performance and efficiency [41]. The model was trained for 200 epochs with an observable plateau in training loss.

To assess the generalizability of individual dataset-specific models to another dataset, we trained on 80% of each dataset and tested each model within and across datasets. That is, the same 20% from each dataset were being evaluated by each model for fair comparison. Training each ethnicity-specific model on an M2 MacBook Air (16GB RAM) took 20-25 minutes, while the cross-cultural model took 60-70 minutes for 200 epochs.

3.3 Shapley explanations and evaluation metrics

SHapley Additive exPlanations (SHAP) is an explainable AI technique that replicates the inputs and outputs of the model to offer insights into the internal workings of the model. After we trained a CNN on 'Confused' versus 'Not Confused' facial expression, we utilized SHAP to provide a reproducible rule set of the CNN model decisions [31, 32]. For each dataset (DevEmo and EA), we trained a CNN model specifically to replicate the input-output relationships. Starting from a base rate of 0.5 (binary classification), SHAP interprets each pixel's importance to raising the probability of 'Confused'. These pixels summarize the model's decision, e.g. 'Confused'. SHAP allows us to understand and interpret the factors contributing to the model's decisions, which enhance our trust towards the model prediction.

To quantify our findings, we randomly selected 50 images from each dataset. We then used SHAP to explain how each model identified confusion, specifically identifying the features that increase or decrease the probability of a person being classified as confused. We focused our analysis on the top features (absolute values that are greater than mean + 2 standard deviations) that most strongly influenced the model's decision. We then plotted the empirical Cumulative Distribution Function (eCDF) to illustrate the distribution differences between DAiSEE, DevEmo, and EA. The distribution differences were validated with two-sided Kolmogorov-Smirnov (KS) test, in which we compared DAiSEE and EA to DevEmo.

3.4 Class used in our analysis

In this study, we focus our analysis on 15 recorded class sessions spanning from February to May in 2019 with a primary focus on designing and prototyping user interfaces. The class had an overall duration of 80 minutes and has 30 students on average. This course consisted of a few modes of teaching: traditional lecturing, instructor-led discussion, group discussion, in-class quiz, video demonstration, and project presentation. We take privacy concerns seriously and have a privacy and security expert on our team. We obtained the approval of the IRB for this work and have extensively reviewed privacy and data protection concerns with the ethics board. Furthermore, our larger team has conducted several studies with such systems in the classroom and has previously reported on the results, including teacher and student impressions.

For each class session, we followed the CNN training procedure, where we used a DeepFace-based model to detect faces, grayscale the image, and finally standardize the images to 64 x 64 pixels prior to CNN model prediction. We take a snapshot every 2 seconds, which would lead to a total of 2,400 time points for an 80-minute session. Each snapshot is then segmented to student faces, resulting in 17 standardized facial images per 2s snapshot on average. For 15 sessions, we have a total of 613,092 images (2,400 time points x ≈ 17 faces per class x 15 sessions). We then processed each image through the CNN model to generate confusion predictions. For an 80-minute session, this end-to-end process—including running the model and performing speech-to-text conversion—took approximately 80-120 minutes on a single NVIDIA GeForce GTX 1080 Ti, with 8 minutes dedicated to the speech-to-text step.

3.5 Difficult concept identification

Upon obtaining each student's emotional state as 'Confused' or 'Not Confused', we binned them into 1 minute intervals. To identify the moments when students struggled most in class, we compare per 1 minute confused students to the session's minimum and chose 2 times the minimum as the threshold. While less than 1 minute does not produce enough transcript for topic identification, more than 2 minutes may wash out confusion signals. Next, we use OpenAI wrapper to transcribe the session's audio and save it as a CSV file for each session. We then pinpoint the minutes where students felt confused (again, twice the minimum) and locate the corresponding transcript. This process allows us to correlate moments of confusion with specific parts of the lecture to locate the difficult concepts discussed in class.

3.6 Baseline comparison model

The purpose of a chance model comparison is to provide a baseline to demonstrate that the confusion classification is not random during deployment. Similar to our confusion classification framework, our chance model also uses the number of detected faces through DeepFace. However, instead of using model generated prediction, the chance model randomly assigns 'Confused' or 'Not Confused' to each identified student. Since randomly assigning affect prediction to students should yield 50% 'Confused' and 50% 'Not Confused', using detected students from one session would be no different from using detected students from another session. We picked the April 16th, 2019 session to have a total student count per every 2-second interval. Again, since the 'Confused' or 'Not Confused' at every 2-second interval should result in a 50 % probability overall. We then averaged the number of 'Confused' students every 1 minute, following the same procedure as with our classifier.

To quantify our findings, we plotted the empirical Cumulative Distribution Function (eCDF) to illustrate the distribution for each of the 15 recorded sessions versus the chance model. The distribution differences were validated with a two-sided Kolmogorov-Smirnov (KS) test, in which we compared each session to the chance model output.

3.7 RAG and dashboard implementation

After obtaining the session transcripts and identifying the minutes when students are feeling most confused, we were inspired by Retrieval-Augmented Generation (RAG) for a grounded response [30]. Specifically, we use the session transcript as the knowledge base and the confused-isolated transcript described previously as the relevant context. This combined data is then fed into ChatGPT. This approach allows us to generate concise and targeted summaries of the difficult topics to help in identifying the areas that caused confusion.

We then deployed this pipeline through streamlit to enable a no-code user interface. Our app 'Understand Me' has a login screen for students/instructors. Once logged in, they can select from a list of courses for which they are registered. With each class, students/instructors can then select the course date, and our chatbot will provide a summary of difficult topics.

4 Results

4.1 Developing a confusion meta-dataset

We utilized two existing datasets online, DAiSEE [18] and DevEmo [33]. Each of the datasets is of a certain ethnicity: South Asian for DAiSEE and Causasian for DevEmo. Facial expression intensity is valued differently in different cultures [8]. Therefore, it is essential to consider cultural variations in emotional expression. We first describe the individual dataset sizes. After balancing the data, the final counts for the DAiSEE dataset were 7.8k for 'Confused' and 4.3k for 'Not Confused' (see Methods, Fig. 1B). For DevEmo, the final counts were 2.8k and 6.9k for 'Confused' and 'Not Confused', respectively (Fig. 1B). For Eastern Asian (EA), we collected 6.3k and 5.3k for 'Confused' and 'Not Confused', respectively (Fig. 1B). For each dataset, we can see that the raw images for each dataset (either 'Confused' and 'Not Confused') contain different lighting, angles, and proximity to the camera (Fig. 1C).

4.2 Ethnic differences in expressing confusion

Previous research has shown that facial expression models trained on specific ethnicities cannot predict another ethnicity [24], likely due to differences in learned muscles movements and philosophical backgrounds [10]. To test this hypothesis in a learning environment, we tested the generalizability of each dataset-specific trained model on itself and the remaining two datasets (see Methods). For DAiSEE model, it reaches an accuracy of 0.68 for its own dataset. These results are consistent with other research, although not in a learning environment, which suggests that recognizing negative emotions, especially for Indian participants, can result in higher error rates [7, 18]. When using DAiSEE model to test on the other two datasets, it performed close to chance (0.38 for DevEmo and 0.53 for EA, Fig.

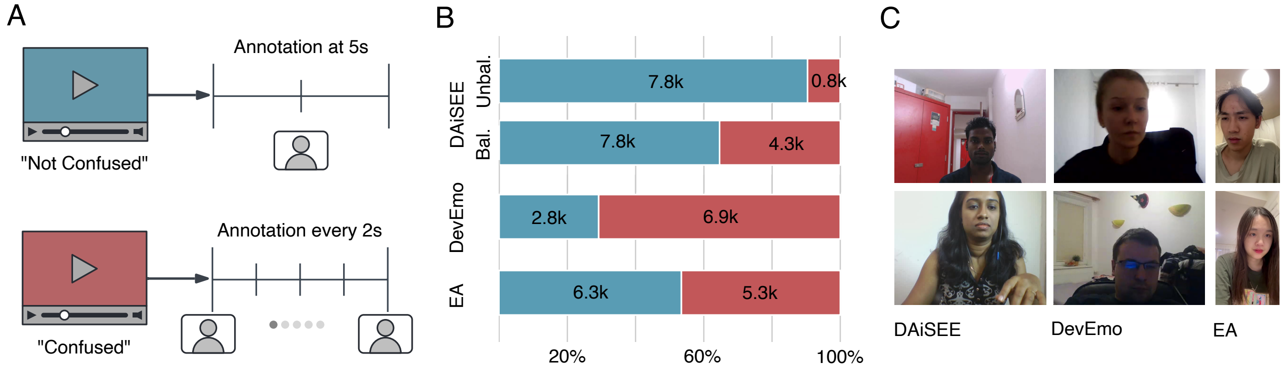


Figure 1: Confusion meta-dataset curation. A) Schematic of our annotation approach for the DAiSEE dataset. We grabbed one frame for students feeling not confused (at 5 second mark) and one frame every 2 seconds for students feeling confused. B) Stacked bar chart for data counts for DAiSEE (top: initial; bottom: upon schematic approach in A), DevEmo and East Asian (EA) datasets C) Two example images from each dataset (columns).

2A). Similarly, the other two models performed better within their dataset, but close to chance outside the ethnic group. (Fig. 2A). To resolve ethnicity-specific differences, we were inspired to include both international core patterns (ICPs) - facial features that are shared amongst cultures, and cultural variant patterns (CVPs) - cultural accents in facial features [10]. To harness the importance for both ICPs and CVPs, we trained our model across ethnicities with relatively balanced counts of 'Confused' vs 'Not Confused'. In our final model, by combining all datasets, we achieved an accuracy of 0.69 for DAiSEE dataset, 0.99 for DevEmo dataset and 0.95 for EA dataset (Fig. 3B). Overall, our cross-cultural model was able to preserve important features from all individual ethnicity models in identifying confused facial expression.

To explore differences in CVPs for each individual dataset, we separately trained a CNN model for South Asians, Caucasians, and East Asians, where each model will only be tasked to identify the corresponding ethnic group. We then applied SHapley Additive exPlanations (SHAP) in an attempt to reproduce the results through establishing a rule-set independent of the classification network (Fig. 2B, see Methods). For instance, if the model predicts a student is confused with a 90 percent probability, SHAP can attribute a portion of this prediction to specific features, such as eye pixels, nose pixels etc.. By examining 2 examples from each cohort, we can see a more crystallized pattern around the eyes and nose in EA and DAiSEE when compared to DevEmo (Fig. 2D).

To quantify our findings, we plotted an empirical cumulative distribution function (eCDF) for the number of most important SHAP pixels (see Methods). The eCDF plot reveals that at the 50th percentile, it takes 7 pixels to identify confusion for DAiSEE learners, 27 pixels to identify confusion for EA learners, and 52 pixels to identify confusion for DevEmo learners (Fig.2C). When compared to DevEmo, both EA's and DAiSEE's distributions were statistically different (DAiSEE $p < 0.001$; EA $p < 0.001$, see Methods for statisical test, Fig. 2C). These results demonstrated that DAiSEE's facial expression, followed by EA's facial expression, are more concentrated in certain areas than DevEmo (Fig. 2C-D).

4.3 CNN built to reproduce confusion annotations in unseen dataset

Using all three datasets (DAiSEE, DevEmo, and our own EA dataset), we trained a CNN model to identify when facial expressions are confused (see Methods). To test the model's generalizability, we plotted ground truth against predicted annotations in the held-out facial images. As shown in the confusion matrix, our CNN model achieved an 89 % for true negatives and 82 % for true positives (Fig. 3A). The facial images that are predicted as 'Confused' and 'Not Confused' include faces from all three datasets (Fig. 3C). The accuracy was slightly better for DevEmo and EA (Fig. 3B). This is consistent with previous research using the same dataset [18].

4.4 Population confusion revealed difficult topics

After we determined that our model could distinguish between 'Confused' and 'Not Confused', we aimed to apply this model in our classes to identify topics that cause confusion. For each session, we took the average 'Confused' students per minute (see Methods). We considered the fact that the number of identified faces could vary within a session; therefore, we adopted the normalized number of confused students per minute, i.e., the number of confused faces divided by the total identified faces. To identify the minutes when students are 'Confused', we compared the proportion of 'Confused' students to the session's minimum. We locate difficult topics by examining the relative increase in per minute 'Confused' students. In other words, if the proportion of 'Confused' students per minute exceeds 2 times the minimum proportion in that session, it will be considered a relatively more difficult topic (Fig. 4A).

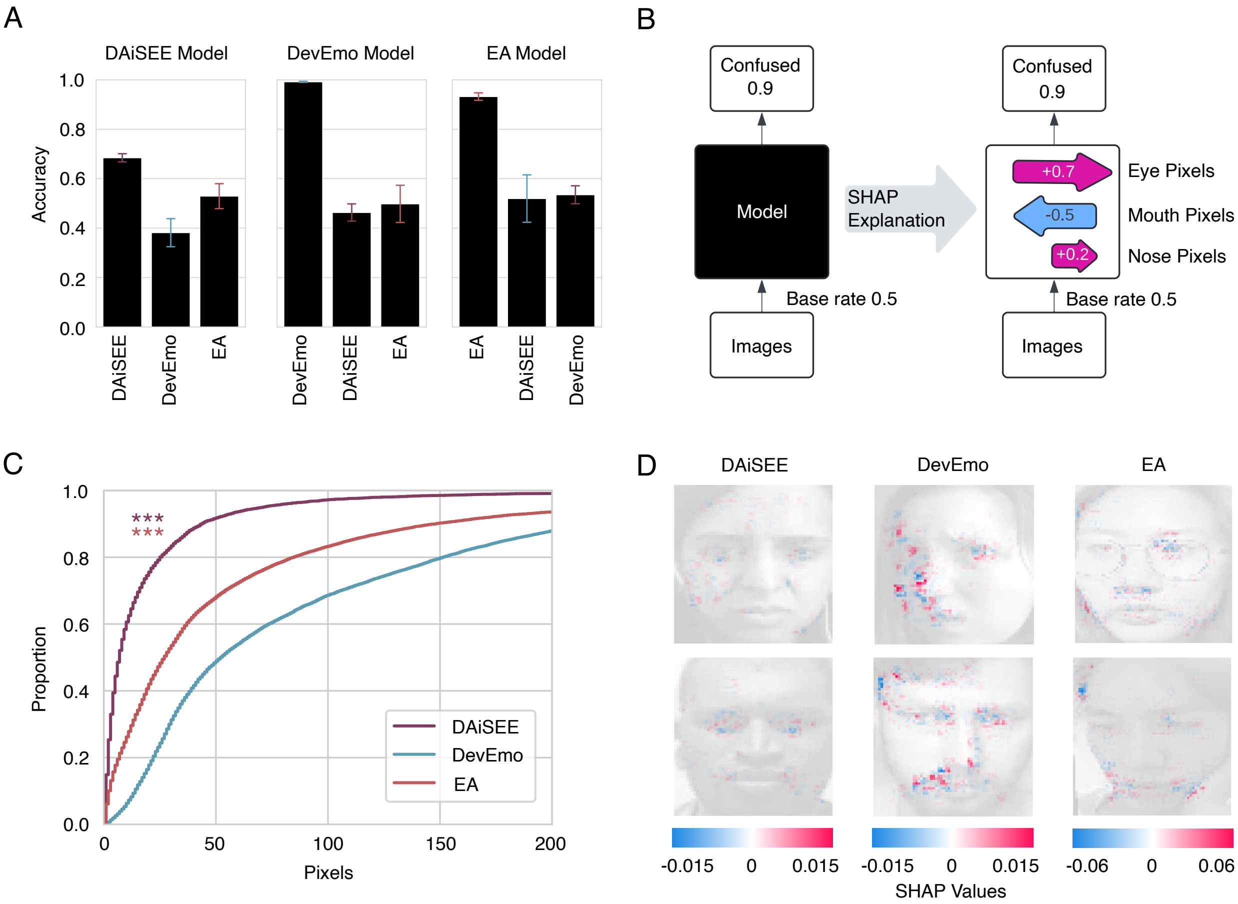


Figure 2: Comparing and contrasting the confusion emotion across all three ethnicities. A) Accuracy bar charts of DAiSEE model on all three datasets (left), DevEmo model on all three datasets (middle), and EA model on all three datasets (right). Error bars represent the standard deviation across 10 random seeds. B) Schematic of SHAP analysis on a Convolutional Neural Network Model. SHAP makes the model transparent by applying certain rules to mimic the output of the model. C) CDF plot the number of top (mean + 2 standard deviations) SHAP pixels for all test images across 10 random seeds: 24010 DAiSEE, 19460 DevEmo, and 23300 EA faces. *** = p < 0.001. D) Confusion SHAP values drawn on images of South Asian (DAiSEE, left), Caucasians (DevEmo, middle) and East Asians (EA, right). The higher the SHAP values (red), the more likely it predicts confusion.

To validate whether our model predict better than chance, we generate the empirical cumulative distribution function (eCDF) and compare each session to a chance model (Fig. 4A, see Methods). We found that all sessions' normalized confused students ratio is significantly above chance model output (see Methods for statistical test, $p < 0.001$). For most of the sessions, there appear to be 6 out of 15 sessions that consist of ≥ 5 minutes confused by our definition (Fig. 4B). To compare our model to the chance model, we first plotted the normalized confused students ratio for one of the sessions (April 16th 2019) over the 80 minutes. Then, we overlaid what a chance model would have given us (Fig. 4C). As shown, the chance model stays just above the 1.0 ratio consistently due to both the minimum and maximum proportion of confused students around 50% per minute. However, in our model, we found five 1-minute windows consisting of twice the proportion of confused students when compared to the easiest part of that session.

4.5 RAG-inspired chatbot to help reiterate concepts

Upon finding the topics that are relatively more difficult within the session (i.e. April 16th 2019), we summarized the 'Confused' topics by generating a word cloud in which the most frequent words were mentioned (Fig. 4D). We hypothesize that if a topic was being introduced, the number of times that terminology would show up will be more than a chance. In fact, the concepts that were introduced on April 16th showed up as common words in a word cloud analysis. In this session, some text that showed up more often

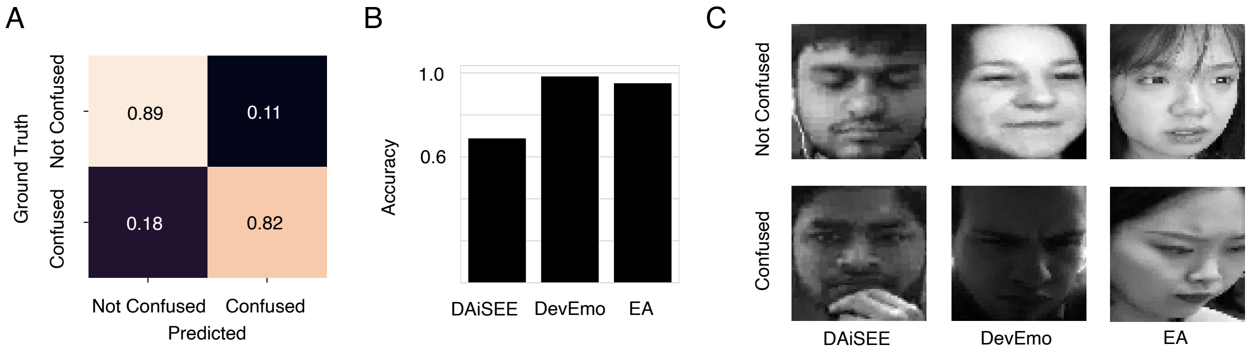


Figure 3: Confusion model generalizes to all three datasets. A) Confusion Matrix on the 20% held-out data. On the rows we have ground truth and on the columns we have what the model predicted. B) Bar chart of prediction accuracy for each of the three datasets (DAiSEE: 0.69; DevEmo: 0.99; EA: 0.95). C) Example facial images of predicted 'Confused' and 'Not Confused' (rows) for each dataset (columns).

during confused time points is 'galvanic response', 'arousal', and 'electrothermal activity' (Fig. 4).

After recognizing the point of confusion in the session transcripts, we find the exact minute where the confusion is taking place and use this as confusion-isolated context. By combining this with the entire session transcript as the relevant context and knowledge base, we combine both in a query to GPT4o (Fig. 5A, see Methods). GPT4o's response should be grounded to individual session transcript, which helps to provide a more accurate and helpful response. We have developed this capability for all sessions and future classes in streamlit (Fig. 5B, see Methods). To provide evidence for how RAG-grounded response reflect the course session better, we then compared the Jaccard similarity between the session transcript with either non-grounded GPT4o's response (ChatGPT) or grounded GPT-4o's response. Across the 15 sessions, the average similarity between RAG-based GPT4o's response to audio transcript almost doubled the similarity between non-RAG-based GPT4o's response (Fig. 5C). Additionally, the responses between RAG and non-RAG were more similar, aligning with the consistency, yet lack in diversity of responses queried at different times with very similar prompt [47].

5 Discussion

We found that our 'Confused' facial expression model captured important features and performed well on held-out datasets. What makes up a 'Confused' facial expression is often disagreed upon, specifically given that observers of different cultures could focus on different aspects of the face [22, 23]. Therefore, it is imperative to incorporate a diverse set of ethnic groups to accurately capture confusion in learning. To this point, we added an East Asian dataset on top of the two openly available datasets (South Asian and Caucasian). Although not in a learning environment, previous work has shown that collectivist cultures (e.g., Indian, Chinese) tend to be less expressive than those from individualist cultures (e.g., Western European) [10, 20, 24]. We found that South Asian learners are more similar to East Asian learners in that the 'Confused', when compared to 'Not Confused', requires pixels around the eyes.

In our work, we have demonstrated the utility in combing culturally specific features to foster inclusivity in education technology. Incorporating diversity in computational models can better support teachers across global settings [26], especially with our low cost requirement. Our solution can help teachers better understand their students without spending additional effort (surveys, one-on-one mentoring, etc.) due to the teacher-student ratio of 1:99 in certain regions [46]. Our approach is also likely to be adopted simply due to the trust between the student and the instructor [35] because they could promote student attendance by including the student in the technology [19, 46].

5.1 Future work

In order to better understand confusion as a feature of student learning, future work might integrate EEG and EDA analysis to monitor alpha waves and sympathetic network activity can provide deeper insight into students' cognitive states [17, 48]. By combining facial recognition with EEG data [48], EDA data [17], posture and gaze tracking [1, 2, 36–38], and finally student feedback [12, 15, 35], we can achieve a more comprehensive understanding of student emotions and improve the overall effectiveness of the model.

5.2 Limitations

Our dataset does not include all ethnicities found in the world. Although we did not include all ethnicities, it is our duty to be conscientious of embracing the differences in facial expression across biological sex, sexual orientation, and ethnicity. Additionally, the faces that we used for training are mostly upright in position; however, the faces are at an angle to the camera in the classroom in which we applied the model. Future study should include training dataset from different angles. Lastly, for the EA dataset, we have a designated annotator for all videos. This could raise some bias issues. It has been reported that expert facial expression annotators often disagree on 'Confused' facial expression [23]. For the two other datasets, they have more than one annotator.

It is very time-consuming and laborious to annotate every frame of the students' facial expression. CNN has enabled us to quickly and

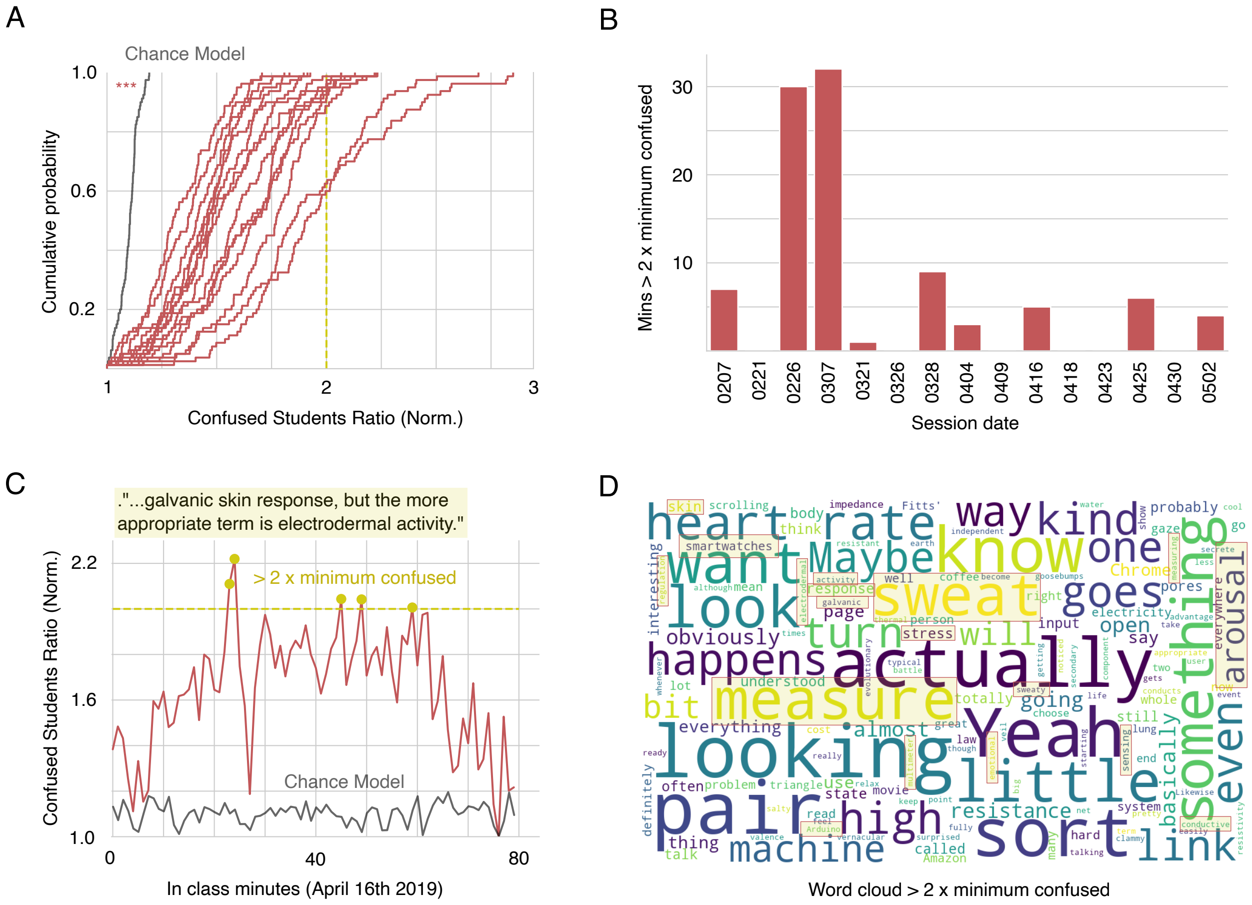


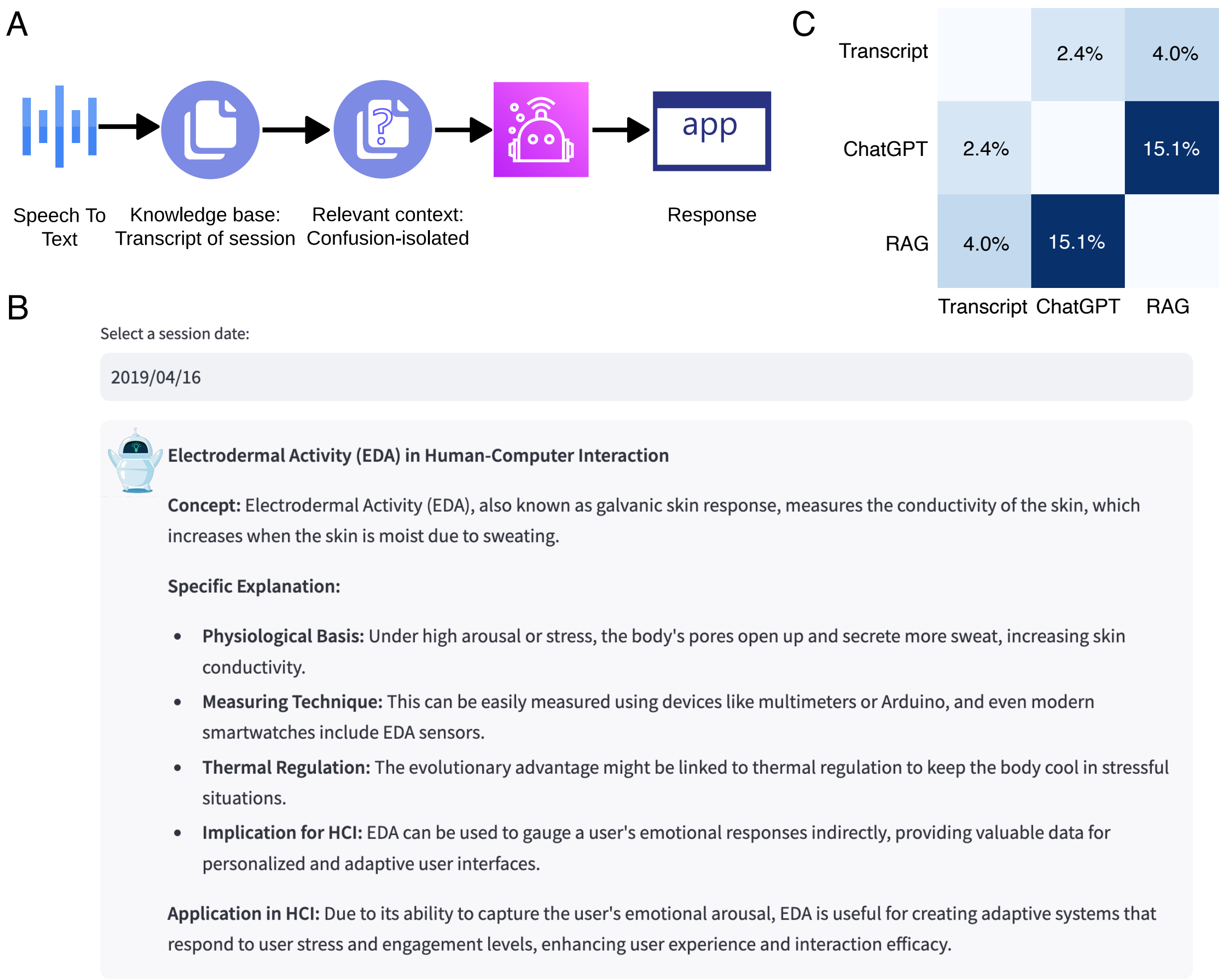
Figure 4: Confusion time points and topics in a sample session. A) Empirical cumulative distribution function of random model (dark gray) versus all 15 sessions using our confusion classification network (red). Yellow vertical dashed-line denotes 2 times the minimum number of confused students. *** p < 0.001. B) Bar plot of the number of total minutes above yellow dashed-line in A. C) Line plot showing number of 'Confused' students binned by every minute using either random model (dark gray) or our confusion classification network (red) on April 16th, 2019. Yellow horizontal dashed-line is the same as A, where red highlights represents all minutes above the dashed-line, shown in B. Yellow box shows a sentence said at one of the red highlighted 1-minute window. D) Word cloud plot of all transcript covered within the red highlighted points in C.

accurately predict the emotions from facial expressions. This comes with the caveat of biasing towards one type of annotation to another ('Confused' vs 'Not Confused') due to training count differences. We have attempted to ameliorate the bias by oversampling the minority class and undersampling the majority class per video. Future work should be attentive to balance the annotations prior to training the model.

Lastly, in the classroom video that we used, we were unable to detect all participating students. The reason for our inability to capture all students could be two fold: first, the face for a student is partially obstructed by another from our camera angle; second, the faces at the edge could be distorted by the fish-eye camera. Future work should consider saving each student's image and annotating based on student feedback to enhance the model's accuracy for facial expressions [4].

6 Conclusion

In a learning environment, East and South Asians use more of their eyes when express their emotions; while Caucasian tend to use their whole face to express emotions. When including all three datasets to train a confusion facial expression model, we were able to include all cultural variant patterns in our final model. Including facial features that are culturally specific has improved generalizability in our results. The classification reached a 82% true positive and 89% true negative on the held-out data set. The overall model performed better on Caucasians and East Asians compared to South Asians, although our South Asian classification performance was similar


A
Speech To Text
Knowledge base: Transcript of session
Relevant context: Confusion-isolated
app
Response
C
B

	Transcript	ChatGPT	RAG
Transcript		2.4%	4.0%
ChatGPT	2.4%		15.1%
RAG	4.0%	15.1%	

Select a session date:

2019/04/16

Electrodermal Activity (EDA) in Human-Computer Interaction

Concept: Electrodermal Activity (EDA), also known as galvanic skin response, measures the conductivity of the skin, which increases when the skin is moist due to sweating.

Specific Explanation:

- **Physiological Basis:** Under high arousal or stress, the body's pores open up and secrete more sweat, increasing skin conductivity.
- **Measuring Technique:** This can be easily measured using devices like multimeters or Arduino, and even modern smartwatches include EDA sensors.
- **Thermal Regulation:** The evolutionary advantage might be linked to thermal regulation to keep the body cool in stressful situations.
- **Implication for HCI:** EDA can be used to gauge a user's emotional responses indirectly, providing valuable data for personalized and adaptive user interfaces.

Application in HCI: Due to its ability to capture the user's emotional arousal, EDA is useful for creating adaptive systems that respond to user stress and engagement levels, enhancing user experience and interaction efficacy.

Figure 5: Retrieval-Augmented Generation design for identified difficult topics in class. A) Customized RAG Pipeline to incorporate session transcripts and students interactions. B) Front-end design showcasing the response for previous demonstrated session in Fig. 4. C) Jaccard similarity matrix showing the response similarity to the audio transcript either using ChatGPT directly or RAG design.

to previous research [18]. Furthermore, as a proof-of-concept, we deployed our model and applied it to a live semester-long class and identified difficult lecture topics in real-time. Through word cloud analysis and RAG, we summarized the concepts that require instructors' attention. Our study shows that recognizing confusion is an important factor in enhancing students learning experience when encountering difficult topics. Confusion is addressed in various cultural backgrounds. Therefore, an important next direction for learning research is to ensure that confusion is adequately addressed for all learners in multicultural classrooms.

Acknowledgments

We thank Dr. Chris Harrison for letting us use his lecture recordings. We also would like to thank Dr. Yuvraj Agarwal and Prasoon Patidar for their help on accessing their video data and setting us up for private server access. Finally, we would like to thank the East Asian students in our study. This work was supported by the National Science Foundation (NSF) under award 1822813.

References

[1] Karan Ahuja, Deval Shah, Sujeath Pareddy, and Franceska Xhakaj. 2021. Classroom digital twins with instrumentation-free gaze tracking. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). https://doi.org/10.1145/3411764.3445711/SUPPL{_}FILE/3411764.3445711{_}VIDEOPREVIEW.MP4

[2] Karan Ahuja, Virag Varga, Eth Zurich, Anne Xie, Dohyun Kim, Franceska Xhakaj, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (9 2019), 1–26. https://doi.org/10.1145/3351229

[3] Sabriya Maryam Alam, Haodi Zou, Reya Vir, and Niloufar Salehi. 2024. SAGE: System for Accessible Guided Exploration of Health Information. www.aaai.org

[4] Mohammed N. Alharbi, Shihong Huang, and David Garlan. 2022. A Probabilistic Model for Effective Explainability Based on Personality Traits. *Lecture Notes in*

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 13365 LNCS (2022), 205–225. https://doi.org/10.1007/978-3-031-15116-3{_}10

[5] Ryan Shaun Baker, Elizabeth B. Cloude, Juliana Andres, and Zhanlan Wei. 2024. The Confrustion Constellation: A New Way of Looking at Confusion and Frustration. (6 2024). https://doi.org/10.35542/OSF.IO/4RGTK

[6] Hannah J. Birnbaum, Nicole M. Stephens, Sarah S.M. Townsend, and Mar Yam G. Hamedani. 2021. A Diversity Ideology Intervention: Multiculturalism Reduces the Racial Achievement Gap. *Social Psychological and Personality Science* 12, 5 (7 2021), 751–759. https://doi.org/10.1177/1948550620938227/ASSET/IMAGES/LARGE/10.1177{_}1948550620938227-FIG2.JPEG

[7] Andrea Bonassi, Tommaso Ghilardi, Giulio Gabrieli, Anna Truzzi, Hirokazu Doi, Jessica L. Borelli, Bruno Lepri, Kazuyuki Shinohara, and Gianluca Esposito. 2021. The Recognition of Cross-Cultural Emotional Faces Is Affected by Intensity and Ethnicity in a Japanese Sample. *Behavioral Sciences 2021, Vol. 11, Page 59* 11, 5 (4 2021), 59. https://doi.org/10.3390/BS11050059

[8] Chaona Chen and Rachael E. Jack. 2017. Discovering cultural differences (and similarities) in facial expressions of emotion. *Current opinion in psychology* 17 (10 2017), 61–66. https://doi.org/10.1016/J.COPSYC.2017.06.010

[9] Elizabeth B. Cloude, Anabil Munshi, J. M.Alexandra Andres, Jaclyn Ocumpaugh, Ryan S. Baker, and Gautam Biswas. 2024. Exploring Confusion and Frustration as Non-linear Dynamical Systems. *ACM International Conference Proceeding Series* (3 2024), 241–252. https://doi.org/10.1145/3636555.3636875

[10] Daniel T. Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2018. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion (Washington, D.C.)* 18, 1 (2 2018), 75–93. https://doi.org/10.1037/EMO0000302

[11] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media* 29, 3 (10 2004), 241–250. https://doi.org/10.1080/1358165042000283101

[12] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2 2014), 153–170. https://doi.org/10.1016/J.LEARNINSTRUC.2012.05.003

[13] Robin Emsley. 2023. ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia 2023 9:1* 9, 1 (8 2023), 1–2. https://doi.org/10.1038/s41537-023-00379-4

[14] Olivier Friard and Marco Gamba. 2016. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution* 7, 11 (11 2016), 1325–1330. https://doi.org/10.1111/2041-210X.12584

[15] Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A spatially anchored census of student confusion for online lecture videos. *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (4 2015), 1555–1564. https://doi.org/10.1145/2702123.2702304

[16] Oleg Aleksandrovich Golev, Michelle Huang, Chanketya Nop, Kritin Vongthongsri, Andrés Monroy-Hernández, and Parastoo Abtahi. 2024. Hapster: Using Apple Watch Haptics to Enable Live Low-Friction Student Feedback in the Physical Classroom. (2024). https://doi.org/10.1145/3613905.3650733

[17] Jamie Gorson, Kathryn Cunningham, Marcelo Worsley, and Eleanor O'Rourke. 2022. Using Electrodermal Activity Measurements to Understand Student Emotions While Programming. *ICER 2022 - Proceedings of the 2022 ACM Conference on International Computing Education Research* 1 (8 2022), 105–119. https://doi.org/10.1145/3501385.3543981

[18] Abhay Gupta, Arjun D ' Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. DAiSEE: Towards User Engagement Recognition in the Wild. (9 2016). https://arxiv.org/abs/1609.01885v7

[19] W. S. Harwood. 1996. The one-minute paper. A communication tool for large lecture classes. *Journal of Chemical Education* 73, 3 (1996), 229–230. https://doi.org/10.1021/ED073P229

[20] Rachael E. Jack, Caroline Blais, Christoph Scheepers, Philippe G. Schyns, and Roberto Caldara. 2009. Cultural Confusions Show that Facial Expressions Are Not Universal. *Current Biology* 19, 18 (9 2009), 1543–1548. https://doi.org/10.1016/j.cub.2009.07.051

[21] Rachael E. Jack, Roberto Caldara, and Philippe G. Schyns. 2012. Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of experimental psychology. General* 141, 1 (2 2012), 19–25. https://doi.org/10.1037/A0023463

[22] Rachael E. Jack, Oliver G.B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences of the United States of America* 109, 19 (5 2012), 7241–7244. https://doi.org/10.1073/PNAS.1200155109

[23] Yang Jiang, Nigel Bosch, Ryan S. Baker, Luc Paquette, Jaclyn Ocumpaugh, Juliana Ma Alexandra L. Andres, Allison L. Moore, and Gautam Biswas. 2018. Expert feature-engineering vs. Deep neural networks: Which is better for sensor-free affect detection? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10947 LNAI (2018), 198–211. https://doi.org/10.1007/978-3-319-93843-1{_}15

[24] Natalya Borisovna Karabuschenko, Aleksandr Vasilevich Ivashchenko, Nina Lvovna Sungurova, and Ekaterina Mihailovna Hvorova. 2016. Emotion Recognition in Different Cultures. *Indian Journal of Science and Technology* 9, 48 (12 2016), 1–17. https://doi.org/10.17485/IJST/2016/V9I48/109085

[25] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (4 2023), 102274. https://doi.org/10.1016/J.LINDIF.2023.102274

[26] Christine Kwon, Darren Butler, Judith Odili Uchidiuno, John Stamper, and Amy Ogan. 2024. Investigating Demographics and Motivation in Engineering Education Using Radio and Phone-Based Educational Technologies. *Conference on Human Factors in Computing Systems - Proceedings* (5 2024). https://doi.org/10.1145/3613904.3642221/ASSETS/HTML/IMAGES/CHI24-331-FIG1.JPG

[27] Agnieszka Landowska, Grzegorz Brodny, and Michal R. Wrobel. 2017. Limitations of Emotion Recognition from Facial Expressions in e-Learning Context. *International Conference on Computer Supported Education* 2 (2017), 383–389. https://doi.org/10.5220/0006357903830389

[28] Blair Lehman, Sidney D'Mello, and Art Graesser. 2013. Who Benefits from Confusion Induction during Learning? An Individual Differences Cluster Analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7926 LNAI (2013), 51–60. https://doi.org/10.1007/978-3-642-39112-5{_}6

[29] Blair Lehman, Sidney D ' Mello, and Art Graesser. 2012. Confusion and complex learning during interactions with computer learning environments. (2012). https://doi.org/10.1016/j.iheduc.2012.01.002

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. [n. d.]. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ([n. d.]). https://doi.org/10.5555/3495724.3496517

[31] Scott M Lundberg, Paul G Allen, and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017). https://github.com/slundberg/shap

[32] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2020 2:1* 2, 1 (1 2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[33] Michalina Manikowska, Damian Sadowski, Adam Sowinski, and Michal R. Wrobel. 2023. DevEmo—Software Developers' Facial Expression Dataset. *Applied Sciences 2023, Vol. 13, Page 3839* 13, 6 (3 2023), 3839. https://doi.org/10.3390/APP13063839

[34] Daniel McDuff, Rana El Kaliouby, Jeffrey F. Cohn, and Rosalind W. Picard. 2015. Predicting Ad Liking and Purchase Intent: Large-Scale Analysis of Facial Responses to Ads. *IEEE Transactions on Affective Computing* 6, 3 (7 2015), 223–235. https://doi.org/10.1109/TAFFC.2014.2384198

[35] Tricia J. Ngoon, David Kovalev, Prasoon Patidar, Chris Harrison, Yuvraj Agarwal, John Zimmerman, and Amy Ogan. 2023. "An Instructor is [already] able to keep track of 30 students": Students' Perceptions of Smart Classrooms for Improving Teaching & Their Emergent Understandings of Teaching and Learning. (7 2023), 1277–1292. https://doi.org/10.1145/3563657.3596079

[36] Xavier Ochoa, Ecuador Federico Domínguez, Ecuador Bruno Guamán, Ecuador Ricardo Maya, Ecuador Gabriel Falcones, Ecuador Jaime Castells, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. 2018. The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and low-Cost sensors. *ACM International Conference Proceeding Series* (3 2018), 360–364. https://doi.org/10.1145/3170358.3170406

[37] Prasoon Patidar, Tricia J Ngoon, Neeharika Vogety, Nikhil Behari, Chris Harrison, John Zimmerman, Amy Ogan, and Yuvraj Agarwal. 2024. Edulyze: Learning Analytics for Real-World Classrooms at Scale. *Journal of Learning Analytics* 11, 2 (8 2024), 297–313. https://doi.org/10.18608/jla.2024.8367

[38] Prasoon Patidar, Tricia J Ngoon, John Zimmerman, Amy Ogan, and Yuvraj Agarwal. 2024. ClassID: Enabling Student Behavior Attribution from Ambient Classroom Sensing Systems. (2024). https://doi.org/10.1145/3659586

[39] Phuong Pham and Jingtao Wang. 2017. Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis. *International Conference on Intelligent User Interfaces, Proceedings IUI* (3 2017), 67–78. https://doi.org/10.1145/3025171.3025186/SUPPL{_}FILE/IUIFP0212-FILE3.MP4

[40] James A. Russell. 1991. Culture and the categorization of emotions. *Psychological Bulletin* 110, 3 (1991), 426–450. https://doi.org/10.1037/0033-2909.110.3.426

[41] Ali Salar and Ali Ahmadi. 2024. Improving loss function for deep convolutional neural network applied in automatic image annotation. *Visual Computer* 40, 3 (3 2024), 1617–1629. https://doi.org/10.1007/S00371-023-02873-3/TABLES/5

[42] Sefik Serengil and Alper Özpınar. 2024. A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. *Bilişim Teknolojileri Dergisi* 17, 2 (4 2024), 95–107. https://doi.org/10.17671/GAZIBTD.1399077

[43] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020* (10 2020). https://doi.org/10.1109/ASYU50717.2020.9259802

[44] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. *7th International Conference on Engineering and Emerging Technologies, ICEET 2021* (2021). https://doi.org/10.1109/ICEET53442.2021.9659697

[45] Maung Thway, Jose Recatala-Gomez, Fun Siong Lim, Kedar Hippalgaonkar, and Leonard W. T. Ng. 2024. Battling Botpoop using GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbots Impact on Learning. (6 2024). https://arxiv.org/abs/2406.07796v2

[46] Judith Uchidiuno, Ken Koedinger, and Amy Ogan. 2021. Teacher Perspectives on Peer-Peer Collaboration and Education Technologies in Rural Tanzanian Classrooms. *Proceedings of 2021 4th ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2021* (6 2021), 14–26. https://doi.org/10.1145/3460112.3471939/ASSETS/HTML/IMAGES/IMAGE5.JPEG

[47] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. (2024).

[48] Keye Xu, Sarah Jo Torgrimson, Remi Torres, Agatha Lenartowicz, and Jennie K. Grammer. 2022. EEG Data Quality in Real-World Settings: Examining Neural Correlates of Attention in School-Aged Children. *Mind, Brain, and Education* 16, 3 (8 2022), 221–227. https://doi.org/10.1111/MBE.12314

[49] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (10 2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342