AffectFusion: A Multimodal Deep Learning Framework for Affective State Recognition

With Applications to Online Education Systems

Samson Oseiwe Ajadalu, Yamoah Attafuah, Hideki Ewan Hill
The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
University of Toronto

***Abstract*—It is difficult for instructors to appropriately adapt lesson elements in a timely manner due to inconsistent and unreliable feedback. We explore a multimodal machine learning approach to estimate a learner's emotional and engagement states from webcam video and microphone audio, allowing for improved lesson adaptation. Our prototypical system has three branches, each predicting one state (emotion or engagement) by observing one modality (video or audio). A 2D convolutional binary classifier of facial emotion is trained by fine-tuning a ResNet-101 model on the DFEW movie dataset, with final test accuracies ranging 76 to 79 percent on each fold. A facial engagement LSTM classifier is trained using landmark methods on the DAiSEE dataset to predict one of four engagement states, with a final test accuracy of 84.96 percent. Finally, a 1D convolutional binary classifier of vocal emotion is trained on the RAVDESS speech and song (audio-only) dataset, with a final test accuracy of 84 percent. We introduce a novel element, late fusion, where the outputs from each model are fused into a high-level "emotion/engagement state" suitable for driving an AI tutor, large language model, or for providing live feedback during human-delivered instruction. Our results indicate that observing a level of engagement and a positive/negative emotion are a promising signal for adaptive tutoring, improving the value of in-class time for students and instructors.**

***Index Terms*—ECE 1513, introduction to machine learning, emotion recognition, engagement detection, multimodal learning, intelligent tutoring, deep learning.**

I. Introduction

The engineering subfield of AI education assistance is a cutting-edge area of human–computer interaction and computer vision. If AI is to be used to provide the best feedback to change lessons according to student needs, it is crucial to have a reliable feed of students' understanding of the course material; the problem is to accurately predict this understanding by perceiving emotion and engagement. Webcam video and microphone audio act as input to our system into three deep learning models: facial emotion, facial engagement, and speech emotion. After the predictions are received, they are fused into a single engagement and satisfaction signal. Non-machine learning solutions rely on inconsistent student feedback such as course surveys, feedback forms, or in-class polls. Existing machine learning solutions usually target the benchmark accuracy of emotion perception on a single dataset, rather than the full integration of the pipeline for educational applications. In this project we demonstrate functional individual models with test accuracies between 75 and 85 percent, and we implement a complete executable prototype demonstrating that the fused model works in real time.

TABLE I
Overview of training datasets

Dataset	Modality	Num. samples	Num. labels (original→new)	Associated Model
DFEW (Part 2)	Video	16,000	7→2	Model 1
DAiSEE	Video	9,000	4→4	Model 2
RAVDESS	Audio	2200	8→2	Model 3
Custom Selfies	video	40	2→2	Model 1 & 2

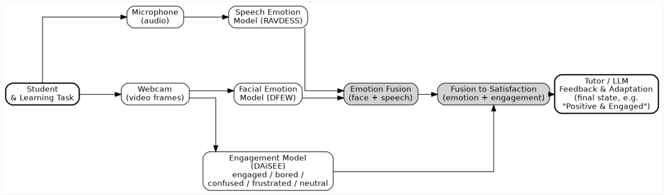


Fig. 1. Pipeline between data, models, fusion, and output

II. System Design

Our primary solution is a multimodal pipeline that fuses predictions from three separately trained models. The webcam provides a continuous stream of facial video, and the microphone provides an intermittent (i.e. only when the subject speaks) stream of speech audio. The video is processed by two models: a facial emotion classifier and a facial engagement classifier. The audio is processed separately by a speech emotion classifier. Their outputs are combined in a simple fusion block that produces a single affective state, which can be fed directly to the education system. Figure 1 depicts this pipeline. Two alternative solutions were considered. One alternative is a single joint audio–visual model that considers both modalities and makes one prediction; this was unfeasible due to the complexity of processing two modalities into one input. Another alternative is a unimodal system that uses only one modality: video or audio; this would not take advantage of all available information (i.e. both modalities) when observing a student. We therefore keep the three-branch late-fusion design which is modular and allows each branch to be improved or replaced independently.

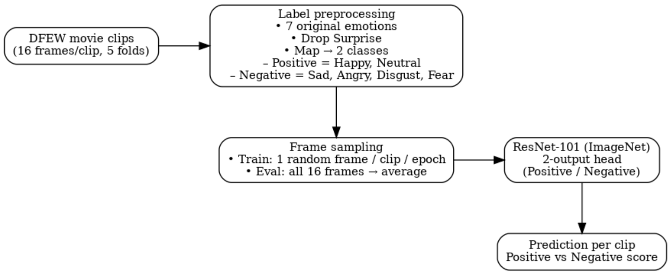

DFEW movie clips
(16 frames/clip, 5 folds)
Label preprocessing
• 7 original emotions
• Drop Surprise
• Map → 2 classes
– Positive = Happy, Neutral
– Negative = Sad, Angry, Disgust, Fear
Frame sampling
• Train: 1 random frame / clip / epoch
• Eval: all 16 frames → average
ResNet-101 (ImageNet)
2-output head
(Positive / Negative)
Prediction per clip
Positive vs Negative score

Fig. 2. DFEW to model 1 pipeline

A. Model 1: Facial emotion model

The facial emotion branch uses the DFEW Part 2 movie dataset [1]. Each clip contains sixteen (224×224) frames and is labeled with one of seven emotions: happy, sad, neutral, angry, surprise, disgust, or fear. For our tutoring scenario we are mainly interested in whether the student looks satisfied or not. We therefore drop all clips labeled surprise and collapse the remaining labels into two classes: positive (happy, neutral) and negative (sad, angry, disgust, fear). This mapping keeps a clear interpretation and leads to almost perfectly balanced positive and negative classes in each split. DFEW provides five train and test folds. For each fold, label mapping is applied as described above, reserving 10% of the train split for validation. We fine-tune a ResNet-101 model pretrained on ImageNet, replacing the final fully connected layer with a two-unit head for positive and negative classification. During training we randomly sample one frame per clip per epoch, reducing compute intensity and serving as temporal data augmentation. During validation and testing we run the model on all sixteen frames of each clip, sum the logits over frames, and one prediction per clip using with the argmax of the softmax. Figure 2 illustrates the pipeline from DFEW to binary prediction with ResNet-101, as well as the label mapping and frame sampling.

B. Model 2: Facial engagement model

The Dataset for Affective States in E-Environments (DAiSEE) is a multi-label video dataset consisting of 10-second clips. To simulate diverse real-world online learning environments, the data was captured through students' webcams, across six distinct locations under three varying illumination settings. The dataset was recorded on 32 females and 80 males aged 18 to 30 years [2]. Each video clip is annotated with four affective states. A notable characteristic of DAiSEE is its severe class imbalance, posing significant challenges for classification. The class distribution of samples is 2268, 853, 8548, and 433, for 'Boredom', 'Confusion', 'Engagement', and 'Frustration' respectively.

1) Preprocessing & feature engineering: The engagement branch uses the DAiSEE dataset [3]. For more information on the dataset, see [2] and [4]. On the DAiSEE authors' advice to simplify the prediction problem and reduce class imbalance [2], the original 4-level ordinal labels (0–3) for each class are first converted into binary labels. For each sample in the dataset, the original levels 0 and 1 are mapped to 0 (low), while levels 2 and 3 are mapped to 1 (high). Every clip sample was then downsampled from 300 frames to 75 frames by selecting every fourth frame to reduce the computational load of processing every single frame while preserving frame order. This resulted in a fixed sequence length of 75 frames per clip. For each frame, facial geometry was extracted using the MediaPipe library, yielding 478 3D (x,y,z) facial landmarks per frame. To ensure continuity, frames in which no face was detected used a forward-filling approach where the most recent valid frame was copied unto the current frame. This correction was only required for 64 out of the total 617,234 frames: a 0.01% fill rate. According to [5], landmark points such as eye corners or the nose tip are "fiducial points" that do not change significantly with different facial expressions and can thus function as reference points in images of faces. To make the model invariant to head position and distance to camera, all landmark points were normalized relative to the nose tip landmark and then scaled using the distance between the outer corners of the eyes, the interocular distance (IOD). Finally, correlation-based feature reduction was applied to remove redundant data (threshold of 0.95), reducing input features from 1,434 to 54 selected features.

2) Data splitting: To ensure reliable evaluation, a subject-independent data split is implemented so that all video clips from any subject were strictly isolated to a single subset of the data. This strategy discourages the model from learning subject-specific facial features instead of generalized emotional expressions. The dataset was partitioned into training (60%), validation (20%), and testing (20%) sets to assess the model's generalization on unseen subjects.

3) Model development: The model selected was a Long Short-Term Memory (LSTM), originally developed by Hochreiter & Schmidhuber [6]. This model has the ability to capture temporal dependencies in each 75-frame sequence of facial features, as well as mitigate the vanishing gradient problem inherent in standard Recurrent Neural Networks. The input shape of the data was 75 frames $\times$ 54 features per sample. The LSTM layer of the model had 64 units, using an input dropout rate of 0.3 to reduce overfitting, followed by a fully connected dense layer of 32 units with ReLU activation and an additional Dropout layer of 0.3. The output layer consists of 4 dense units with sigmoid activation, enabling multi-label binary classification where each emotion is predicted independently for each clip. Two approaches were employed to address class imbalance. First, a weighted binary cross entropy loss was implemented to inverse frequencies of the classes as the loss function's weights to heavily penalize misclassifications of minority classes. Second, a custom data generator was implemented to over-sample minority classes at the start of each epoch, ensuring training batches maintain a balanced class distribution throughout learning.

4) Model training: The model was trained for 40 epochs using the Adam optimizer with an initial learning rate of 0.0005. Additionally, checkpoints were used to save model weights that minimize validation loss, learning rate reduction

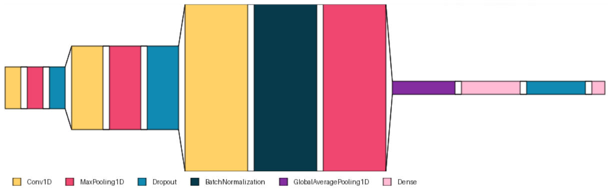


Fig. 3. 1d convolutional audio model 3 architecture

on plateau was used with a factor of 0.5 and patience of 5 to push performance when stalled, and early stopping to halt training if validation performance decreases consistently after 12 epochs. Finally, rather than using the default sigmoid decision threshold of 0.5 for each class, thresholds were tuned per-class post-training by selecting the threshold value that yielded the best performance on the validation set. This ensured the model's sensitivity was tuned to the specific prevalence of each emotion.

C. Model 3: Speech emotion model

To predict emotion from speech, convolutional neural networks were trained on the RAVDESS speech and song audio-only datasets [7]. Loading and preprocessing the data was completed with reference to [8]. For the same reasons as in Section II-A, each of the 8 emotion classes were mapped to one positive or negative emotion, since individual emotions are not explicitly helpful in predicting a student's affective state. Thus we assign each emotion to one of two classes and we retain an acceptable sampling balance. Initially, we attempted to train a 2d convolutional neural network on transformed waveform images of size ranging 48×48 to 512×512 with varying sampling rate. The validation accuracies did not reflect effective learning, and due to the image sizes, the models quickly became too large (beyond 100 MB in size). It was decided that waveforms are not an efficient use of 2d convolution, since there is always a significant part of the image that is empty. This significantly wastes preprocessing time and compute power, explaining the bloated size of the models and inability for the model to learn in any meaningful way. Instead, we switched to a 1d convolutional model that made better use of the audio features without analyzing empty space. To ensure consistency across the data, all files were sampled at 22.050 kHz and normalized to 5 seconds by padding the audio features with zeros. Despite the empty space that we believe was a major flaw in training the 2d convolutional model, this form of padding does not require nearly the same compute power to perform on a large scale. Since the dataset is relatively small of size 2260, we allocated 80% of the dataset to training with an overall $80 - 10 - 10$ split. With an Adam optimizer of learning rate 0.001 and no early stopping callback, we set checkpoints at each epoch while monitoring validation loss and observed a minimum of 0.35. Since the train size is small, we construct a small model that does not overfit quickly: as seen in Figure 3 [9], the architecture consists of three convolutional blocks beginning with a high filter count to observe a large quantity of features and ending with a fully connected layer out to a single neuron with sigmoid activation. The sigmoid output can be interpreted as a confidence in the output being negative (since negative was assigned to class 1).

D. Multimodal fusion and demo

At runtime the system consumes webcam video and microphone audio. The facial emotion model outputs a positive or negative score. The engagement model outputs one of engaged, bored, confused, or frustrated. When speech is detected, the audio model outputs a positive or negative score based on intonation. We maintain a simple satisfaction score driven mainly by the facial emotion branch. The speech prediction serves to boost the facial score in its direction: if both face and voice are positive, the final positive confidence increases. If they conflict, the confidence is reduced. The live demo overlays facial valence, engagement label, and speech-adjusted satisfaction on top of the webcam video.

E. Comparison to alternative solutions

We have discussed up to this point our primary solution of late fusion and will explore two alternative solutions. First, one could have trained each model as a binary classifier on the same emotional classes. This would have been a simpler solution, but also more shallow, not providing enough information and feedback. In this case, the step of late fusion would involve consulting each model separately of their prediction, not making efficient use of the three models. By picking one prediction, information is lost from the others, wasting compute. A second alternative one could explore is a single joint audio-video model: we can not speak to the efficacy or practicality of this solution, as it is an advanced and complicated method that this project did not allow time for. We picked the primary solution because it allows us to use all available modalities while keeping model training and workflow straightforward.

III. Results

A. Model 1: Facial emotion results

For the facial emotion model we report classification accuracy on the official DFEW fold-specific test split. After dropping 'surprised' labels and performing label mapping, the two classes are nearly perfectly balanced in every split, so a single accuracy value is a sufficient metric. In summary, the model has a mean test accuracy of 77% after the 2-class mapping: approximately 76%, 77%, 77%, 79%, and 78% with respect to each of the five folds. Figure 4 shows the training and validation curves for DFEW fold 1. The training loss decreases smoothly and the validation accuracy plateaus after roughly 15 to 20 epochs, indicating that the model converges without severe overfitting. We also test the same model on a small custom selfie dataset of our own faces, roughly balanced between positive and negative expressions. the model reaches about 75% accuracy on this dataset. The model tends to struggle with subtle expressions and poor lighting.

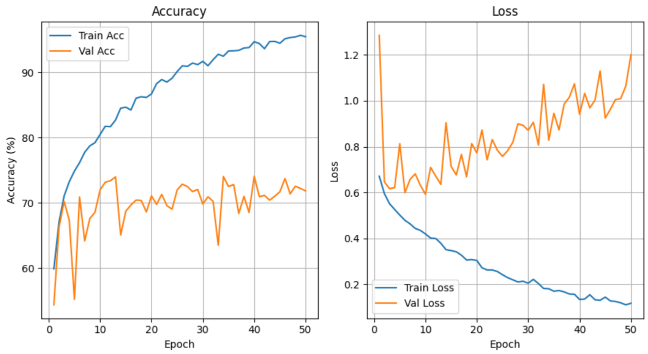


Fig. 4. Model 1: accuracy and loss curves on DFEW fold 1

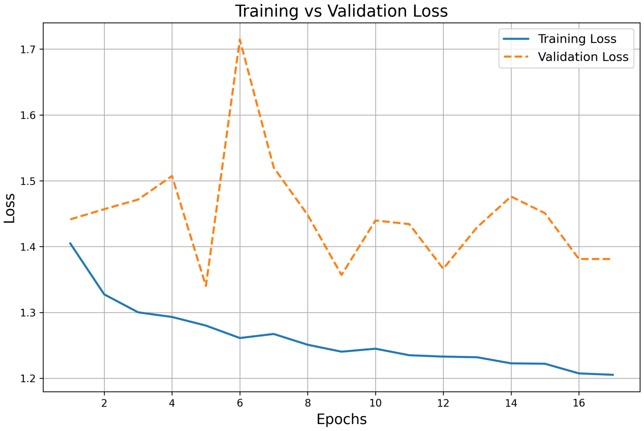


Fig. 5. Model 2: loss curves on DAiSEE

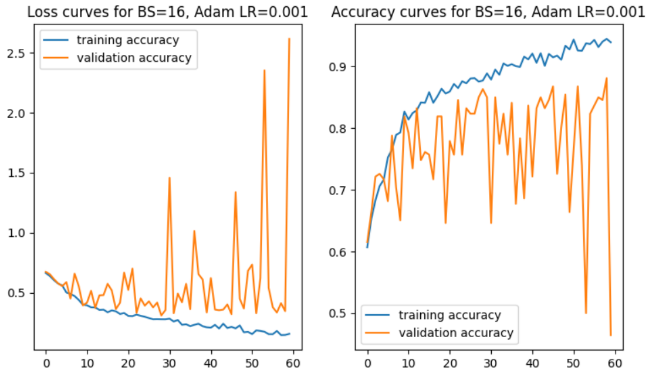
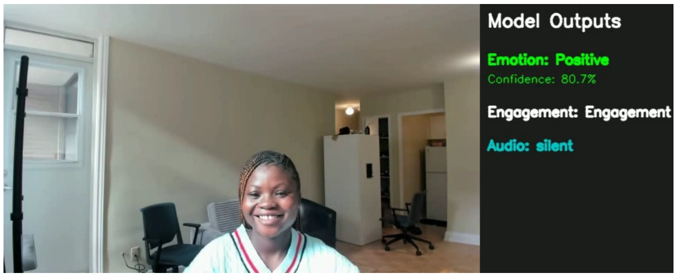


Fig. 6. Model 3: accuracy and loss curves on RAVDESS



Model Outputs

Emotion: Positive
Confidence: 80.7%

Engagement: Engagement

Audio: silent

Fig. 7. Correct affective prediction, presented 27 Nov. 2025.

B. Model 2: Facial engagement results

The results of the model trained on the DAiSEE dataset are detailed in this section. The model achieved an accuracy of 84.96% on the test set; Figure 5 shows the training and validation loss curves of the model until epoch 17, when the training was halted by early stopping. In terms of computational efficiency, the final model architecture is compact, consisting of only 32,676 trainable parameters and a small storage footprint of approximately 419 KB, making it suitable for real-time inference in resource-constrained environments.

C. Model 3: Speech Emotion Results

While training the audio model over 60 epochs, a batch size of 16 was selected due to the small size of the dataset. Accuracy and loss were tracked on the training and validation sets, see Figure 6. Notice that training accuracy is consistently high and training loss is consistently low: these are expected results as the model is learning as a result of observing these data. However, toward the 50th epoch, there is a sharp decrease in validation accuracy, and an equally sharp increase in validation loss. While not desirable, this is also an expected result: with a “sliding window” analysis along the epochs of the validation metrics, it is clear that validation loss increases while accuracy decreases at the 45th epoch. A final test accuracy of 84% was observed. The 2d convolutional model did not perform this well: despite architecture changes, the 2d convolutional model did not outperform random selection.

D. Qualitative analysis of model behaviour

In the full system we observe sensible qualitative behaviour. When the subject smiles and speaks cheerfully, the facial emotion model predicts positive with high confidence, the engagement model predicts engaged, and the audio model predicts positive, producing a strongly positive and engaged state in the fusion output. When the subject looks away, slouches, or deliberately acts bored, the engagement model often switches to bored or confused and the facial emotion score moves toward negative. Short angry or sarcastic utterances can temporarily lower the satisfaction score even if the face remains neutral, showing that the audio branch can influence the fused result. The model’s ability to generalize was apparent in the live demo video presented in class. Figure 7 is a screenshot from this demo, where the webcam feed is overlaid with the current facial valence prediction, the engagement label, and the speech-adjusted emotional satisfaction score.

IV. Concluding Remarks

We have developed a full unified model that predicts affective state, by fusing three individually trained models. Now that the model is trained to detect emotion and engagement, they can be implemented. We have also demonstrated the usage of the model, and this performance can now be used to implement appropriate live lesson changes as desired. Note that since there is a continuous stream video but not audio, the demo of our system detects whether a student is speaking, reducing compute waste, further increasing the efficiency of our system as compared to the other solutions.

References

[1] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, “Dfew: A large-scale database for recognizing dynamic facial expressions in the wild,” 2020. [Online]. Available: https://arxiv.org/abs/2008.05924

[2] A. Gupta, A. D’Cunha, K. Awasthi, and V. Balasubramanian, “Daisee: Towards user engagement recognition in the wild,” 2022. [Online]. Available: https://arxiv.org/abs/1609.01885

[3] ——, “DAiSEE: Dataset for Affective States in E-environments,” Indian Institute of Technology Hyderabad, n.d., accessed: 2025-12-05. [Online]. Available: https://people.iith.ac.in/vineethnb/resources/daisee/index.html

[4] A. Kamath, A. Biswas, and V. Balasubramanian, “A crowdsourced approach to student engagement recognition in e-learning environments,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[5] S. U. Oya Çeliktutan and B. Sankur, “A comparative study of face landmarking techniques,” EURASIP Journal on Image and Video Processing, 2013, accessed: 2025-12-05. [Online]. Available: http://jivp.eurasipjournals.com/content/2013/1/13

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Massachusetts Institute of Technology Press, 1997, accessed: 2025-12-05. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6795963

[7] S. R. Livingstone and F. A. Russo, “Ravdess emotional speech audio,” 2019. [Online]. Available: https://www.kaggle.com/dsv/256618

[8] A. Zanette, “Speech emotion recognition with ConvNet and MFCCs,” Kaggle, Sep. 2021. [Online]. Available: https://www.kaggle.com/code/alessandrozanette/speech-emotion-recognition-with-convnet-and-mfccs

[9] P. Gavrikov and S. Patapati, “visualkeras,” https://github.com/paulgavrikov/visualkeras, 2020.