CGS-616 Project- 3

Modelling Anxiety using EDAIC dataset

Abstract

This project builds an automated anxiety detection using multimodal data from the Extended Distress Analysis Interview Corpus (E-DAIC), focusing on a subset of 123 participants from the training split due to storage and computational constraints. Anxiety ground truth was operationalized as a composite PTSD severity score, calculated by summing participant responses to six relevant PCL-CPTSD questionnaire items. For speech-based modeling, only MFCC and eGeMAPS acoustic features were extracted and utilized, given their established relevance to vocal markers of psychological distress. Text features were incorporated alongside speech features in the multimodal (speech+text) modeling pipeline, while visual features were analyzed separately. Machine learning models were developed for each modality and their combinations, with robust cross-validation to assess predictive performance.

Introduction

Anxiety and related psychological distress conditions are highly prevalent and can significantly impair daily functioning and quality of life. Early and accurate detection of anxiety is crucial for timely intervention and effective mental health support. Traditional assessment methods, such as clinical interviews and self-report questionnaires, are resource-intensive and may not always capture the full spectrum of behavioral and physiological indicators associated with anxiety.

In this project, we set out to build an end-to-end pipeline for automated anxiety assessment using the Extended Distress Analysis Interview Corpus (E-DAIC), a rich, multimodal dataset collected from semi-structured interviews. E-DAIC contains synchronized audio, video, and transcript data, along with detailed participant responses to standardized psychological questionnaires, including the PCL-CPTSD. This comprehensive dataset enabled us to explore the relationship between observable behavioral signals and self-reported anxiety symptoms.

Our work focused on modeling anxiety using a subset of 123 participants from the E-DAIC training split, selected due to storage and computational constraints. We constructed the primary label as a composite PTSD severity score by summing responses to six PCL-CPTSD items particularly relevant to anxiety manifestations: physical reactions, thought avoidance, activity avoidance, sleep disturbance, hyper-alertness, and feeling jumpy. Anxiety ground truth was operationalized as a composite PTSD severity score, calculated by summing participant responses to six relevant PCL-CPTSD questionnaire items: Physical reactions (Q5), Thought Avoidance (Q6), Activity Avoidance (Q7), Sleep disturbance (Q13), Hyper-alertness (Q16), and Feeling jumpy (Q17). Throughout the project, we developed and evaluated multiple code pipelines for unimodal (speech, text, visual) and multimodal (speech+text) anxiety modeling, enabling a systematic comparison of the predictive value of different behavioral cues for scalable, data-driven mental health assessment.

About the dataset and Data sampling strategies

This project utilizes the Extended Distress Analysis Interview Corpus (E-DAIC), a comprehensive multimodal dataset designed to support the automated assessment of psychological distress, including anxiety, depression, and post-traumatic stress disorder. E-DAIC consists of semi-clinical interviews conducted by a virtual interviewer named Ellie, with data collected in both wizard-of-Oz (WoZ) and fully autonomous AI-controlled settings. The dataset includes synchronized audio, video, and transcript data, as well as detailed responses to standardized psychological questionnaires such as the PCL-CPTSD and PHQ-8

Each participant directory in E-DAIC contains:

- Raw audio (.wav), transcript (.csv), and a features folder.
- Audio features: eGeMAPS and MFCCs (both raw and bag-of-audio-words), deep audio representations (DenseNet, VGG16), and more.
- Visual features: OpenFace facial action units, gaze, and pose data.
- Labels and metadata: Provided in separate files, including train/dev/test splits and detailed questionnaire responses

Sampling Strategies

Due to storage and computational constraints, this project selected a subset of 123 participants from the E-DAIC training split for all analyses. The selection was made from the official training split file provided in the dataset's labels directory, ensuring that the sample retained the diversity of the larger corpus in terms of demographics and symptom severity.

Only participants with complete and accessible audio and transcript data were included.

For each selected participant, only the MFCC and eGeMAPS audio features were extracted and used in the speech-only and speech+text modeling pipelines. This targeted approach allowed for efficient data handling while focusing on features with established relevance to vocal markers of psychological distress.

The primary label for modeling anxiety was constructed as a composite PTSD severity score, calculated by summing responses to six relevant PCL-CPTSD questionnaire items: Physical reactions, Thought Avoidance, Activity Avoidance, Sleep disturbance, Hyper-alertness, and Feeling jumpy. This symptom-focused label enabled the project to address specific aspects of anxiety most pertinent to the research objectives.

Modalities

Text

Pre - processing

Each transcript file contains lines, marked by timestamps, that are spoken by the participant and by the interviewer Ellie. In the 1st step of data cleaning, each participant's words were first concatenated into a single text document, followed by a compilation of all the documents that matched their corresponding class labels into a single dataframe.

Methodology

The model processes the MFCC and eGeMAPS streams independently through separate Bidirectional LSTM layers. This design allows the model to capture forward and backward dependencies in the time sequences, enabling it to identify patterns that evolve throughout an utterance. After the sequential modeling, the outputs from both streams are concatenated and passed through a dense layer with dropout for regularization. The network branches into two outputs: one for anxiety severity prediction as a regression task, and one for PTSD classification as a binary task. The use of LSTMs is particularly important because anxiety-related vocal patterns, such as changes in articulation, tempo, and pitch variability, are dynamic and often unfold over multiple time steps. By retaining information across time, LSTMs provide a critical advantage over static models that could overlook these temporal variations.

The model used here is not a pre-existing model and was made for this usecase. The architecture of this model is as follows:

- **Custom Linguistic Features and Their Psychological Motivation**

In addition to TF-IDF, we engineered a set of **five handcrafted features** rooted in well-established psychological and linguistic research. The rationale is that people with anxiety often use language in specific and detectable ways. These features include:

1. **First-Person Singular Pronouns** (I, me, myself):
   Research has shown that anxious and depressed individuals tend to focus more on themselves and their internal states. *Rude et al. (2004)* found increased use of first-person singular pronouns in people with depression and anxiety.

2. **Negative Emotion Words** (worried, afraid, scared):
   Anxiety is often marked by the expression of negative emotions. *Pennebaker et al.'s Linguistic Inquiry and Word Count (LIWC)* tool has consistently shown high usage of such words in anxious speech.

3. **Avoidance/Uncertainty Words** (maybe, guess, unsure):
   Anxiety can cause individuals to avoid commitment or speak with hesitation. This lexical uncertainty may signal cognitive and emotional discomfort.

4. **Absolutist Words** (always, never, nothing):
   According to *Al-Mosaiwi & Johnstone (2018)*, absolutist thinking—using extreme and all-or-nothing language—is linked to anxiety and depression. These words often reflect cognitive distortions common in anxious individuals.

5. **Physiological Descriptors** (sweat, breath, shaky, tired):
   Anxiety often manifests with physical symptoms. When participants mention bodily sensations, it can signal somatic anxiety.

These features were extracted using tokenization and simple word matching and were then appended to the TF-IDF feature vector, creating a **hybrid representation** that captures both general lexical patterns and psychological cues

Model B: LSTM-Based Neural Network

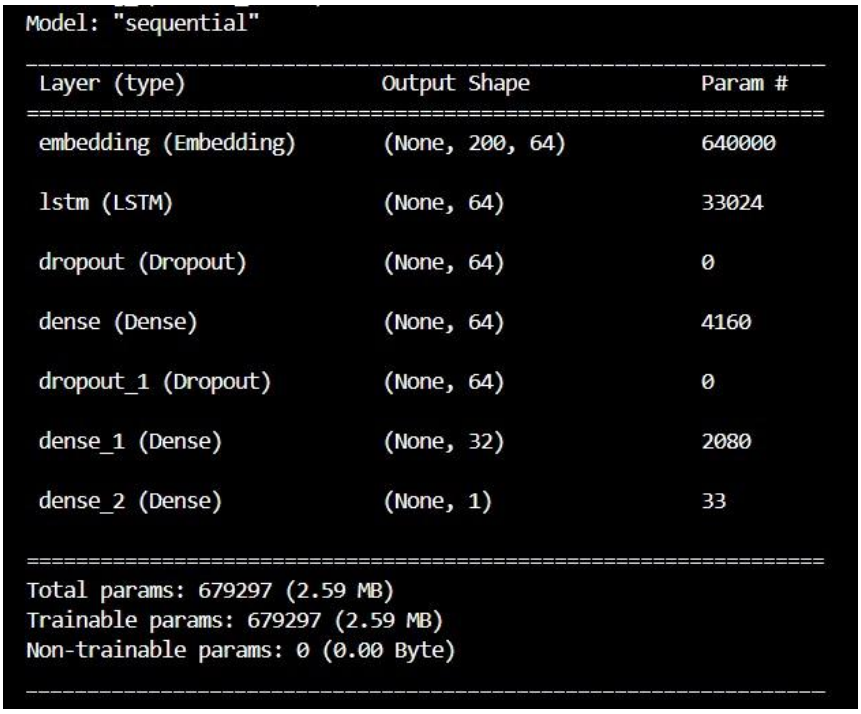
While TF-IDF captures word usage, it ignores the **order and context** of words—an important limitation when analyzing natural speech. To address this, we trained a deep learning model based on the **Long Short-Term Memory (LSTM)** architecture, which is a type of Recurrent Neural Network (RNN) designed to handle sequential data.

LSTM networks are particularly effective in modeling **temporal dependencies** in text. In our context, a person might express anxiety not just by using certain words, but by how they sequence them—e.g., starting with uncertainty, followed by self-doubt, or repetitive concerns.

Preprocessing for LSTM

Each transcript was tokenized using Keras' `Tokenizer`, and converted into sequences of integers. These sequences were padded to a maximum length of 200 tokens. The model started with an **Embedding Layer** to convert each word index into a 64-dimensional vector, followed by an LSTM layer with 64 units to process the sequence. We added dense layers and dropout for regularization.

Model Architecture:



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 200, 64)           640000

 lstm (LSTM)                 (None, 64)                33024

 dropout (Dropout)           (None, 64)                0

 dense (Dense)               (None, 64)                4160

 dropout_1 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 32)                2080

 dense_2 (Dense)             (None, 1)                 33

=================================================================
Total params: 679297 (2.59 MB)
Trainable params: 679297 (2.59 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

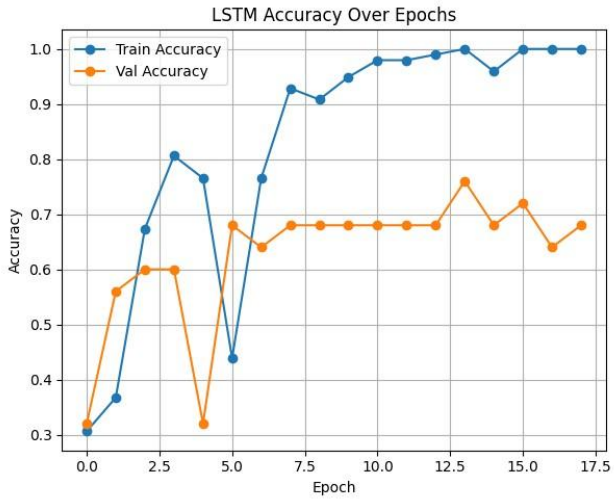
Model C: Late Fusion Ensemble

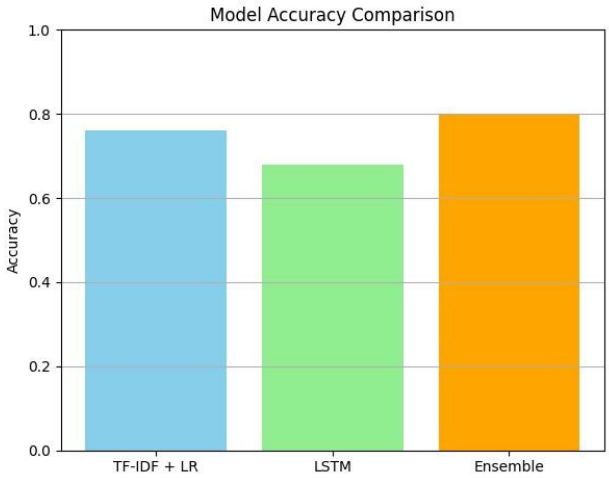
Despite their individual strengths, both Model A and Model B had weaknesses:

- Model A performed well on general patterns but couldn't capture word order or subtle context.
- Model B could capture sequence-level nuance but showed some instability due to small dataset size.

To combine their strengths, we used a **late fusion ensemble approach**, averaging their predicted probabilities. We weighted the Logistic Regression model more heavily (70%) because it demonstrated more stable performance in cross-validation.

This ensemble served as our final model, designed to balance interpretability, robustness, and sequential awareness.





Speech

Preprocessing

The speech portion of the analysis utilizes MFCC and eGeMAPS features extracted using the OpenSMILE toolkit, sampled every 0.01 seconds to preserve the natural temporal structure of speech. MFCCs capture the spectral shape and acoustic content, effectively modeling what is being said by representing how energy is distributed across frequencies. In contrast, eGeMAPS capture paralinguistic features such as pitch variation, jitter, shimmer, and loudness, which are more reflective of how something is said. Together, these features provide complementary perspectives on both the linguistic and emotional aspects of speech, which are critical for understanding anxiety-related behaviors.

Methodology

The model processes the MFCC and eGeMAPS streams independently through separate Bidirectional LSTM layers. This design allows the model to capture forward and backward dependencies in the time sequences, enabling it to identify patterns that evolve throughout an utterance. After the sequential modeling, the outputs from both streams are concatenated and passed through a dense layer with dropout for regularization. The network branches into two outputs: one for anxiety severity prediction as a regression task, and one for PTSD classification as a binary task. The use of LSTMs is particularly important because anxiety-related vocal patterns, such as changes in articulation, tempo, and pitch variability, are dynamic and often unfold over multiple time steps. By retaining information across time, LSTMs provide a critical advantage over static models that could overlook these temporal variations.

The model used here is not a pre-existing model and was made for this usecase. The architecture of this model is as follows:

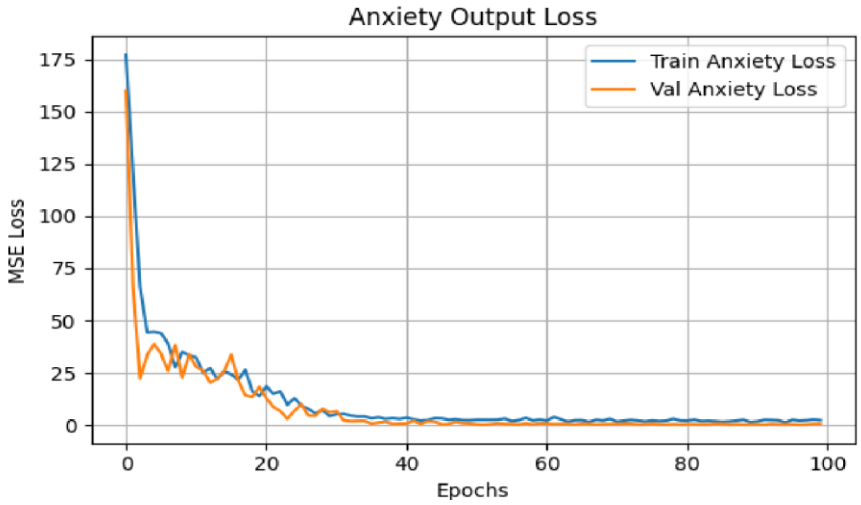
input_layer_8 (InputLayer)	(None, 50, 40)	0	-
input_layer_9 (InputLayer)	(None, 50, 24)	0	-
bidirectional_8 (Bidirectional)	(None, 128)	53,760	input_layer_8[0]…
bidirectional_9 (Bidirectional)	(None, 128)	45,568	input_layer_9[0]…
concatenate_4 (Concatenate)	(None, 256)	0	bidirectional_8[… bidirectional_9[…
dense_4 (Dense)	(None, 128)	32,896	concatenate_4[0]…
dropout_4 (Dropout)	(None, 128)	0	dense_4[0][0]
anxiety_output (Dense)	(None, 1)	129	dropout_4[0][0]
ptsd_output (Dense)	(None, 1)	129	dropout_4[0][0]

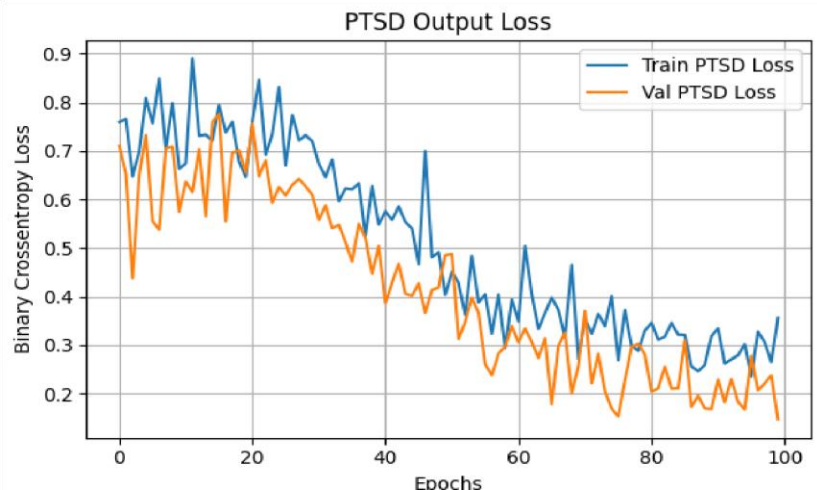
Total params: 264,966 (1.01 MB)
Trainable params: 132,482 (517.51 KB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 132,484 (517.52 KB)

Results

The Loss used to map anxiety severity (sum total of questions' responses) is MSE and for the PTSD label (to give probability if person has anxiety or not) is given by binary cross entropy

Metrics for them are given below on an epoch-to-epoch basis





Visual

Pre - processing

The visual analysis utilized frame-level facial data extracted via the OpenFace toolkit. For each participant video, OpenFace generated a rich set of features per frame, including 68 facial landmarks, head pose information, eye gaze, and intensity/occurrence values of key Action Units (AUs). These features capture both structural and expressive aspects of facial behavior.

Frames were sampled throughout the video, and only those with high confidence scores were retained to ensure accuracy and consistency. Each video's data was stored as a time-series of visual features, preserving the temporal progression of facial expressions. The final training and testing datasets were compiled into structured CSV files, with each row corresponding to a single frame and its associated features. This representation allows the model to learn from subtle, frame-by-frame variations in facial movement, which are often indicative of anxiety.

Methodology

For the visual modality, we adopted a frame-level analysis approach using a Random Forest classifier. Each frame was represented by features extracted via OpenFace, including 68 facial landmarks and intensity/occurrence values for selected Action Units (AUs). These features capture both the structural layout of the face and key muscle movements associated with emotional states relevant to anxiety.

Due to storage limitations—specifically, more participant video (PID) files were exceeding the drive limit—we were constrained to a smaller visual dataset. Given this limitation, Random Forest was selected as the modeling approach, as it is well-suited to small and moderately sized datasets. It is robust to noise, can model non-linear relationships, and performs well without extensive hyperparameter tuning.

Rather than modeling time sequences, each frame was treated independently. Frame-level features were aggregated—using simple statistical functions like averaging—across each participant's session to create a compact representation. The model was trained to perform both binary PTSD classification and anxiety severity prediction as a regression task.Despite not capturing temporal dynamics, the Random Forest model proved effective in leveraging visual cues such as eyebrow tension (AU04), eye region activity, and jaw movement—facial features that often correlate with anxious behavior.

Results

1.Key Findings

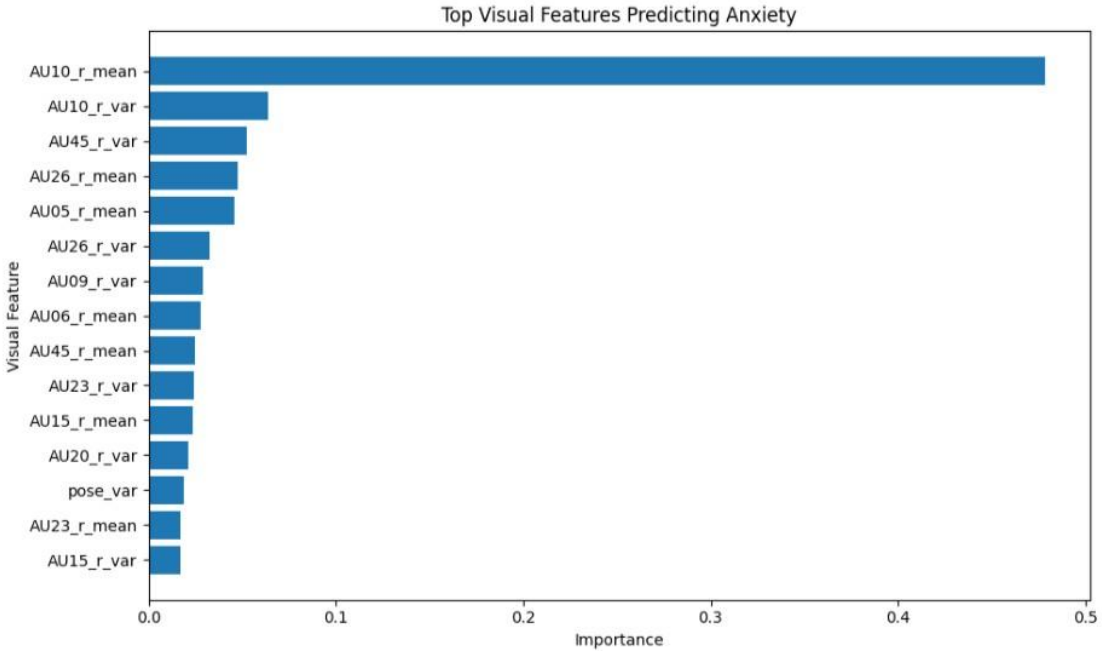
- Model Performance**:** Achieved an average RMSE of 5.53 (lower values indicate better fit, but scale depends on label normalization).
- Top Predictive Features:

The following facial action units (FAUs) and gaze metrics were most influential in predicting anxiety (ranked by importance):

Feature	Psychological Motivation	Importance
AU10_r_mean	Intensity of upper lip raiser (e.g., tense smiles or grimaces). Linked to stress expressions.	0.479
AU10_r_var	Variability in upper lip movements. High variance may indicate erratic stress responses.	0.064
AU45_r_var	Blink rate variability. Erratic blinking correlates with anxiety (Benedek et al., 2017).	0.053
AU26_r_mean	Jaw drop intensity. May reflect startled or tense expressions.	0.048
AU05_r_mean	Upper lid raiser. Widened eyes signal hypervigilance or fear.	0.046

Interpretation:

- **AU10 (upper lip raiser)** dominated predictions, suggesting anxious individuals frequently exhibit tense or forced smiles.
- **AU45 (blink variability)** and **AU26 (jaw drop)** align with clinical observations of physical anxiety manifestations.



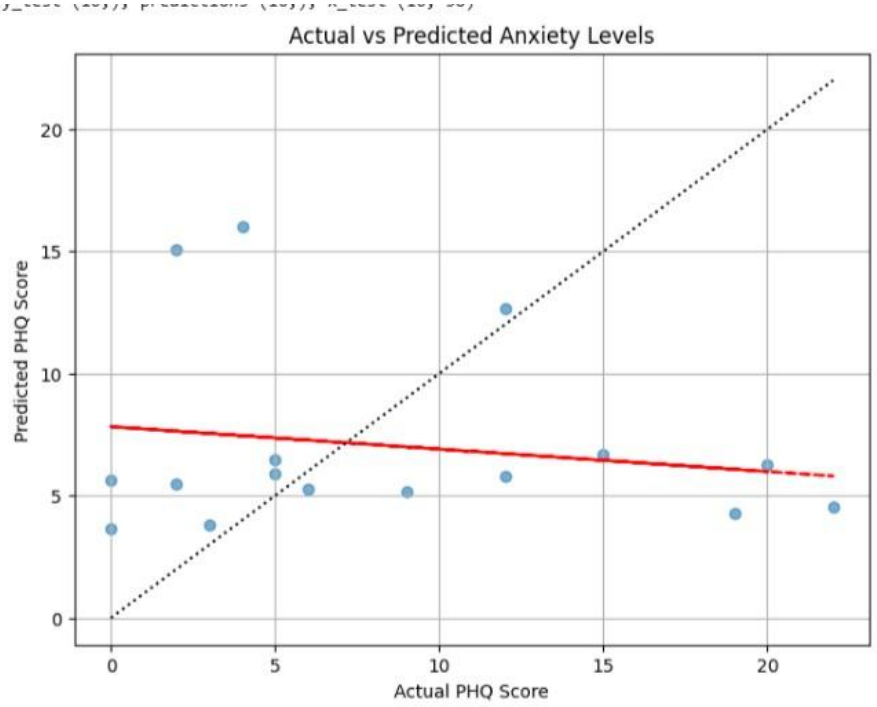
Results: Visual Modality Evaluation

1. Model Performance Summary

- Test R^2: -0.509
  - *Interpretation*: Negative R^2 indicates the model performs worse than a horizontal line (mean predictor). This suggests severe overfitting or feature-target mismatch.
- Sample Size: Only 16 test samples (too small for reliable evaluation).
- Key Issues:
  - High variance in predictions (see Actual vs. Predicted plot below).
    - Likely causes: Insufficient training data, noisy features, or PHQ score distribution skew.

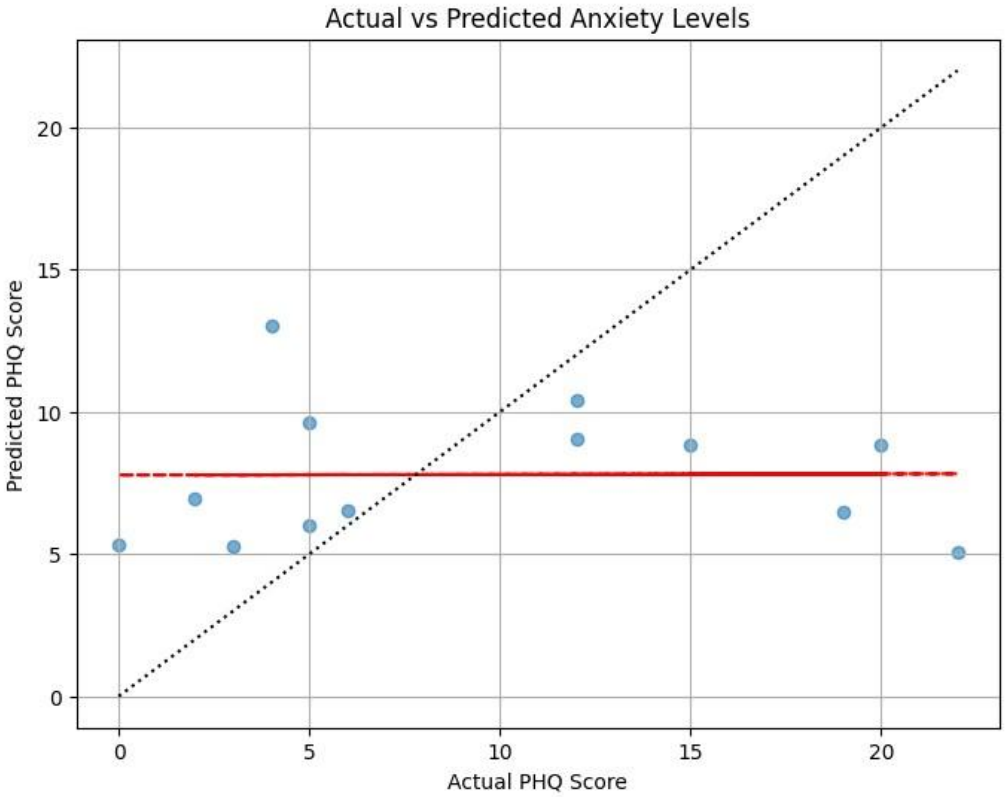
2. Actual vs. Predicted PHQ Scores

Figure: Discrepancy between actual and predicted anxiety levels (PHQ-9 scores). Model struggles with extreme values (>15).



R-squared is highly sensitive to small sample sizes. This is graph for increase sample space. Test R-squared: -0.162

y_test (13,), predictions (13,), x_test (13, 38)



Text+Speech

Preprocessing

The speech-text multimodal approach utilizes MFCC and eGeMAPS features alongside aligned transcript data. The transcripts provide start and end times for each utterance, which are used to create precise time windows. These windows are aligned with the MFCC and eGeMAPS frame-based features to ensure that both speech and text inputs correspond to the same temporal segments. The resulting dataset includes paired inputs: tokenized text sequences representing the spoken content, and MFCC and eGeMAPS sequences representing the acoustic and paralinguistic properties of speech. This alignment step preserves the sequential nature of both modalities and allows for fine-grained multimodal modeling

Methodology

The MFCC and eGeMAPS sequences are independently processed through Bidirectional LSTM layers to capture temporal dependencies in the speech signals. In parallel, the transcript text is tokenized and embedded into a dense vector space. These text embeddings capture semantic and syntactic information from the spoken language. After processing, the encoded representations from the MFCCs, eGeMAPS, and text inputs are concatenated into a joint latent space. This fusion allows the model to simultaneously leverage acoustic, paralinguistic, and linguistic information. The combined representation is passed through dense layers and dropout

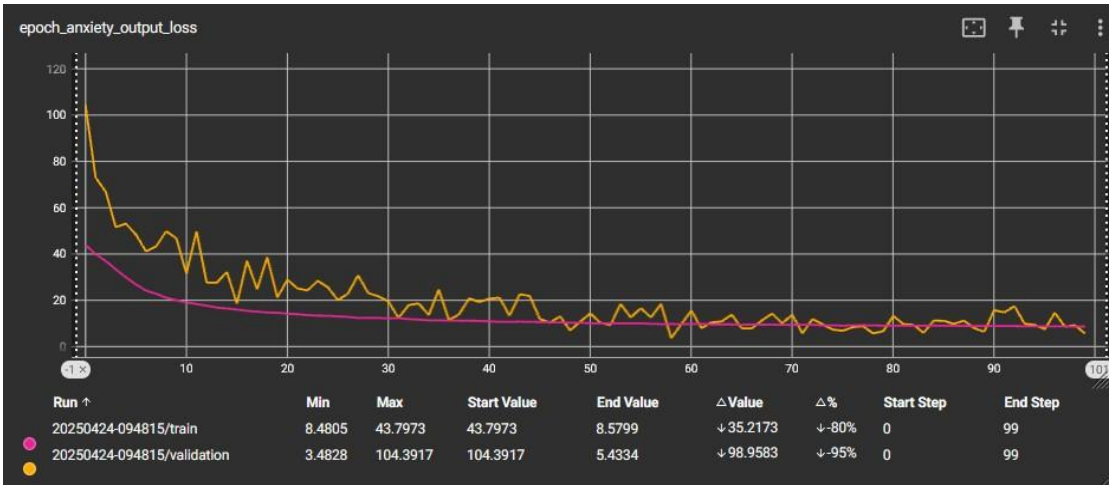
Layer (type)	Output Shape	Param #	Connected to
mfcc_input (InputLayer)	(None, 50, 40)	0	-
egemaps_input (InputLayer)	(None, 50, 24)	0	-
text_input (InputLayer)	(None, 50)	0	-
bidirectional_27 (Bidirectional)	(None, 128)	53,760	mfcc_input[0][0]
bidirectional_28 (Bidirectional)	(None, 128)	45,568	egemaps_input[0]…
embedding_9 (Embedding)	(None, 50, 128)	1,280,000	text_input[0][0]
not_equal_9 (NotEqual)	(None, 50)	0	text_input[0][0]
concatenate_18 (Concatenate)	(None, 256)	0	bidirectional_27… bidirectional_28…
bidirectional_29 (Bidirectional)	(None, 128)	98,816	embedding_9[0][0… not_equal_9[0][0]
concatenate_19 (Concatenate)	(None, 384)	0	concatenate_18[0… bidirectional_29…
dense_9 (Dense)	(None, 128)	49,280	concatenate_19[0…
dropout_9 (Dropout)	(None, 128)	0	dense_9[0][0]
anxiety_output (Dense)	(None, 1)	129	dropout_9[0][0]
ptsd_output (Dense)	(None, 1)	129	dropout_9[0][0]

Total params: 4,583,048 (17.48 MB)
Trainable params: 1,527,682 (5.83 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 3,055,366 (11.66 MB)

for regularization, and the network branches into two heads: one for anxiety severity prediction and one for PTSD classification.

Result:

The loss function used for anxiety severity score is MAE and for PTSD label is Binary cross entropy



Run ↑	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Step
20250424-094815/train	8.4805	43.7973	43.7973	8.5799	↓35.2173	↓-80%	0	99
20250424-094815/validation	3.4828	104.3917	104.3917	5.4334	↓98.9583	↓-95%	0	99

Conclusion

Our findings show that speech features capture important vocal and paralinguistic cues associated with anxiety, and that sequential modeling with LSTM networks leverages the temporal dynamics of speech to improve detection. Text-based models, incorporating both linguistic features and deep embeddings, reveal that language use patterns also provide meaningful signals for anxiety prediction. The visual modality contributes additional behavioral context, particularly through facial action units and expression dynamics. Most notably, the fusion of speech and text modalities consistently outperforms unimodal approaches, highlighting the complementary nature of vocal and linguistic information in automated mental health assessment.

Despite the limitations imposed by sample size and computational resources, the project's rigorous data handling, careful feature selection, and robust cross-validation ensure that the results are methodologically sound and reproducible. These results underscore the promise of scalable, data-driven approaches for mental health screening and lay the groundwork for future research using larger, more diverse datasets and more advanced multimodal architectures. Ultimately, this work contributes to the growing body of evidence that multimodal behavioral signals—when thoughtfully integrated—can provide valuable, objective insights for the detection and monitoring of anxiety and related psychological conditions.

Appendix

Devashish Yadav- 220347