DES646 Group PROJECT

Depression Detection Using Various Features

Group Members

Name	Roll No.	Email
Aditya Khandelwal	230067	adityakh23@iitk.ac.in
Capriate Yadav	230309	cyadav23@iitk.ac.in
Laksh Bansal	230595	lakshb23@iitk.ac.in
Mihir Mandloi	230650	mihirm23@iitk.ac.in
Tanishq Saraf	231077	stanishq23@iitk.ac.in

Indian Institute of Technology Kanpur

09 November 2025

Contents

1. Dataset Used and Resources - https://dcapswoz.ict.usc.edu/wwwedaic/

About the dataset:

•Total of 275 participants: 170 male and 105 female, aged 18-69 of which 66 and 87 are diagnosed with depression and PTSD (Post Traumatic Stress Disorder) respectively.

•Common data contains participant_id, gender, age, question-wise scores on Depression and PTSD scales and whether an individual is diagnosed for depression or PTSD.

•Multimodal visual features, audio .wav files and participant response transcripts available.

1) For Heuristic based methods = Pose_gaze_AUs.csv, metadata mapped.csv all participants
2) For Text + Audio = transcript.csv, egemaps.csv , mfcc.csv, metadata mapped.csv for all participants
3) For Facial = Pose_gaze_AUs.csv, mfcc.csv, generated transcripts (Using Whisper V3, metadata mapped.csv) for 105 participants
4) MultiModal (Facial + Text + Audio) = The above used data was converted to frame level features and segments were aggregated to combine the data containing all 3 modalities

2. Attempt to Detect Depression through Heuristic Analysis

Firstly, we attempted to investigate whether facial visual features extracted from the dataset could be indicative of depressive tendencies by applying simple heuristics based on psychological literature on facial expressions. For this analysis, we referred to the paper *Automatic Identification of Depression Using Facial Images with Deep Convolutional Neural Networks* [1] for guidance and inspiration.

2.1. Methodology

We used Action Units (AUs) and gaze data from the facial features CSV files in the DAIC-WOZ dataset. The heuristics were formulated based on empirical patterns observed in psychological research on depression-related facial behavior in the referred paper.

The following heuristics were tested:

Visual Trait	CSV Feature(s)	Interpretation / Heuristic Rule
Mouth angle downward (' Ω ')	AU15 r (Lip corner depressor)	High AU15 r → Suggests sadness
Tight eyebrows	AU04 r, AU01 r	High values → Emotional distress or frustration
Dull/vacant eyes	gaze angle X, gaze angle Y	Low movement variance → Reduced focus, disen-gagement
Reduced blinking	AU45 r (Blink)	Low frequency → Fatigue or emotional numbing
Tearful expression	AU01 r, AU04 r, AU45 r	Co-activation → Tearfulness, sadness
Reduced smiling	AU12 r (Lip corner puller)	Low activity → Decreased positive affect
Composite sadness score	Combination of above AUs	High AU15, AU01, AU04 + low AU12 → Indicative of depressive state

Table 1: Heuristic Framework for Depression Detection Using Facial Visual Features

2.2. Various Approaches to test heurestics

2.2.1. Approach 1: Visual Analysis via Plots

We implemented feature extraction and visualized the relevant heuristic indicators across all participants to assess their validity. However, the results showed no consistent or distinguishable visual patterns that could effectively differentiate between depressed and non-depressed individuals.

2.2.2. Approach 2: N-Condition Thresholding

This method identifies depressive indicators in each frame using seven facial cues:

If at least N of these conditions are met in a frame, it is marked as depressed. A participant is classified as depressed if more than 40% of frames are depressed. We tested for $N = 2, 3, 4$.

2.2.3. Approach 3: Inverse Logic

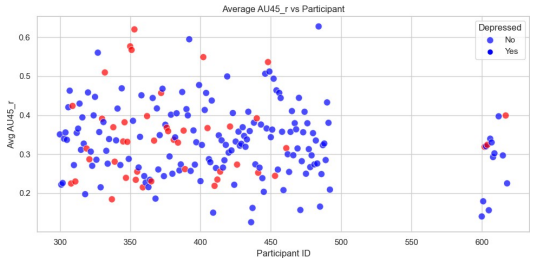
This stricter method assumes that the absence of depressive cues (e.g., smiling, blinking, lively gaze) indicates non-depression. A frame is marked as depressed if it satisfies none of the non-depressed conditions. A participant is labeled as depressed if more than 30% of frames are considered depressed.

2.3. Evaluation

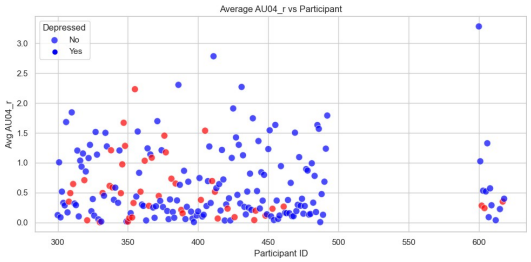
Below is the confusion matrix for the result of approach 2.

Table 2: Confusion Matrix for Depression Detection

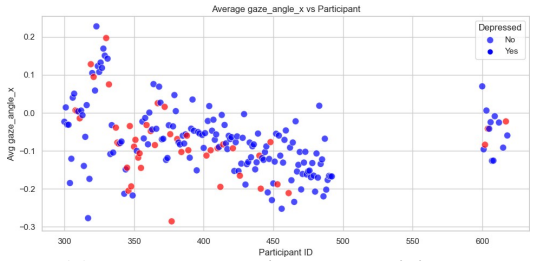
Actual \Predicted	**Not Depressed (0)**	**Depressed (1)**
Not Depressed (0)	114	43
Depressed (1)	35	10



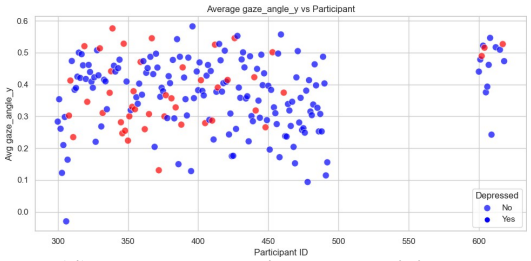
(a) Blink frequency



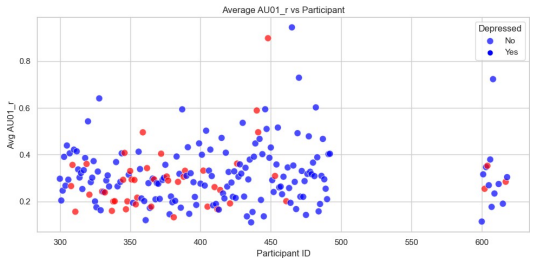
(b) Brow Lowerer



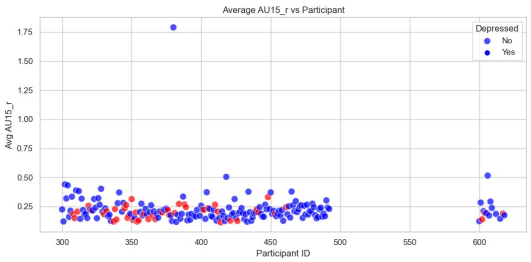
(c) Avg. Gaze Angle X vs Participant



(d) Avg. Gaze Angle Y vs Participant



(e) Inner Brow Raiser



(f) Lip Corner Depressor

Figure 1: Visualizations of heuristic features extracted for each participant. No distinct or consistent trends were observed between depressed and non-depressed groups.

2.4. Conclusion

While conceptually sound, these facial heuristics alone do not yield robust predictive power for depression detection in the DAIC-WOZ dataset.

3. Text-Based Depression Detection Report

3.1. Goal

This part of the project focuses on detecting depression in individuals using conversational transcripts derived from interview settings. By analyzing the linguistic patterns, word usage, and semantic content of spoken language, the model aims to identify markers indicative of depressive symptoms. Emphasis is placed not only on classification accuracy but also on interpretability , enabling insights into which language features contribute to predictions.

3.2. Feature Selection Strategy

We are using text-only features that are grouped into 3 broad categories:

- a. Empath Lexicon Features

The Empath Lexicon is a tool used to analyze the emotional content of text by assigning normalized scores to various semantic categories based on word usage. In the context of depression detection, four specific Empath categories are especially important:

1. **sadness** - Measures how frequently the speaker uses words related to grief, despair, or unhappiness.
2. **positive_emotion** - Captures the use of optimistic or joyful language; lower scores may correlate with depressive states.
3. **anger** - Reflects expressions of irritation, rage, or hostility; high levels can be linked to emotional dysregulation.
4. **death** - Tracks mentions of mortality, dying, or related topics, which can be strong indicators of depressive or suicidal thoughts.

How It Works:

- Empath uses a predefined lexicon of words grouped into semantic categories.
- These raw counts are normalized, so that scores can be compared across individuals even if their transcripts differ in size.
- The resulting scores give a quantitative representation of the participant's emotional tone.

Why It Matters for Depression Detection:

People with depression may unconsciously use more language related to sadness and death and less related to positive emotion. By quantifying these tendencies, Empath provides interpretable indicators of mental health status.

- b. Lexical and Syntactic Features

Lexical and syntactic features focus on the structure, complexity, and expressiveness of a person's language. These features are important because depression often affects cognition and language use, and analyzing them can reveal subtle signs of mental distress.

Here's a breakdown of the key features:

- **1. Average Word Length**

- **What it is:** The average number of characters (or syllables) per word. Longer words often reflect a more sophisticated or complex vocabulary.
- **Why it matters:** People experiencing depression might use simpler or shorter words due to cognitive fatigue, reduced focus, or lower mental energy.

- **2. Type-Token Ratio (TTR)**

$$\text{TTR} = \frac{\text{Total number of words (tokens)}}{\text{Number of unique words (types)}} \quad (1)$$

A high TTR means the speaker is using a diverse vocabulary indicating flexibility and richness in thinking.

- **Why it matters:** Depression is often associated with repetitive or narrow language use, resulting in a lower TTR.

- **3. Average Utterance Length**

- **What it is:** The average number of words per spoken segment or sentence. Longer and logical speech usually suggests fluent thinking and organized speech.
- **Why it matters:** Short or fragmented utterances may reflect difficulty in organizing thoughts, low motivation, or emotional strain, i.e., common symptoms in depression.

Together, these features provide a linguistic window into a person's cognitive state. For example:

- Someone who is depressed may show low TTR, short utterances, and simpler vocabulary, all pointing to cognitive slowing or emotional distress.
- In contrast, higher scores could reflect mental clarity, engagement, and emotional well-being.

- c. Sentiment Analysis using VADER

As part of this project, features were extracted from participant interview transcripts to help identify linguistic markers of depression. One such feature set was derived using VADER (Valence Aware Dictionary and sEntiment Reasoner) — a rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media and conversational text.

VADER is particularly well-suited for this application due to its ability to handle:

- Informal language, including slangs
- Polarity shifters such as negations ("not good") and intensifiers ("very happy")

VADER provides four sentiment scores for any given text:

- **pos** – Proportion of positive sentiment in the text
- **neu** – Proportion of neutral sentiment
- **neg** – Proportion of negative sentiment
- **compound** – A normalized, weighted composite score ranging from -1 (most negative) to +1 (most positive), reflecting overall sentiment

These scores capture the emotional tone of language at both the sentence and document level. In the context of depression detection, transcripts characterized by consistently low compound scores and higher proportions of negative or neutral sentiment may indicate symptoms of low mood or emotional disengagement. By quantifying emotional polarity through VADER, we gain a scalable and interpretable metric that contributes to both real-time classification and post-hoc analysis of language use.

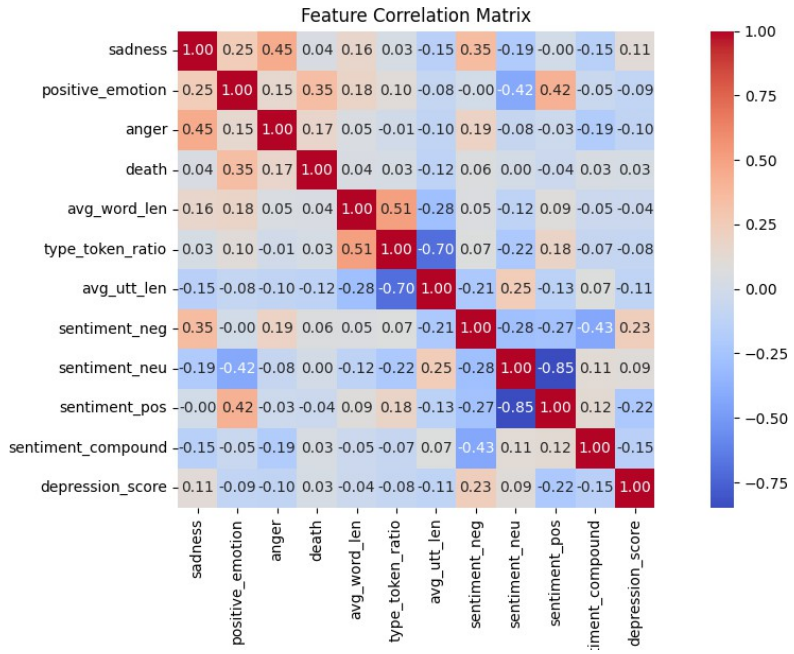


Figure 2: Co-Relation data of selected features

- d. Depression Label (Target)

This is the ground truth label used for supervised learning, based on the PHQ-8 score, a clinically validated questionnaire for measuring depression severity.

- **PHQ-8** $\geq 10 \rightarrow$ **Depressed** $\rightarrow$ **Label = 1**
- **PHQ-8** $< 10 \rightarrow$ **Not Depressed** $\rightarrow$ **Label = 0**

This binary label is used as the target variable when training machine learning models to classify or predict depression based on linguistic and emotional features.

3.3. 3. Approach and Methods

- Step 1: Preprocessing

Participant transcripts are merged into a single document per person. Features are extracted using startegy given above , missing values are imputed, and all features are standardized.We have used **all the participants data** for this model.

- Step 2: Class Imbalance Handling

In our dataset, there exists a significant imbalance between the number of depressed and non-depressed participants, with the depressed class being underrepresented. This imbalance can lead to biased model predictions that favor the majority class and result in poor recall for the minority (depressed) class.

To mitigate this issue, we applied SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic examples of the minority class by interpolating between existing data points. This helps in balancing the dataset and providing the model with more representative samples of depressed participants.

Before applying SMOTE, the model exhibited low recall for depressed cases, indicating a failure to correctly identify many instances of depression. After applying SMOTE, we observed improvements in both recall and overall classification accuracy, demonstrating the effectiveness of this technique in handling class imbalance and enhancing model performance.

- Step 3: Model Selection

A Logistic Regression model was ultimately chosen for this task due to its strong interpretability, which is crucial for understanding the contribution of individual linguistic features to depression prediction.

To handle class imbalance, the parameter classweight='balanced' was used, ensuring that the model gives appropriate attention to the minority (depressed) class during training. Hyperparameter tuning was performed using GridSearchCV, where the regularization strength (C) and penalty type (L1, L2) were optimized. A 5-fold cross-validation strategy was adopted to ensure robust and unbiased estimation of model performance.

Prior to selecting logistic regression, we experimented with more complex models, including XGBoost, Random Forest, and other ensemble-based methods. However, these models yielded lower accuracy and offered limited interpretability in our setting, making logistic regression the more suitable choice for both performance and explainability.

- Step 4: Evaluation

Model performance is assessed using accuracy, precision, recall, F1-score, and the confusion matrix. These metrics help evaluate both overall performance and balance between false positives and false negatives.

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.74	0.76	23
1	0.45	0.50	0.48	10
accuracy			0.67	33
macro avg	0.61	0.62	0.62	33
weighted avg	0.68	0.67	0.67	33

Figure 3: Evaluation Results obtained after testing on 85-15 data split

4. Audio-Based Depression Detection Report

• Goal

To detect depression in participants based on their speech characteristics. We go beyond simple classification by focusing on:

- **Interpretability:** Features are psychologically and acoustically meaningful.
- **Real-time feasibility:** Lightweight, averaged per-user.
- **Clinical alignment:** Captures changes in speech patterns linked to mental health.

2. Feature Selection Strategy

• Files Used

audio_file = f"E:/CAP/processed_data/{id}_audio/{id}_OpenSMILE2.3.0_egemaps.csv" audio_file1 = f"E:/CAP/processed_data/{id}_audio/{id}_OpenSMILE2.3.0_mfcc.csv"

These are extracted using OpenSMILE 2.3.0.

a. Features from egemaps.csv (important features)

The following features are selected from the eGeMAPS feature set due to their potential to reflect vocal and emotional changes related to depression:

- Loudness_sma3**:** This feature measures the perceived energy of the voice. Depression is often associated with a low energy level, which may be reflected in a reduced loudness. Individuals with depression tend to have a flat affect, and a low loudness could be an indicator of this diminished emotional expression.
- **alphaRatio_sma3:** This feature captures the ratio of energy between low and high frequency bands. A higher alpha ratio indicates a brighter, more energetic voice, while a lower ratio is linked to a duller or more nasally voice, which can occur in depressed individuals. Depression can alter the typical brightness of the voice, making this feature crucial for detecting such changes.
- **hammarbergIndex_sma3:** This feature measures the spectral balance between low and high frequency energy. A strained or breathy voice often correlates with depression, as individuals may speak with less vocal power or clarity when experiencing depression. The Hammarberg index helps to quantify these changes in vocal tone, providing insight into emotional distress.
- **slope0-500 sma3:** The slope of the spectrum from 0 to 500 Hz indicates vocal tension or a lack of vocal variation. Depressed speech may lack the usual modulation in pitch and tone, often resulting in a monotonous or flat voice. This feature can be a strong indicator of depression-related vocal changes.
- **spectralFlux_sma3:** Spectral flux measures how much the audio spectrum changes over time. A depressed voice may have less variation and a more monotonous sound, which is reflected in a lower spectral flux. This feature is essential for detecting the temporal stability or variation in speech, which is often impaired in depression.
- **mfcc1_sma3, mfcc2_sma3, mfcc3_sma3, mfcc4_sma3:** Mel-frequency cepstral coefficients (MFCCs) capture the timbre and resonance properties of the voice. Depression can affect the voice's resonance, leading to alterations in these coefficients. MFCCs are crucial for modeling voice quality and identity, making them highly relevant in detecting emotional and mental health changes.
- **FosemitoneFrom27.5Hz sma3nz:** This feature measures the fundamental frequency (F0) of the voice, normalized to avoid interference from background noise. A lower or more monotone pitch is commonly associated with depression, as individuals with depression often exhibit a lack of vocal variability and a subdued tone. Monitoring changes in pitch provides valuable insights into emotional states.
- **jitterLocal_sma3nz:** Jitter measures short-term variations in pitch. High jitter values are indicative of irregular vocal patterns, often linked to vocal fatigue or emotional stress. Depression can lead to subtle yet noticeable irregularities in voice pitch, making jitter a relevant feature for detecting such conditions.
- **shimmerLocaldB_sma3nz:** Shimmer measures variations in the amplitude of the voice, reflecting vocal instability. High shimmer is associated with a less stable voice, often due to fatigue or emotional distress. Depression can lead to vocal instability, making shimmer a key feature for recognizing such disturbances.
- **HNRdBACF sma3nz:** The harmonics-to-noise ratio (HNR) measures the ratio of harmonic sound to noise. A low HNR indicates breathiness or hoarseness, which are common in depressed speech patterns. A depressed individual may exhibit a more breathy or unclear voice, making this feature an important indicator of vocal distress.

Each of these features was chosen for its direct relevance to the acoustic manifestations of depression. By capturing vocal characteristics such as pitch, loudness, resonance, and variability, these features can provide meaningful insights into an individual's emotional state, helping to detect depression based on speech.

b. Features from mfcc.csv (importantfeatures1)

The following features are extracted from the MFCC (Mel-frequency cepstral coefficients) file. These features provide insights into the voice quality, articulation, and dynamics, which are crucial for detecting emotional and mental states such as depression:

important_features1 =
['pcm_fftMag_mfcc[1]' to '[5]',
'pcm_fftMag_mfcc_de[1]' to '[5]',
'pcm_fftMag_mfcc_de_de[1]' to '[5]'
]

- **Feature Descriptions**
  - **pcm_fftMag_mfcc[1-5]:** These are the first 5 MFCCs derived from the FFT magnitude spectrum. MFCCs are widely used for analyzing voice quality and are sensitive to changes in vocal tone and articulation. In the context of depression, alterations in these coefficients may reflect reduced emotional expressiveness and vocal modulation. The first 5 MFCCs are particularly relevant for capturing the timbre of the voice, which can be less dynamic or less resonant in depressed speech.
  - **pcm_fftMag_mfcc_de[1-5]:** These represent the first derivatives (Δ) of the MFCCs, which measure the rate of change of the voice features over time. The presence of high derivative values indicates rapid variations in the voice, such as increased expressiveness or excitement. In contrast, depression is often characterized by a flat or monotonous speech pattern with minimal changes, leading to low derivative values. Monitoring these changes in voice dynamics can provide insights into emotional fluctuations and stability.
  - **pcm_fftMag_mfcc_de_de[1-5]:** These are the second derivatives ($\Delta\Delta$) of the MFCCs, which capture acceleration or sudden changes in vocal tone and energy. Depression may lead to a more subdued and stable voice with fewer dramatic shifts in pitch and intensity. Reduced acceleration in speech, as indicated by low second derivatives, can be a sign of a flattened emotional state often observed in depressed individuals. Therefore, these features help to detect subtle changes in the pacing and intensity of speech that may signal underlying emotional distress.

c. Depression Label (Target)

Labels come from PHQ-8 scores:
\ From: PHQ8_labels.csv

- PHQ-8 $\geq 10 \Rightarrow$ Label = 1 (Depressed)
- PHQ-8 $< 10 \Rightarrow$ Label = 0 (Not Depressed)

3. Approach and Methods

• Step 1: Data Preprocessing

In this step, we process the audio feature data and prepare it for training:

- **Feature Extraction:** We extract the key features, such as Loudness, MFCCs, and jitter values, from the CSV files. The selected features are stored in a list, important_features, to focus on speech characteristics relevant to detecting depression.
- **Feature Averaging:** To reduce noise and dimensionality, we compute the mean of the selected features over time. This provides a simplified representation of each participant's speech patterns.
- **Label Assignment:** Depression labels are derived from the PHQ-8 scores. A label of 1 is assigned to participants with PHQ-8 scores equal to or greater than 10 (indicating depression), and a label of 0 is assigned to others.

- **Step 2: Handling Class Imbalance**

Given the likely imbalance between depressed and non-depressed participants, we apply **SMOTE (Synthetic Minority Over-sampling Technique)** to generate synthetic samples for the minority class. This ensures that both classes are adequately represented in the training set, improving the model's ability to learn from both classes effectively.

- **Step 3: Model Training**

The following steps outline the training process for the depression detection model:

- **Feature Scaling:** We use StandardScaler to standardize the features, ensuring that all features have a mean of zero and a standard deviation of one. This improves the model's performance and convergence during training.
- **Train-Test Split:** The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing. Stratified sampling ensures that both classes (depressed and non-depressed) are proportionally represented in both sets.
- **SMOTE Resampling:** The training data is resampled using SMOTE to balance the class distribution. This creates synthetic samples for the minority class, providing a more balanced dataset for model training.
- **Logistic Regression:** A Logistic Regression model is used due to its interpretability and effectiveness in binary classification. The class_weight='balanced' parameter is set to handle class imbalance, and the model is trained with a maximum of 1000 iterations to ensure convergence.
- **Other ML models tried:** Although we primarily use Logistic Regression, a Random Forest classifier as well as XGBoost classifier is also trained but Logistic Regression gives better results.

- **Step 4: Model Evaluation**

After training the model, we evaluate its performance using several key metrics:

- **Confusion Matrix:** The confusion matrix is generated to visualize the true positives, false positives, true negatives, and false negatives. This helps assess how well the model distinguishes between the depressed and non-depressed classes.
- **Classification Report:** A detailed classification report is generated, which includes metrics such as accuracy, precision, recall, and F1-score. These metrics are essential for understanding the model's performance in terms of both overall accuracy and its ability to correctly identify each class.
- **Additional Evaluation Metrics:** Although optional, the ROC-AUC score may be computed to assess the model's ability to discriminate between classes.

5. Multimodal Model : Audio + Text

5.1. Introduction

The individual models for audio and text features were developed and evaluated separately in earlier work. Here, we leverage both types of data simultaneously in a single model to improve the predictive performance.

5.2. Data Preprocessing

Feature

Standardization

The audio and text features are standardized independently using the StandardScaler from scikit-learn. This ensures that each feature has zero mean and unit variance, which helps the model perform better by preventing any one feature from dominating due to differences in scale.

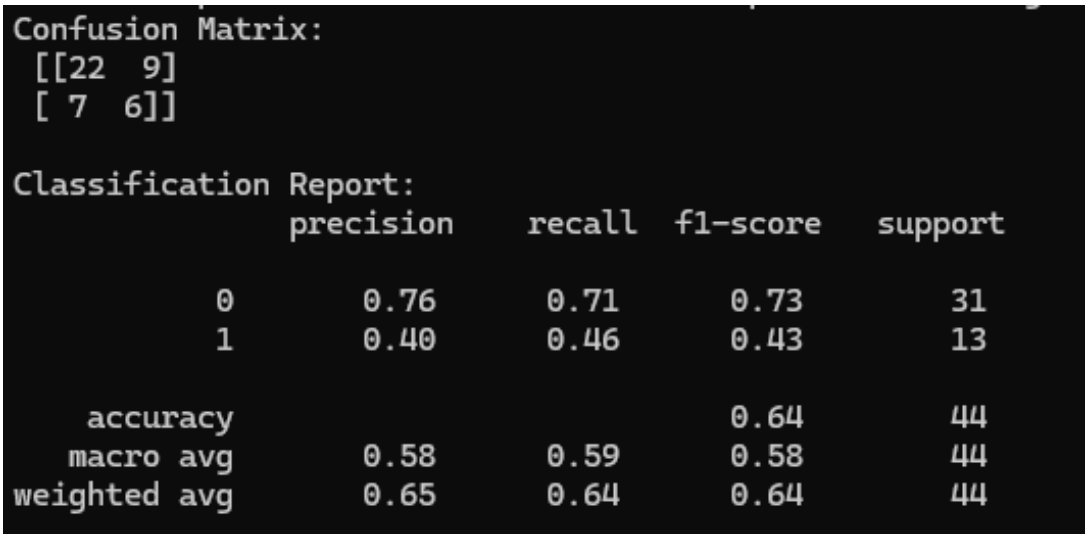


Figure 4: Evaluation Results obtained after testing on 80-20 data split

Combining Features

Once the features are standardized, the audio and text features are concatenated to form a single feature matrix, which will be used for training and prediction.

5.3. Model and Training

For training the model, we use **XGBoost**, a powerful gradient boosting classifier, which has shown to be effective in a wide range of classification tasks. The dataset is split into training and test sets, with 80% used for training and 20% for testing.

Handling Class Imbalance

To address any class imbalance in the dataset, we apply **SMOTE** (Synthetic Minority Over-sampling Technique) to the training set. SMOTE generates synthetic samples of the minority class to ensure the classifier is not biased toward the majority class.

Hyperparameter Tuning

To optimize the model's performance, we perform hyperparameter tuning using **RandomizedSearchCV**. This allows us to search for the best values for several hyperparameters, including:

- n_estimators: the number of trees in the model
- max_depth: the maximum depth of each tree
- learning rate: the learning rate for gradient boosting
- subsample: the fraction of samples used to train each tree
- colsample bytree: the fraction of features used to train each tree

5.4. Performance Metrics

The model's performance metrics are as follows:

Metric	Value
Training Accuracy	**1.0000** (The model perfectly fits the training data, suggesting possible overfitting.)
Test Accuracy	***0.8182 (The model performs well on unseen data, with an accuracy of approximately 82%.)***
Best Hyperparameters	
colsample bytree	0.8515
learning rate	0.1101
max depth	3
n estimators	541
subsample	0.9812

Table 3: Model Performance and Best Hyperparameters

6. Facial and Multimodal (Text + Audio + Facial)

6.1. Shared Data Preprocessing

6.1.1. Data Sources

- **Facial Features:** Extracted using OpenFace, yielding FAU (Facial Action Units) CSV files.
- **Audio Features:** Extracted using OpenSMILE, focusing on MFCC features.
- **Transcripts:** Generated using Whisper V3 (small model) for more accurate and complete segmentation than the original dataset transcripts.

6.1.2. Transcript Segmentation

Generated transcripts were divided into segments of 15 utterances with an overlap of 4 utterances to preserve semantic continuity. Each segment serves as a temporal unit for synchronized feature extraction across all modalities.

6.1.3. Facial Modality

From OpenFace CSV files, relevant feature columns were selected based on prior analysis and significance in depression detection. For each segment:

- Feature values were max-pooled over the segment duration. (Table 4)
- Data shaped to (participants × segments × facial features).

6.1.4. Audio Modality

Similarly, for OpenSMILE MFCC files:

- Relevant MFCC columns were selected. (Table 5)
- Max pooling was performed over each segment.
- Data shaped to (participants × segments × audio features).

6.1.5. Text Modality

Using SBERT embeddings:

- Each transcript segment was encoded using SBERT to obtain a 384-dimensional embedding.
- Resulting data shape: (participants × segments × 384).

Feature	Type	Depression-Relevant Cue
pose Rx, Ry, Rz	Head rotation	Downward, tilted head indicating sadness or fatigue
pose Tx, Ty, Tz	Head translation	Physical withdrawal or disengagement
gaze 0, gaze 1 vectors	Eye gaze direction	Reduced eye contact, avoidance behavior
gaze angle x/y	Gaze angle	Looking down or away, low social engagement
AU04 r	Brow lowerer	Sadness, worry
AU05 r	Upper lid raiser	Low alertness, flattened emotion
AU07 r	Lid tightener	Tension, anxiety
AU14 r	Dimpler	Less social/expressive interaction
AU15 r	Lip corner depressor	Sadness indicator
AU20 r	Lip stretcher	Flat affect, low expressiveness
AU25 r	Lips part	Reduced speech or response delay
AU26 r	Jaw drop	Apathy, lack of engagement
AU45 r	Blink	Abnormal blinking due to disengagement or fatigue

Table 4: Facial features and their relevance to depression detection

Feature	Domain Captured	Depression-Relevant Cue
MFCC[2], MFCC[4]	Vocal tract shape	Reduced articulation, lower energy
MFCC[7], MFCC[9]	Fine spectral detail	Lack of pitch modulation, flat prosody
Delta MFCCs (de)	Rate of change	Monotony, slowed speech
Delta-Delta MFCCs (de de)	Acceleration of change	Smoothness/rigidity in transitions, hesitation

Table 5: Selected MFCC features and their relevance to depression detection

6.2. Facial-Only CNN-BiLSTM

- **Feature Input:** Each video segment was processed to extract a sequence of facial features using a pre-trained model (e.g., OpenFace), capturing frame-level descriptors such as Action Units (AUs), gaze direction, and head pose. The input to the model was therefore a 3D tensor of shape batch_size, time_steps, num_facial_features, where each time step represents a video frame.
- **CNN Layer:** A 1D Convolutional Neural Network (CNN) layer was employed initially to extract local temporal patterns across consecutive frames. The convolution operation was performed along the time dimension with multiple filters of kernel size 3, enabling the model to detect short-term temporal dependencies and micro-expressions.
- **Batch Normalization and Dropout:** To enhance generalization and prevent overfitting, the CNN output was passed through a Batch Normalization layer followed by a Dropout layer (with a rate of 0.3). Batch Normalization ensured that the network remained stable during training by normalizing intermediate outputs,

while Dropout randomly deactivated neurons to reduce co-adaptation.

- **BiLSTM Layer:** A Bidirectional Long Short-Term Memory (BiLSTM) layer with 32 hidden units was used to capture longer temporal dependencies from both past and future directions. This layer processed the sequential features extracted by the CNN and summarized them into a fixed-length context vector representing the entire facial expression sequence.
- **Dense Layers and Output:** The LSTM output was fed into a fully connected Dense layer with 16 units and ReLU activation, followed by another Dropout layer. Finally, a Dense layer with a single neuron and sigmoid activation was used to output the binary class probability — indicating whether the participant is likely depressed or not.
- **Training Performance:** The model was trained using the binary cross-entropy loss function and the Adam optimizer. It achieved an overall classification accuracy of **71%** on the validation set. While this demonstrates some predictive power, it also indicates that facial features alone may not capture the full complexity of depression indicators, especially in cases with subtle or masked expressions.

6.3. Multimodal CNN-BiLSTM

To capture the complex and multifaceted nature of depression indicators, we developed a multimodal deep learning architecture that integrates facial, audio, and textual features into a unified classification model. Each modality captures unique aspects of human expression and behavior—facial expressions reveal non-verbal cues, audio conveys prosody and tone, and text provides semantic content. Combining these modalities allows the model to benefit from complementary sources of information, leading to more robust and accurate predictions.

- **Multimodal Feature Extraction and Fusion:** For each participant's video, we segmented the data into fixed-length segments to ensure consistent input size across samples. For every segment, we extracted three distinct types of features:

  1. **Facial Features:** Frame-level features were obtained using OpenFace, which detects facial landmarks, head pose, gaze direction, and Action Units (AUs) corresponding to muscle activations. These features help quantify micro-expressions and facial movement patterns over time.
  2. **Audio Features:** From the audio track, we used the OpenSMILE toolkit to extract prosodic, spectral, and voice quality descriptors such as pitch, energy, jitter, shimmer, and Mel Frequency Cepstral Coefficients (MFCCs). These features are known to reflect affective and psychological states.
  3. **Textual Features:** Transcripts of the participant's speech were first generated using OpenAI's Whisper V3 model. These transcripts were then embedded into dense vector representations using a pre-trained Sentence-BERT model, which captures semantic meaning and context.

  All three sets of features were synchronized (aligned temporally per frame or segment), normalized, and then concatenated along the feature axis. The resulting 3D tensor had the shape (participants × segments × combined feature length), where each segment contained fused multimodal descriptors.

- **CNN Layer:** The combined feature sequence for each segment was passed through a 1D Convolutional Neural Network (CNN) layer. This layer consisted of multiple filters with a kernel size of 3, operating over the temporal dimension. The goal was to detect local patterns in the time series—such as sudden changes in pitch, short bursts of facial expressions, or context shifts in text—that may signal emotional distress or hesitation.

- **BiLSTM Layer:** The output from the CNN layer was then fed into a Bidirectional Long Short-Term Memory (BiLSTM) layer with 64 units. This recurrent layer was crucial for learning long-term dependencies and temporal context across the multimodal sequence. By incorporating both forward and backward temporal passes, the BiLSTM effectively captured the evolution and transitions of affective states, hesitation, and engagement, which are key markers of depression.

- **Dense Layers and Output:** The BiLSTM output was projected through fully connected Dense layers with ReLU activations to form high-level feature abstractions. Dropout regularization (e.g., 0.3) was applied to prevent overfitting. The final output layer used a sigmoid activation to predict a binary classification probability: 1 indicating "depressed" and 0 indicating "not depressed."

- **Training Details and Performance:** The model was trained using the Adam optimizer with a learning rate of 0.001 and the binary cross-entropy loss function. Training was performed over 15 epochs with a batch size of 32. The network achieved a training accuracy of 94%, substantially outperforming the unimodal (facial-only) model. This performance gain underscores the value of multimodal fusion—especially in a complex domain like affective computing, where subtle behavioral indicators may only be visible in certain modalities. Given the small size of the dataset (only 105 participants), a typical train-validation-test split was not applied, as the resulting validation set would have been too small to reliably represent the overall data distribution. Instead, the model was trained and evaluated on the full dataset to maximize data utilization.

- **Interpretation:** The improved accuracy highlights how depression manifests not only in facial expressions but also in voice tone and language content. For instance, a participant may maintain a neutral facial expression while exhibiting low vocal energy or using negative semantic content. The multimodal model captures such cases more effectively by leveraging inter-modal correlations and cross-channel redundancies.

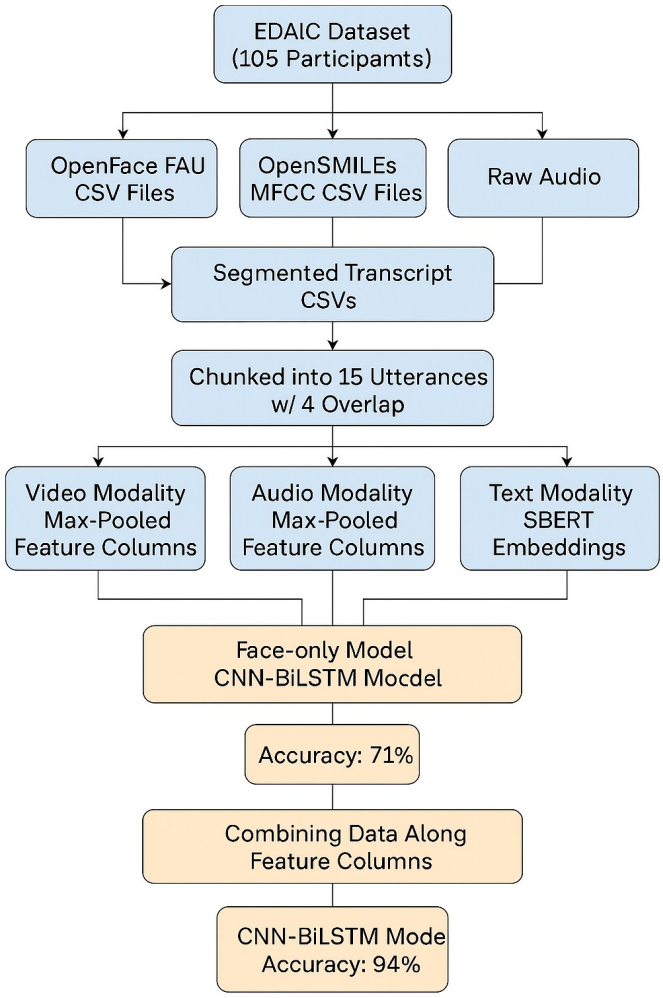


Figure 5: Workflow of the Facial and Multimodal Depression Detection Pipeline.

7. Problems Faced

Low Disk Storage Space for Handling Large Datasets

The EDAIC dataset, when processed across multiple modalities such as video (OpenFace), audio (OpenSMILE), and raw audio recordings, results in a significant increase in storage requirements. With 105 participants and separate CSVs, audio files, and generated transcripts, disk space quickly became a limiting factor. This constraint hindered smooth experimentation, requiring frequent manual cleanup, recompression, and careful memory management during preprocessing and model training phases.

GPU Memory Limitations Prevent Batch Processing of Full Dataset

Deep learning models like CNN-BiLSTM are computationally intensive, especially when handling multimodal data across many segments and long temporal sequences. Due to limited GPU memory, it was not feasible to load and train on the entire dataset at once. This led to the need for mini-batching, splitting datasets, and reducing sequence lengths, which potentially affected training stability and model generalization. Efficient scaling became a challenge in the training pipeline.

Lack of Raw Facial Features and Poor-Quality Transcripts

While the dataset provided OpenFace CSV files, raw facial videos or high-resolution features were not available, which limited the expressivity and richness of the visual modality. Furthermore, the pregenerated transcript files included in the EDAIC dataset were poorly constructed – they only contained the participants' responses and often ended abruptly, omitting the context provided by the interviewer. This significantly affected the linguistic modality. To address this, Whisper V3 was used to regenerate high-quality transcripts from raw audio, but aligning these transcripts with feature timestamps across modalities was time-consuming and computationally expensive.

Noisy Audio Recordings

Listening to the raw audio files revealed that many contained background noise, static, or microphone artifacts, which complicated downstream audio feature extraction. This noise negatively impacted the quality of the MFCC features generated by OpenSMILE and may have introduced inaccuracies in transcript generation as well. It also reduced the reliability of the acoustic modality as a standalone predictor for depression detection.

8. Future Scope

Scalable and Real-Time Multimodal Depression Detection Models

A promising direction for future work involves developing models that are optimized for real-time and large-scale deployment. These models should be capable of processing streaming multimodal data efficiently, using techniques like online learning, low-rank approximations, or attention-based mechanisms to handle long-term dependencies without overwhelming memory. Cloud-based platforms or distributed training systems could also be explored to scale computation and storage. Integrating noise-robust audio models and real-time facial feature extraction would help further improve accuracy and usability in clinical or telehealth settings.

Enhanced Preprocessing Pipelines and Data Augmentation

Future work can benefit from more sophisticated preprocessing pipelines, such as advanced noise filtering for audio, data augmentation for limited facial features, and automatic alignment tools across modalities. Improved transcript generation models that understand both speakers' roles can significantly enrich the context and quality of textual inputs.

9. Conclusion

As a group, we systematically explored depression detection using all available participants across individual modalities — audio, text, facial — as well as a combined audio-text model. For the final multimodal detection model integrating audio, text, and facial features, we utilized data from all 105 participants. This comprehensive approach allowed us to assess each modality's contribution and evaluate the effectiveness of full multimodal fusion in identifying depressive patterns.

All the codefiles, are given in the zip file

References

1. P. Ching, L. Xie, D. Zhang, et al. "Automatic Identification of Depression Using Facial Images with Deep Convolutional Neural Networks." *Med Sci Monit*, vol. 28, 2022. https://pmc.ncbi.nlm.nih.gov/ articles/ PMC9281460/pdf/medscimonit-28-e936409.pdf

2. R. Khan, F. Pirbhulal, A. Majumder. "Multimodal Depression Estimation via Gated CNN-BiLSTM with Fusion Strategies." arXiv preprint, arXiv:2504.01767, 2024. https://arxiv.org/pdf/2504.01767
3. M. Pampouchidou, C. Simantiraki, P. Charisis, et al. "Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text." arXiv preprint, arXiv:1909.01417, 2019. https://arxiv.org/pdf/1909.01417
4. L. I. Rincón, M. B. I. Reátegui, D. Cabrera, et al. "Multimodal Deep Learning for Depression Detection: A Review." *Sensors*, vol. 24, no. 2, 2024. https://www.mdpi.com/1424-8220/24/2/348
5. H. Shanavas, H. X. Zhang, M. Riegler, et al. "Fusion Strategies for Multimodal Depression Detection." Master's Thesis, Umeå University, 2024. https://umu.diva-portal.org/smash/get/diva2:1838706/FULLTEXT01.pdf