**1. Dataset Description**

**Source**: The analysis is based on the **DAIC-WOZ (Distress Analysis Interview Corpus - Wizard of Oz)** dataset, particularly focusing on the **“Depression”** subfolder.

**Data Composition**:

* + Each subfolder corresponds to a participant and includes:
    - A .wav audio recording of the participant’s interview.
    - Precomputed acoustic feature files (e.g., COVAREP.csv).
  + Labels are extracted from combined\_PHQ8\_scores.csv, which contains PHQ-8 (Patient Health Questionnaire-8) scores for each participant.

**Labeling Scheme**:

* + PHQ-8 scores were binarized for classification:
    - **Depressed**: PHQ-8 score ≥ 10 → label 1
    - **Not Depressed**: PHQ-8 score < 10 → label 0

**Features Used**:

* + **COVAREP Features**: Precomputed frame-level acoustic descriptors like pitch, creak, harmonic model parameters, and glottal source features.
  + **MFCCs**: In later experiments, Mel-Frequency Cepstral Coefficients were extracted using torchaudiodirectly from the waveform.

**Voice Activity Filtering**:

* + In the original attempt, only **voiced frames** were retained using a VUV (voiced/unvoiced) flag in the COVAREP features, under the assumption that voiced segments contain more emotionally relevant signals.

**2. Modeling and Experiments**

**A. Reproducing the Original LSTM**

* The initial LSTM model was reproduced using:
  + Input: COVAREP features (after VUV-based voiced frame filtering).
  + Architecture: A single LSTM layer followed by dense and dropout layers.
* **Result**: This setup yielded a **test accuracy of ~67%**, slightly outperforming the originally reported 60%.

**B. Extended Modeling Attempts**

* Multiple variations were tested:
  + **CNN + BiLSTM + Attention** using MFCC features extracted with torchaudio.
  + Variants with:
    - **Dropout (0.5)** and **L2 Regularization** to mitigate overfitting.
    - Smaller LSTM sizes (e.g., 32 or 16 units).
    - Comparison with traditional classifiers: **Random Forest**, **SVM**, **Logistic Regression** applied to flattened feature representations.
* **Observations**:
  + MFCC-based models did not outperform the COVAREP-based LSTM, possibly due to signal loss in the MFCC transformation or model complexity.
  + Traditional classifiers plateaued around **55–60% accuracy**, lacking the temporal modeling power of LSTMs.

**3. Best Performing Model**

**A: Input Processing**:

* Audio signals were converted into fixed-length feature arrays with **100 time steps** per sample using COVAREP.
* Samples shorter than 100 steps were zero-padded; longer samples were truncated.

**B:** **Data Split**:

* The dataset was split into **80% training and 20% testing** using train\_test\_split with a fixed random seed for reproducibility.

**C: Batch Size**:

* Training was performed using a **batch size of 4**, meaning the model updated weights after processing every 4 samples.
* A small batch size helps stabilize training on limited datasets and introduces useful gradient noise to reduce overfitting.

**D.** **Training Epochs and Early Stopping**:

* Models were trained for **up to 50 epochs** with **early stopping** (patience = 5) to avoid overfitting.

**E. Class Imbalance Handling**:

* **Class weights** were computed and applied during training to correct for imbalance between depressed and non-depressed samples.

**F: Model Evaluation**:

* Accuracy was measured on the 20% held-out test set.
* LSTM using raw COVAREP (without over-regularization) achieved **~70% accuracy** without signs of overfitting.
* **Why it Worked**:
  + **No overfitting was observed** despite the high training accuracy, as the validation accuracy plateaued at a similar level and loss curves did not diverge.
  + Using the full sequence of COVAREP frames preserved more contextual information than filtering out unvoiced segments.
  + The relatively **light regularization** avoided underfitting, which was seen in deeper or more heavily regularized models.