# Multimodal Depression Detection on the DAIC-WOZ interview dataset

João Mata, ist1102444 Liane Carolina, ist1103065 Tiago Menezes, ist11122249

## I. Introduction

Depression and PTSD are clinical conditions affecting up to 1 billion people globally. [Our World in Data, 2023] The number of cases has risen substantially in recent years, especially since Covid, making this topic extremely relevant. [La-Dépression.org, 2023] [of Mental Health, 2023] [Organisation mondiale de la Santé, 2023] [État Dépressif, 2023] Data collection is still mostly done through individual assessment and non-standardised interviews, which is costly, time consuming, and introduces interviewer-dependent bias. There's a need for a unified, objective, and time-efficient way to assess mental health. This would let researchers focus more on data analysis, enhancing R&D applications.

A promising solution are recent ML and DL models. We applied them to the "E-DAIC Depression Database," replicating the work of [Sadeghi et al., 2024a] and [Hassan et al., 2025]. The multimodal dataset includes video features and audio from 219 interviews assessing depression and PTSD, later integrated with the PHQ-8 questionnaire.

We adopted a multimodal approach to improve diagnostic accuracy. By focusing not only on vocal cues or facial expressions, but also on the interview transcriptions. The transcripts allow our system to capture explicit language and cognitive patterns, which better mirrors traditional clinical assessments. This integration allows for an automated and more objective evaluation of conditions such as depression and PTSD, while minimizing the risk of interviewer bias.

We will herby proceed to describe the dataset, provide an overview of the models we will use, of the different fusion strategies used by one of the considered papers [Hassan et al., 2025] and of the evaluation metrics that will be used to assess the classification task. We will also illustrate the preprocessing procedures we will apply to the three separately elements of our dataset (video, audio and text elements) to be later analysed.

## II. Problem Formulation

### A. Dataset Description

The **Extended Distress Analysis Interview Corpus (E–DAIC)**[1] contains semi-structured clinical interviews conducted by the virtual agent *Ellie* to screen for depression, anxiety, and post-traumatic stress disorder (PTSD). In the subset used here, interviews were run in a *wizard-of-Oz* (WoZ) set-up: a human clinician in another room tele-operated Ellie, ensuring natural dialogue flow while preserving the appearance of autonomy.

The data is organized into three stratified splits that preserve the gender, age and PHQ-8 scores distributions. The training set has 163 participants, while the development and test sets each contain 56 participants, respectively.

The following data is available for each session:

- **Audio**
  - 16 kHz WAV recording of the entire interview;
  - Automatic transcripts generated by Google Cloud Speech-to-Text;
  - Pre-computed embeddings, not used in our pipeline (see Sec. III-D).
- **Visual:** Due to privacy concerns, raw video frames are not available. Instead, our dataset includes:
  - Deep representations from CNN ResNet [He et al., 2016] and CNN VGG [Simonyan and Zisserman, 2014] for each frame;
  - OpenFace features [Baltrušaitis et al., 2016]:
    - ✓ 3D facial landmarks, $x_1$ (68 three-dimensional points);
    - ✓ Gaze direction vectors of each eye, $x_2$;
    - ✓ Head-pose translation and rotation vectors, $x_3$;
    - ✓ Intensities of 17 Facial Action Units (FAUs) with confidence scores, $x_4$.
- **Clinical labels** Post-interview self-reports provide the **PHQ-8** (0-24) for depression and the **PCL-C** (17-85) for PTSD, with PHQ-8 $\geq$ 10 and PCL-C $\geq$ 44 indicating depression and PTSD, respectively.

### B. Classification Task

We predict both targets as regression problems, predicting the continuous PHQ-8 (0–24) and PCL-C (17–85) scores instead of fixed diagnostic labels. This preserves clinical nuance: PHQ-8 values of 0–4, 5–9, 10–14 and 15+ mark none, mild, moderate and moderately-severe depression, while PCL-C scores below 38, 38–43 and 44+ correspond to sub-clinical, possible and probable PTSD.

Our pipeline first turns audio, text and visual data into fixed-length embeddings, and we test two temporal granularities. In the global variant, the entire interview is encoded once, producing an embedding for each modality for the session. In the segment variant, we use Whisper's word-level time-stamps to cut the recording into short utterances, and transform each into a fixed-length vector separately. However, Whisper does not label speakers, so some utterances still mix Ellie and the participant.

## III. System Architecture

### A. Model Overview

Our multimodal methodology takes advantage of the three types of data available in the E-DAIC dataset: text, audio and visual. For each modality we propose the use of *fixed-length embeddings* $\mathbf{e}_t \in \mathbb{R}^{d_t}$, $\mathbf{e}_a \in \mathbb{R}^{d_a}$ and $\mathbf{e}_v \in \mathbb{R}^{d_v}$ respectively.

[1] Form at https://dcapswoz.ict.usc.edu/extended-daic-database-download/

In this report, sections III-C, III-D, III-E detail the encoding process for each one.

**Fusion strategies.** To fuse the different modality embeddings, we will take two approaches:

- *Early (data–level) fusion.* Here, we consider a concatenation of all the embeddings $\mathbf{e} = [\mathbf{e}_t; \mathbf{e}_a; \mathbf{e}_v] \in \mathbb{R}^d$ ($d = d_t + d_a + d_v$) which is then passed through a fusion model to get a final representation of the data, $\tilde{\mathbf{e}}$ (we will try MLPs, CNNs, BiLSTMs, Transformers). To get PHQ-8 and PCL scores, a MLP regressor is trained over these transformed features (Figure 1).
- *Intermediate (feature–level) fusion.* Each modality passes through its *own* feature extractor and the resulting feature maps are subsequently concatenated. The resulting vector is then fed to a simple regressor as previously mentioned (Figure 2).

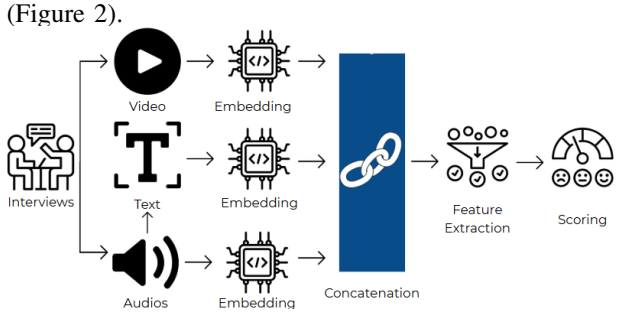


Fig. 1. Early fusion strategy

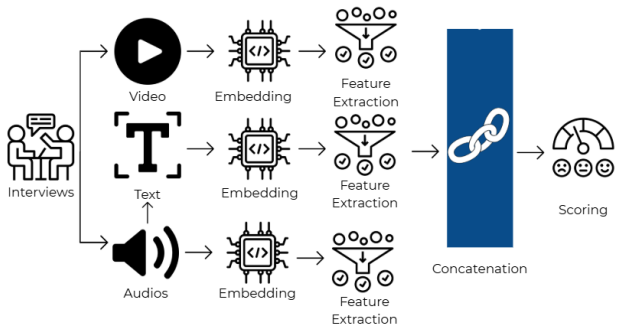


Fig. 2. Intermediate fusion strategy

### B. Evaluation metrics

We will use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for PTSD (PCL-C) scores and Depression (PHQ-8) scores:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

with $y_i$ and $\hat{y}_i$ representing the true value of the score given and the predicted score.

### C. Automated Text Processing and Feature Extraction

The automated transcripts (.csv files) in the dataset are often incomplete or inaccurate, so the first step will be to use Advanced Automatic Speech Recognition (ASR) systems (e.g., Whisper by OpenAI) to generate high-quality transcripts from the raw audio. Text processing will then be performed using large language models to clean the interview data, partitioning it into question/answer pairs, removing unmatched fragments, and correcting grammar for clarity and completeness.

From the corrected transcripts, two sets of features will be extracted using different processes.

- **Features from a Fine-tuned RoBERTa model for depression severity classification**. Using DepRoBERTa, a fine-tuned RoBERTa model pre-trained on depression-related Reddit posts, transcripts are classified as "not-depression", "moderate", or "severe" [Poświata and Perełkiewicz, 2022]. Additional finetuning may improve performance for this interview format. The model outputs class probabilities, which are then used as features. [Sadeghi et al., 2024b].
- **Features from the LLM-driven question based method**. An LLM designs 11 questions from sample interviews to distinguish individuals with and without depression. It answers these for each transcript using constrained responses (e.g., 'YES', 'NO', 'To Some Extent'), which are then converted into an 11-dimensional numerical feature vector added to the DepRoBERTa output. [Sadeghi et al., 2024b]

These two sets of textual features are then concatenated to form a combined feature-vector for each interview file. [Sadeghi et al., 2024b].

### D. Audio Features

In addition to the transcripts, raw audio also carries relevant cues about a person's mental health conditions; for example, monotonous prosody and reduced vocal energy are well-known correlates of depression [Menne et al., 2024], [Di et al., 2024].

We will consider three encoders:

- Wav2Vec2 in addition to 1D-CNN and attention pooling layers, as proposed by [Zhang et al., 2024]. We can also consider Wav2Vec2-BERT 2.0 for increased robustness to noise and sensitiveness to paralinguistics [Barrault et al., 2023].
- A Bi-LSTM encoder trained from scratch, following the implementation of [Chen et al., 2025] for feature extraction.
- A CNN encoder trained from scratch, following [Kim et al., 2023].

For each encoding approach, it is necessary to convert the input waveform into Mel Spectograms images using `librosa` package. If proved necessary by preliminary results, additional audio features from the GeMAPS set [Eyben et al., 2015] can be incorporated (such as statistics of loudness, pitch, jitter, shimmer, etc.). We note that the architectural choices made will also be influenced by computational costs, given the lack of powerful infrastructure we face.

### E. Visual Features

As previously mentioned, the video frames were not made available in the dataset, which could pose challenges when training models; however, features from OpenFace are available. In this work, our aim is to implement the approach in [Wang et al., 2025] that takes advantage of such features to generate embeddings.

Additionally, the dataset contains two kinds of deep representations of the frames extracted from CNN ResNet [He et al., 2016] and CNN VGG [Simonyan and Zisserman, 2014], which we will consider in our pipeline.

# REFERENCES


[Baltrušaitis et al., 2016] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

[Barrault et al., 2023] Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duppenthaler, M., Duquenne, P.-A., Ellis, B., Elsahar, H., Haaheim, J., et al. (2023). Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

[Chen et al., 2025] Chen, Z., Wang, D., Lou, L., Zhang, S., Zhao, X., Jiang, S., Yu, J., and Xiao, J. (2025). Text-guided multimodal depression detection via cross-modal feature reconstruction and decomposition. *Information Fusion*, 117:102861.

[Di et al., 2024] Di, Y., Rahmani, E., Mefford, J., et al. (2024). Unraveling the associations between voice pitch and major depressive disorder: a multisite genetic study. *Molecular Psychiatry*.

[Eyben et al., 2015] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

[Hassan et al., 2025] Hassan, A. A., Ali, A. A., Fouda, A. E., Hanafy, R. J., and Fouda, M. E. (2025). Leveraging embedding techniques in multimodal machine learning for mental illness assessment. *arXiv preprint arXiv:2504.01767*. Preprint, accessed May 12, 2025.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Kim et al., 2023] Kim, A. Y., Jang, E. H., Lee, S.-H., Choi, K.-Y., Park, J. G., and Shin, H.-C. (2023). Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. *Journal of medical Internet research*, 25:e34474.

[La-Dépression.org, 2023] La-Dépression.org (2023). La dépression en chiffres. Consulté en mai 2025.

[Menne et al., 2024] Menne, F., Dörr, F., Schrader, J., et al. (2024). The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry*, 24:794.

[of Mental Health, 2023] of Mental Health, N. I. (2023). Depression. https://www.nimh.nih.gov/health/topics/depression.

[Organisation mondiale de la Santé, 2023] Organisation mondiale de la Santé (2023). Dépression. Consulté en mai 2025.

[Our World in Data, 2023] Our World in Data (2023). Population growth. Consulté en mai 2025.

[Poświata and Perełkiewicz, 2022] Poświata, R. and Perełkiewicz, M. (2022). Deproberta-large-depression. https://huggingface.co/rafalposwiata/deproberta-large-depression. Hugging Face.

[Sadeghi et al., 2024a] Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., Berking, M., and Eskofier, B. M. (2024a). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*.

[Sadeghi et al., 2024b] Sadeghi, M., Richer, R., Egger, B., Schindler-Gmelch, L., Rupp, L. H., Rahimi, F., Berking, M., and Eskofier, B. M. (2024b). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3(66).

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Wang et al., 2025] Wang, X., Xu, J., Sun, X., Li, M., Hu, B., Qian, W., Guo, D., and Wang, M. (2025). Facial depression estimation via multi-cue contrastive learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

[Zhang et al., 2024] Zhang, X., Zhang, X., Chen, W., Li, C., and Yu, C. (2024). Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, 14(1):9543.

[État Dépressif, 2023] État Dépressif (2023). Histoire et épidémiologie de la dépression. Consulté en mai 2025.