Decoding Emotional Nuances: A Multimodal Approach to Detecting Depression through Audio, Video, and Text

Decoding Emotional Nuances: A Multimodal Approach to Detecting Depression through Audio, Video, and Text

Siddhant Bikram Shah[1], Shubham Garg[2], Nikolaos Kouis[3] and Aikaterini Bourazeri[3]
[1]Department of Electrical and Computer Engineering, Northeastern University
[2]Tandon School of Engineering, New York University
[3]School of Computer Science and Electronic Engineering, University of Essex

Abstract

Early detection of depression is crucial to prevent serious consequences, such as chronic fatigue, substance abuse, and worsening mental health. Traditional diagnostic methods often rely on self-reported questionnaires, which can be influenced by a patient's willingness to disclose information, or on unimodal approaches that may not capture the full range of depressive symptoms. To address these limitations, we present LUNA (Listen, Understand, Nurture, Advise), a unified multimodal application-based framework designed to emulate real-world mental health assessments by integrating video, audio, and text inputs. LUNA employs individual modules for each modality, combining their results to provide a comprehensive analysis of the user's mental state. Our findings show that each module can independently and effectively screen for depression, and their combined scores yield a more comprehensive and accurate regression score based on the PHQ-8 scale for each user session. Benchmarking against state-of-the-art depression detection models using the DAIC-WOZ dataset demonstrates that LUNA performs comparably to or better than existing validated models. To ensure privacy, no user data is stored post-assessment. Furthermore, the system features an interactive avatar to enhance user engagement and comfort. LUNA represents a significant advancement in the early detection of depression by providing a robust, privacy-conscious and user-friendly diagnostic tool.

Keywords: Deep Learning, Depression Detection, Mental Health Assessment, Multimodal Analysis, PHQ-8.

1. Introduction

Major Depressive Disorder (MDD) is one of the most common mental health disorders, affecting approximately 280 million people worldwide [1]. It is diagnosed when a patient consistently exhibits symptoms such as low energy, lack of concentration, decreased enthusiasm, loss of temper, and feelings of worthlessness [2]. If left untreated for extended periods, depression can lead to severe consequences, including chronic fatigue, substance abuse, permanent neurological damage, and even suicide [3]. Therefore, early detection of depression is crucial to preventing these outcomes.

Despite depression being treatable with early diagnosis, treatment is often constrained by a lack of awareness, limited access to resources, and various other socio-economic factors [4]. Although clinical intervention is the gold standard for depression treatment, patients may refuse treatment due to feelings of vulnerability, the cost of treatment, or the social stigma associated with mental health problems. While less invasive methods for mental health assessment exist, they mainly rely on self-reported questionnaires [5], which may not always be reliable. Such methods depend on voluntary disclosure, which can be influenced by a person's mental state [6], and often fail to capture the subtle complexities of emotional distress, as they focus solely on verbal expression.

Machine learning (ML) has demonstrated stellar performance in various critical domains such as cybersecurity [7] and healthcare [8], often outperforming human experts [9]. However, detecting an individual's mental health remains a significant research challenge [10], and diverse methods have been employed, including the analysis of social media posts [11], speech [12], and responses to questionnaires [13]. Given the complexity of depression, effective assessment cannot be achieved through a single modality, making depression detection inherently a multimodal challenge. Leveraging multiple streams of complementary data enhances the decision-making capabilities of an ML model, and these modalities can compensate for each other's limitations. Research has shown that multimodal methods significantly outperform unimodal approaches in depression detection [14]. Nonetheless, this research area remains underdeveloped due to the challenges associated with

processing, annotating, and disseminating sensitive data [15]. Processing multimodal data requires integrating diverse signals (e.g., facial expressions, vocal tone, and linguistic content), each of which may exhibit inconsistencies or noise due to variations in lighting, background noise, or speaking styles [16]. Annotating depression datasets is also particularly challenging, as labeling depression severity requires expert clinical assessment, and self-reported measures may introduce bias or inconsistency [17]. Furthermore, disseminating sensitive mental health data poses ethical and privacy concerns, as depression-related datasets often contain identifiable patient information [18]. Due to these challenges, many existing methods lack large-scale, high-quality datasets, limiting their ability to generalize across populations. Additionally, many approaches lack a user-friendly interface, making them inaccessible to non-experts [19].

Motivated by these challenges, we propose **L**isten, **U**nderstand, **N**urture, and **A**dvise (LUNA), an interactive application that leverages transfer learning with multimodal data to assess the severity of depression and provide recommendations, all within the comfort of individuals' own homes. LUNA processes user inputs in real-time to generate an immediate mental health assessment and is intentionally designed as a stateless, privacy-preserving screening system that does not retain historical user data or adapt its models over time, prioritizing user trust, ethical deployment, and accessibility in non-clinical and at-home settings. LUNA's design is guided by its core functionalities, where it "listens" through audio analysis, "observes" via video-based emotion recognition, and "interprets" text sentiment analysis. These capabilities align with the intuitive and user-friendly nature of LUNA's screening process. To address the challenge of limited sensitive multimodal data, three independent modules were implemented to process video, audio, and text inputs. By leveraging transfer learning on foundation models, meaningful and distinct neural representations were created, bypassing the need for large amounts of labeled data [20]. This approach is particularly advantageous for depression detection, where task-specific data is often scarce or sensitive [21].

LUNA's modularity allows users to select from three interaction modes, ensuring accessibility across diverse technical and privacy preferences. Users can opt for (1) unimodal text input via keyboard, (2) multimodal audio and text input via a microphone, or (3) full multimodal input incorporating video, audio, and text via webcam. This flexibility ensures a personalized mental health assessment experience tailored to individual needs. The proposed framework integrates video (ResNet-50) [22], audio (Wav2Vec2) [23],

and text (BERT) [24] to enhance depression detection accuracy. While these models have been utilized individually in prior research, the contribution of this study lies in their structured integration, feature alignment, and interpretable multimodal fusion strategy, enabling a more comprehensive and robust mental health assessment. By leveraging complementary signals from different modalities, the framework effectively overcomes the limitations of unimodal approaches and maps the results onto the PHQ-8 scale [25]. The PHQ-8 is a modified version of the PHQ-9 depression screening questionnaire, derived by excluding the item on suicidal ideation, and is considered a more efficient instrument for population-level screenings focused on depressive symptoms rather than acute suicide risk [26].

This study also investigated how to process user inputs in real-time to generate an immediate mental health assessment. To achieve this, an application was developed to engage with and evaluate patients in a manner closely resembling the approach of a mental health professional. The application provides general recommendations to enhance users' mental health based on their assigned mental health scores and tracks the scores after each use. User data is not stored after being securely processed by the application. While not intended to replace professional mental health intervention, the application serves as a private, accessible, and convenient tool for early identification of mental health disorders. Based on these motivations, the contributions of this study are as follows:

- We introduce LUNA, a multimodal framework that integrates video, audio, and text processing modules, leveraging transfer learning to independently evaluate each modality and combine the results for a more accurate and comprehensive mental health assessment.
- We integrate high-performing models for video, audio, and text into a unified system, demonstrating how their combined use improves prediction accuracy over unimodal setups.
- We design a user-friendly interface featuring an empathetic avatar to enhance engagement, guiding users through the PHQ-8 questionnaire and providing a seamless assessment experience.

Accordingly, this paper is structured as follows. Section 2 reviews existing unimodal and multimodal approaches to depression detection, highlighting their limitations and setting the stage for the novel contributions of

LUNA. Section 3 describes the datasets used in LUNA, including FER-2013 for video and DAIC-WOZ for audio and text, detailing the preprocessing steps necessary for model training. Section 4 explains the architecture and functionality of the LUNA framework, describing how it processes video, audio, and text inputs to assess depression and assigns a PHQ-8 score, while Section 5 presents the performance results of the individual video, audio, and text models, as well as the combined multimodal approach used in LUNA. Section 6 discusses the implications of the experimental results, emphasizing the advantages of multimodal data integration for more accurate and comprehensive depression detection. Finally, Section 8 summarizes the key contributions of the paper, focusing on LUNA's effectiveness and potential future research directions in multimodal mental health assessment.

2. Related Work

Machine learning for depression detection has received significant attention in recent years, resulting in the development of numerous approaches to tackle this pressing issue [27]. In this section, we examine recent advances in unimodal, multimodal, and application-based depression detection methods, highlighting the limitations that our proposed work aims to address. Table 1 provides a comparative overview of the methods discussed in the literature.

Table 1: A Comparative Summary of This Study and Other Papers on Unimodal, Multimodal, and Application-Based Methods for Depression Detection

Method	**Type**	**Data Used**	**Modality**			**Interactive**	**Data Privacy**
			Text	**Audio**	**Video**		
Stolicyn et al. [28]	Unimodal	Custom Dataset	✗	✗	✓	✗	✓
Sardari et al. [12]		DAIC-WOZ	✗	✓	✗	✗	✓
Cai et al. [11]		SWDD	✓	✗	✗	✗	✓
Nguyen et al. [17]		RSDD, eRisk2018, and TRT	✓	✗	✗	✗	✓
Othmani et al. [29]		RECOLA and DAIC-WOZ	✓	✗	✗	✗	✓
Yoon et al. [30]	Multimodal	D-Vlog	✗	✓	✓	✗	✓
Fang et al. [31]		DAIC-WOZ	✓	✓	✓	✗	✓
Park and Moon [14]		DAIC-WOZ	✓	✓	✗	✗	✓
Xie et al. [32]		Private Dataset	✓	✗	✗	✗	✓
Kaywan et al. [10]	Application-based	Private Dataset	✓	✗	✗	✓	✗
Anmella et al. [33]		Private Dataset	✓	✗	✗	✓	✗
Rathnayaka et al. [34]		No data used	✓	✗	✗	✓	✗
Jiang et al. [35]		Private Dataset	✓	✗	✗	✓	✗
He et al. [36]		No data used	✓	✗	✗	✓	✗
Jang et al. [37]		No data used	✓	✗	✗	✓	✗
Our study	All	DAIC-WOZ and FER-2013	✓	✓	✓	✓	✓

2.1. Unimodal Methods

Unimodal methods for detecting depression can be beneficial in cases where participants prefer to communicate using only one data modality. These methods typically utilize either video, audio, or predominantly text data. For example, Stolicyn et al. [28] developed an efficient method for detecting symptoms of depression by tracking the face and eye movements of participants while they performed cognitive tasks. Sardari et al. [12] proposed an end-to-end framework that uses a convolutional autoencoder to process audio data. Cai et al. [11] developed a framework to detect depressive symptoms by treating social media posts as a multivariate time series, curating the Sina Weibo Depression Dataset (SWDD), a large-scale annotated dataset based on the tweet history of 3,711 depressed users and 19,526 non-depressed users. Nguyen et al. [17] leveraged depressive symptoms described in the PHQ-9 questionnaire to enhance the generalizability of depression detection models in real-world scenarios. They developed nine symptom detection models corresponding to the nine questions in the PHQ-9 questionnaire and evaluated their method against datasets of social media posts. Othmani et al. [29] introduced EmoAudioNet, a network designed to predict valence, arousal, and depression in voice signals, trained on the RECOLA and DAIC-WOZ datasets. Despite the convenience of unimodal systems, their performance remains inferior to multimodal methods in depression detection [38]. Unimodal approaches often fail to capture the full spectrum of depressive symptoms, as they rely on a single data type, which may not provide a comprehensive view of a user's mental state. This gap highlights the growing interest in multimodal approaches, which our work addresses by integrating multiple data modalities to enhance the accuracy and comprehensiveness of depression detection.

2.2. Multimodal Methods

Multimodal approaches have gained prominence in depression detection due to their ability to capture a broader range of behavioral and emotional cues, which single-modality methods may miss. With the rise of multimodal machine learning [39], significant research has been dedicated to exploring these approaches for detecting depression. These methods often combine video, audio, and text data. Yoon et al. [30] proposed a novel model that leverages visual and acoustic features to detect depression in individuals, marking a significant step in integrating non-verbal cues into depression detection. Similarly, Fang et al. [31] expanded on this approach by developing

MFM-Att, a multimodal regression model that fuses video, audio, and text features to generate a PHQ-8 score as an indicator of depression severity. Park and Moon [14] further advanced the field by proposing a multimodal framework that combines a BERT-CNN model for text processing with a CNN-BiLSTM model for audio processing to classify depression. An attention mechanism was applied to mitigate gradient exploding and information loss in the representation vectors, enhancing the model's robustness. Xie et al. [32] addressed the limitations of the Self-Reported Anxiety Scale and Self-Reported Depression Scale by developing a multimodal model that analyzes facial features as patients respond to these tests. Visual features were encoded and combined with numerical scores from the two tests to create a unified representation for the joint classification of depression and anxiety. Although numerous multimodal methods have been developed over the years, they are rarely integrated into deployable applications for practical use [40]. This is often due to challenges related to data privacy, data scarcity, and the complexity of managing and processing multimodal data streams in real-world scenarios. Our work addresses these limitations by implementing a robust framework that integrates multiple data modalities while ensuring user privacy and providing a practical, deployable solution for real-world applications.

2.3. Application-based Methods

While much of the research on depression detection focuses on data gathered within controlled environments, this approach often fails to account for the extensive variability present in real-world data. However, several methods have been developed to address this gap by applying practical solutions in real-world scenarios. Kaywan et al. [10] proposed a mass screening Depression Analysis (DEPRA) chatbot for early depression detection by analyzing text responses from patients. Their conversational agent, created using DialogFlow, is integrated with Facebook Messenger, which acts as an interface between the user and the chatbot. Similarly, Anmella et al. [33] developed Vickybot, a chatbot deployed on handheld devices to detect and monitor anxiety-depressive symptoms in healthcare workers using the PHQ-9 and GAB-7 self-assessment tests. Vickybot not only suggests strategies for improving mental health but also issues reminders for weekly goals and biweekly reassessments. It incorporates crisis intervention features, including a suicide alert, urgent notifications, and access to emergency resources. In another approach, Rathnayaka et al. [34] developed Bunji, a chatbot designed

to administer Behavioral Activation by reinforcing positive behavior through activity scheduling. The application generates mood scores from text inputs and keeps a record of historical scores, while also facilitating journaling to help patients reflect on events that may have influenced their mood. Expanding beyond text-based chatbots, Jiang et al. [35] leveraged behavioral and psychological cues extracted from video, audio, text, and photoplethysmography (PPG) signals, collected through diverse consumer devices during remote interviews to evaluate mental health. He et al. [36] proposed XiaoE, a mental health chatbot, and conducted a trial to assess its impact on the PHQ-9 scores of 148 Chinese college students after a one-week intervention. They further designed seven modules to foster multiple aspects of mental health and psychology. Additionally, Jang et al. [37] developed Todaki, a mobile app-based chatbot that provides cognitive behavioral therapy to attention-deficit adults. This study was pioneering in investigating the benefits of chatbots for patients with attention deficiencies, comparing these interventions with traditional methods such as informative books. While these application-based methods demonstrate effective performance in practical scenarios, they also present notable shortcomings, including the potential for less robust predictions due to unimodal processing and significant risks associated with storing sensitive data. Moreover, these approaches often lack the flexibility to allow users to choose their preferred modality based on their needs and technical capabilities.

Our proposed framework, LUNA, addresses these limitations by utilizing multiple input streams to provide a more accurate diagnosis than unimodal methods. LUNA combines independent modules that process video, audio, and text modalities, tackling the limited availability of multimodal data in this domain. Additionally, we integrate an empathetic avatar to support user engagement in mental health assessments without storing any user data. Finally, LUNA allows users to choose between text, audio, and video inputs according to their convenience, improving prediction robustness in low-resource settings while safeguarding user privacy in real-world applications.

3. Datasets

3.1. Video

The FER-2013 dataset [41] was used to train the video processing module. The FER-2013 dataset contains 32,298 grayscale images each with a resolution of 48×48 pixels, encompassing 7 facial emotions: 'Neutral', 'Angry',

'Disgust', 'Fear', 'Sad', 'Happy', and 'Surprise'. The dataset is divided into a training set of 28,709 samples and a test set of 3,589 samples, following the default train-test split. FER-2013 was selected due to its large dataset size and well-annotated emotional expressions, enabling effective transfer learning for facial emotion recognition. Since depression is often characterized by specific emotional expressions, leveraging a robust emotion recognition model can help infer depression-related facial cues. While several depression-specific datasets such as AVEC 2013 [42] exist, they often have limitations such as restricted access, smaller sample sizes, or missing modalities. Additionally, many domain-specific video datasets are typically recorded in laboratory environments or contain encrypted visual data, limiting their generalizability. FER-2013, on the other hand, contains diverse real-world instances of obstructed, partial, and low-resolution facial images, contributing to a model that is more robust to common real-world irregularities in facial emotion recognition [43]. Training on grayscale images also enhances model performance in low-light conditions commonly encountered in real-world scenarios [44]. To further improve robustness against variations in face angle and partial occlusions, facial images were converted into facial landmarks [45]. The F1-Score was used to evaluate image classification, considering the class imbalance in the FER-2013 dataset.

Following prior research in depression detection [46, 47], emotions such as sadness, fear, and disgust have been associated with depressive symptoms. While some studies have linked neutral expressions to emotional blunting and apathy in individuals with depression [48], neutral expressions can also occur in non-depressed individuals who are simply in a resting or unexpressive state. To minimize misclassification and ensure a more conservative labeling approach, 'Neutral' expressions were categorized as non-depressive in this study. In contrast, positive emotions such as happiness and surprise are generally not linked to clinical depression and are categorized as non-depressive. Given these distinctions, we adopted the methodology of Kumar et al. [49] to convert the multi-class emotion labels in the dataset into a binary classification task, grouping depressive and non-depressive emotions accordingly.

Emotions from the 'Angry', 'Disgust', 'Fear', and 'Sad' classes were categorized as depressive, while 'Neutral', 'Happy' and 'Surprise' classes were categorized as non-depressive. This adjustment ensures that neutral expressions are not inherently classified as depressive, aligning better with existing psychological studies that associate depression with more distinct emotional

markers rather than emotional neutrality. Samples labeled as 'Angry', characterized by lowered eyebrows that are drawn together and tense facial muscles, were classified as depressive, reflecting the potential manifestation of depressive symptoms through irritability and low temper. 'Disgust' samples, marked by wrinkled nostrils and raised cheeks, were also classified as depressive due to evidence suggesting that depressed individuals exhibit increased sensitivity to negative emotions. 'Fear' samples, identified by raised eyebrows and widened eyes, were placed in the depressive category because depression is commonly associated with heightened anxiety and hesitation toward new stimuli. Conversely, 'Neutral' samples were categorized as non-depressive to prevent the misclassification of emotionally neutral expressions as indicative of depression. 'Happy' samples, characterized by slightly squinted eyes and an upward-turning mouth into a smile, were assigned to the non-depressive category, as non-depressed individuals are more likely to express joy in response to prompts from the application. Similarly, 'Surprise' samples, featuring raised eyebrows, widened eyes, and parted lips, were categorized as non-depressive, since non-depressed individuals generally respond more actively to new stimuli, such as questions posed by the application's avatar.

3.2. Audio

Audio data from the DAIC-WOZ dataset [50] was used to train the audio processing module. For each patient P, all utterances were segmented as separate samples $U_P = \{U_1, U_2, ...U_n\}$ using the speech timestamps provided in the DAIC-WOZ dataset. The problem was modeled as a regression task, with each patient's PHQ-8 score serving as the label for the corresponding samples. This process resulted in an utterance-level dataset, consisting of a training set with 16,908 samples and a test set with 8,809 samples. The audio processing module was trained as a regression model, predicting continuous PHQ-8 scores rather than classifying users into discrete categories. Performance was evaluated using the Root Mean Square Error (RMSE) metric. The audio processing module was designed to identify underlying depressive indicators solely from an individual's speech signal.

3.3. Text

Transcripts from the DAIC-WOZ dataset were used to train the text processing module. The word count of the text samples for each patient ranged from 167 – 4,709. Since one word is approximately equal to one token in language transformer models, many of these samples greatly exceeded the

maximum positional token limit of architectures such as BERT, DistilBERT, and RoBERTa, which can accommodate up to 512 tokens [51]. To address this limitation, each participant's utterances were segmented into text groups containing fewer than 512 tokens, thus preventing the loss of information that would result from truncation. This process yielded a training set of 368 samples and a test set of 188 samples, with each sample containing fewer than 512 word tokens after preprocessing. The effectiveness of the text regression task was evaluated using the RMSE metric. The text processing module was designed to identify the risk of depression based on a user's responses to questions about their mood and habits in the recent past.

4. Methodology

This section describes each component of the proposed framework LUNA. The study was conducted and reported in accordance with the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) guidelines, and a completed CLAIM checklist is provided as supplementary material [52]. Users interact with the front-end application that captures video, audio, and text data through their webcam and microphone. These inputs are preprocessed and fed into their respective classification models. The system is based on the PHQ-8 questionnaire, with users answering eight questions about their recent mental health. The visual, audio, and text processing modules of the proposed framework each assign a unimodal score $s_{unimodal} \in [0, 1]$, with 0 indicating no signs of depression and 1 indicating the highest risk of depression. Therefore, the classification framework operates as a combination of three regression methods, allowing input data in three modes: text-only, audio+text, or video+audio+text according to the users' convenience, preference, or willingness. Similar to the PHQ-8 scale, the total score for each question ranges from 0 - 3; therefore, when all eight questions are combined, the maximum possible multimodal score is $s_{mm} \in [0, 24]$. This process is visualized in Figure 1.

4.1. Video

Facial affect has been closely linked to symptoms of depression [48]. The LUNA framework leverages this correlation by using a facial recognition model to analyze the user's captured video. The video is divided into frames $f = \{f_1, f_2, ..., f_n\}$, and the emotion for each frame is predicted. To achieve near real-time efficiency, ten frames per second of video are extracted as a

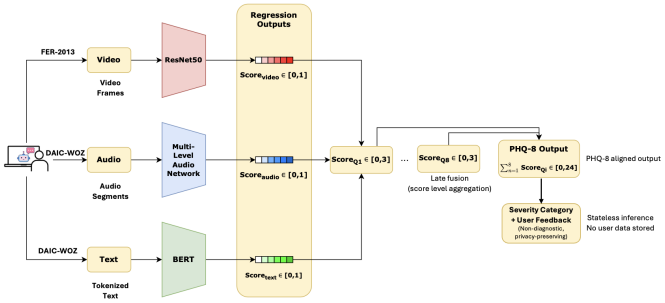


Figure 1: An overview of the proposed LUNA framework and end-to-end screening pipeline. Video, audio, and text inputs are obtained from the user and processed using ResNet50, a multi-level audio network, and BERT, respectively. For each question asked by LUNA, a score between 0 and 1 is assigned for each modality. These scores are then aggregated for each question to produce a score between 0 and 3 using a late-fusion strategy at the score level. Across all 8 questions, the total score is aggregated to yield a final score ranging from 0 to 24, following the PHQ-8 scale.

trade-off between latency and temporal coverage. Each frame is resized to 48×48 pixels before classification. A pre-trained ResNet50 model [22] is fine-tuned as the backbone for the classification task. ResNet-50 was selected due to its strong representation learning capabilities, particularly in extracting hierarchical features from images. Its relatively deep architecture allows it to capture subtle facial affect patterns that are relevant for emotion recognition and depression-related analysis.

Additionally, ResNet-50 offers a good balance between model complexity and computational efficiency, making it suitable for real-time applications without compromising accuracy. For each frame, the facial images are converted into a set of 3D facial landmarks using the Mediapipe library's Face Mesh model [53], extracting 478 facial landmarks across three dimensions, resulting in a feature vector $x_{landmarks} \in \mathbb{R}^{3\times478}$. Among these, 468 landmarks represent the entire facial structure, while 10 landmarks represent the irises. These landmarks are then fed into the ResNet-50 model, which generates an intermediate representation $x_{visual} \in \mathbb{R}^{1000}$.

Each frame f_i is converted into facial landmarks $x^{(i)}_{landmarks} \in \mathbb{R}^{3\times478}$, which are passed through a ResNet-50 backbone to extract a feature representation. A dense layer with sigmoid activation then maps this representa-

tion to a probability of depressive affect:

$$c(f_i) = \sigma\big(w_v^\top f_{\text{ResNet}}(x^{(i)}_{landmarks};\, \theta_v) + b_v\big), \quad c(f_i) \in [0, 1] \tag{1}$$

where $f_{\text{ResNet}}(\cdot;\, \theta_v)$ denotes the ResNet-50 transformation with parameters θ_v and $\sigma(\cdot)$ is the sigmoid function. Here, $c(f_i)$ is interpreted as the probability that frame f_i is either depressive (+1) or non-depressive (+0). Depressive emotions include sadness, fear, and disgust, while non-depressive emotions include neutral, happy, and surprise. The final unimodal video score, s_{visual}, is obtained by averaging the frame-level probabilities across all frames:

$$s_{visual} = \frac{1}{n} \sum_{i=1}^{n} c(f_i) \tag{2}$$

where n is the total number of analyzed frames. The model is trained to differentiate between these two categories using facial affect recognition techniques, aligning with prior studies on emotion-based depression detection [46].

4.2. Audio

The manifestation of depression can affect a person's speech and articulation [54]. Conversely, audio waves in speech contain valuable information that can be effectively used to detect depression [55]. The raw audio signal is preprocessed by splitting it into three components:

1. **Low-Level Features:** Low-Level Descriptors (LLDs) are extracted from the raw audio signal using the eGeMAPSv02 [56] feature set through the OpenSMILE library [57]. This process produces a set of features $f_{lld} \in \mathbb{R}^{l \times 25}$, where l depends on the length of the sequence. These features include loudness, shimmer, jitter, and other time-domain and frequency-domain characteristics. To create uniform tensors for training, a max-pooling-like operation is applied by selecting the maximum values across l for each audio sample, resulting in a final LLD tensor $x_{lld} \in \mathbb{R}^{1 \times 25}$.
2. **Mel-Spectrogram:** The audio samples are converted into mel - spectrograms using the librosa library [58]. This transformation maps all frequencies in the audio sample to the mel scale, representing signal amplitude across time and frequency. The number of mels is fixed at

128, and the mel - spectrogram is resized using torchvision to produce a mel-spectrogram feature vector $f_{mel} \in \mathbb{R}^{128\times128}$. These features are then passed through a pre-trained ResNet-50 vision model [22] to be converted into an intermediate representation $x_{mel} \in \mathbb{R}^{1000}$.

3. **Foundation Model Representation:** The pre-trained Wav2Vec2-Base model [23], with frozen parameters, is used to convert the raw audio signal into high-level representations that encompass both time-domain and frequency-domain features. The model outputs an intermediate representation $x_{w2v2} \in \mathbb{R}^{768}$.

The choice of ResNet-50 for the mel-spectrogram data was based on its proven effectiveness in image classification tasks, which directly applies to spectrogram analysis given the similar nature of these data types. ResNet-50's depth allows it to capture complex patterns in the frequency domain that are crucial for distinguishing between depressive and non-depressive speech patterns. Wav2Vec2-Base was selected due to its state-of-the-art performance in speech representation learning, particularly its ability to capture nuanced audio features across different time scales.

The representations x_{lld}, x_{mel}, and x_{w2v2} are then converted into low-dimensional projections $x_{lld}^{proj}, x_{mel}^{proj}, x_{w2v2}^{proj} \in \mathbb{R}^{24}$ using individual dense layers L_{lld}, L_{mel}, and L_{w2v2} respectively. Empirical findings suggest that a tensor of size $\mathbb{R}^{24}$ effectively encapsulates the information from all components into a low-dimensional embedding without sacrificing performance.

$$x_{lld}^{proj} = L_{lld}(x_{lld}) \tag{3}$$

$$x_{mel}^{proj} = L_{mel}(x_{mel}) \tag{4}$$

$$x_{w2v2}^{proj} = L_{w2v2}(x_{w2v2}) \tag{5}$$

These projections are then concatenated into a single multi-level audio representation $x_{audio} \in \mathbb{R}^{72}$:

$$x_{audio} = x_{lld}^{proj} \oplus x_{mel}^{proj} \oplus x_{w2v2}^{proj} \tag{6}$$

where $\oplus$ denotes the concatenation operation. This unified feature vector is then passed through a regression layer with parameters (w_a, b_a), which maps the embedding into the unimodal audio score:

$$s_{audio} = w_a^\top x_{audio} + b_a \tag{7}$$

Here, $s_{audio} \in \mathbb{R}$ is subsequently normalized to the PHQ-8 scale, providing a clinically interpretable measure of depression severity from the audio modality.

4.3. Text

Language is one of the most widely used tools for depression detection as it captures the emotions that individuals choose to convey [59]. Additionally, the text modality avoids many of the privacy concerns associated with video and audio data. The recent rise in transformer-based NLP networks has highlighted their proficiency in handling complex tasks such as emotion recognition and depression detection [60]. In the proposed framework, a pretrained language transformer, BERT [24], is fine-tuned to process the textual content of the input. BERT was selected for its state-of-the-art performance in natural language understanding tasks, particularly its ability to capture contextual information and handle long sequences of text. Its transformer architecture, with its self-attention mechanism, excels at capturing the nuances of language, making it ideal for detecting linguistic indicators of depression. Additionally, BERT's versatility and transfer learning capabilities enable it to generalize effectively across different text inputs, enhancing the robustness of the proposed model.

After the text is preprocessed, BERT's tokenizer is used to convert the text into tokens, mapping words to their corresponding IDs in BERT's vocabulary. The tokenizer also adds the special tokens '[CLS]' at the beginning and '[SEP]' at the end of the text. BERT then converts the tokenized sequence $t_{tokenized}$ into an intermediate representation $x_{text} \in \mathbb{R}^{768}$.

$$x_{text} = BERT(t_{tokenized}) \tag{8}$$

The text representation is then passed through a regression layer with parameters (w_t, b_t) to predict the unimodal text score:

$$s_{text} = w_t^\top x_{text} + b_t \tag{9}$$

The predicted score, $s_{text} \in [0, 1]$, provides a continuous measure of depressive markers in the text, where higher values indicate stronger depressive cues.

4.4. Combined Prediction

To validate LUNA's effectiveness, its predictions were benchmarked against published models using the DAIC-WOZ dataset, a clinically labeled resource widely used in depression detection research. Comparison with prior multimodal frameworks demonstrates that LUNA performs at a competitive level, reinforcing its reliability. LUNA combines information from visual, acoustic, and linguistic modalities through a modular multimodal fusion framework designed to balance predictive performance, interpretability, and deployment practicality. At the representation level, each modality is processed through a dedicated feature extraction pipeline tailored to its signal characteristics, preserving modality-specific information prior to integration. At the system level, the outputs of these modality-specific components are aggregated to generate the final PHQ-8 severity estimate and depression classification outcome. This design was intentionally chosen to maintain transparency, robustness to missing or degraded modalities, and ease of deployment in privacy-sensitive real-world settings, while still allowing complementary multimodal cues to contribute to the final decision. The framework therefore emphasizes structured multimodal integration within an interpretable system architecture suitable for practical mental health screening.

To estimate the PHQ-8 score, the outputs from the video, audio, and text models are aggregated to produce a final depression severity score. By aggregating modality-specific outputs, this multimodal fusion approach captures relevant depression indicators from different behavioral and linguistic sources. In the current implementation the final PHQ-8 score is computed through an interpretable score-level late-fusion strategy, following the principle that depression symptoms may manifest differently across modalities and that their combined effects contribute to overall severity. This approach is consistent with multimodal depression detection studies, which suggest that integrating different behavioral signals can improve predictive performance compared to unimodal models [61, 62]. While this implementation assumes equal weighting across modalities at the final aggregation stage, we acknowledge that their individual predictive contributions vary. The equal-weighting strategy was deliberately chosen in the current version for simplicity, interpretability, and fairness in modality weighting, while preserving modularity and robustness to missing or low-quality modalities, properties that are particularly important for real-world, user-facing mental health screening. The final prediction process is formalized as follows. Let $\mathcal{M}$ denote the set of modalities enabled by the user (e.g., $\{visual, audio, text\}$), where $N = |\mathcal{M}|$.

The combined score for a single question is defined as the sum of the unimodal scores across available modalities:

$$S_{question} = \sum_{j \in \mathcal{M}} s_j, \quad S_{question} \in [0, N] \tag{10}$$

Since each unimodal score $s_j \in [0,\ 1]$, the combined score for a single question lies in the range [0, N], where N is the number of available modalities. The total raw score for a session of 8 questions is then obtained by summing across all responses:

$$S_{raw} = \sum_{i=1}^{8} S_{question_i}, \quad S_{raw} \in [0, 8N] \tag{11}$$

The PHQ-8 score is computed as a continuous value ranging from 0 to 24, with higher scores indicating a greater severity of depression. To ensure consistency across different input modalities, the score is adjusted while maintaining the $0 - 24$ range. For interpretability, these scores are mapped to standard PHQ-8 severity levels: Non-Depressed $(0 - 4)$, Mild Depression $(5 - 9)$, Moderate Depression ($10 - 14$), Moderately Severe Depression $(15 - 19)$, and Severe Depression $(20 - 24)$. To ensure consistency of prediction when users select one or two modalities instead of all three, the final PHQ-8 score is normalized. The maximum possible raw score is $8 \times N$, where $N \in \{1, 2, 3\}$ denotes the number of modalities used. To map this to the standard 0-24 range, we use the following normalization:

$$S_{final} = S_{raw} \times \frac{3}{N} = \left(\sum_{i=1}^{8} S_{question_i} \right) \times \frac{3}{N} \tag{12}$$

This normalization ensures that the final PHQ-8 score remains within the standard $0 - 24$ range, independent of the number of modalities provided. This classification method aligns with clinical PHQ-8 guidelines, making the results interpretable within established mental health assessment frameworks [63, 64, 65].

The models were trained using appropriate loss functions for their respective tasks. For the regression-based PHQ-8 score prediction in the audio and text modalities, the Mean Squared Error (MSE) Loss was used, which minimizes the difference between predicted and actual scores. For the video-based

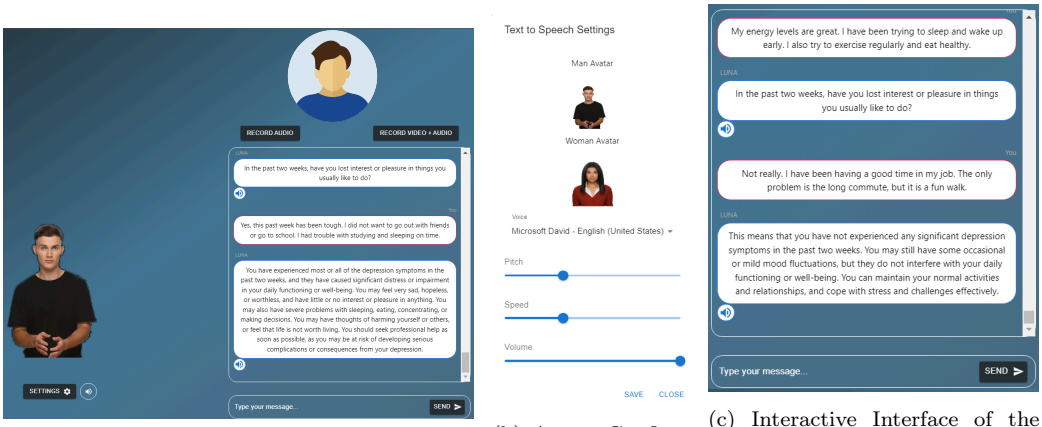
classification model, Cross-Entropy Loss was employed, which is well-suited for binary classification tasks.

4.5. User-Centric Interface

In the domain of mental health support, the development of a web-based application with an interactive avatar serves as a promising avenue for fostering engagement and well-being [66]. The core concept of the user interface involved an empathetic digital companion (i.e., avatar) that poses questions to users, facilitating self-reflection and expression [67]. The success of such an application depends not only on the sensitivity of its content but also on the simplicity and intuitiveness of its user interface (UI). The primary goal in designing the UI was to ensure accessibility and ease of use. The layout was intentionally kept clean and uncluttered, with a minimalist design that uses clear fonts and well-defined elements to promote a calm and straightforward user experience. While no formal usability evaluation was conducted in this study, the design was informed by best practices in mental health technology interfaces.

The avatar design plays a crucial role in fostering a safe and supportive environment. Rather than serving as a direct "inner projection" of the user, the avatar functions as a neutral and empathetic digital companion, designed to facilitate engagement and encourage self-reflection. To provide a degree of personalization while maintaining simplicity, users can select between male and female avatars and adjust voice characteristics, enhancing a sense of ownership. Real-time feedback from the avatar, acknowledging user responses with empathy and understanding, strengthens user engagement. Visual representations of LUNA's user interface are shown in Figure 2. By prioritizing simplicity and intuitiveness, the design ensures a digital space where technology acts as a gentle guide, fostering a positive user experience.

At the end of each session, users receive guidance based on their assigned score, helping them interpret their assessment results and identify appropriate mental health support options. For individuals in the 'non-depressed' category, the system indicates that no significant depressive symptoms are present and recommends maintaining normal functioning. For 'mildly depressed' individuals, it is noted that slight depressive symptoms may slightly affect their daily mood but not their responsibilities. For those assigned to the 'moderate depression' category, seeking professional help is suggested, as they exhibit noticeable symptoms that may impair daily functioning and well-being. Individuals with 'moderately severe depression' are recommended



(a) User Interface layout of the LUNA System

(b) Avatar Configuration Settings.

(c) Interactive Interface of the LUNA Depression Detection Chatbot.

Figure 2: Overview of the LUNA System Interface and Key Features.

to seek professional intervention, as they may experience significant distress from depressive symptoms. Finally, for individuals classified as 'severely depressed', urgent professional help is strongly urged, as they may be at risk of extreme distress or a potential crisis. While the wording across categories may appear subtle, the recommendations are aligned with the established PHQ-8 clinical severity scale and are intentionally phrased to match the level of urgency and functional impact associated with each range. For example, the escalation from "suggested" to "recommended" to "strongly urged" corresponds to increasing symptom severity and risk. These distinctions are designed to remain general, empathetic, and non-diagnostic, preserving user privacy and safety in a self-screening context.

5. Experimental Results

To determine the most effective architecture for each modality within the LUNA framework, a comprehensive benchmarking analysis was conducted. For each data type, a range of models was evaluated, from traditional machine learning algorithms to various state-of-the-art deep learning architectures. This section presents the comparative results that informed the final model selections for the video, audio, and text modules. A primary motivation for developing LUNA was the lack of publicly available datasets that encompass visual, audio, and text modalities. To address this challenge, three separate models were trained, each designed to handle a single modality. The entire

framework was implemented using Pytorch 2.1.0 and trained on an NVIDIA P100 GPU with 16 GB of dedicated memory. A batch size of 16 was used for the image and audio processing modules, while a batch size of 8 was applied for the text processing module. The learning rates were set to $1 \times e^{-5}$ for the image and audio models, and $5 \times e^{-4}$ for the text model.

5.1. Unimodal Experimental Results: Video

Table 2 and Figure 3 present the results of the video processing module as a classification task. Figure 3 provides a visual comparison of model performance, complementing the detailed analysis below. The complex facial images were decomposed into facial landmarks and fed them into the vision models. The random forest model struggled to process the high-dimensional 3D face meshes, resulting in poor performance. Similarly, the ANN failed to capture the spatial relationships in images, leading to suboptimal results. The CNN performed significantly better than the previous methods, but it still struggled to capture higher-level features and was prone to overfitting due to its relatively shallow depth. While CNN-based models such as VGG and MobileNet performed reasonably well, they may lack the depth or architectural design to extract higher-level or more abstract facial features relevant for depression detection, as seen in the slightly lower F1-scores compared to ResNet-50 or transformer-based models. The VGG16 architecture, a deep 16-layer CNN, improved performance compared to the shallower CNN. The MobileNet and EfficientNet-B3 models offered a trade-off between performance and computational efficiency, with slightly reduced accuracy. The ViT-B/16 transformer surpassed the previous models, demonstrating its strong capability in image classification. Notably, traditional deep-learning models achieved competitive performance compared to the vision transformer model, which may be due to the relatively straightforward nature of facial emotion recognition. The ResNet-50 model marginally achieved the best performance among all the models tested, and we implemented it to process the individual frames of the captured video. The results from the video processing module suggest that while modern architectures like ViT-B/16 offer strong performance, more traditional deep-learning models like ResNet-50 remain highly competitive for specific tasks like facial emotion recognition. This finding highlights the efficiency of ResNet-50 in handling the facial landmark data, which is beneficial for modeling subtle facial affect patterns observed in the dataset. ResNet-50 was selected for the LUNA framework due to its superior ability to extract the complex, hierarchical features necessary for this task. This

benchmarking establishes ResNet-50 as the most effective and reliable architecture for the visual modality, balancing high accuracy with computational efficiency. The overall effectiveness of the video module in isolation, however, is limited by its focus on visual cues alone, underscoring the need for a multimodal approach to capture a more comprehensive picture of an individual's mental state.

Table 2: F1-Score results for the binary classification task on the FER-2013 dataset. The classification task distinguishes between "depressive" and "non-depressive" facial expressions. Depressive expressions include emotions such as sadness, fear, and disgust, while non-depressive expressions include neutral, happy, and surprise.

Method	F1-Score
Random Forest	70.72
ANN	74.41
CNN	80.97
VGG-16	84.15
MobileNet	82.21
EfficientNet-B3	81.64
ViT-B/16	84.38
ResNet-50	**84.57**

5.2. Unimodal Experimental Results: Audio

Table 3 and Figure 4 present the results of the audio processing module as a regression task, with Figure 4 providing a clear visual comparison of unimodal and multi-level models. Individual models were tested at different levels of audio data before combining the best-performing models for each level into multi-level models. For signal-based models, the Wav2Vec2 transformer outperformed the BiLSTM and 1-D CNN models, likely because these latter architectures struggle to model long-term dependencies in sequences. For processing LLDs, the DNN and random forest algorithms exhibited similar performance. However, the higher RMSE for both models suggests that LLDs cannot fully capture the complexity of audio patterns in compressed feature descriptors. The mel-spectrogram data provided richer information, as evidenced by the improved regression performance of the spatial models, with the ResNet-50 and ViT-B/16 models performing similarly in this task. Ultimately, the Wav2Vec2, DNN, and ResNet-50 models were selected for

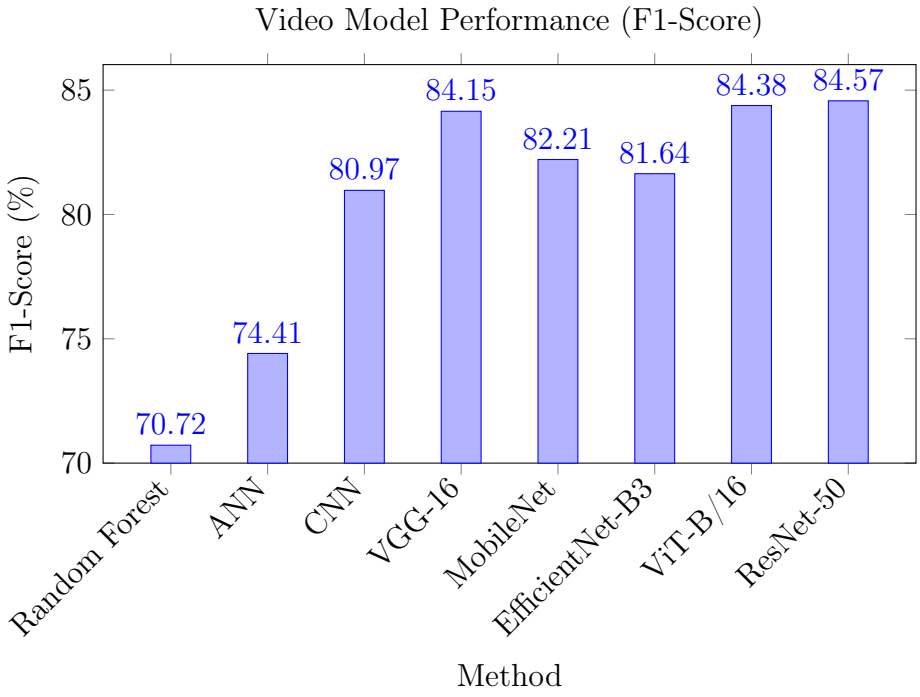


Figure 3: Comparison of F1-Scores for various video classification models on the FER-2013 dataset.

the multi-level models due to their superior performance with each respective data type.

Among the multi-level models, the signal + mel-spectrogram and signal + LLD models were outperformed by the LLD + mel-spectrogram model, indicating that the information contained in LLDs and mel-spectrograms is complementary to each other and has fewer redundancies, leading to better multi-level regression performance. The three-level model, which integrates signal, LLD, and mel-spectrogram data, produced the best regression results overall, suggesting that these diverse data types combine effectively to capture the complex nature of audio data for more nuanced classification. Based on these findings, the top-performing model from each data type was combined into a multi-level neural network architecture to process the audio data. The audio processing results emphasize the value of integrating multiple audio features, as the multi-level models consistently outperform unimodal approaches. This reinforces the importance of multifaceted strategies in capturing the complex nature of depressive symptoms, which manifest

differently in various modalities. By effectively combining signal, LLD, and mel-spectrogram data, the model is better equipped to detect subtle changes in speech that may indicate depression, contributing to more reliable and nuanced mental health assessments. These benchmarking results demonstrate that multi-level integration of signal, LLD, and mel-spectrogram features provides the most robust approach for audio-based depression assessment.

Table 3: Audio regression results on the audio samples of the modified DAIC-WOZ dataset. RMSE stands for Root Mean Square Error.

Data Type	Method	RMSE
Signal	Wav2Vec2	6.35
	BiLSTIM	6.40
	1-D CNN	6.37
LLD	Random Forest	6.51
	DNN	6.50
M-S	VGG-16	6.37
	ResNet-50	6.34
	ViT-B/16	6.34
Signal + M-S	Wav2Vec2 + ResNet50	6.20
Signal + LLD	Wav2Vec2 + DNN	6.21
LLD + M-S	DNN + ResNet50	6.16
Signal + LLD + M-S	Multi-level DNN	**6.09**

5.3. Unimodal Experimental Results: Text

Table 4 and Figure 5 present the results of the text processing module as a regression task. Both naive deep-learning models and state-of-the-art NLP transformers were evaluated. The 1D-CNN model produced the highest RMSE score, primarily due to the sparse representations of text created by one-hot encoding and the CNN architecture's limitations in modeling long-range dependencies. The Word2Vec model performed slightly better due to its embeddings capturing the semantic relationships between words. However, Word2Vec treated each token independently and failed to fully capture the sequential nature of long text sequences, limiting its performance. The LSTM model improved upon these methods, as its memory mechanism allowed it to capture relationships within large sequences of words. The BiLSTM model performed even better by considering both past and future context for each word. Despite these improvements, both LSTM and

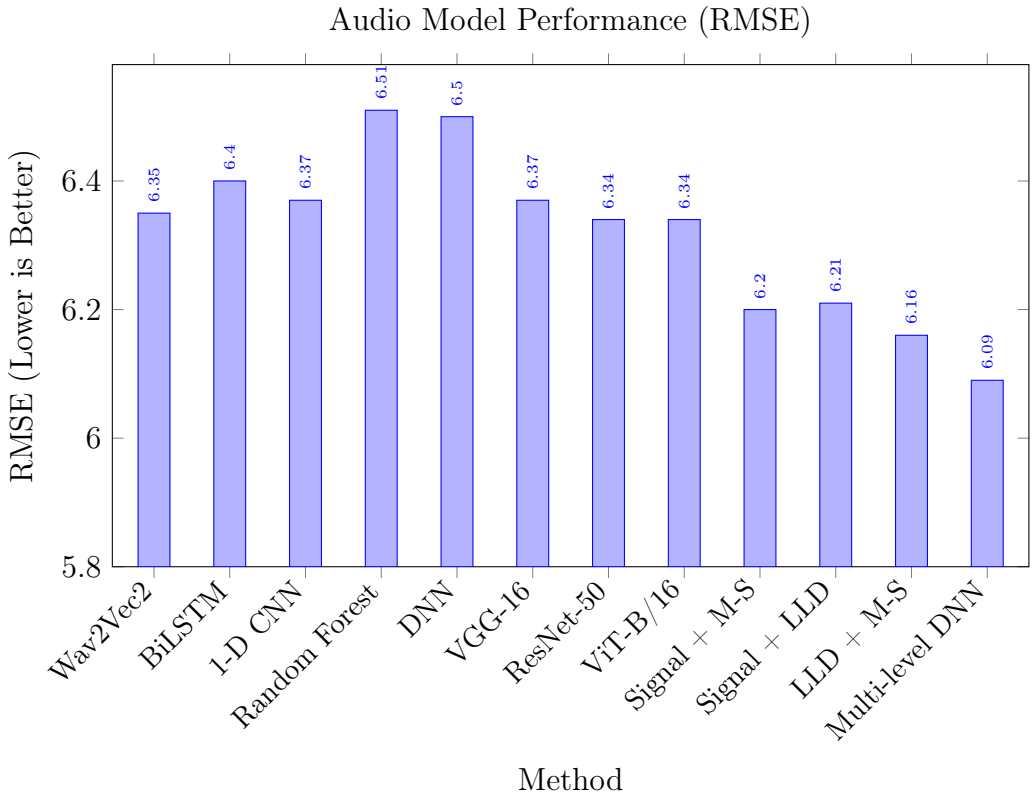


Figure 4: Comparison of RMSE for various audio regression models and feature combinations on the DAIC-WOZ dataset.

BiLSTM models were significantly outperformed by the transformer models. The DistilBERT model showed a notable improvement in RMSE over the traditional deep-learning models but was outperformed by other transformer models due to its small size. The RoBERTa and ALBERT models showed similar performance with a marginal improvement over DistilBERT. Among the transformer models from the BERT family, the baseline BERT model achieved the best RMSE, indicating that its size was optimal for capturing the complex linguistic dependencies present in the long transcripts of the DAIC-WOZ dataset. These findings are further illustrated in Figure 5, which provides a clear visual comparison of the performance gap between traditional deep-learning models and transformer-based architectures.

The text processing results highlight the significant advantage of using transformer-based models for natural language processing tasks, particularly

in the context of depression detection. BERT's superior performance indicates that its architecture was optimally sized to capture the complex linguistic dependencies in the DAIC-WOZ transcripts, justifying its selection for the LUNA framework. These findings support the broader trend in NLP towards leveraging transformers for complex language tasks. Moreover, the results reaffirm the limitations of relying solely on unimodal text data, as even the best-performing text models cannot fully encapsulate the emotional and behavioral nuances that might be captured through audio and video modalities. Overall, benchmarking across traditional and transformer-based models confirms BERT as the most effective architecture for text, achieving the lowest RMSE and providing the most reliable linguistic representations.

Table 4: Text regression results on the transcripts of the modified DAIC-WOZ dataset. RMSE stands for Root Mean Square Error.

Method	RMSE
1-D CNN	7.06
Word2Vec	6.91
LSTM	6.80
BiLSTM	6.71
DistilBERT	6.26
RoBERTa	6.25
ALBERT	6.25
BERT	**6.23**

5.4. Multimodal Experimental Results

In addition to unimodal experiments, the full multimodal version of LUNA was evaluated, where video, audio, and text modalities were combined to predict PHQ-8 scores. The results presented in Table 5 indicate that multimodal integration yields the best overall performance across both regression and classification settings, indicating that complementary information from multiple m

The multimodal configuration achieves an RMSE of 5.42 for PHQ-8 score prediction, which is lower than either unimodal approach. Relative to the strongest unimodal regression baseline (audio, RMSE = 6.09), this corresponds to an absolute improvement of 0.67 and an approximately 11.0% relative reduction in error. Similarly, in classification tasks, the F1-score

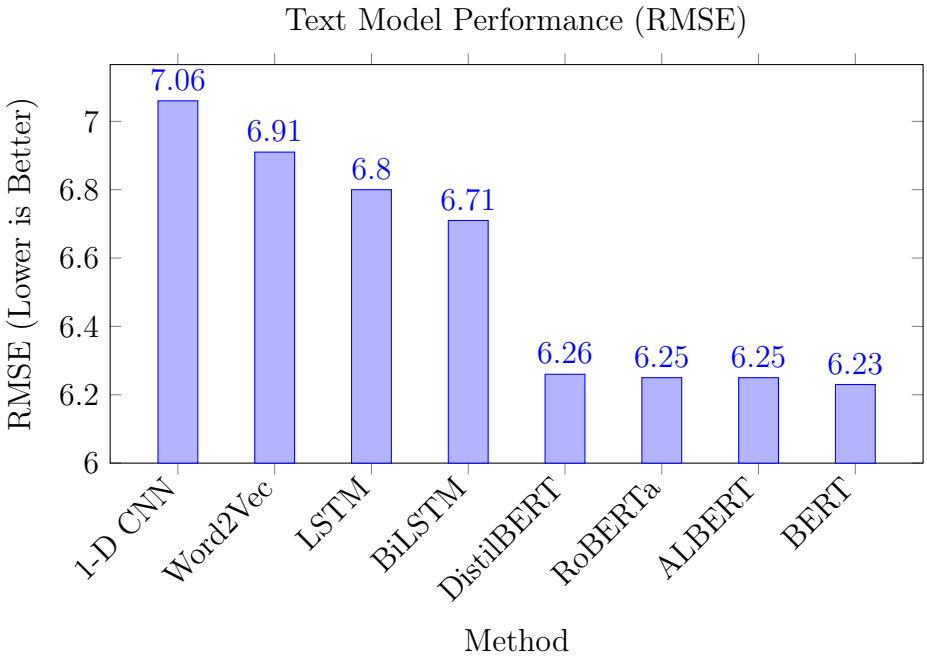


Figure 5: Comparison of RMSE for various text regression models on the DAIC-WOZ dataset.

increases from 84.57% (video-only) to 86.12%, corresponding to a gain of 1.55 percentage points. While these gains are modest in absolute terms, they suggest that multimodal integration provides a measurable and consistent performance benefit in this setting. This improvement can be attributed to the ability of multimodal learning to capture different aspects of depression symptoms. The text modality identifies linguistic indicators, the audio modality captures variations in vocal tone, and the video modality detects facial expressions. By integrating these sources, the model produces more robust assessments, reducing the likelihood of misclassification based on a single modality. Rather than relying on any single modality alone, the framework combines heterogeneous behavioral, vocal, and linguistic evidence relevant to depression screening. These findings support the use of multimodal approaches for automated depression assessment, while also highlighting important limitations. While the framework achieves competitive performance, some limitations remain: dataset separation (with video and audio/text trained independently) restricts cross-modal learning, and

Table 5: Comparison of Unimodal and Multimodal Performance for PHQ-8 Score Prediction on the DAIC-WOZ dataset, illustrating the complementary contribution of multiple modalities.

Modality	**RMSE**	**F1-Score (Classification)**
Text Only	6.23	-
Audio Only	6.09	-
Video Only	-	84.57%
Multimodal (Audio + Text + Video)	**5.42**	**86.12%**

the equal-weighted fusion strategy, while interpretable, may not fully leverage the differing predictive strengths of each modality.

6. Discussion

The results across video, audio, and text modalities emphasize the individual strengths of each modality in detecting depressive symptoms, while also highlighting the critical value of multimodal integration within the LUNA framework. The findings suggest that multimodal approaches improve depression detection compared to unimodal methods, as evidenced by lower RMSE and higher F1-scores in the multimodal setup.

The ResNet-50, multi-level audio network, and BERT models performed well within their respective domains, effectively capturing the nuances of each modality. The combination of these high-performing models into a unified system highlights the potential of multimodal frameworks to provide a more holistic understanding of a user's mental state. This observation is consistent with prior studies showing that facial affect, speech characteristics, and linguistic cues provide complementary information for depression screening when jointly modeled [39, 14, 31].

Video processing, powered by the ResNet-50 model, effectively captures facial affect patterns relevant to emotion recognition, while audio processing captures nuanced speech characteristics through the multi-level DNN model, which combines signal, LLD, and mel-spectrogram data for improved performance. The text module, using BERT, excels at analyzing linguistic features associated with depression. These modality-specific results highlight how different behavioral signals contribute distinct information to the overall screening process. Furthermore, even within individual modalities, the results underscore the importance of combining multiple feature types.

For instance, the multi-level approach in audio processing, which integrates signal, LLD, and mel-spectrogram data, outperformed unimodal methods, reinforcing the value of multi-representation modeling for depression screening.

Beyond aggregate performance, the unimodal and multimodal results also provide insight into the role of each modality within the fusion framework. The audio modality yielded the strongest standalone regression performance, suggesting that vocal characteristics capture a substantial portion of the signal relevant to PHQ-8 severity estimation. The text modality performed similarly, indicating that linguistic content also provides informative depressive markers. In contrast, the visual modality was most informative in the classification setting, where affective facial cues may be more useful for discriminating depressive versus non-depressive patterns than for estimating continuous symptom severity.

The improvement observed in the multimodal setting therefore appears to arise from the complementary integration of partially distinct vocal, visual, and linguistic cues, rather than from simple redundancy across modalities. From a fusion perspective, this suggests that the value of multimodal integration in depression screening lies not only in maximizing benchmark accuracy, but also in enabling the system to combine heterogeneous behavioral evidence in a way that is more robust and potentially more clinically relevant than any single modality alone.

The framework is grounded in a privacy-by-design philosophy that emphasizes non-retentive data processing. All user inputs, including any combination of video, audio, and text, are processed exclusively in real time and held transiently in memory only for the duration required to compute assessment scores for that session. Once the session is complete and the final PHQ-8 score is displayed, all associated raw data and intermediate feature representations are immediately and irrevocably purged. No user data is ever written to disk, stored in a database, or used to retrain the models. This stateless architecture ensures maximum privacy and confidentiality, distinguishing LUNA from systems that rely on persistent storage or longitudinal analysis. Such design principles are essential for fostering user trust and comfort in a self-screening context.

To further validate LUNA's effectiveness, we compared its PHQ-8 predictions against established depression detection models that use the DAIC-WOZ dataset. The results in Table 6 highlight LUNA's performance relative to other multimodal frameworks. These results show that LUNA's audio

model achieves a lower RMSE (6.09) than existing models, and its text model performs on par with state-of-the-art methods. While the individual unimodal components of LUNA may not outperform the latest state-of-the-art models in isolated settings, this work does not aim to push the performance boundaries of single-modality depression detection. Rather, the contribution lies in the structured integration of video, audio, and text modalities into a unified, PHQ-8-aligned, real-time framework designed for user-facing deployment. This integration allows LUNA to leverage complementary information across modalities, as shown by the improved performance in the multimodal configuration. Moreover, the framework emphasizes practical usability, modular input selection, and accessibility, key features that are often overlooked in highly specialized unimodal research. Additionally, its video model (F1-score = 84.57%) aligns with recent vision-based depression detection techniques. This benchmarking supports LUNA's reliability for automated depression screening.

Beyond predictive performance, the findings underscore the importance of integrating multiple data modalities with a user-friendly interface for advancing early detection and support in depression screening. Prior work has shown that accessibility and engagement are critical factors for adoption of digital mental health tools, particularly in non-clinical and at-home settings [66, 67]. In this context, LUNA demonstrates how multimodal systems can improve screening accuracy while remaining accessible and engaging for users.

Table 6: Evaluation of LUNA's Performance in Relation to Prior Depression Detection Frameworks.

Model	Dataset	Best RMSE	F1-Score
LUNA	DAIC-WOZ	6.09 (Audio), 6.23 (Text)	84.57% (Video)
MFM-Att (Fang et al., 2023)	DAIC-WOZ	6.23	N/A
Park & Moon (2022)	DAIC-WOZ	6.25	N/A
Xie et al. (2022)	Private Dataset	6.34	N/A

These findings reinforce the potential of combining video, audio, and text modalities for a more comprehensive assessment of depressive symptoms. The benchmarking results also suggest that LUNA performs competitively with established depression detection models, further supporting its effectiveness as a real-world screening tool. While a direct clinical validation study has not yet been conducted, LUNA's performance on DAIC-WOZ is comparable to that of established depression detection models reported in

the literature. These results suggest that LUNA constitutes a promising and reliable automated screening framework, motivating future clinical validation in controlled settings.

While LUNA shows promising potential, several aspects warrant further investigation. In particular, processing multiple data streams may introduce computational overhead in resource-limited environments, and broader real-world testing is required to assess robustness across diverse usage conditions. These considerations are common challenges in multimodal mental health systems and have been highlighted in prior work [15, 16].

7. Limitations and Future Work

While the LUNA framework demonstrates promising results and supports the feasibility of multimodal depression assessment, several limitations remain that are common challenges in this research domain and suggest directions for future work. A primary limitation is the absence of a single, publicly available dataset that simultaneously includes high-quality video, audio, and text modalities for depression analysis. As a result, each unimodal module was trained on separate, specialized datasets (FER-2013 for video, DAIC-WOZ for audio and text), which prevents the model from learning cross-modal correlations within the same subjects. Access to, or the development of, integrated multimodal datasets would substantially enhance the effectiveness of fusion strategies.

Another limitation lies in the current fusion approach. The present framework uses an equal-weighted aggregation of unimodal scores for simplicity and interpretability. However, model performance varied across modalities, reflecting that individuals may convey depressive symptoms differently; some primarily through facial affect, others through speech or text. This highlights the potential benefit of more sophisticated fusion mechanisms, such as attention-based weighting, which could dynamically adjust modality importance according to input quality or contextual cues. Additionally, ablation studies on different fusion weighting strategies were not feasible under the current dataset separation, but such experiments represent an important direction for future research once integrated multimodal datasets become available.

In addition, modality-specific challenges such as poor video quality, background noise in audio, or short and ambiguous text responses can reduce

prediction reliability. Addressing these issues through improved preprocessing, robust feature extraction, and noise-aware modeling represents another direction for future work. A further limitation concerns the evaluation strategy: while standard train–test splits from benchmark datasets were used to ensure comparability with prior work, this limits the assessment of generalizability. Future studies should incorporate cross-validation protocols and evaluation on additional multimodal datasets to provide stronger evidence of robustness in real-world settings.

Finally, the framework received initial expert review from a licensed Cognitive Behavioural Psychotherapist and Clinical Psychologist, as well as a specialist Psychotherapist. Their feedback provided qualitative validation of the framework's psychological appropriateness, ethical positioning, suitability, usability, and potential comfort for patient-facing use, and confirmed its clinical relevance. At the same time, the feedback reinforced the need for formal quantitative and qualitative clinical validation as an essential next step to establish real-world applicability. Within this clinical context, extensions that support secure, consent-driven longitudinal monitoring may further enhance assessment fidelity. Overall, these limitations should not be regarded as barriers but as opportunities for extending this line of research. Future studies that integrate richer multimodal datasets, adopt adaptive fusion strategies, and enhance robustness to input variability are likely to yield further improvements in the reliability and clinical applicability of multimodal depression assessment systems.

8. Conclusion

This work introduced LUNA, a unified multimodal framework that integrates video, audio, and text inputs with an empathetic interactive avatar to screen for depressive symptoms in users, assigning scores based on the PHQ-8 scale. While each of the individual models used (ResNet-50, Wav2Vec2, BERT) has been explored in prior research, the key contribution of this work lies in their structured integration, PHQ-8 score alignment, and real-time deployment within a user-friendly avatar-based interface. By supporting flexible modality input and real-time depression scoring, LUNA offers a practical and accessible approach to automated mental health assessment.

In conclusion, LUNA represents a significant step forward in depression detection, demonstrating the powerful potential of multimodal frameworks to improve screening accuracy, user engagement and early intervention. This

work provides a strong foundation for future developments in mental health technology, aiming to make care more effective, empathetic and widely accessible.

References

[1] Institute of Health Metrics and Evaluation. *Global Health Data Exchange (GHDx)*. Web page: https://vizhub.healthdata.org/gbd-results/. Accessed September 2023.

[2] Jodi Allen and Corrine M Djuric. Major depressive disorder. *The 3P's for Advanced Healthcare Providers-E-Book: The 3P's for Advanced Healthcare Providers-E-Book*, page 54, 2024.

[3] B Adroa Afiya. Interconnection between depressive disorders and persistent diseases. *Res Invention J Res Med Sci*, 3(1):45–51, 2024.

[4] SAGS Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, WT Chiu, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological medicine*, 48(9):1560–1571, 2018.

[5] Alan B Shafer. Meta-analysis of the factor structures of four depression questionnaires: Beck, ces-d, hamilton, and zung. *Journal of clinical psychology*, 62(1):123–146, 2006.

[6] Sven Alfonsson, Pernilla Maathz, Timo Hursti, et al. Interformat reliability of digital psychiatric self-report questionnaires: a systematic review. *Journal of medical Internet research*, 16(12):e3395, 2014.

[7] Asmaa Halbouni, Teddy Surya Gunawan, Mohamed Hadi Habaebi, Murad Halbouni, Mira Kartiwi, and Robiah Ahmad. Machine learning and deep learning approaches for cybersecurity: A review. *IEEE Access*, 10:19572–19585, 2022.

[8] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[9] Yu Liu, Hantian Zhang, Luyuan Zeng, Wentao Wu, and Ce Zhang. Mlbench: benchmarking machine learning services against human experts. *Proceedings of the VLDB Endowment*, 11(10):1220–1232, 2018.

[10] Payam Kaywan, Khandakar Ahmed, Ayman Ibaida, Yuan Miao, and Bruce Gu. Early detection of depression using a conversational ai bot: A non-clinical trial. *Plos one*, 18(2):e0279743, 2023.

[11] Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538, 2023.

[12] Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, and Peter Eklund. Audio based depression detection using convolutional autoencoder. *Expert Systems with Applications*, 189:116076, 2022.

[13] Jihoon Oh, Kyongsik Yun, Uri Maoz, Tae-Suk Kim, and Jeong-Ho Chae. Identifying depression in the national health and nutrition examination survey data using a deep learning algorithm. *Journal of affective disorders*, 257:623–631, 2019.

[14] Junhee Park and Nammee Moon. Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability*, 14(6):3569, 2022.

[15] Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. Researching mental health disorders in the era of social media: systematic review. *Journal of medical Internet research*, 19(6):e228, 2017.

[16] Gustave Udahemuka, Karim Djouani, and Anish M Kurien. Multimodal emotion recognition using visual, vocal and physiological signals: a review. *Applied Sciences*, 14(17):8071, 2024.

[17] Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*, 2022.

[18] Samar Samir Khalil, Noha S Tawfik, and Marco Spruit. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns*, 5(7), 2024.

[19] Rachel Kornfield, Jonah Meyerhoff, Hannah Studd, Ananya Bhattacharjee, Joseph Jay Williams, Madhu Reddy, and David C Mohr. Meeting users where

they are: user-centered design of an automated text messaging tool to support the mental health of young adults. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[20] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. *Advances in Neural Information Processing Systems*, 36:71242–71262, 2023.

[21] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[25] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

[26] Cheolmin Shin, Seung-Hoon Lee, Kyu-Man Han, Ho-Kyoung Yoon, and Changsu Han. Comparison of the usefulness of the phq-8 and phq-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. *Psychiatry investigation*, 16(4):300, 2019.

[27] Shuang Gao, Vince D Calhoun, and Jing Sui. Machine learning in major depression: From classification to treatment outcome prediction. *CNS neuroscience & therapeutics*, 24(11):1037–1052, 2018.

[28] Aleks Stolicyn, J Douglas Steele, and Peggy Seriès. Prediction of depression symptoms in individual subjects with face and eye movement tracking. *Psychological medicine*, 52(9):1784–1792, 2022.

[29] Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid. Towards robust deep neural networks for affect and depression recognition from speech. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 5–19. Springer, 2021.

[30] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234, 2022.

[31] Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561, 2023.

[32] Wanqing Xie, Chen Wang, Zhixiong Lin, Xudong Luo, Wenqian Chen, Manzhu Xu, Lizhong Liang, Xiaofeng Liu, Yanzhong Wang, Hui Luo, et al. Multimodal fusion diagnosis of depression and anxiety based on cnn-lstm model. *Computerized Medical Imaging and Graphics*, 102:102128, 2022.

[33] Gerard Anmella, Miriam Sanabra, Mireia Primé-Tous, Xavier Segú, Myriam Cavero, Ivette Morilla, Iria Grande, Victoria Ruiz, Ariadna Mas, Inés Martín-Villalba, et al. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: Development, feasibility, and potential effectiveness studies. *Journal of medical Internet research*, 25:e43293, 2023.

[34] Prabod Rathnayaka, Nishan Mills, Donna Burnett, Daswin De Silva, Damminda Alahakoon, and Richard Gray. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. *Sensors*, 22(10):3653, 2022.

[35] Zifan Jiang, Salman Seyedi, Emily Lynn Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O Cotes, and Gari D Clifford. Multimodal mental health assessment with remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *medRxiv*, pages 2023–09, 2023.

[36] Yuhao He, Li Yang, Xiaokun Zhu, Bin Wu, Shuo Zhang, Chunlian Qian, and Tian Tian. Mental health chatbot for young adults with depressive symptoms

during the covid-19 pandemic: single-blind, three-arm randomized controlled trial. *Journal of Medical Internet Research*, 24(11):e40719, 2022.

[37] Sooah Jang, Jae-Jin Kim, Soo-Jeong Kim, Jieun Hong, Suji Kim, and Eunjoo Kim. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *International journal of medical informatics*, 150:104440, 2021.

[38] Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 11–20, 2013.

[39] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[40] Summaira Jabeen, Xi Li, Muhammad Shoib Amin, Omar Bourahla, Songyuan Li, and Abdul Jabbar. A review on methods and applications in multimodal deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–41, 2023.

[41] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[42] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.

[43] Goutam Kumar Sahoo, Jayakrishna Ponduru, Santos Kumar Das, and Poonam Singh. Deep leaning-based facial expression recognition in fer2013 database: An in-vehicle application. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6. IEEE, 2022.

[44] Tanoy Debnath, Md Mahfuz Reza, Anichur Rahman, Amin Beheshti, Shahab S Band, and Hamid Alinejad-Rokny. Four-layer convnet to facial emotion

recognition with minimal epochs and the significance of data diversity. *Scientific Reports*, 12(1):6991, 2022.

[45] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3486–3496, 2019.

[46] Rupali Gill, Jaiteg Singh, Susheela Hooda, and Durgesh Srivastava. Delineating emotional differences between depressed and non-depressed individuals using a novel multimodal framework. *Multimedia Tools and Applications*, pages 1–22, 2024.

[47] Jun-Teng Yang, Guei-Ming Liu, and Scott C-H Huang. Emotion transformation feature: Novel feature for deception detection in videos. In *2020 IEEE international conference on image processing (ICIP)*, pages 1726–1730. IEEE, 2020.

[48] Antonia Vehlen, Antonia Kellner, Claus Normann, Markus Heinrichs, and Gregor Domes. Reduced eye gaze during facial emotion recognition in chronic depression: Effects of intranasal oxytocin. *Journal of Psychiatric Research*, 159:50–56, 2023.

[49] Gajendra Kumar, Tanaya Das, and Kuldeep Singh. Early detection of depression through facial expression recognition and electroencephalogram-based artificial intelligence-assisted graphical user interface. *Neural Computing and Applications*, 36(12):6937–6954, 2024.

[50] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik, 2014.

[51] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874, 2024.

[52] John Mongan, Linda Moy, and Charles E Kahn Jr. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers, 2020.

[53] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

[54] James R Williamson, Diana Young, Andrew A Nierenberg, James Niemi, Brian S Helfer, and Thomas F Quatieri. Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech & Language*, 55:40–56, 2019.

[55] Yizhuo Dong and Xinyu Yang. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing*, 441:279–290, 2021.

[56] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[57] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[58] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24, 2015.

[59] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

[60] Yazhou Zhang, Yu He, Lu Rong, and Yijie Ding. A hybrid model for depression detection with transformer and bi-directional long short-term memory. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2727–2734. IEEE, 2022.

[61] Karla Maria Valencia-Segura. Detection of signs of depression based on a multimiodal approach. 2024.

[62] Wei Zhang, Kaining Mao, and Jie Chen. A multimodal approach for detection and assessment of depression using text, audio and video. *Phenomics*, 4(3):234–249, 2024.

[63] Qingxin Ye, Zhenming Xie, Hao Sun, Luyao Xin, Youwen Chen, Jian Song, and Yen-Wei Chen. Enhancing deep learning-based depression level estimation based on multi-task learning. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, pages 59–65, 2024.

[64] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 53–59, 2017.

[65] Jue Gong, Gregory E Simon, and Shan Liu. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PloS one*, 14(9):e0222665, 2019.

[66] Jillian Lane Warren. Digital and interactive technologies for children's mental health and socio-emotional wellbeing: Exploring potential, gaps, and design opportunities. 2023.

[67] Qiaolei Jiang, Yadi Zhang, and Wenjing Pian. Chatbot as an emergency exist: Mediated empathy for resilience via human-ai interaction during the covid-19 pandemic. *Information processing & management*, 59(6):103074, 2022.