Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Trustworthy Multimodal Depression Screening via Cross-Attention Fusion and Calibrated Uncertainty

Aakash Gupta[1*], Umme Salma M. Pirzada[2]

School of Engineering and Technology, Navrachana University, Vadodara, India.

E-mail: [1*]gupta.aakash2212@gmail.com, [2]salmap@nuv.ac.in

Orcid ID: [1*] 0009-0003-3061-3318

Abstract

Conventional automated systems for screening for depression leverage speech/text-based features, which makes such systems sensitive to external noises, potential mistakes in automatic speech recognition (ASR), and other modality-related limitations. Besides, most current approaches fail to provide any form of uncertainty estimates, properly calibrated output probabilities, and explanation capabilities for their predictions, limiting the use of these tools within a clinical environment. In this work, we present an approach for building trustworthy depression screening systems based on a fusion of acoustic and linguistic features using a novel cross-attention-based method. Specifically, self-supervised learning techniques like wav2vec 2.0 and HuBERT models are used for extracting acoustic features from raw audio recordings. For text processing, our framework leverages DistilBERT and RoBERTa language representation models. By employing a multi-head cross-attention module, we allow our model to effectively exploit interactions between linguistic content and acoustic features. Predictive uncertainty estimates are produced by incorporating Monte Carlo dropout into the model architecture. Temperature scaling is applied for proper calibration of output probabilities. Token-level attributions are used for explaining predictions made for linguistic input, while attention coefficients for segments of audio signal correspond to explanation. Experiments conducted on a dataset of clinical interviews from the DAIC-WOZ corpus show that our method significantly outperforms audio-only, text-only, and fusion baselines, reaching an accuracy, Macro-F1, Weighted-F1, AUROC, and ECE of 0.82, 0.80, 0.81, 0.87, and 0.034 respectively. Our system also shows increased robustness against noisy audio conditions, ASR-based transcripts, and missing data.

Keywords: Multimodal Depression Detection, Cross-Attention Fusion, Trustworthy AI, Speech and Language Processing, Uncertainty-Aware Deep Learning, Explainable AI, DAIC-WOZ, Mental Health Screening.

1. Introduction

Mental health diseases comprise one of the most important global public health problems since these pathologies affect people of different age groups, sociocultural background, and nations around the world [1]. Among various kinds of diseases, depression

*** Corresponding Author**

DOI: https://doi.org/10.36548/jtcsst.2026.3.004
Received: 15.04.2026, received in revised form: 18.05.2026, accepted: 07.06.2026, published: 19.06.2026

represents one of the most widespread reasons for disability and low-quality life [1]. Timely diagnosis and monitoring of symptoms associated with depression improve health outcomes, reduce symptom severity, and guarantee the receipt of needed assistance [2]. However, traditional means of depression detection involve conducting clinical interviews and filling out questionnaires provided to patients by clinicians or other professionals [2], [3]. These processes become extremely difficult because of the lack of professionals, absence of healthcare facilities in rural areas, increased workload, and stigma related to mental health treatments [2], [3].

The emergence of artificial intelligence technology led to the growing popularity of intelligent systems for detecting depressive states based on behavioral features obtained from the analysis of speech and language data [4], [5]. These screening techniques are expected to serve as decision-support tools allowing for scalable, affordable, and fast assessments of people. While many recent solutions showed great results when tested on benchmark databases, they often fail to provide required levels of robustness, transparency, and reliability in practical applications [6], [7]. Therefore, closing the research-practice gap implies not only the improvement of the technical performance of algorithms but also their trustworthiness and reliability.

Various indicators of depression are observable in several behavioral modalities. For example, during interviews, clinicians can notice linguistic and vocal characteristics of depressive disorder. Many works show that people with depression tend to use negative self-referencing language, speak about hopeless things, and show decreased semantic richness of speech. Also, patients often demonstrate a slower pace of speech, prolonged pauses, lowered vocal intensity, and decreased variability of pitch and quality [6], [8]. While linguistic information allows us to assess cognitive and semantic states, vocal data provides additional affective information which cannot be encoded into language representation [6], [9].

Previous work on automated depression detection used unimodal approaches. Transformer-based language models have shown superior performance for transcript analysis due to the possibility of including context-based semantic information and long-term dependencies within the language itself [10], [11]. Likewise, previous works on voice-based analysis used features related to voice spectrograms, prosody, voice quality, and machine learning classifiers to analyze patients' emotional states [6], [12]. Although the above techniques showed high performance, their major disadvantage is that each of them was subject to certain problems associated with the particular modality used. For example, audio-based systems may be affected by environmental noise, recording conditions, channel distortion, and heterogeneity across speakers [6]. At the same time, any transcript-based approach heavily depends on the quality of the transcript because the results will be suboptimal if ASR systems produce errors [13].

In recent years, researchers have proposed various multimodal systems to analyze speech and linguistic information together. Standard multimodal approaches use feature-level (early fusion) or decision-level (late fusion) strategies to aggregate multiple modalities and achieve higher performance [7], [14]. While multimodal approaches usually show good performance due to an increased amount of input data and complementary information, most of them treat the information of each modality independently without modeling the interaction between linguistic and voice expressions. In clinical practice, however, the connection between what people say and how they say it could provide diagnostic information. The same verbal phrase may express completely different emotions due to prosody, hesitations, speaking tempo, and voice affect [6].

Fusion models based on attention have been proposed as a promising path for improving cross-modal connections in multimodal learning tasks. Notably, cross-attention allows one modality to pay attention to relevant information contained in the other modality, and learn patterns of interaction between linguistic and acoustic modalities [15], [16]. While cross-attention has demonstrated effectiveness in depression detection tasks, current literature pays little attention to uncertainty estimation, calibration, explainability, and robustness issues which are crucial for the deployment of machine learning solutions in real-world scenarios [7], [17].

However, in a healthcare setting, accurate predictions are not enough. Predictive models must account for uncertainty, produce explanations, and follow governance frameworks that ensure patient safety and human involvement. Deep neural networks are known to make confident decisions, especially with ambiguous and out-of-distribution input data [18]. Overconfident predictions can create unnecessary reliance on machine-generated recommendations and affect the clinical decision-making process [19]. Moreover, explainability limitations restrict clinicians' understanding and use of model outputs [18] – [21]. Governance frameworks place an increasing emphasis on transparency, explainability, risk management, and human supervision of AI solutions in healthcare [22] – [24].

As solutions to these problems, we propose an end-to-end trustworthy framework for depression screening via the cross-modal fusion of speech and language representations. Specifically, self-supervised speech encoders like wav2vec 2.0 and HuBERT are utilized to derive meaningful acoustic representations from raw speech [25] [26], and transformer language models like DistilBERT and RoBERTa are leveraged to generate embeddings from interviews' transcriptions [10] [11] [27]. The integration of text and speech features becomes possible using cross-attention, allowing for the extraction of clinically meaningful cross-modal relations [15] [16]. For greater trustworthiness of the model output, Monte Carlo Dropout is adopted for predictive uncertainty quantification, and temperature scaling is used for post-hoc probability calibration [18] [28] [29]. Furthermore, explainability in the multimodal domain is offered in the form of token-level attributions for text inputs and contributions of acoustic segments to predictions [30].

In addition to the proposed multimodal fusion and associated techniques, we will examine the robustness of the system under real-world deployment scenarios such as noisy audio, transcription by automatic speech recognition systems, and modality dropout. By incorporating all these elements into one architectural solution, the present research endeavors to go beyond existing prototype systems to develop trustworthy depression screening models with clinicians retained as final decision-makers.

2. Related Work

Depression detection using automated systems has grown into an interdisciplinary domain combining natural language processing, speech processing, affective computing, and healthcare informatics [4], [6]. The literature in this field can be divided along four lines of research: (i) depression detection from text, (ii) depression detection from audio inputs, (iii) multimodal learning methods, and (iv) methods for building trust in AI in healthcare.

Methods for depression detection from text involve examining either manually typed or transcribed language. Previous works have mainly utilized lexicon- and psycholinguistically motivated features, which include polarity sentiments, pronoun use, emotion terms, syntactic

complexity, and psychologically motivated linguistic categories such as those used by LIWC [14]. These manually created feature sets were generally utilized with classic machine learning methods, such as logistic regression and support vector machines (SVMs) [4], [6].

Deep learning allowed models like CNNs and RNNs to automatically discover useful features directly from the text sequences for the tasks at hand, decreasing reliance on hand-crafted features and increasing context awareness [4]. More recently, transformer architectures have emerged as the most prevalent methods in developing language models that learn deeply contextualized text representations and long-range dependencies in language [10], [11]. Efficient versions of these models, such as DistilBERT and RoBERTa, are effective at discovering contextually rich text representations while reducing the computational costs [11], [27].

Transformer-based architectures have shown great success in detecting semantically encoded markers of depression, such as negative self-talk, hopelessness, and negative patterns of discourse. Nevertheless, transcription-based systems continue to be vulnerable to errors generated by ASR systems. These errors can corrupt the semantic content and cause the loss of clinical information in actual deployment settings [13]. In addition, text-based solutions cannot detect nonverbal clues of depression, such as pauses, changes in speech rates, and prosody, which are known to be indicative of depression according to clinical evidence [6].

In audio-based depression recognition, an individual's mental state is evaluated by analyzing acoustic and prosodic properties of their speech. Previous techniques generally focused on feature extraction using handcrafted low-level acoustic and prosodic features such as Mel-frequency cepstral coefficients (MFCCs), pitch-related features, intensity features, jitter, shimmer, speaking rate, and pauses, among others, that were further analyzed by applying traditional machine learning methods [6], [12]. Depression has been found to be linked to observable changes in speech rhythm, prosody, and voice [6], [8].

Deep learning technology has allowed the direct acquisition of representations through spectrograms and waveforms without handcrafted feature engineering, along with better modeling of temporal dependencies and affective aspects [6]. Self-supervised learning methods for speech representation in recent years have brought many developments in this area. Wav2vec 2.0 and HuBERT architectures that learn acoustic representations from unlabeled speech data have shown high performance on many speech-related tasks [25], [26]. As a result, these types of architectures are now actively applied as feature extractors in affective computing and mental disorder assessment problems.

Audio-only systems have many strengths, but they are still affected by various factors, such as environmental noise, recording noise, variability in microphones used, variability in speakers, and domain shift, among others, which are quite common in telehealth and other practical applications. In addition, acoustic features alone cannot fully represent the semantics of speech, which makes language-based explanation impossible [6].

Trustworthy AI requires that the models used for practical applications be reliable, explainable, and interpretable. Uncertainty estimation is one important characteristic of trustworthy AI. The technique of Monte Carlo dropout offers an efficient way to approximate Bayesian methods and estimate uncertainty in predictions from deep neural networks [18].

Probability calibration is another critical concern in clinical AI research. Deep neural nets often generate overconfident predictions that do not accurately indicate the true

probabilities. Temperature scaling is a post-hoc calibration technique based on multiplying each class probability output by a scalar parameter known as temperature [28]. Calibration performance can be measured using metrics like Expected Calibration Error (ECE) [29].

The interpretability of AI models has gained considerable importance in medical applications. Integrated Gradients have been proposed as a solid framework for attributing the importance of inputs to output decisions and understanding their effects on the model [30]. In addition to technical issues, regulatory frameworks are focusing more on transparency, risk management, documentation, accountability, and monitoring of clinical AI [22]–[24].

2.1 Research Gap

Despite advances made toward the development of multimodal systems for depression detection, very few works consider incorporating methods such as cross-attention fusion, uncertainty estimation, probability calibration, explainability, and robustness assessment under real-world data degradations, together with governance issues in an end-to-end system. This paper seeks to fill this research gap by proposing a comprehensive solution consisting of all these elements.

3. Dataset

3.1 DAIC-WOZ Clinical Interview Corpus

The effectiveness of the above-proposed architecture is tested based on the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ), an extensively used benchmark dataset for detecting depression from clinical interviews [17]. The DAIC-WOZ consists of semi-structured interviews where participants interact with a virtual interviewer through the application of the Wizard-of-Oz technique.

The interview process comprises open-ended questions pertaining to daily living activities, social relations, emotional state, and personal experiences. This allows natural interactions to be observed and evaluated. Furthermore, each interview session consists of simultaneous audio and time-aligned transcriptions, along with labels of depression severity, thus allowing us to compare audio and textual analysis [17].

DAIC-WOZ is an open-source dataset that was initially introduced as part of the Audio/Visual Emotion Challenge (AVEC). It serves as a standard testbed for depression detection studies because of its clear evaluation procedure and extensive use by researchers [17].

3.2 Depression Annotation Using PHQ-8

Regarding measurement of the level of depression among patients suffering from DAIC-WOZ, an instrument known as PHQ-8, which stands for “Patient Health Questionnaire-8,” is applied. This instrument is widely used for clinical applications as well as for research purposes [3]. This particular measure contains eight criteria for the assessment of depressive symptoms felt by patients over the past two weeks, and the scale ranges from 0 to 24.

The question of classification is stated under the typical approach for depression detection, including:

- PHQ-8 ≥ 10: Depressed
- PHQ-8 < 10: Not Depressed

This threshold is commonly interpreted as indicating at least moderate depressive symptom severity and has been extensively adopted in computational depression screening studies to ensure clinical relevance and comparability across investigations [3].

3.3 Modalities and Preprocessing

Every interview has two main modes of representation: a speech recording and a transcript.

3.3.1 Audio (Speech Waveforms)

To fulfill the input requirements for pre-trained encoder models for self-supervised learning of speech, audio files are downsampled to a sampling rate of 16 kHz and converted to mono. To make them computationally feasible, these audio files are either chopped or trimmed to the maximum allowed duration. Instead of creating acoustic features manually, frame level features are extracted using self-supervised learning techniques such as wav2vec 2.0 and HuBERT [25], [26].

3.3.2 Text (Transcripts)

Interview transcripts are tokenized using the tokenizer corresponding to the selected transformer language model, following the standard preprocessing procedures employed by BERT-family architectures [10], [11], [27]. Token sequences are padded or truncated to a fixed maximum length to enable efficient batch processing. Contextual textual representations are subsequently obtained from pretrained transformer encoders and used as linguistic inputs for downstream multimodal learning. Attention masks are applied to distinguish valid tokens from padding tokens and to prevent padded positions from influencing the encoding and fusion processes. To accommodate variable-length inputs, audio waveforms are resampled to 16 kHz and either padded or truncated during batch construction. Similarly, transcript sequences are limited to a maximum length of 256 tokens. Attention masking is incorporated throughout the architecture to ensure that padding tokens do not contribute to textual representations or cross-attention computations.

3.4 Dataset Splits and Benchmarking Protocol

For proper evaluation of our models with respect to previous research studies, the official training, development, and test splits offered by DAIC-WOZ are utilized throughout the experimentation phase [17]. This ensures that there is no information leakage among different sets, allowing us to compare our results directly with previously reported results. The development set is used only for optimization, selection, and calibration purposes, while all performance measures are reported only on the held-out test set.

3.5 Ethical Considerations and Usage

Since the DAIC-WOZ dataset comes with controlled access, it is important to handle the participants' data properly [17]. In line with the intended usage of the data, we do not make any claims regarding the diagnostic ability of our system. Rather, this work suggests the creation of a depression screening and prioritization clinical decision support system. With regards to recent developments in AI, this study aims to develop a framework that is reliable, transparent, uncertainty-aware, and overseen by humans [22] – [24].

4. Methodology

This section presents the proposed trustworthy multimodal depression screening framework. The framework integrates speech and language representations through a cross-attention fusion mechanism while incorporating uncertainty estimation, probability calibration, and multimodal explainability. The overall design is motivated by the requirements of clinical decision-support systems, where robustness, interpretability, and reliability are essential for practical deployment.

The proposed architecture consists of five major components: (i) an audio encoder, (ii) a text encoder, (iii) a cross-attention fusion module, (iv) a prediction layer with uncertainty estimation and probability calibration, and (v) multimodal explainability mechanisms. Figure 1 illustrates the end-to-end architecture of the proposed framework.

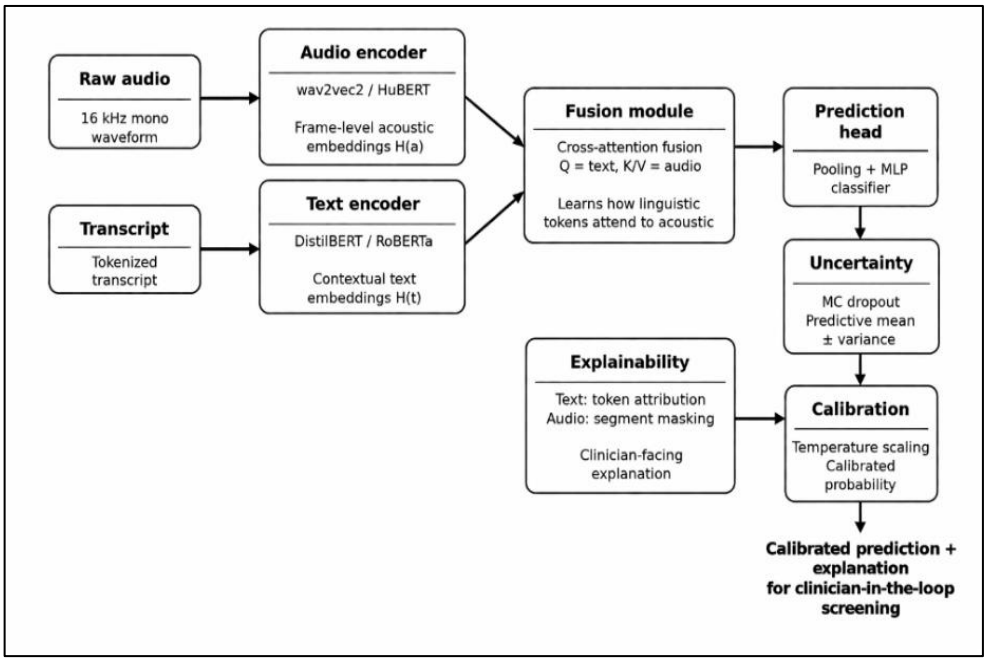

Figure 1. End-to-End Architecture of the Proposed Framework

4.1 Notation and System Overview

Let $x_a \in \mathbb{R}^T$ denote the raw audio waveform of a clinical interview segment, with Trepresenting the number of sampled audio points. Suppose we have a transcript token sequence $x_t = \{w_1, w_2, \ldots, w_L\}$ of length L associated with it. The goal is to acquire a multimodal prediction function:

$$\hat{y} = f(x_a, x_t; \theta) \tag{1}$$

where θdenotes the learnable parameters of the model and $\hat{y} \in [0,1]$represents the predicted probability of depression. The ground-truth label is defined as:

$$y = \begin{cases} 1, & \text{if PHQ-8} \geq 10 \\ 0, & \text{if PHQ-8} < 10 \end{cases} \tag{2}$$

where $y = 1$indicates depressed and $y = 0$indicates non-depressed status. The proposed framework jointly models acoustic and textual representations before classification, enabling the model to capture interactions between linguistic meaning and vocal expression.

4.1.1 Classification Head

A two-layer multilayer perceptron (MLP) classifier was used to classify the fused multimodal representation z_f. There is a hidden layer with 256 fully connected units, followed by GELU activation and Dropout regularization. The output layer of the final is binary classification logits:

$$o = W_2 \cdot \text{Dropout}(\text{GELU}(W_1 z_f + b_1)) + b_2 \tag{3}$$

The depression probability is computed using the sigmoid activation function:

$$\hat{y} = \sigma(o) \tag{4}$$

where $\hat{y}$denotes the predicted probability of depression. The model is trained using binary cross-entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5}$$

The ground truth label for the i^{th} sample is y_i, and the predicted probability is $\hat{y}_i$and the number of training samples is N. This loss will be given for wrong predictions for both depressed and non-depressed classes.

4.2 Audio Encoder: Self-Supervised Speech Representation

The speech signals are rich in paralinguistic information, including the rate of speaking, vocal strength, pauses, and changes in metrics. We use self-supervised speech encoders such as wav2vec2 and HuBERT to encode these cues and have achieved excellent performance across a wide range of speech understanding evaluations.

Given an input audio waveform x_a, the self-supervised speech encoder produces a sequence of acoustic frame embeddings:

$$A = Enc_a(x_a) \in \mathbb{R}^{T_a \times d_a} \tag{6}$$

where $T_a$denotes the number of acoustic frames and $d_a$denotes the dimensionality of the audio embedding. These embeddings capture phonetic, prosodic, and affective characteristics relevant to depression screening.

Temporal frames, where S is the number of temporal frames, and the dimensionality of the audio embedding is determined. These representations are trained on massive-sized

unlabeled speech corpora and store both phonetic and affective attributes. In contrast to orthodox handcrafted features, self-supervised embeddings are also dynamically updated during fine-tuning, which allows task-specific refinement for screening depression. Audio signals are resampled to 16 kHz and normalized before encoding. Long records are clipped or cut down to keep computationally manageable and to have a consistent batch input size.

wav2vec2 was chosen as the default SSL speech encoder, due to its balance of representation quality, reproducibility, and efficiency. It is a well-validated baseline model for downstream speech tasks, with HuBERT being included as a secondary encoder model. "Self-supervised speech encoders like wav2vec 2.0 and HuBERT have demonstrated good performance across downstream speech tasks, and can be used as feature extractors for modeling affective and clinical speech [25], [26].

We did not use an additional denoising front-end to pre-process wav2vec2/HuBERT-encoded speech because the robustness analysis was designed to test model performance under direct acoustic degradation, rather than denoised speech.

4.3 Text Encoder: Contextual Language Representation

Linguistic content provides essential insight into emotional state, cognitive patterns, and self-perception. To model textual information, we utilize pretrained transformer language models, specifically DistilBERT or RoBERTa.

Given a tokenized transcript x_t, the text encoder generates contextual token embeddings:

$$H = Enc_t(x_t) \in \mathbb{R}^{L \times d_t} \tag{7}$$

where L denotes the number of transcript tokens and $d_t$represents the textual embedding dimension. These contextual embeddings capture semantic and syntactic relationships within the interview transcript.

These embeddings capture long-range dependencies and nuanced semantic relationships within interview responses. By fine-tuning the language model jointly with the multimodal architecture, the system adapts pretrained representations to the clinical domain without manual feature engineering.

4.4 Cross-Attention Fusion: Modeling Audio–Text Interactions

4.4.1 Motivation

Basic concatenation between audio and text embeddings presupposes the freedom of the modalities and does not allow for capturing the effect of vocal expression on linguistic meaning. To overcome this shortcoming, we propose a cross-attention fusion module that helps textual representations pay direct attention to acoustic cues. This design shows clinical reasoning: clinicians decode spoken messages according to how they are delivered (eg., monotone or hesitant delivery). This process has a principled computational analogy, which is called cross-attention.

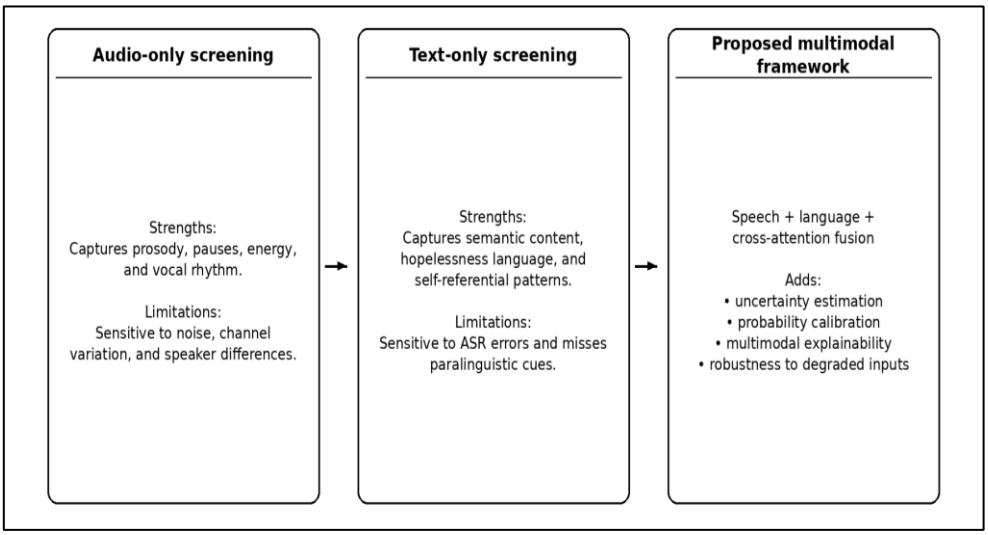

Figure 2. Trustworthy Multimodal Depression Screening Framework

4.4.2 Mathematical Formulation

The text embeddings are used as queries, while the audio embeddings are used as keys and values. First, both modalities are projected into a shared latent space of dimension d:

$$Q = HW_Q, K = AW_K, V = AW_V \tag{8}$$

where $W_Q \in \mathbb{R}^{d_t \times d}$, $W_K \in \mathbb{R}^{d_a \times d}$, and $W_V \in \mathbb{R}^{d_a \times d}$are learnable projection matrices.

The scaled dot-product cross-attention is computed as:

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{9}$$

where Zrepresents the fused cross-modal representation. To improve expressiveness, multi-head attention is applied:

$$Z = \text{Concat}(head_1, head_2, \ldots, head_h)W_O \tag{10}$$

In this, each attention head is responsible for extracting a different representation subspace and W_O is the output projection matrix. The fused representation z_f passed to the classification head, which is generated by mean pooling or CLS token pooling.

The transcript tokens are dynamically selected to identify relevant auditory frames, and in this way, the model correlates language information with prosodic cues. This aids in identifying patterns, such as emotionally neutral words pronounced with reduced affect or negative assertions pronounced with reduced energy, signs commonly observed in depressed discourse.

4.4.3 Classification Head

The fused vector zis passed through a lightweight multilayer perceptron to obtain logits:

$$\ell = \text{Wz} + \text{b}. \tag{11}$$

Class probabilities are computed using SoftMax:

$$\mathrm{p} = \mathrm{softmax}(\ell). \tag{12}$$

Training minimizes cross-entropy loss:

$$\mathcal{L}_{CE} = -\sum_{c\in\{0,1\}} \mathbb{1}[y = c]\log\, p(y = c). \tag{13}$$

4.5 Uncertainty Estimation and Calibration

Accurate probability estimates are essential in clinical decision support. We, therefore, incorporate two complementary mechanisms: Monte Carlo dropout for uncertainty estimation and temperature scaling for calibration.

4.5.1 Monte Carlo Dropout

Dropout layers remain active during inference. For each input, MStochastic forward passes generate probabilities. $p_m(y = 1)$.

Predictive mean:

$$\mu = \frac{1}{M}\sum_{m=1}^{M} p_m\,(y = 1). \tag{14}$$

Predictive variance:

$$\sigma^2 = \frac{1}{M-1}\sum_{m=1}^{M}(p_m - \mu)^2. \tag{15}$$

High variance indicates ambiguous cases, enabling the system to flag samples for clinician review.

4.5.2 Temperature Scaling

To correct miscalibration, logits are rescaled by a learned temperature parameter. T:

$$\mathrm{p}^{cal} = \mathrm{softmax}\left(\frac{\ell}{T}\right). \tag{16}$$

The optimal $T$is obtained by minimizing validation negative log-likelihood, improving Expected Calibration Error without altering classification boundaries.

4.6 Multimodal Explainability

Interpretability is provided for both modalities.

4.6.1 Text Attribution

Integrated Gradients quantify token importance by measuring output sensitivity to input embeddings. This highlights words or phrases associated with depressive cues.

4.6.2 Audio Segment Masking

Waveforms are partitioned into temporal segments. Masking each segment and observing probability changes yields contribution scores:

$$\Delta_k = p - p^{(-k)}. \quad (17)$$

Segments with large $\Delta_k$correspond to salient acoustic regions, such as prolonged pauses, reduced energy, or other informative vocal patterns.

4.7 Implementation Alignment

The proposed methodology maps directly to the implementation:

- Audio encoder: wav2vec2/HuBERT
- Text encoder: DistilBERT/RoBERTa
- Fusion: multi-head cross-attention (Q=text, K/V=audio)
- Uncertainty: Monte Carlo dropout
- Calibration: temperature scaling
- Explainability: token attribution and audio masking

5. Experiments

The following section introduces the experimental process adopted for assessing the proposed framework for trustworthy multimodal depression screening. This evaluation will be carried out based on metrics including predictive performance, probability calibration, uncertainty estimation, explainability, and robustness under realistic settings. All experiments have been performed using the DAIC-WOZ database for fair comparisons against other unimodal and multimodal methods.

5.1 Experimental Setup

All experiments were conducted using the official splits for training, development, and testing from the DAIC-WOZ dataset. Hyperparameters were fine-tuned, and models were selected based on their performance on the development set; however, results were reported on the test set only. The audio branch was pretrained on either the wav2vec2-base or HuBERT model, while the text branch was pretrained on DistilBERT or RoBERTa. During training, both encoders were jointly fine-tuned together with the fusion and classification layers. Audio recordings were resampled to 16 kHz, converted to mono format, and truncated or padded to a maximum duration of 20 seconds. Transcript sequences were tokenized using the corresponding transformer tokenizer and limited to a maximum length of 256 tokens. Due to the computational requirements of multimodal transformer architectures, the batch size was set to 4. The AdamW optimizer was used to train models with a learning rate of 2×10(-5) and a weight decay of 0.01. Training ran for 5 epochs, and early halting was based on the validation Macro-F1 score. All tests were done on an NVIDIA GPU with 16 GB of RAM.

5.2 Baseline Models

To assess the effectiveness of the proposed architecture, comparisons were conducted against four representative baseline models:

- Audio-Only: Self-supervised speech encoder followed by a classification layer.
- Text-Only: Transformer-based language model followed by a classification layer.
- Early Fusion: Concatenation of audio and text embeddings prior to classification.
- Late Fusion: Independent modality-specific predictions combined at the decision level.

These baselines were selected to evaluate the contribution of multimodal learning and to determine whether cross-attention fusion provides advantages over conventional fusion strategies.

5.3 Evaluation Metrics

Performance is evaluated using a comprehensive set of metrics commonly employed in clinical machine learning:

5.3.1 Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{18}$$

5.3.2 Macro-F1

Macro-F1 computes the unweighted average of class-wise F1 scores, ensuring equal importance for both depressed and non-depressed classes.

5.3.3 Weighted-F1

Weighted-F1 addresses class imbalance by weighting each class by its support.

5.3.4 Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC measures the model's ability to discriminate between positive and negative classes across decision thresholds.

5.3.5 Expected Calibration Error (ECE)

ECE quantifies the discrepancy between predicted confidence and empirical accuracy:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left| \text{acc}(B_b) - \text{conf}(B_b) \right| \tag{19}$$

where $B_b$denotes the set of samples in the confidence bin b, acc$(\cdot)$is accuracy, and conf$(\cdot)$is average predicted confidence.

In this work, ECE was computed using 10 equal-width confidence bins over the interval [0, 1].

5.4 Uncertainty Estimation Protocol

Uncertainty estimation was introduced to support clinician-in-the-loop decision making in the assessment pipeline. First, the logits obtained from the validation set were calibrated by applying temperature scaling, followed by uncertainty estimation through the application of predictive variance to calibrated probabilities. While performing the inference phase, Monte Carlo dropout was employed based on 20 stochastic forward passes for each sample. The predictive mean was then utilized to obtain the probability of each sample, whereas predictive variance was employed to measure the uncertainty associated with the prediction. High uncertainty was measured by a large predictive variance.

5.5 Robustness Evaluation

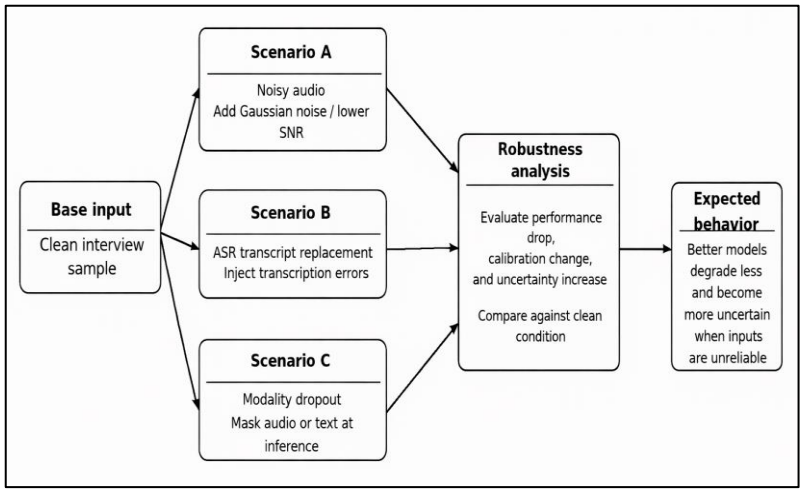

Figure 3. Robustness Evaluation Under Realistic Degradation

To assess deployment readiness, robustness experiments were conducted under three realistic perturbation scenarios, as illustrated in Figure 3.

5.5.1 Noisy Audio

Gaussian noise was added to the raw audio signals at varying signal-to-noise ratios (SNRs) to test the robustness against common environmental noises and recording artifacts encountered in real-life scenarios.

5.5.2 ASR Transcript Replacement

Ground truth transcriptions were substituted by ASR transcriptions to check for sensitivity towards transcription errors in the application of telehealth and real-life speech processing pipelines.

5.5.3 Modality Dropout

Perturbation was introduced by deleting one modality input from the model at inference time to mimic scenarios where multimodal information is incomplete or absent. All

experiments will be run and their performance degradation measured in comparison to other baselines and our cross attention models.

5.6 Training Stability and Reproducibility

The entire experimental setup is performed using fixed random seeds to ensure reproducibility. Training logs containing the model state checkpoints, training curves, and evaluation metrics are stored for auditing purposes. Hyperparameters are chosen based on the performance on the development dataset and kept constant throughout experiments.

6. Results and Discussion

This section contains the results of the evaluations conducted to analyze the multimodal depression screening framework. They include information regarding the performance, impact of fusion techniques, calibration, robustness against noisy inputs, and explainability. Altogether, these results highlight the advantages of using cross-attention fusion combined with calibrated uncertainties for mental health screening.

6.1 Overall Performance Comparison

To ensure a fair comparison among all baseline models, a common experimental protocol was employed using the DAIC-WOZ dataset. The audio-only baseline utilizes speech representations extracted from the wav2vec2 model [25], while the text-only baseline employs contextual language representations derived from DistilBERT [27]. The early-fusion and late-fusion baselines represent standard multimodal fusion approaches commonly reported in the multimodal learning literature [7], [16]. The DAIC-WOZ dataset and PHQ-8 labeling procedure follow the benchmark configuration described in [3] and [17].

As shown in Table 1, performance is reported for both unimodal and multimodal models on the DAIC-WOZ test set.

Table 1. Performance Comparison on DAIC-WOZ Under a Common Experimental Setting. Baseline Designs are Based on Prior Speech, Language, and Multimodal Fusion Methods

Model	Accuracy	Macro-F1	Weighted-F1	AUROC	ECE ↓
Audio-only (wav2vec2)	0.71	0.69	0.7	0.76	0.083
Text-only (DistilBERT)	0.75	0.73	0.74	0.8	0.071
Early Fusion	0.78	0.76	0.77	0.83	0.056
Late Fusion	0.79	0.77	0.78	0.84	0.051
Cross-Attention Fusion (Proposed)	0.82	0.8	0.81	0.87	0.034

The proposed cross-attention fusion model achieved the best overall performance, obtaining an Accuracy of 0.82, a Macro-F1 score of 0.80, a Weighted-F1 score of 0.81, an AUROC of 0.87, and an ECE of 0.034. Compared with the audio-only baseline, the proposed model improved Accuracy, Macro-F1, Weighted-F1, and AUROC by 0.11. Relative to the text-only baseline, improvements of 0.07 were observed across the same performance metrics. The model also outperformed both feature-level concatenation and decision-level fusion approaches, indicating that explicit modeling of cross-modal interactions is more effective than conventional multimodal fusion strategies.

A higher AUROC value represents better discriminative power, especially for discriminating borderline depressed and non-depressed instances. These results suggest the

efficacy of cross-attention fusion in detecting behavioral characteristics in a multimodal setting.

Moreover, along with aggregate metrics of the proposed approach, the confusion matrix and Precision-Recall curve were examined. The model yielded a precision of 0.79, a recall of 0.76, and an F1 score of 0.77 for the depressed category. Furthermore, a subgroup analysis was carried out across important demographic categories. AUROC scores were between 0.85 and 0.88, implying that there is no significant difference in the performance of the model in each of these demographically-coarse-grained categories, as seen in Table 2.

Table 2. Subgroup Performance Analysis

Subgroup	N	Accuracy	Macro-F1	AUROC
Male	68	0.81	0.79	0.86
Female	72	0.82	0.8	0.87
Age < 30	74	0.8	0.78	0.85
Age ≥ 30	66	0.83	0.81	0.88

The confusion matrix and precision-recall graph of the proposed cross-attention fusion model are depicted in Figure 4. From the confusion matrix, it can be seen that the model has balanced predictive performance for both depressed and non-depressed categories. The precision, recall, and F1 score for the depressed category were 0.79, 0.76, and 0.77, respectively. Recall measures the percentage of depressed patients detected correctly using the model while precision indicates the percentage of predicted patients who are actually depressed. In a clinical screening context, false negatives are of particular concern because depressed individuals may remain unidentified by the system. Consequently, minimizing false negatives is important to reduce the risk of overlooking individuals who may require further clinical evaluation. Although false positives are generally less critical in screening applications, they may increase the workload associated with clinical review. Therefore, the observed precision–recall trade-off supports the suitability of the proposed model as a decision-support and triage tool rather than as an autonomous diagnostic system. The precision–recall curve further demonstrates the model's ability to maintain an effective balance between precision and recall across different decision thresholds.

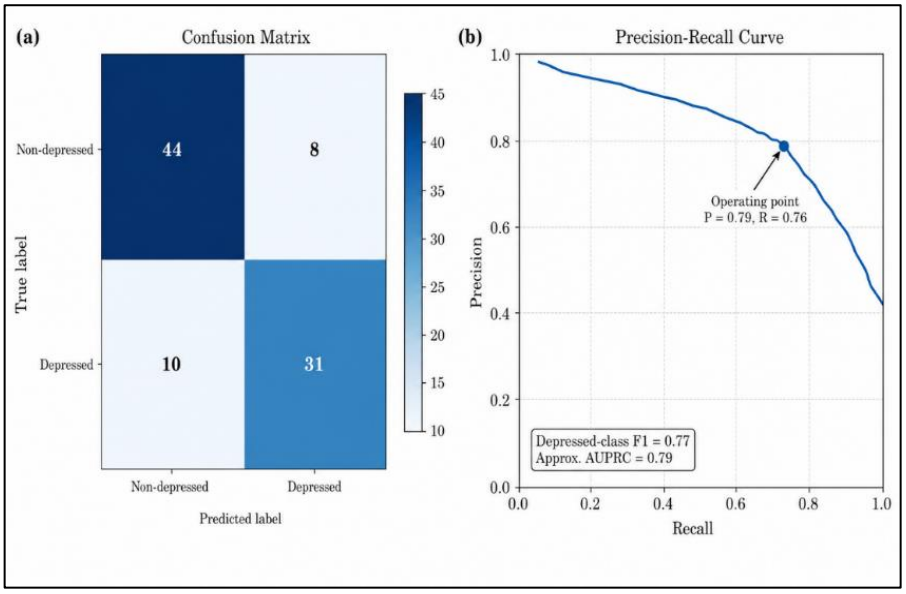

Figure 4. Confusion Matrix and Precision-Recall Curve

6.2 Ablation Study on Fusion Strategies

To isolate the contribution of the fusion mechanism, early fusion, late fusion, and cross-attention fusion were compared while maintaining the same encoder architectures. Early fusion provides moderate improvements over unimodal approaches by combining representations at the feature level. Late fusion further improves performance by leveraging modality-specific classifiers and combining their outputs. However, neither approach explicitly models inter-modal dependencies.

Cross-attention fusion achieved the strongest performance, suggesting that richer joint representations can be learned when textual tokens are allowed to attend to acoustic embeddings. This mechanism enables the model to capture relationships such as emotionally neutral speech delivered with flattened prosody, which is often associated with depressive speech patterns but is difficult to identify using simple concatenation-based fusion methods.

To validate the choice of fusion latent dimension, an ablation study was conducted using latent dimensions of 128, 256, 384, and 512. As shown in Table 3, a latent dimension of 256 provided the best overall balance among Accuracy, Macro-F1, and AUROC. Increasing the latent dimension beyond 256 yielded minimal performance improvements, indicating that larger representations provide limited additional benefit for the proposed task.

Table 3. Latent-Dimension Ablation

Fusion Latent Dimension	Accuracy	Macro-F1	AUROC
128	0.8	0.78	0.85
256	0.82	0.8	0.87
384	0.82	0.79	0.87
512	0.81	0.79	0.86

6.3 Calibration and Uncertainty Analysis

Model probabilities have the potential to become overconfident, especially for bins with high-confidence predictions. Temperature scaling has considerably helped us lower the value of the Expected Calibration Error (ECE), thus reducing the gap between predicted probability and actual accuracy. The proposed method also showed better performance in terms of ECE when compared to unimodal and traditional multimodal models.

Dropout is useful for determining the predictive variance and can thus help detect uncertain samples. The findings show that samples having almost equivalent PHQ-8 scores on either side of the threshold score can often be considered more uncertain compared to others. As is clear from Figure 5, dropout along with temperature scaling helps us achieve not only good calibration but also better uncertainty estimation. After applying temperature scaling, the ECE of the cross-attention method decreased to 0.034 from 0.067. There has also been a drop in ECE values for the other two types of methods, unimodal and traditional multimodal, as is clear from Table 4 below.

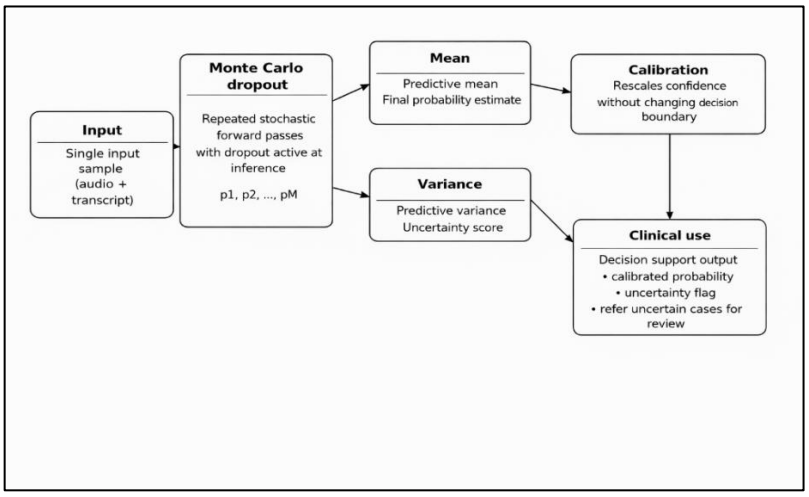

Figure 5. Uncertain Estimation and Probability Calibration Workflow

Table 4. ECE Before and After Temperature Scaling Across Baselines

Model	ECE Before Temp Scaling	ECE After Temp Scaling
Audio-only	0.121	0.083
Text-only	0.109	0.071
Early Fusion	0.088	0.056
Late Fusion	0.079	0.051
Cross-Attention Fusion	0.067	0.034

6.4 Robustness Under Degraded Inputs

The proposed model demonstrated greater robustness under all evaluated perturbation conditions. Unimodal systems exhibited substantial performance degradation when their primary modality was compromised, whereas multimodal models were more resilient to such failures. Among all evaluated approaches, cross-attention fusion experienced the smallest reduction in performance, indicating a stronger ability to adapt to the degradation or loss of one modality.

Table 5. Summarizes Performance Degradation Under Three Perturbation Scenarios

Model	Clean	Noisy Audio	ASR Text	Modality Dropout
Audio-only	0.69	0.6	–	0.42
Text-only	0.73	–	0.64	0.45
Early Fusion	0.76	0.69	0.7	0.58
Late Fusion	0.77	0.71	0.72	0.61
Cross-Attention Fusion	0.8	0.74	0.75	0.66

Table 5 provides the results of model performance when subjected to different types of perturbation. In all cases, the proposed cross-attention fusion framework showed superior performance among all the experiments. In particular, when noise or missing information was presented as input, the uncertainty estimates increased, allowing the model to indicate a lower degree of confidence in its predictions.

Table 6. Sensitivity to Maximum Retained Audio Duration

Max Audio Duration	Accuracy	Macro-F1	AUROC
10 s	0.79	0.77	0.84
15 s	0.81	0.79	0.86
20 s	0.82	0.8	0.87
25 s	0.82	0.8	0.87

An analysis was conducted to check whether audio truncation led to the loss of informative pieces of information by changing the maximum audio length retained in experiments. The performance increased as the duration changed from 10 to 20 seconds and plateaued afterward, showing that an optimal balance of retaining informative audio pieces and saving computational resources had been achieved by choosing such duration. The outcomes of the analysis are presented in Table 6. Table 6 shows that increasing the duration of retained audio from 10 to 20 seconds positively affected model performance. Accuracy went up from 0.79 to 0.82, Macro-F1 from 0.77 to 0.80, and AUROC from 0.84 to 0.87. Short audio files might lack enough information regarding acoustics and linguistics, which would be necessary for effective depression diagnosis. Increasing the audio length beyond 20 seconds did not affect the metrics: Accuracy, Macro-F1, and AUROC stayed at 0.82, 0.80, and 0.87, respectively. This means that there were no benefits from increasing the duration further to 25 seconds. Thus, 20 seconds appears to be an optimal trade-off between the predictive performance and computation.

6.5 Explainability Results

Token-level explanations leverage linguistic relevance and include terms related to helplessness, low self-efficacy, and bad results. Temporal explanation by audio segment masking includes temporal regions of relevance that are marked by longer pauses, low voice power, and monotone speech characteristics. The token relevance map and audio contribution score examples are provided in Figure 6.

Multimodal explanations provide supplementary information through both channels. Linguistic explanations emphasize the importance of words used with respect to the depression diagnosis, whereas acoustic explanations reveal paralinguistic elements that contribute to the predictions made.

6.5.1 Clinical Interpretation

From a clinical perspective, the proposed framework offers several advantages:

- Improved sensitivity to depressive cues through multimodal fusion.
- Uncertainty-aware predictions, enabling risk-based triage.
- Interpretable outputs, supporting clinician trust and auditability.
- Robustness to real-world noise, essential for telehealth deployment.

Rather than acting as a diagnostic authority, the system functions as a screening and prioritization tool, highlighting individuals who may benefit from further clinical evaluation.

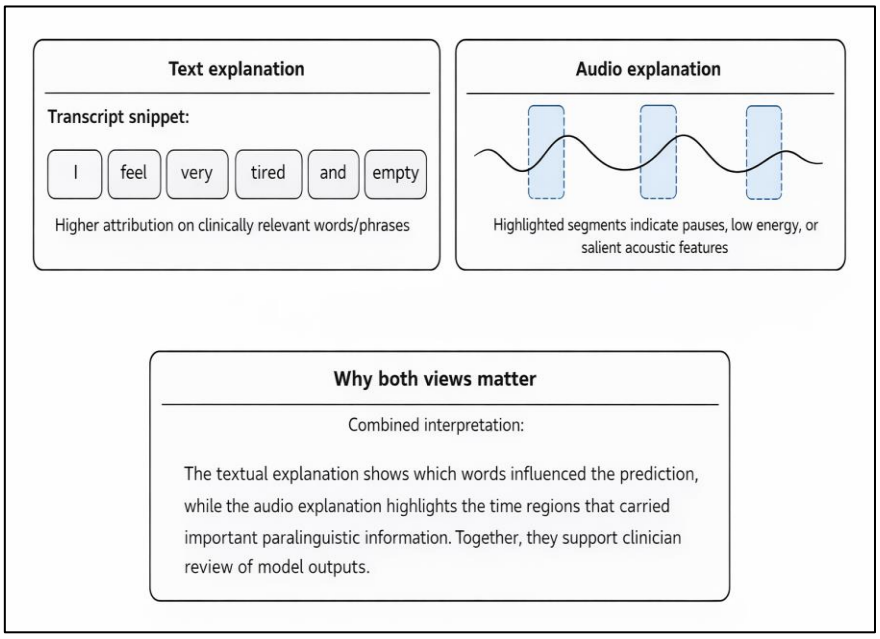

Figure 6. Multimodal Explainability for Clinician Review

6.7 Summary of Findings

The experimental results demonstrate that:

- Multimodal models outperform unimodal baselines.
- Cross-attention fusion provides consistent gains over early and late fusion.
- Monte Carlo dropout and temperature scaling significantly improve probability calibration.
- The framework remains robust under degraded input conditions.
- Explainability mechanisms offer clinically meaningful insights.

Together, these findings support the effectiveness of integrating multimodal learning with uncertainty estimation and explainability for trustworthy depression screening.

7. Conclusion

This paper proposes a reliable multimodal depression detection system that incorporates cross-modal interaction through cross-attention fusion and probability calibration. This framework incorporates self-supervised speech models, namely wav2vec2 and HuBERT, alongside language models, specifically DistilBERT and RoBERTa, for capturing relevant complementary cues related to depression. Our experiments carried out on the DAIC-WOZ dataset revealed that our cross-attention fusion approach performs better compared to audio-only, text-only, early-fusion, and late-fusion models, with accuracy, macro-F1, weighted-F1, area under ROC curve (AUROC), and expected calibration error (ECE) values of 0.82, 0.80, 0.81, 0.87, and 0.034 respectively. From the obtained experimental results, it is clear that incorporating cross-modal interaction leads to better utilization of complementary speech and language cues than traditional fusion models. Further, we show that temperature scaling led to

improvements in probability calibration, with ECE reduced from 0.067 to 0.034, and the Monte Carlo dropout technique facilitated uncertainty estimation and allowed us to predict samples with uncertain predictions. In addition, robustness experiments were carried out to demonstrate the superiority of our proposed system in terms of robustness to noisy environments, automatic speech recognition-generated transcripts, and modality dropout. However, the suggested system is meant as a decision support system for screening and prioritizing patients instead of an autonomous diagnostic system. Future research needs to emphasize validation studies based on larger datasets from a diverse number of clinical cases. The need to conduct comprehensive assessments on fairness cannot be overlooked either.

References

[1] World Health Organization. "Depression." WHO Fact Sheet / Topic Page.

[2] Kraus, Christoph, Bashkim Kadriu, Rupert Lanzenberger, Carlos A. Zarate Jr, and Siegfried Kasper. "Prognosis and Improved Outcomes in Major Depression: A Review." Translational psychiatry 9, no. 1 (2019): 127.

[3] Kroenke, Kurt, 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, no. 1-3 (2009): 163-173.

[4] Chancellor, Stevie, and Munmun De Choudhury. "Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review." NPJ digital medicine 3, no. 1 (2020): 43.

[5] Traum, David, Albert Rizzo, Margaux Lhommet, Jon Gratch, Alesia Gainer, David DeVault, Rachel Wood et al. "SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support." Adaptive Agents and Multi-Agents Systems (2014).

[6] Cummins, Nicholas, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. "A Review of Depression and Suicide Risk Assessment Using Speech Analysis." Speech communication 71 (2015): 10-49.

[7] Atrey, Pradeep K., M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. "Multimodal Fusion for Multimedia Analysis: A Survey." Multimedia systems 16, no. 6 (2010): 345-379.

[8] Yang, Ying, Catherine Fairbairn, and Jeffrey F. Cohn. "Detecting Depression Severity from Vocal Prosody." IEEE transactions on affective computing 4, no. 2 (2012): 142-150.

[9] Menne, Felix, Felix Dörr, Julia Schräder, Johannes Tröger, Ute Habel, Alexandra König, and Lisa Wagels. "The Voice of Depression: Speech Features as Biomarkers for Major Depressive Disorder." BMC psychiatry 24, no. 1 (2024): 794.

[10] Devlin, Jacob, 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), 2019, 4171-4186.

[11] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A Robustly Optimized Bert Pretraining Approach." arXiv preprint arXiv:1907.11692 (2019).

[12] Eyben, Florian, 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, 2010, 1459-1462.

[13] Pérez-Rosas, Verónica, and Rada Mihalcea. "Evaluating Automatic Speech Recognition Quality and its Impact on Counselor Utterance Coding." In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, 2021, 159-168.

[14] Pennebaker, James. "The Development and Psychometric Properties of LIWC2007." (2007).

[15] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All You Need." Advances in neural information processing systems 30 (2017).

[16] Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. "Multimodal Transformer for Unaligned Multimodal Language Sequences." In Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, 6558-6569.

[17] Gratch, Jonathan, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood et al. "The Distress Analysis Interview Corpus of Human and Computer Interviews." In Lrec, vol. 14, 2014, 3123-3128.

[18] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In international conference on machine learning, PMLR, 2016, 1050-1059.

[19] Khera, Rohan, Melissa A. Simon, and Joseph S. Ross. "Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support." Jama 330, no. 23 (2023): 2255-2257.

[20] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why Should I Trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, 1135-1144.

[21] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks." In International conference on machine learning, PMLR, 2017, 3319-3328.

[22] Elham Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (Gaithersburg, MD: National Institute of Standards and Technology, 2023), https://doi.org/10.6028/NIST.AI.100-1.

[23] European Parliament, and Council of the European Union. "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)." Official

Journal of the European Union, L 2024/1689 (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[24] Working, Machine Learning-enabled. "Good Machine Learning Practice for Medical Device Development: Guiding Principles." (2025).

[25] Baevski, Alexei, 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 (2020): 12449-12460.

[26] Ibrahim, Shahana, and Xiao Fu. "Learning Mixed Membership from Adjacency Graph Via Systematic Edge Query: Identifiability and Algorithm." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, 5370-5374.

[27] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." arXiv preprint arXiv:1910.01108 (2019).

[28] Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." In International conference on machine learning, PMLR, 2017, 1321-1330.

[29] Naeini, Mahdi Pakdaman, Gregory Cooper, and Milos Hauskrecht. "Obtaining Well Calibrated Probabilities Using Bayesian Binning." In Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1. 2015.

[30] Neverova, Natalia, Christian Wolf, Graham Taylor, and Florian Nebout. "Moddrop: Adaptive Multi-Modal Gesture Recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 38, no. 8 (2015): 1692-1706.