Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Multimodal Depression Detection System

Dr. Bhargavi Peddireddy, Sree Sai Chandana Kunta, Archana Nagaram, Saishruthi Kethireddy
Dept. of Computer Science & Engineering, Vasavi College of Engineering, Hyderabad, India

Abstract - According to the World Health Organization, depression is one of the most common mental disorders globally. Approximately 5.7% of the adult population worldwide suffers from depression, and it is typically more prevalent in women than in men. Standard methods of screening for depression rely solely on patient self-report using instruments such as the Patient Health Questionnaire (PHQ-8); however, they can fail to capture the full range of an individual's emotional state. The proposed project develops a multimodal depression detection system. This system combines three complementary sources of information for detecting depression: (1) a validated eight-item questionnaire (PHQ-8), (2) sentiment analysis of free-text responses, and (3) acoustic analysis based on a short voice recording. Each of these three modalities will be processed independently using modern machine learning algorithms (i.e., transformer-based natural language processing [NLP] for text; signal processing features combined with a support vector regression [SVR] model for audio; and sentiment analysis of free-text responses), after which the resulting scores will be combined to produce a final depression index, with 32 being the maximum score.

The proposed technology innovations of the multimodal depression detection system include a secure Fernet encryption mechanism for all sensitive inputs, standardising audio samples via preprocessing with FFmpeg (to create a common sample rate), and using environment variables to store and retrieve API keys and encryption secret(s). Results from a pilot test with 15 volunteer participants demonstrate that the multimodal index identifies patterns of distress that are not detectable with single-modality tools Using the audio SVR (Support Vector Regression) model with the RAVDESS (RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song) was able to produce an R^2 of approximately 70%. Furthermore, the correlations found between scores from PHQ-8 and both text and audio SVM (Support Vector Machine) models were moderate positive (text $r \approx 0.52$ & audio $r \approx 0.46$). All three modalities are complementary as opposed to redundant.

Keywords: Multimodal Depression Detection, PHQ-8, Sentiment Analysis, Acoustic Features, Support Vector Regression, RAVDESS, Transformer NLP, Fernet Encryption, Django.

1. INTRODUCTION

Depressive disorder is a syndrome where the person has depressed feelings, loss of interest in activities that are enjoyable and low energy levels. The World Health Organization estimates that about 5.7% of all adults around the world suffer from depression, with a much higher rate in women than men. Untreated depression can have devastating effects on productivity, interpersonal relationships and risk of self-harm that makes early identification and intervention a critical aspect in improving outcomes down the line.

Despite being very easy to administer, as conventional screening tools do (eg, Patient Health Questionnaire [PHQ-8]), the low sensitivity value of PHQ-9 deploying these questionnaires in the clinical setting is common practice. But they depend solely on the individual's self-reported responses. Stigma, fluctuation in mood or misunderstanding the questions can also lead to over-reporting or under-reporting of symptoms by respondents. In parallel, the field of affective computing has found that speech signal patterns (pitch, energy, stability), as well as emotional tone in written text provide useful information about a person's mood state that questionnaires alone might underestimate. Thus, by merging various modalities it is possible to obtain a more general view of individual emotion.

The rationale behind this project lies in three major necessities. Firstly, there is an interest in a holistic method that not only includes self-reported measures but also uses subconscious cues expressed via language and vocal qualities. Secondly, convenience and accessibility are considered essential since web-based software run within a browser enables individuals to do testing at home, which is particularly helpful when in-person appointments become impossible. Lastly, personal data associated with the mental state is highly confidential, so privacy and security are crucial when working with such material.

To address these requirements, a complete stack web app was built that facilitates the filling of the PHQ-8 questionnaire,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

allows for the provision of free text descriptions, and analyzes a brief speech sample. The system implements a transformer-based module that analyzes the sentiment of the text and utilizes an audio processing pipeline that first downsamples the uploaded audio sample to 16 kHz mono WAV and then extracts audio features including MFCC, pitch, jitter, and shimmer. All three scores – from the questionnaire, text, and audio samples – are normalized and combined into one depression score along with personalized self-care recommendations. Related work and system details can be found in Sections 2 and 3. Experimental results are given in Section 4; future directions are presented in Section 5.

2. RELATED WORK

There have been three major generations in terms of research on automated depression and mood detection. The first one made use of keyword and rule-based techniques, scoring documents on their word frequency. The second generation used latent techniques like Latent Semantic Analysis, Word2Vec, or GloVe to learn representations. Finally, the third generation uses transformer models, employing the technique of self-attention to represent context within the whole input.

In the beginning, depression detection was done using text analysis, where the focus was on the analysis of sentiment and linguistic patterns in social media or patient narratives. Such work relies on the techniques of natural-language processing for measuring mood on the basis of word choices, syntax, and polarity. However, there are certain disadvantages in text-based models that are associated with the interpretation of sarcasm, slangs, or cultural differences and that necessitate the presence of big labelled data.

In addition, researchers investigated the possibilities of using audio-based solutions, as the pitch, energy, jitter (frequency cycle-to-cycle variability), and shimmer (amplitude variability) changed according to the level of depression severity. Clinical experiments showed a correlation between depression scores and changes in those acoustic markers with depressed people displaying increased jitter and shimmer, decreased energy, monotonous intonation, and slow speech. Dense retrieval passage models proved the efficacy of a combined usage of sparse and dense retrieval signals over each other.

Finally, there is an increasing tendency in recent research to utilise multilingual and multimodal approaches. Machine-learning technologies allow the automatic analysis of such datasets to discover non-linear dependencies, invisible otherwise. M3L framework proves that pre-trained multimodal models like Whisper (speech) and XLM-RoBERTa (text) can be successfully adapted for depression detection in any language. Other studies reveal that multimodal models have higher F1 scores (up to 0.94) and better accuracy compared to models based on a single modality. Besides, it is shown that eight-item PHQ-8 questionnaire has similar diagnostic power to nine-item one, having identical AUC values (0.76) and no statistically significant differences in sensitivity or specificity.

3. SYSTEM ARCHITECTURE

The proposed architecture is highly modular with six steps and clear separation of input/output, allowing replacement of subsystems independently without breaking another modules. The stages are the following: data collecting and ingestion, PHQ-8 score calculation, text sentiment analysis, audio feature extraction and regression, score fusion and visualization with AI-generated recommendation.

Security module covers all processes encrypting the data and managing secrets with environment variables.

3.1 Data Collection and Documents Ingestion Upon registration, a user enters the web application based on Django and fills out the form with responses to PHQ-8 questions (from 0 to 3). Additionally, a user can provide free-form texts about his feelings towards each question, respectively. Finally, the user can record a short voice note. Every input is automatically encrypted using Fernet symmetric encryption before storing. The assessment data regarding user's state are stored as an Assessment object in Django ORM containing eight PHQ-8 answers, eight text descriptions (optional), depression score, and time when it was created.

3.2 PHQ-8 Questionnaire Scoring

In order to find raw PHQ-8 score, it is necessary to map PHQ-8 descriptive responses ("Not at all," "Several days," "More than half the days," "Nearly every day") to numeric values within the range 0-3, then sum it up and obtain the raw PHQ-8 score from [0-24]. The higher the score, the more clinically serious the depression problem is. The raw score determines the severity of the condition in one of five classes: minimal (0-4), mild (5-9), moderate (10-14),
moderately severe (15-19), and severe (20-24).

3.3 Text Sentiment Analysis

To extract information regarding sentiments in texts provided

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

by users, we employ a pre-trained transformer-based machine learning model from the Hugging Face library with its pipeline API wrapper. The probability of text to belong to the category of negativity is predicted for each input text, after that, it is averaged among all texts and multiplied by 4 (and capped to [0, 4] interval) to make it equally important to audio. This way, such text nuances as hopelessness, frustration, and indifference can be taken into account in calculating final score.

3.4 Voice Processing Pipeline and Feature Extraction

All voice notes submitted by users are converted to 16 kHz Mono WAV files via pydub and FFmpeg. Next, with the help of the librosa library, the following acoustic features are extracted from every audio clip: MFCC (Mel-frequency cepstral coefficient) mean and std with 13 dimensions; chroma energies (12 dimensional representation describing energy distribution between 12 pitch classes); mel-spectrogram energy; spectral centroid, bandwidth and roll-off coefficients (spectral characteristics); root-mean-square energy; zero-crossing rate; mean pitch. Using Parselmouth package, the following voice quality features are calculated with the use of Praat voice analyser: local jitter and local shimmer parameters. They are standardized using sklearn.preprocessing.StandardScaler and further used to train Support Vector Regressor (SVR) model.

3.5 SVR Model and Dataset

Audio features obtained in Stage 4 are used to predict PHQ-8 score with the SVR model with Gaussian kernel on the dataset named RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). It contains 1440 clips of 24 professional actors (12 males, 12 females) performing 8 emotions – neutral, calm, happy, sad, angry, fearful, disgusted, and surprised – at both low and high intensity levels. All the voices are recorded with 16 bit-per-sample, sampling rate 48 kHz. The predictions of SVR model are multiplied by 4 and capped to [0-4] range to be equivalent to other modalities.

3.6 Query Processing and Score Fusion

Final PHQ-8 score is calculated via simple additive fusion: **Final Score = PHQ-8 Score (0-24) + Text Score (0-4) + Audio Score (0-4)**, hence the maximum of the final score is 32. While PHQ-8 remains to define clinical severity class of depression, the latter will contain additional nuances. For each of three scores, a color-filled gauge is provided to the user via Chart.js library. On the result page, there is a disclaimer stating that this app should only be used for awareness purposes and not diagnosis.

Besides, recommendations on self-care activities are obtained via an external Groq API for LLaMA 3.3 70B large language model. If no connection to external API is found, general recommendations are provided to the user. API keys are provided in environment variables with python-dotenv.

4. EXPERIMENTAL RESULTS

4.1 Setup and Dataset

A pilot experiment involving 15 volunteers completing the entire survey was performed to assess performance. The textual responses were annotated independently by two annotators to validate the sentiment scores with disagreements discussed until consensus reached. A separate evaluation of the audio classifier on the RAVDESS dataset used an 80/20 train/test split.

4.2 Audio Model Performance

On the held-out portion of the RAVDESS dataset, the support vector regression (SVR) using the radial basis function (RBF) kernel trained on the 33D feature vectors generated from standardised audio input achieved $R^2 \approx 0.70$. This translates into the ability to explain ~70% of the variance in emotions with the selected features — MFCC, spectrum statistics, pitch, jitter, and shimmer — despite being acted rather than clinically recorded. This classifier type was selected due to its strong performance on medium to small datasets and producing smooth output.

4.3 Modality Score Distribution

Table 1 contains the distribution of the modalities' scores obtained in the pilot experiment.

Table 1: Retrieval Performance - Modality Score Distribution (n = 15, mean over all users)

Modality	Mean	SD	Min	Max
PHQ-8 (0–24)	8.1	3.7	3	17
Text (0–4)	2.3	1.2	0.0	4.0
Audio (0–4)	1.6	0.9	0.4	3.8
Final (0–32)	12.0	5.1	5	24

4.4 Ablation — Relationship Between Modalities

Despite design considerations aiming at independence,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

some level of correlation between the three modalities was observed. The PHQ-8 score showed positive correlations with both text score ($r \approx 0.52$) and audio score ($r \approx 0.46$). This demonstrates that subjects with a higher symptom severity tend to use negatively toned language and have depressed speech characteristics. However, these correlations were not sufficiently high to imply redundancy. For instance, there were users scoring low on the PHQ-8 but having high scores on either audio or text modalities, suggesting that other modalities can pick

up on indicators of depression undetected in the self-report questionnaire. Table 2 presents a qualitative comparison of approaches.

Table 2: Qualitative Comparison of Approaches

Approach	Semantic	Hallucin.	Explain.
PHQ-8 Only	Low	N/A	Minimal
NLP Only	High	Low	Partial
Audio Only	Medium	Low	Partial
Standalone LLM	High	High	None
Proposed	High	Low	Full

4.5 Example Result and Interpretation

As an example of interpreting a result, consider a pilot participant scoring PHQ-8 7 (mild range), text score 4.0 (strongly negative language usage), and audio score 1.59 (moderate vocal affect). The final index obtained by the method is 12.59 (out of 32) with the label "Mild" inferred based on the PHQ-8 classification. On the results page, the user gets their overall score shown together with the breakdown of modalities with color-coding and progress bars, and relevant suggestions — establish regular sleep pattern, divide large tasks into smaller parts, and combat negative thinking. As a disclaimer, the tool informs that its output is meant for awareness purposes only and encourages consulting a healthcare professional.

End-to-end latency of the entire process was measured in the pilot experiments. The median response time turned out to be about 2.1 seconds with the Groq LLM generation step (~1.7 seconds) accounting for the majority of latency. The local portion including embedding conversion, audio feature extraction, and prediction in SVR took under 200 milliseconds.

5. FUTURE WORK

Directions contributing to the improvement of the proposed solution's efficiency are reviewed.

First of all, it is crucial to use clinical speech samples to train models. The current audio encoder model was trained on the RAVDESS dataset with professionally recorded actors. This approach may prevent

achieving the most accurate representation of depression-related characteristics of the voice because of the nature of the used samples. Therefore, it is necessary to use a clinical dataset such as DAIC-WOZ or EATD to train the models on spontaneous interviews.

The next promising direction of improvement concerns end-to-end deep learning. Rather than working with manually selected features, the idea is to use the power of neural networks that work directly with the raw inputs of speech and text samples. Speech representations can be extracted through Wav2Vec 2.0, while BERT encoder family pretrained on large text corpora can be fine-tuned to process text. As was demonstrated previously, M3L framework allows building multilingual models in multiple modalities. Also, one can try using an attention mechanism to fuse information from all modalities rather than simply adding scores together.

Multilinguism and cross-cultural validation are among the highest priorities. In order to solve cross-lingual issues, the use of pretrained multilingual language models such as Whisper, XLM-RoBERTa, and mBERT is highly recommended. Moreover, it is essential to validate PHQ-8 values applicable to other cultures. Adding facial expression (CNN model) or physiological signals (wearable) will bring extra value related to understanding patient's condition. Development of native applications for mood tracking longitudinally along with using FHIR-based electronic health record integration will make the developed solution suitable for clinical usage.

6. CONCLUSION

The aim of this paper is to introduce a multi-modal depression detector that transforms the existing passive self-assessment process into a more complex one, comprising three signals for evaluation. In terms of our solution's novelty, we claim that the proposed system offers the following three features at once, none of which can be achieved with any single-component system before: (1) semantic analysis of users' intent via transformer sentiment analysis rather than via keyword matching; (2) generative synthesis of personalized self-care recommendations based on users' feedback; and (3)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

transparency of the explainability component with clear modality breakdown.

With the integration of a standardized PHQ-8 survey test, transformer sentiment analysis of users' text responses, and an SVR prediction model trained on the RAVDESS speech data set, the proposed system produces clinically grounded depression severity assessment and a depression index ranging from 0 to
32 with personalized self-care suggestions. The audio-based SVR model demonstrated an excellent performance, with $R^2 \approx 0.70$. Moderate correlations between modalities ($r \approx 0.46$ - 0.52) prove that each of the three modalities is a valuable and non-redundant source of information. A pilot evaluation study confirmed that those respondents whose PHQ-8 score was lower sometimes had a higher index in either text or audio modality.

According to our empirical results, the average end-to-end latency time amounts to 2.1 sec, with LLM-based generation being the most time-consuming part, local retrieval and feature extraction being executed faster than 200 ms. The use of Fernet scheme, environment-variable secrets management, and temporary audio storage prove that it is possible to make a privacy-conscious software system using only a standard Django web application. Future steps include integration of speech corpus training, deep learning model with end-to-end prediction, multilingual functionality, new modalities and mobile app deployment.

REFERENCES

[1] M. Sadeghi, R. Richer, B. Egger, L. Schindler-Gmelch, L. H. Rupp, F. Rahimi, M. Berking, and B. M. Eskofier, “Harnessing multimodal approaches for depression detection using large language models and facial expressions,” npj Mental Health Research, vol. 3, Art. no. 66, Dec. 2024.

[2] World Health Organization, "Depression Fact Sheet," 2023.

[3] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, 2018.

[4] M.-S. Seifpanahi et al., "The Association between Depression Severity, Prosody, and Voice Acoustic Features in Women with Depression," The Scientific World Journal, 2023.

[5] H. Abedi et al., "Depression Detection Methods Based on Multimodal Fusion of Voice and Text," IEEE Access, 2025.

[7] B. McFee et al., "Librosa: Audio and Music Signal Analysis in Python," SciPy Conference Proceedings, 2015.

[8] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python Interface to Praat," Journal of Phonetics, vol. 71, 2018.

[9] Hugging Face, "Transformers Pipeline API," 2024. [Online]. Available: www.huggingface.co/docs

[10] Groq, "Groq API Documentation."

[11] Python Cryptography Authority, "Fernet Specification," 2013.

[12] Django Software Foundation, "Django Web Framework Documentation," 2024.