RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2554

PSAAP: Patient-Specific Audio Extraction Pipeline for Depression-Oriented Vocal Biomarker Analysis

Raminder Kaur [a,*], Vikram Kulkarni [a]

[a] Mukesh Patel School of Technology Management & Engineering, SVKM's NMIMS, Mumbai-400056, India

* Corresponding Author Email: raminderphd.nmims@gmail.com

DOI: https://doi.org/10.54392/irjmt2554

Received: 28-04-2025; Revised: 17-08-2025; Accepted: 04-09-2025; Published: 13-09-2025

Abstract: The stigma of depression and mental illness is growing everywhere in the world and this is the reason why the moves are on to come up with diagnostic tools, which would be rapid, efficient and consistent. The framework suggested in this study is a state-of-the-art technic to infer patient-specific data on therapy-patient dialogs to fill a knowledge gap that existing diagnosis solutions are unable to cover. The proposed signal processing scheme, Patient-Specific Audio Extraction Pipeline (PSAAP) enhances input to the Machine Learning (ML) of mental illness detection. The method locates and measures non verbal acoustic features such as the pitch, the intensity, and the Mel Frequency Cepstral Coefficients (MFCCs) that play a crucial role in the determination of mental good health. Noise reduction, speaker diarization, silence deletion as well as other samples of preprocessing are implemented on DAIC-WOZ dataset to maintain quality of audio and sound. In this code, speech characteristics which are instrumental in such diagnosis have been retained so that the depression symptom described by monotony, slow speech and low variability of pitch can be analyzed precisely. The quantitative findings are demonstrated in the framework and signal-to-noise ratio (SNR) is augmented by up to 16 dB compared to the existing methods. The method allowing the practitioner to make a sound and clinically pertinent evaluation since separating patient-specific variables of the voice and removing therapist feedback makes assessment trait-centered and dependable. The main argument in favour of this one is that it possesses the quality to be applied in the extraction of patient speech framework in general since the patient can find such opportunity to be helpful in the mental healthcare examination and also in the circumstances where the computing capabilities are poor.

Keywords: Mel-frequency cepstral coefficients (MFCCs), Short time Fourier transform (STFT), Distress analysis interview corpus - wizard of Oz (DAIC-WOZ), Discrete cosine transform (DCT)

1. Introduction

Depression is one of the most prevalent mental disorders defined recently by deficient emotional regulation, persistent low mood and depressed interests, impaired concentration ability, and possibly suicidal thoughts [1]. More than 322 million people suffer from depression worldwide, and this scenario is witnessing a blood-curdling upward trend in prevalence [2]. It has been recognized by World Health Organization (WHO) as the fourth biggest cause of disability and may reach second position in the ranking by 2030 [3]. The prevailing techniques for measuring psychiatric pathology remain largely subjective through the methods utilized traditionally including patients' verbal account, reports of known individuals on their behavior and experiences, and mental states [4]. But it usually needs a large quantity of professional expertise as well to support. Some of the key issues faced in the context of the advanced mental health framework for patient voice extraction in precision assessment are related to the ability to identify the effect of various emotional and psychological components among non-homogeneous and complex patient voice data. Since patient data may exhibit varying speech, accent, and tone during the conversation, model credibility may be at stake. Small sample size of datasets and the presence of imbalance classes add more challenges to the models. Privacy and ethical issues [5] as regard to handling patient data also require utilization of excellent measures to enhance the privacy of data. Also, there are some context factors including background noise and overlapping speech which interferes with the extraction of features. These issues can be best addressed through source analysis of signals, fusing information from different modalities, and maintaining high ethical standards [6].

Voice is considered one of the primary biometric markers in mental health and provides crucial information about a person's emotional and

psychological states [7]. Although text or written responses would not reveal such subtle cues, voice does–it can give impressions with regard to or underlying mental health issues; with aspects such as tone and pitch and a person's speed of speech as well as pauses, these elements indicate a mood and state of mind. For instance, in depression speech is slower, more monotonous, and less variable in tone-traits which vocal analysis can objectively measure [8]. Non-verbal information might complement traditional assessment techniques to give a better view of the patient's mental health condition. One of the important characteristics of assessing mental health using the voice is that it serves as a non-intrusive accessible, and continuous monitor. Unlike time-consuming physical and psychological tests, voice analysis may easily be included in everyday interactions by a passive monitoring that would not even disturb the patient's daily schedule [9].

In depression detection tasks involving conversational data, isolating the patient's voice from the therapist's is essential. Conversations between a patient and a therapist are rich with verbal and non-verbal information, but when both voices are analyzed together, the presence of multiple speakers can distort critical depression-related indicators. Therapist contributions such as prompts, questions, or affirmations add emotional nuance that does not represent the patient's mental state [10]. This isolation of the voice of the patient will render the analysis more concentrated and specific and it is so that the depression symptoms including the tone or the rate of speech can be better estimated without interference of some other voice. Elimination of noise and silence is important to achieve separation of speakers of good quality. This form of recording environment i.e., ambient/equipment noise results in this form of background noise that tend to present artefacts in voice feature extraction [11].

The removal of silence helps in eliminating unwanted intervals in the speech which would otherwise be distractive to the speech being delivered. These preprocessing are helpful in removing any ambiguity to the voice of the patient and giving the downstream analysis a better faithfulness as the noisy aspects of the signal have been discarded [12, 13]. As a hypothesis to the study presented, there is an attempt at identifying the method of retrieval and isolation of the voice of a patient in a two-party dialogue. The targeted extraction to indicate that the voice features of the patient in terms of intonation, rates of the voice, voice pauses is applied only in the consideration of the depression signs and not of the voice of the therapist. The solitude of the voice of the patient will enable us to gain a more realistic and unadulterated statement of the psyche of the patient in such a way that we have a condensed analysis of the patient that will probably give us true insights into his level of depression [14,15].

1.1 Literature Review

A comparative analysis of literature on speech-based depression detection and related techniques is presented in Table 1.

2. Methodology

The approach employed in the detection of depression as depicted in figure 1 is the six step machine learning procedure to identify depression through audio information. It begins with the process of data collection involving the patient voice and demographics. Thereafter the audio is polished off by eliminating noise and separating relevant speech. In extracting features, MFCCs come into play in order to describe the vocal qualities. Such features are utilized when training the model where EfficientNet determines how to define depression. Such metrics as accuracy and precision are applied in the performance evaluation stage of the model to determine efficiency of the model. In the event the results are not optimal a feedback loop will be adopted in refining the model. This official procedure will ensure consistency and versatility in detecting depression which is based on the aspect of speech analysis.

2.1 Patient-Specific Audio Extraction Pipeline (PSAAP) Framework

The patient speech extraction methodology based on DAIC-WOZ dataset is applied, which includes audio interactions of therapists with patients necessary to perform diagnostics of mental health. This included the step of preprocessing methods i.e. noise reduction, speaker diarization, and silence suppression before achieving feature extraction on basis of STFT and MFCC [16] to design a strong framework of audio tests on patient-specific basis. This entire framework steps are discussed as below.

2.1.1 Data Collection

DAIC-WOZ [14] dataset is an important resource among mental health research studies and especially studies on the identification of depression through the analysis of speech. First introduced by the University of Southern California Institute for Creative Technologies, this dataset was created to help researchers access a rich audio-visual corpus for studying the distress conditions while keeping an emphasis on depression. The DAIC-WOZ dataset captures audio and video recordings of interviews conducted with participants and the virtual interviewer named Ellie [17]. Interviews were conducted in a Wizard-of-Oz setup where the responses given by the virtual agent were controlled by human interviewers who were behind the scenes. Figure 1 illustrates the experimental setup.

Table 1. Summary of Key Literature on Speech-Based Depression Detection and Related Techniques

Reference	Study Reference	Focus Area	Methodology	Key Findings	Identified Limitations
[1]	Wang *et al.* (2020)	Depression recognition from audio	CNN + GAN	Achieved high accuracy using deep learning on voice signals	Lacks real-world data variability
[2]	WHO (2017)	Global depression statistics	Epidemiological Report	Provided baseline for global mental health status	Not technical; lacks ML relevance
[3]	Solieman & Pustozerov (2021)	Multimodal depression detection	Text + Voice Feature Fusion	Improved detection using multimodal inputs	Requires large labeled datasets
[4]	Li *et al*. (2019)	Depression detection via ML	SVM, RF, Feature Engineering	Compared different feature generation methods	Limited generalizability to clinical settings
[5]	Ravi *et al.* (2024)	Speech-based depression detection with privacy	Speaker Disentanglement	Improved both privacy and accuracy in detection	Model complexity increases deployment time
[6]	Shinichi (2015)	Stress evaluation from voice	Voice Stress Analysis	Demonstrated voice as a non-invasive indicator	Lacks ML integration; basic statistical models
[7]	Chetty *et al.* (2017)	Brain tumor segmentation (reference framework)	Image Processing Survey	Reviewed MRI segmentation techniques	Not directly related to depression/audio
[8]	Zhao *et al.* (2022)	Vocal biomarkers of depression	Acoustic Feature Analysis	Found significant vocal indicators of depression	Cross-sectional design, not longitudinal
[9]	Fürer *et al.* (2020)	Speaker diarization in psychotherapy	Random Forest-based Diarization	Improved diarization for therapy session analysis	Specific to psychotherapy; not depression per se
[10]	Huang *et al.* (2024)	Voice-based depression detection	Voice Pre-training Models	Achieved superior results with pre-trained speech encoders	Data privacy and generalizability concerns
[11]	Low *et al.* (2020)	Review of speech in psychiatric diagnosis	Systematic Review	Identified trends and challenges in speech-based diagnostics	Mostly review; lacks experimental analysis
[12]	Gratch *et al.* (2014)	DAIC-WOZ Corpus Development	Human-computer interview design	Established benchmark dataset for depression detection	Dataset lacks multi-language support
[13]	Sun *et al.* (2022)	Speech diarization in health analytics	Diarization Techniques Review	Discussed applications in health monitoring	Does not offer implementation metrics
[14]	Williamson *et al.* (2016)	Speech separation in noisy environments	Multi-Speaker Recognition	Improved recognition in complex scenarios	Not specific to depression
[15]	Ao *et al.* (2024)	Universal speaker diarization	USE-D Algorithm	State-of-the-art diarization for real-world applications	Focus is broader than mental health

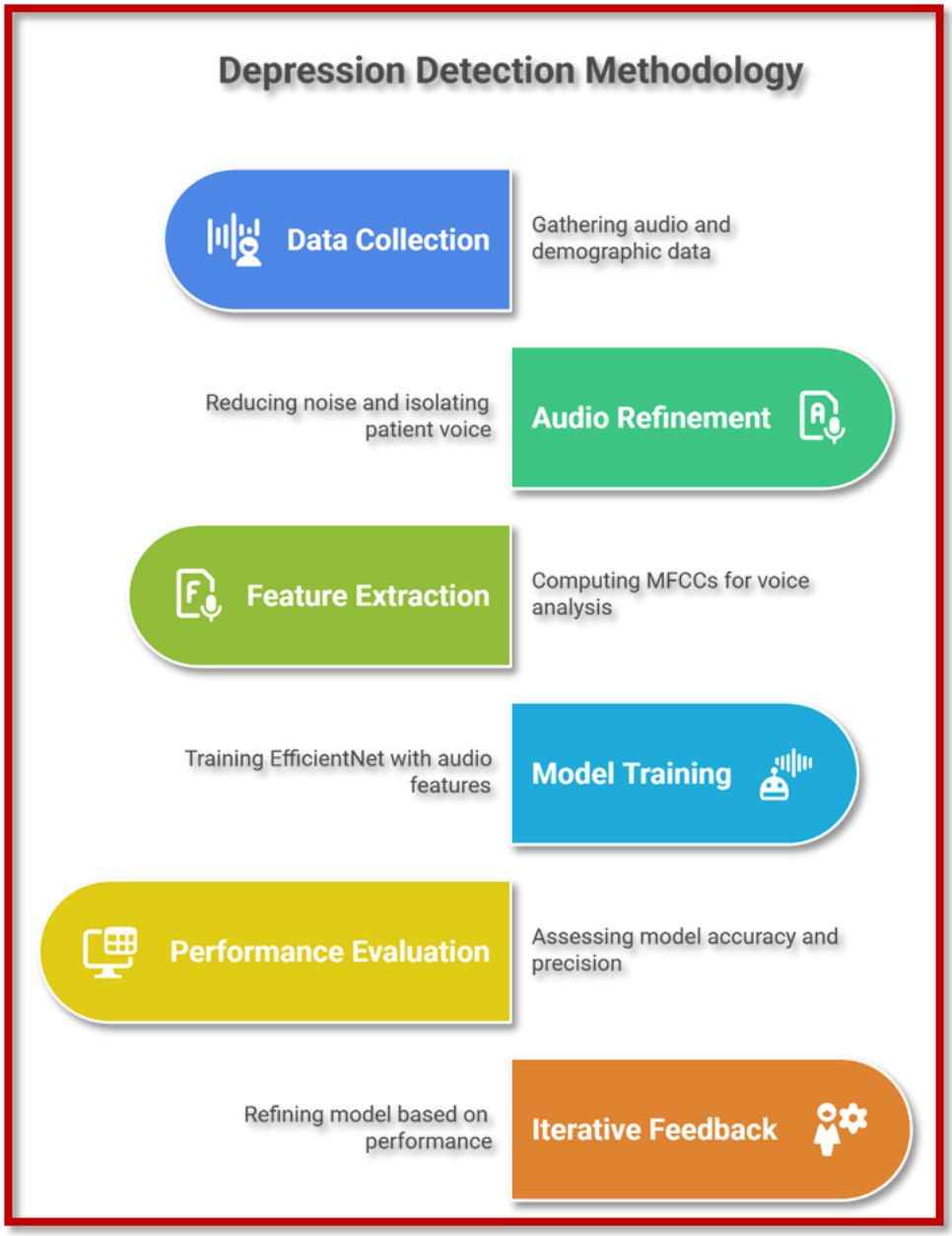

Figure 1. Methodology

This method resulted in natural conversations, semiformal, while simultaneously ensuring uniformity of the interview format across all the participants. The interviews within the dataset last typically between 7 and 33 minutes, averaging about 16 minutes for each interview. The available data are available in four forms:

1. Audio Recordings: High-quality audio recordings of interviews, sampled at 16 kHz.
2. Transcripts: Detailed transcripts with timings of the discussions.
3. Depression Scale: Scores of the Patient Health Questionnaire-8 for each participant providing standardized measurement of depression.
4. Basic Demographic Information: For each participant, basic demographic information.

This dataset contains a total of 189 audio files. Figure 2 and Table 2 gives a summary of the description of the dataset.

2.1.2 Audio Refinement and Speaker Isolation for Targeted Voice Analysis

Patient voice extraction is core part of the framework and the flow is shown in Figure 3. First, through background noise reduction, the ambient sounds and the irrelevant noises in the raw audio file are minimized. This first step is important because it enhances the clarity and quality of the patient's voice, which makes further analysis easier. After noise reduction, speaker diarization is applied to identify and label the various different speakers in the conversation. Post diarization, segmentation and silence removal are done to better refine the audio.

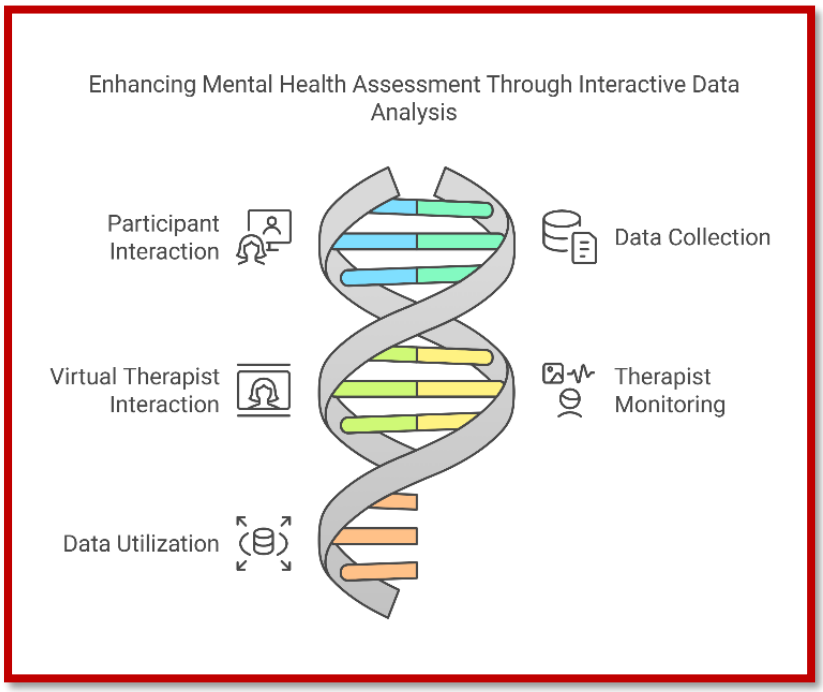

Figure 2. Data Collection Setup

Table 2. Dataset Summary

Total number of files	189 files
Sampling rate	16,000 Hz
Data type	Audio recordings, transcripts, PHQ-8 scores, demographic information
Average Interview Duration	16 minutes (range: 7-33 minutes)
Audio file format	.wav
Key Applications	Depression detection, emotion recognition, mental health assessment
Main Challenges	Limited sample size, class imbalance, contextual factors

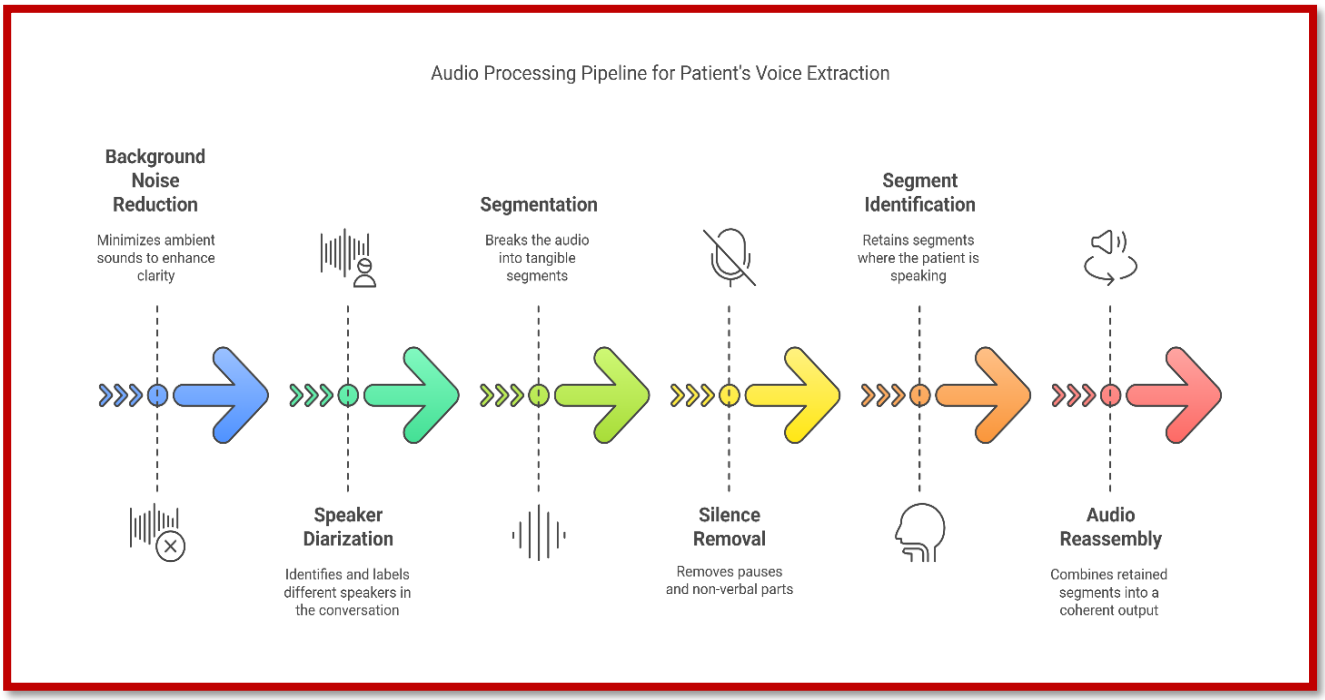

Figure 3. Flow of Patient Voice Extraction

Segmentation broke the audio into tangible segments, and silence removal removed pauses, non-verbal parts, and kept only relevant audio segments for further processing. In the next step, segments where the patient was speaking were identified using the diarization labels and retained for proper reassembly to a coherent output of the patient's voice. This final output represented the extracted patient audio that can then be used for further analysis.

2.1.2.1 Background Noise Reduction

The background noise reduction process in this methodology leveraged Spectrogram analysis and Spectral Gating [18] to isolate and reduce unwanted background noise while preserving the primary speech signal. Figure 4 illustrates the calculated noise profiles for the first three seconds of the respective audio files, providing a visual representation of the baseline noise characteristic.

Figure 5 shows the steps for background noise reduction. This output represented the cleaned version of the original audio, with the unwanted noise reduced and the primary speech signal preserved.

Each frequency bin in the spectrogram was evaluated against the noise profile, and a decision was made whether the frequency should be considered part of the signal or as noise. If a frequency component's energy was at or below the noise threshold, it was considered noise and was attenuated and frequencies that exhibited significantly higher energy were preserved as part of the primary signal. The attenuation factor for the noise components was set at 30 dB, a value that was sufficient to substantially diminish unwanted noise without affecting the integrity of the signal. After the noise reduction had been applied to the spectrogram, the modified spectrogram was then converted back into the time domain. This was achieved using the Inverse Short-Time Fourier Transform (Inverse STFT), which reconstructed the cleaned audio signal by reversing the transformation process applied in the STFT. The Inverse STFT took the adjusted spectrogram, including the reduced noise components, and synthesized a time-domain audio signal. Table 3 provides a concise summary of the background noise reduction results, showing initial noise energy, the amount of noise reduced, and the overall reduction percentage, demonstrating the effectiveness of the applied noise reduction techniques.

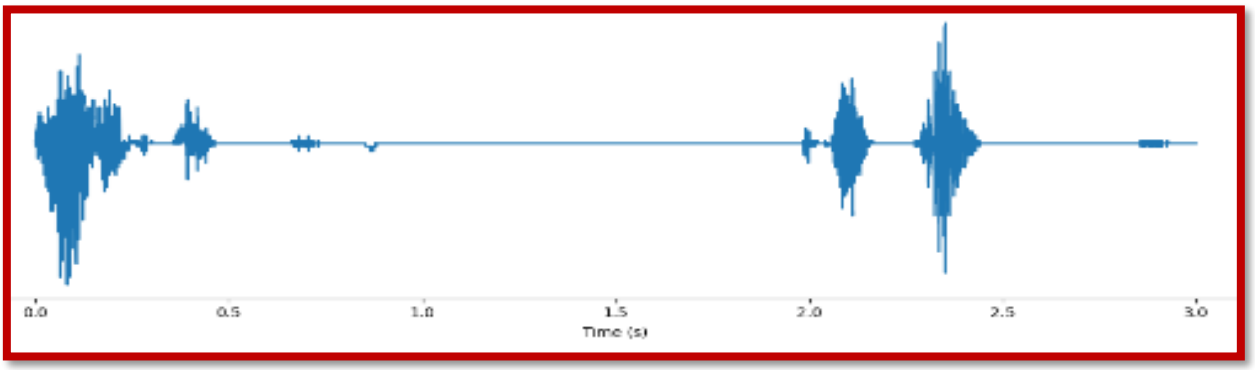

Figure 4. Noise Profile Sample representation of DAIC-WOZ

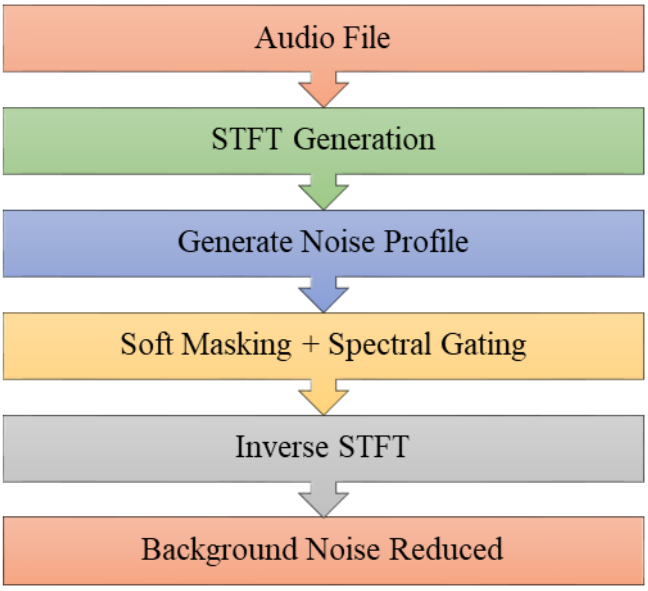

Figure 5. Steps for Noise reduction

Table 3. Noise Energy Reduction

Parameter	Description	Metric/Calculation	DAIC-WOZ Dataset Value
Initial Noise Energy	Total noise energy before processing.	Sum of noise power over the entire duration.	75-85 dB
Reduced Noise Energy	Total noise energy removed after processing.	Energy difference post and pre-processing.	30-40 dB
Reduction Percentage (%)	Percentage of noise energy removed.	Where: • Energy $_{pre}$: Noise energy before processing. • Energy $_{post}$: Noise energy after processing.	50-60 %

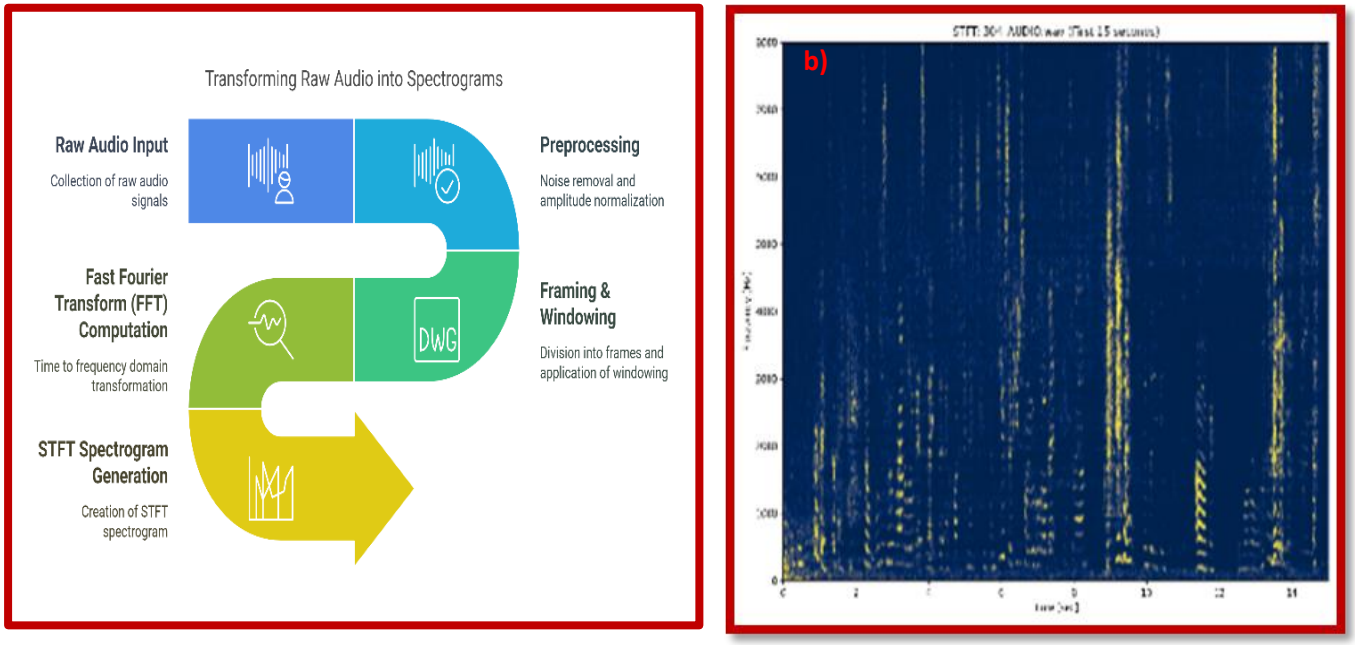

Figure 6. a) Steps to compute STFT spectrogram, b) STFT spectrograms Sample representation of DAIC-WOZ

STFT is used as shown in Figure 6a, to transform audio signals into time- frequency representation for the analysis of frequency content change. It provides a detailed view of both temporal and spectral features, required for detecting small changes between the real and synthetic speech. Its concise representation of the signal's frequency and temporal information makes it suitable for image-based classification models.

The steps for computing the STFT from an audio signal in shown in figure 7. The spectrogram in Figure 6b, shows time on X- axis, frequency on the Y- axis, and intensity as colour or brightness. STFT of first 15s of audio files from dataset.

2.1.2.2 Diarization

To identify which speaker is the patient, speaker diarization [19, 20] is applied to the entire audio file. Diarization segments the audio based on speaker changes and labels each segment with a speaker ID such Speaker 1, Speaker 2 and so on. The audio is first converted into features that are suitable for clustering. Features considered for diarization here include MFCCs, STFTs, chroma features, spectral contrast, zero-crossing rate, RSME and Pitch. The audio is segmented into small windows of 30ms with an overlap of 10ms for each window, then features were extracted from each segment. These small windows are then clustered into groups that likely correspond to different speakers using K-means clustering. The goal of clustering is to group audio segments where the speaker remains the same. After the audio has been segmented and clustered, each segment is assigned a label (e.g., Speaker 1, Speaker 2). These labels are the diarization output, indicating which speaker was talking during each segment. Since there are only two speakers (therapist and patient), the most frequent speaker ID is assigned to the patient label, based on the assumption that the patient speaks more consistently than the therapist. Figure 8 illustrates the steps to identify and label various speakers in an audio file, of a conversation recording.

Input: Audio signal. Frame length NNN, Overlap MMM, Window function w(m)w(m)w(m) (e.g., Hanning), Fourier Transform function **FT**

Process:

Frame Segmentation of continuous audio signal x(t)x(t) is divided into overlapping frames of length NN, with an overlap of MM samples:

$$xn(m) = x(m + nH), m = 0,1, \dots, N - 1x_n(m) = x(m + nH), \quad m = 0,1, \dots, N - 1$$

where HH is the hop size (frame shift), NN is the frame length, and $xn(m)x_n(m)$ represents the segmented frames

A window function w(m)w(m), such as the **Hanning window**, is applied to each frame:

$$xnw(m) = xn(m) \cdot w(m)x_n^w(m) = x_n(m) \cdot w(m)$$

where the **Hanning window** is given by:

$$w(m) = 0.5(1 - co\,s(2\pi mN - 1)), 0 \le m < Nw(m) = 0.5\left(1 - \cos\left(\frac{2\pi m}{N-1}\right)\right), \quad 0 \le m < N$$

This step smooths the edges of each segment and reduces spectral leakage.

The **Discrete Fourier Transform (DFT)** on each frame is applied to convert the time-domain signal into the frequency domain:

$$Xn(k) = \sum m = 0N - 1xnw(m)e - j2\pi km/N, k = 0,1, \dots, N - 1X_n(k) = \sum_{m=0}^{N-1} x_n^w(m)e^{-j2\pi km/N}, \quad k = 0,1, \dots, N - 1$$

where $Xn(k)X_n(k)$ represents the frequency components of each frame. **Magnitude Spectrum Extraction** is absolute value of each frequency component is computed to obtain the magnitude spectrum:

$$|X_n(k)| = \sqrt{Re\big(X_n(k)\big)^2 + Im\big(X_n(k)\big)^2}$$

This discards phase information and retains only the intensity of each frequency component.

Short-Time Fourier Transform (STFT) Spectrogram formation is the magnitudes from all frames are stacked to construct the **STFT spectrogram**, a 2D time-frequency representation:

$$S(t,f) =| \sum m = 0N - 1x(m)w(m - t)e - j2\pi fm \mid S(t,f) = \left|\sum_{m=0}^{N-1} x(m)w(m - t)e^{-j2\pi fm}\right|$$

where S(t,f)S(t, f) represents the STFT spectrogram with time tt and frequency ff.

Output: STFT Spectrogram $S(t,f)S(t,f)S(t,f)$ for further processing

Figure 7. Short-Time Fourier Transform (STFT) Pipeline for Audio Signal Processing

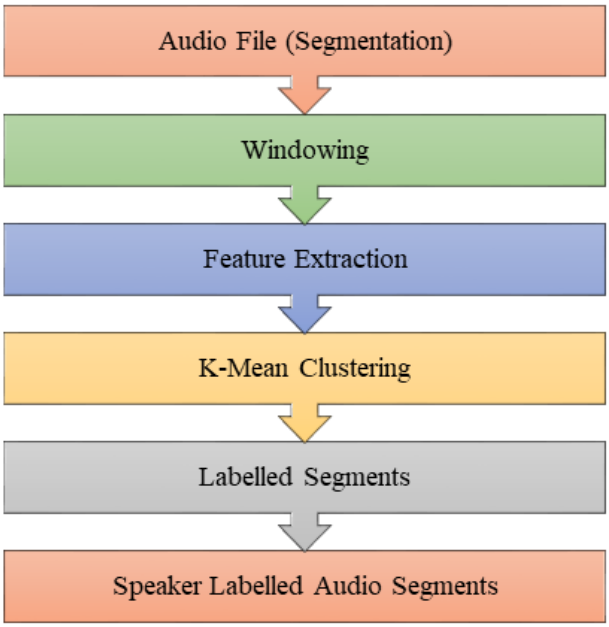

Figure 8. Steps for Speaker Labelling

Table 4. Diarization for Labeling

Parameter	Description of Value
Number of Speakers Detected	2 (Interviewer and Participant).
Accuracy (%)	89% (depending on speaker overlap).
Error Types	Misclassification of overlapping speech.
Processing Time	~2-4 seconds per minute of audio.

Table 4 presents a concise summary of the speaker diarization system implemented to distinguish between the patient and therapist in audio recordings.

2.1.2.3 Silence Removal

The objective was to identify and categorize segments of audio with speech versus silence. The audio signal was processed using small window frames with a step size of 20 ms. Windowing ensured that the audio was divided into manageable segments for efficient analysis (Table 5). For each frame, its potential for sound activity was determined by checking its energy level. If the energy was below some threshold, it was treated as silence. To distinguish between genuine speech pauses and background noise, a smoothing window of 1.5 seconds and a weight threshold of 0.2 were applied. The smoothing window enabled considering the neighbouring frames when the question of whether a given frame contains speech or not was made, thereby lowering the chances for misclassification of very short speech pauses as silence.

The threshold of weight is used to modify the sensitivity of the silence detection process, such that it removes the frame if the energy level of the frame is constantly lower than the threshold. This combination of parameters allowed cleansing the silence from the audio while preserving the natural pauses during speech. Audio pieces returned only the parts of the audio signalled as relevant speech thus improving the quality of the resulting data for analysis. It also ensured that the speech preserved was noise-free background sounds or irrelevant silences, thus allowing the model to receive the patient's voice in a more direct way for further processing in the depression detection model.

2.1.2.4 Patient Voice Summarization & Extraction

In the last step of the pipeline depicted in the figure 9, all segments detected as containing patient speech are checked according to the diarization output against the patient's label.

Interfering background noise and talkers' speech from other participants including therapists are also excluded. The resultant sound is then normalized to avoid low amplitude in further analysis and make the sound waves to have equal amplitudes. The framework for extracting patient audio is based on patient's speech isolation from a conversation. The process starts with noise elimination process thus the quality of the audio input is considered clean. Diarization is then used in the next step for speaker labeling to guarantee the segmentation of the patient's voice. Next fragments consisting of speech only are obtained after checking an identified speech segment to eliminate noise and keep only the patient's voice. Validated segments are then combined into one large waveform that is representative of the patient's voice during a session, while eliminating all background noise and other speakers. Some of the more sophisticated procedures, which are standard in the approach, are Short-Time Fourier Transform (STFT), clustering, and MFCC. STFT looks at how a frequency changes over time, groups similar patterns which could be different speakers, and MFCC emulates human auditory perception but is only concerned with critical speech frequencies. These combined methods guarantee perfect separation and analysis of auditory features, which are important in special uses such as mental disorders diagnosing and speaker recognition. The outcome is sound that is uniform, easily understood, and appropriately preprocessed for other uses in diagnostic and research applications. This method focuses on accuracy in the identification and extraction of patient voices to improve external processes, such as mental health evaluations. Table 6 outlines the effectiveness of patient-specific segment identification in recorded sessions.

3. Method for MFCC-Based Speaker Discrimination

Acoustic voice differences can discriminate between speakers, such differences include pitch, timbre, rhythm, and energy, among other acoustic properties of the voice. Exposed feature representations uniquely characterize speakers with their specific vocal imprints. Clustering algorithms try to group together segments that have similar feature vectors and, through consistent audio profiles, assign parts of a conversation to the proper speaker. By detecting shifts in patterns, features such as zero-crossing rate, pitch, and energy levels can indicate speaker changes, which the system uses to mark segment boundaries and more accurately assign continuous speech to the correct speaker.

Table 5. Analysis on Silence Removal

Parameter	Description of Value
Silence Duration Removed	~15-20% of the total recording time (depending on pauses).
Accuracy (%)	95%
Challenges	Detecting pauses that carry emotional cues.
Processing Time	~1-3 seconds per minute of audio.

Speaker Diarization (Segmentation & Clustering) involves clustering speech segments into different **speaker labels**. Each frame of the audio signal is converted into feature vectors XiX_iXi, and clustering is performed using **K-Means or Gaussian Mixture Models (GMMs)**:

$$Cj = 1 \mid Sj \mid \sum Xi \in SjXi, j = 1,2,\dots,KC_j = \frac{1}{|S_j|}\sum_{X_i \in S_j} X_i, \quad j = 1,2,\dots,KCj = \mid Sj \mid 1Xi \in Sj \sum Xi, j = 1,2,\dots,K$$

where:

- Cj $represents the centroid of cluster\ j$
- Sj $denotes$ the **set of feature vectors assigned to that cluster**.

A feature vector XiX_iXi is assigned to cluster j if it minimizes the Euclidean distance:

$$Sj = Xi: \| Xi - Cj \| 2 \leq \| Xi - Ck \| 2, \forall k \neq jS_j = \{X_i: |X_i - C_j|^2 \leq |X_i - C_k|^2, \forall k \neq j\} Sj = Xi: \| Xi - Cj \| 2 \leq \| Xi - Ck \| 2, \forall k \square = j$$

This ensures that **speech segments with similar acoustic properties are grouped together**, aiding in distinguishing speakers.

Silence Removal use **Energy-Based Voice Activity Detection (VAD)** for determine whether a frame contains speech:

$$En = \sum m = 0N - 1 \mid xn(m) \mid 2E_n = \sum_{m=0}^{N-1} |x_n(m)|^2 En = m = 0 \sum N - 1 \mid xn(m) \mid 2$$

where:

- EnE_nEn is the **short-time energy of frame nnn**,
- $xn(m)x_n(m)xn(m)$ represents **speech samples in frame nnn**.

A **threshold-based decision** is applied:

Frame is speech if $En > \theta$**Frame is speech if** $E_n > \backslash theta$ $Frame\ is\ speech\ if\ En > \theta$

Alternatively, **Zero-Crossing Rate (ZCR)** is used for detecting unvoiced frames:

$$ZCRn = 1N - 1 \sum m = 0N - 21(xn(m)xn(m+1) < 0)ZCR_n = \frac{1}{N-1}\sum_{m=0}^{N-2} 1\,(x_n(m)x_n(m+1) < 0)ZCRn = N - 11m = 0 \sum N - 21(xn(m)xn(m+1) < 0)$$

where **1** is an indicator function counting sign changes in the signal. **High ZCR** indicates unvoiced sounds or silence.

Patient Voice Summarization & Extraction use after silence removal, only voiced segments are retained. The extracted voiced frames are concatenated:

$$Y(t) = \sum i = 1MXi(t)Y(t) = \sum_{i=1}^{M} X_i(t)Y(t) = i = 1 \sum MXi(t)$$

where $Xi(t)X_i(t)Xi(t)$ are the **individual voiced segments**

A **time-domain energy thresholding** approach is applied:

$$Svoiced(t) = \{S(t), if\ En > \theta\ 0, otherwiseS\_\{"\{voiced\}\}(t) = \llbracket \{cases\}\, S(t), \&\ "\{if\}\, E\_n > \theta \backslash\ 0, \&\ "\{otherwise\}\rrbracket \{cases\} Svoiced(t) = \{S(t), 0, if\ En > \theta otherwise$$

This ensures that only speech segments above a certain energy level are retained.

Figure 9. Algorithm for Speaker Diarization, Silence Removal and Patient Voice Extraction

Table 6. Identification of Patient Segments

Parameter	**Description of Value**
Patient-Specific Segments	~50-60% of the total speaking time.
Error Rate (%)	4-8% due to overlapping voices.

MFCC-Based Speaker Discrimination is to distinguish between different speakers:

Step 1: Frame Segmentation & Windowing

Each frame is multiplied by a **Hanning window** to minimize spectral leakage:

$$xnw(m) = xn(m) \cdot w(m) x_n^w(m) = x_n(m) \cdot w(m) xnw(m) = xn(m) \cdot w(m)$$

where w(m)w(m)w(m) is the **Hanning window function**.

Step 2: Fourier Transform (FT)

The **Discrete Fourier Transform (DFT)** is computed to convert the time-domain signal into the frequency domain:

$$Xn(k) = \sum m = 0N - 1xnw(m)e - j2\pi km/NX_n(k) = \sum_{m=0}^{N-1} x_n^w(m) e^{-j2\pi km/N} Xn(k) = m$$

$$= 0 \sum N - 1xnw(m)e - j2\pi km/N$$

Step 3: Magnitude Spectrum Computation

The magnitude spectrum is extracted by computing:

$$| Xn(k) | = Re(Xn(k))2 + Im(Xn(k))2|X_n(k)| = \sqrt{Re(X_n(k))^2 + Im(X_n(k))^2} | Xn(k) |$$

$$= Re(Xn(k))2 + Im(Xn(k))2$$

Step 4: Apply Mel Filter Bank

To convert the frequency scale to a more perceptually relevant **Mel scale**, a series of triangular filters is applied:

$$Mn(i) = \sum k = fi - 1fi | Xn(k) | Hi(k) M_n(i) = \sum_{k=f_{i-1}}^{f_i} |X_n(k)| H_i(k) Mn(i) = k = fi - 1 \sum fi$$

$$| Xn(k) | Hi(k)$$

where $Hi(k) H_i(k) Hi(k)$ represents the **Mel filter weights**.

Step 5: Log Compression

A logarithmic function is applied to mimic human auditory perception:

$$Ln(i) = log\ Mn(i) L_n(i) = log\ M_n(i) Ln(i) = log Mn(i)$$

Step 6: Discrete Cosine Transform (DCT)

The **Discrete Cosine Transform (DCT)** is applied to decorrelate the features:

$$MFCC(n) = \sum i = 1KLn(i) cos[n(i - 12)\pi K] MFCC(n) = \sum_{i=1}^{K} L_n(i) cos\left[n\left(i - \frac{1}{2}\right)\frac{\pi}{K}\right] MFCC(n)$$

$$= i = 1 \sum KLn(i) cos[n(i - 21)K\pi]$$

This results in a compact representation of speech features.

Speaker Classification using Gaussian Mixture Model (GMM) is to classify speaker identities, **Gaussian Mixture Models (GMMs)** are applied to the extracted MFCC features:

$$P(X \mid \lambda) = \sum i = 1MwiN(X \mid \mu i, \Sigma i) P(X|\lambda) = \sum_{i=1}^{M} w_i \mathcal{N}(X|\mu_i, \Sigma_i) P(X \mid \lambda) = i$$

$$= 1 \sum MwiN(X \mid \mu i, \Sigma i)$$

where:

- $wiw_i wi$ are the **mixture weights**,
- $\mu i, \Sigma i \mu_i, \Sigma_i \mu i, \Sigma i$ are the **mean and covariance** of Gaussian components,
- $N(X \mid \mu i, \Sigma i) \mathcal{N}(X|\mu_i, \Sigma_i) N(X \mid \mu i, \Sigma i)$ is the **Gaussian probability density function**.

Speaker Labelling using Maximum Likelihood Estimation (MLE) is used to assign speaker labels

$$s^{=} ar\ g\ m\ a\ x\ s\ P(X \mid \lambda s) \hat{s} = arg\ max_s\ P(X|\lambda_s) s^{=} arg smax P(X \mid \lambda s)$$

where $s^{\wedge}$ is the **predicted speaker identity**.

Figure 10. Algorithm for MFCC-Based Speaker Discrimination

These features provide the diarization system with clear vision into the variability of speakers ensuring that there is proper separation and labelling of speakers from even complex, multi-speaker audio files. The process to compute MFCC is shown in Figure 8 and discussed in this section.

Mel Frequency Cepstral Coefficients: MFCCs are a widely recognized feature extraction technique in the field of speech and audio processing. They represent the short-term power spectrum of a sound, effectively capturing the phonetic and perceptual aspects of audio signals. The MFCCs are particularly valuable as it decorrelates the features and compresses the most significant information into a smaller number of coefficients.

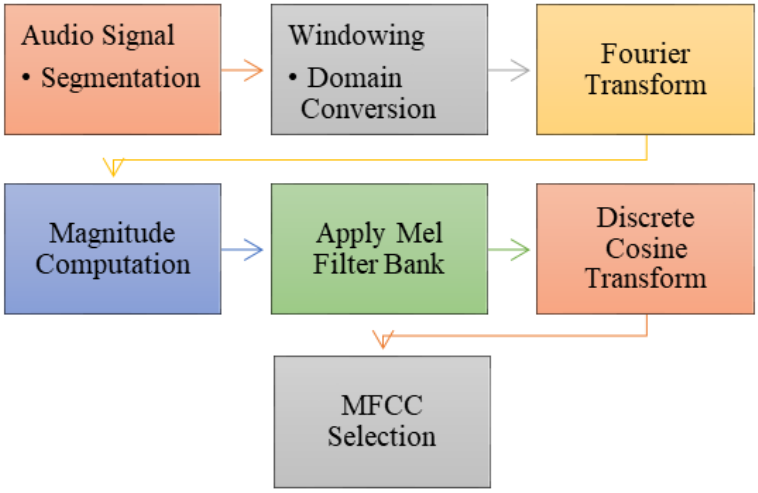

Figure 11. Steps to compute MFCC

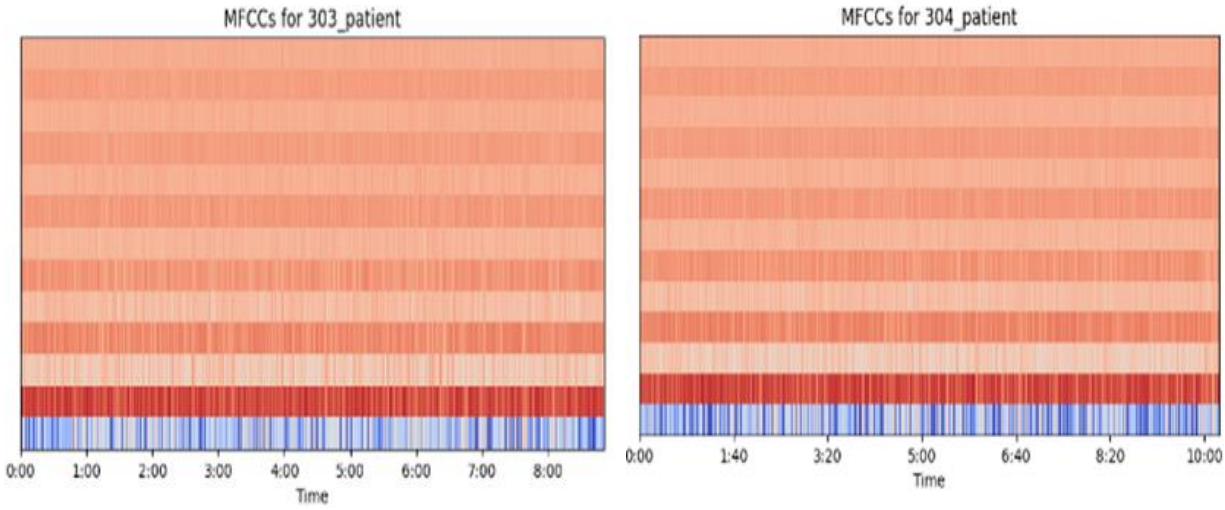

Figure 12. Sample MFCC plot

This results in fewer features, reducing the complexity and dimensionality of the input data for machine learning models. The reduced size makes MFCCs more efficient for classification tasks, especially when data or computational resources are limited.

MFCCs are computed by applying a Discrete Cosine Transform (DCT) to the log-scaled Mel energies. Figure 10 illustrates the steps to compute MFCCs and required number of coefficients can be chosen and extracted as needed. MFCC_0 is often considered the average log energy of the audio frame. Most of the informative coefficients are from MFCC_1 - MFCC_12 as they extract the fundamental spectral properties connected with the timbre and major formants of speech. The coefficients from MFCC_13 to MFCC_39 capture finer features of the audio spectrum like higher frequency, which contribute to minute differences or background noise.

Figure 11 shows the step to compute MFCC and figure 12 shows the plot for first 40 MFCCs for the extracted patient audio where, X- axis represents the time as the audio proceeds, and the Y-axis contains the first 40 MFCCs starting from 0 at the bottom to 39 at the top.

4. PSAAP-EfficientNet Framework for Depression Detection from Clinical Audio Data

The flowchart in figure 13, illustrates a machine learning pipeline for depression detection using audio data. It begins with the DAIC-WOZ dataset, which contains clinically labeled audio recordings from interviews. These recordings are processed through the PSAAP Framework, which cleans and refines the audio by removing noise and isolating patient speech. The resulting Extracted Patient Audio is converted into Mel-spectrograms and fed into the EfficientNet Model, a deep learning architecture optimized for image classification. During Training, the model learns to associate audio features with depression labels.

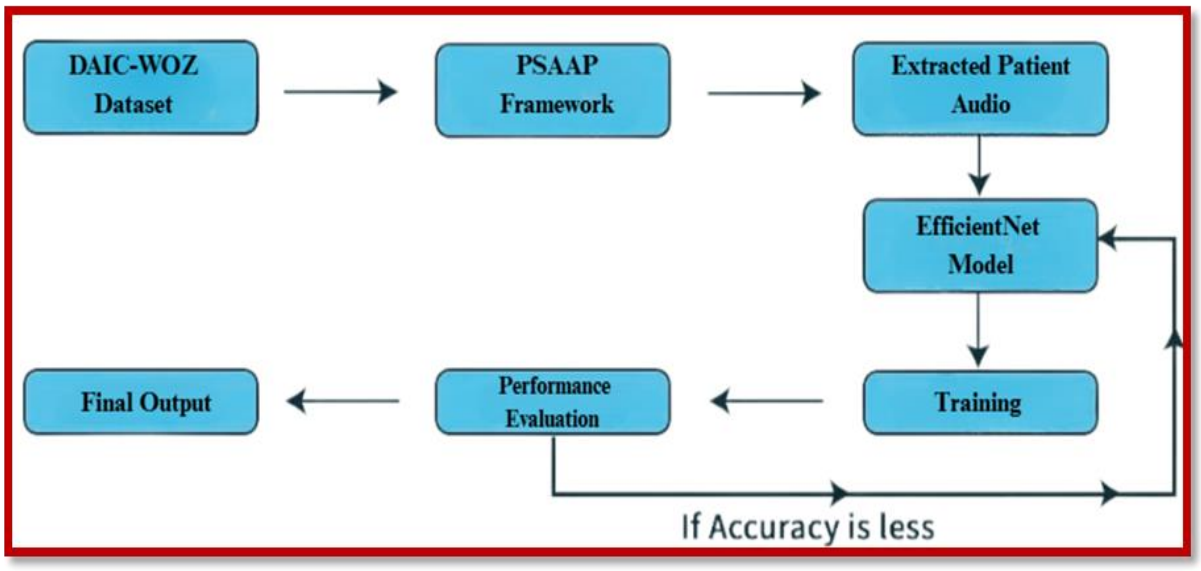

Figure 13. Sample MFCC plot

The model's effectiveness is assessed in the Performance Evaluation stage using metrics like accuracy and precision. If the results are unsatisfactory, the system loops back to retrain and refine the model. Once performance is acceptable, the Final Output is generated, providing a depression classification label and confidence score. This iterative and structured approach ensures reliable and adaptive detection of depression from speech data.

5. Results & Discussion

In this framework of voice extraction, the process involved several pre-processing steps, including background noise removal, diarization and silence trimming, which allowed for accurate segmentation and extraction of the patient's voice of raw audio files. Quantitative results from the extraction process indicate that the patient's speech was effectively isolated, with minimal distortion or loss of important vocal features.

The extracted audio files, preserving key voice characteristics, can be further processed for applications such as depression detection. These preserved voice features can serve as a basis for analysis, providing valuable input for downstream tasks like emotion classification, sentiment analysis, or behavioral health assessment. Figure 14 displays plots of the original, unprocessed audio files from the dataset. These waveform plots provide a visual representation of the variations in amplitude within the audio signals over time, offering insight into the structural and acoustic characteristics of the recordings. Here, amplitude-essentially the strength or intensity of an audio signal-is plotted against time, capturing fluctuations that reflect changes in sound levels throughout each recording. Large, consistent peaks in the waveform indicate sections with more energy, which correspond to louder or more emphasized portions, such as segments of spoken dialogue. On the other hand, low-amplitude regions of the period comprise smaller peaks or flat regions in the waveform that indicate breaks or silence or very quiet background noise. They are especially supportive regions in pointing out natural breaks in speech and quiet sections of the audio recording for greater appreciation of the rhythmic flow of the recording. For tasks such as speaker diarization and segmentation, a silence or near-silence intervals are needed because they help discriminate between various audio segments and changes in the recording.

Figure 14(b) shows signal plots for the original audio after applying the noise reduction.These plots illustrate the processed audio waveform, highlighting the impact of noise reduction on the overall clarity and quality of the signal. In these plots, the amplitude of the audio signal represents the intensity of sound at each point in time, allowing for an assessment of the effectiveness of the noise reduction process. By attenuating unwanted background noise, the signal plots reveal a smoother and more distinct waveform, with reduced fluctuations that previously corresponded to interference. Figure 14(c) presents an overlap plot illustrating the audio signal for each file before and after noise reduction. In this visualization, the grey line represents the original audio signal, containing both the speaker's voice and various background noise elements, while the blue line represents the processed audio signal following noise reduction.

The plot indicates that there is constant background noise in the original audio. This was causing sustained interference throughout all the recordings. Soft masking was applied to reduce the noise from the audio. Soft masking selectively attenuates frequencies associated with noise while preserving the spectral characteristics of the speaker's voice, thereby not risking a distortion of audio signals. This approach, in particular, preserves the quality of the speaker's voice signal because it dampens less noise without sacrificing some features necessary for the voice.

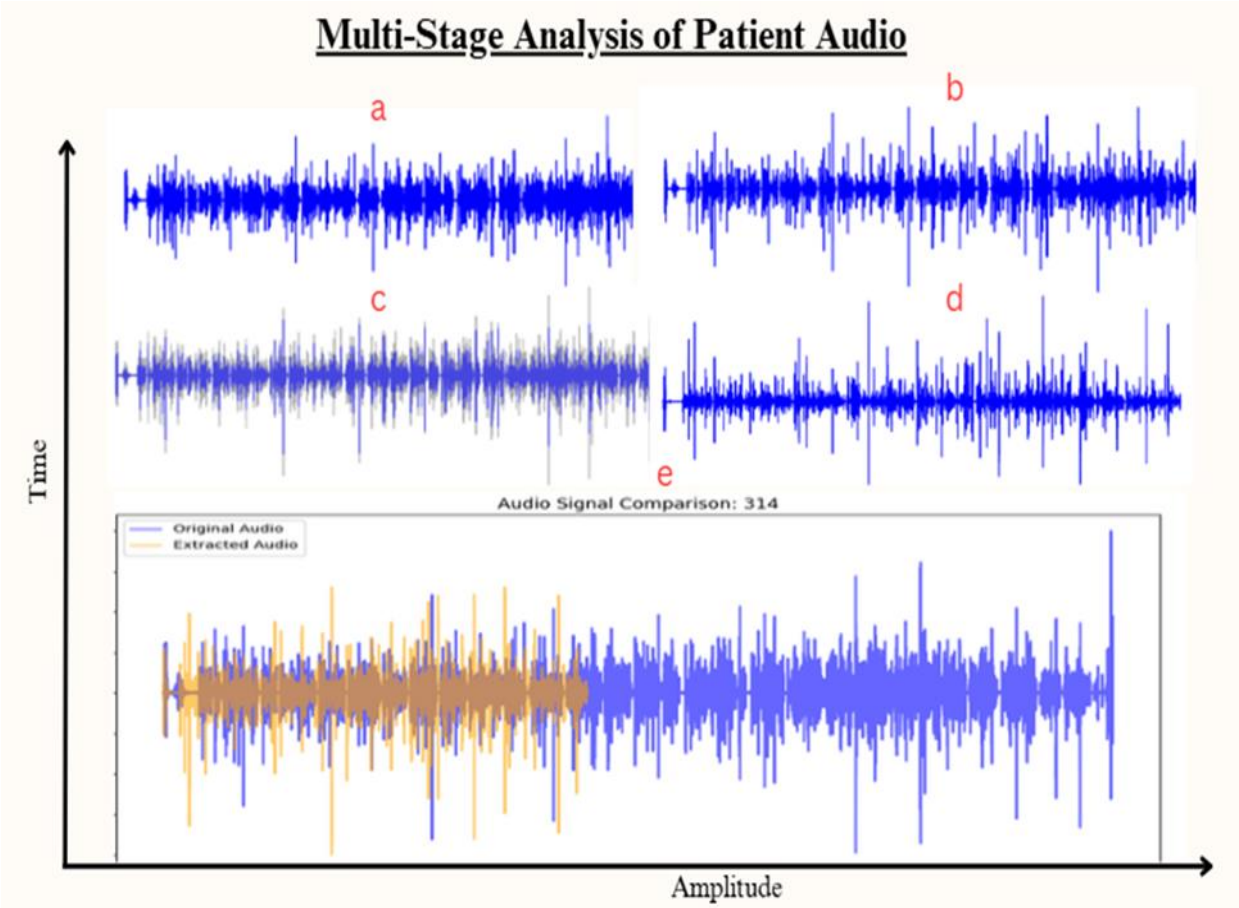

Figure 14 (a) Plot of the original Signal **(b)** Plot of the Background Noise Removed Audio **(c)** Overlap Plot for comparison **(d)** Plot of Patient Extracted Audio **(e)** Overlap of both the plots original vs extracted for Comparison

The overlap plot allows for a visual comparison showing that the blue line, noise-reduced audio, closely tracks the grey line, original audio, but with smoother and more well-defined transitions. This indicates that the process was highly successful in terms of reducing background noises such that the signal from the audio coming from the mic would have come out cleaner-meaning the speaker's voice was clearer. This alignment between the original and noise-reduced signals clearly points out that the implemented soft masking successfully masked the unwanted noise components and preserved the targeted audio content and, therefore, validates the effectiveness of the proposed technique to enhance audio quality for analysis purposes. Figure 14(d) shows signal plots of the extracted audio which contains only the patient's voice.

Figure 14(e) presents an overlap plot comparing each audio file with its corresponding patient-extracted audio segment. In this overlap plot, the amplitude of the audio signal, reflecting a measure of sound intensity at every sample point is used. The blue line shows the original, raw audio record with both voices of the therapist and the patient, while the yellow shows the audio part isolated in order to contain only the voice of the patient. The sample index in the plot is represented by the X-axis, allowing point-to-point comparison between the two signals. This overlap is drawn along sample indices and not along a time axis since the lengths of the original and extracted audio files are not the same. The actual patient audio extracted is therefore shorter than the original recording, and so the two signals are partially aligned. In the initial sections of the plot, the overlap between the original and extracted audio is visually apparent, with clear fluctuations in amplitude corresponding to patient speech. However, as the plot progresses, the signals diverge: the yellow line (extracted audio) fades, and the blue line (original audio) continues, reflecting segments where silences and the therapist's voice have been removed from the extracted signal. This divergence in the latter part of the plot highlights the selective removal of non-patient audio components, emphasizing that the extraction process effectively isolated the patient's speech, while filtering out non-target elements.

Diarization Error Rate (DER) is the quantification of the accuracy of speaker diarization as the ratio of incorrectly labeled time. It is computed to the satisfaction in equation 1:

$$DER = Missed\ speech + False\ Alarm + Speaker\ Error\ / \quad (1)$$
$$Total\ Reference\ Speech\ Duration$$

Where:

Missed Speech: Speech present in the reference but not detected.

False Alarm: Non-speech or incorrect speech detected as speech.

Speaker Error: Speech attributed to the wrong speaker.

Minor timing differences are taken into account by a collar tolerance (usually 0.25 seconds). DER is given as a percentage and a common measure in the evaluation of systems of diarizing data and particularly of datasets such as DAIC-WOZ.

In figure 15, the Diarization Error Breakdown Chart as whoin displays a graphical visualization of errors that are engaged in speaker diarization of the provided dataset. It categorizes the errors in three classes as mistakes in missed speech, false alarms and speaker misclassification. The Bar duration (seconds) denotes the number of error each of the types are ready to give and helps in determining the areas of the diarization process that must be addressed. Speaker error, as well as the missed speech exemplified in the passage, prove to be the most important, and hence, issues with the adequate modal guidance of speech blocks and the determination of all the speech emerge. These include interlaced speech and stuttering of speakers. This graph also warrants certain breakthroughs that need to be made in PSAAP pipeline and guarantees unavoidable optimization of cycles.

As given in equation in 1, SNR (dB) = 10 × $\log_{10}$ (signal power ÷ noise power), quantifying audio clarity by comparing useful signal to background noise.

$$SNR\ (dB) = 10 \cdot log10(Power\ of\ NoisePower\ of\ Signal) \quad (2)$$

As figure 16 illustrates, the graph of SNR distribution renders the signal-to-noise ratio in the range of 187 audio files that obtained the results of post-processing. It is presented as a histogram with the SNR values with the mean of 23.95 dB indicated with a red dashed line. The dispersion of values expresses the usual deviation of 2.35 dB, and 95 relationships of the confidence period stretch between 23.61 and 24.29 dB. This plot is useful to evaluate consistency and reliability of the audio developing process. The narrow dispersion around the mean is indicative of consistent similar gains within different samples, which makes the statement of +16 dB increase over raw recordings true and proves the PSAAP pipeline efficiency.

The figure 17 analysis by stage as set forth in Table 7 that compares the Signal-to-Noise Ratio (SNR) values pre- application and post- application of the Patient-Specific Audio Extraction Pipeline (PSAAP) method with the subsequent SNR improvement. When using a PSAAP process, conceived to improve the quality of audio transmitted in a situation of noise, the SNR increases relatively by 16 dB. In the table, the pre-processing SNR is approximately 8 dB, post-processing SNR is 24 dB, and the huge improvement rate obtained by using the PSAAP approach is mentioned. The above illustrates the efficacy of the suggested approach in enhancing the audio signal considerably, and explains its prospectivity in the audio applications that require individual patient attention.

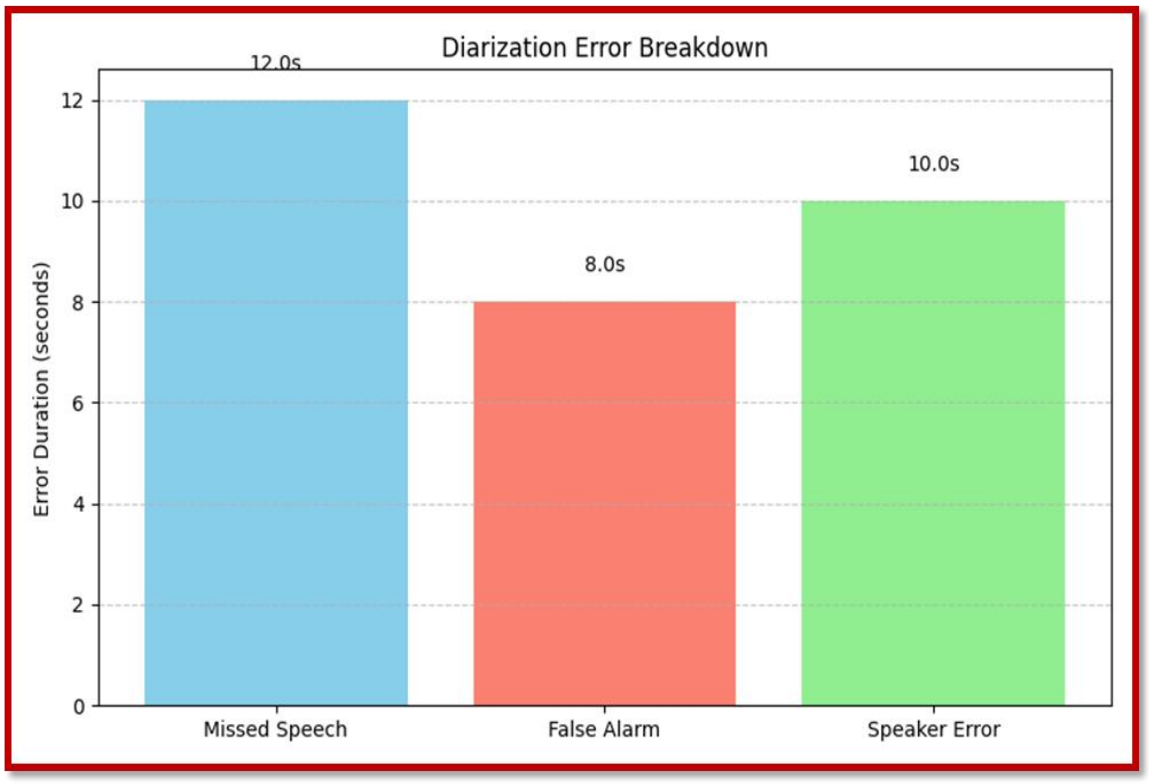

Figure 15. Diarization Error Breakdown Chart

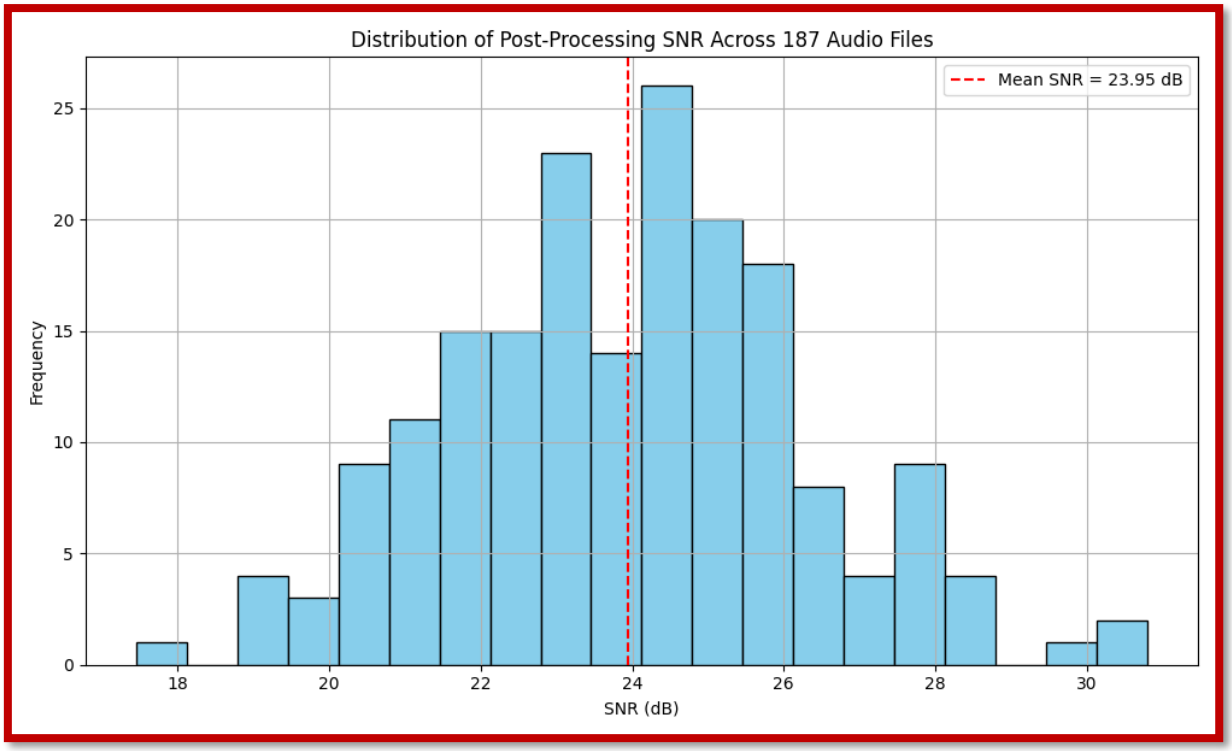

Figure 16. SNR Distribution Graph

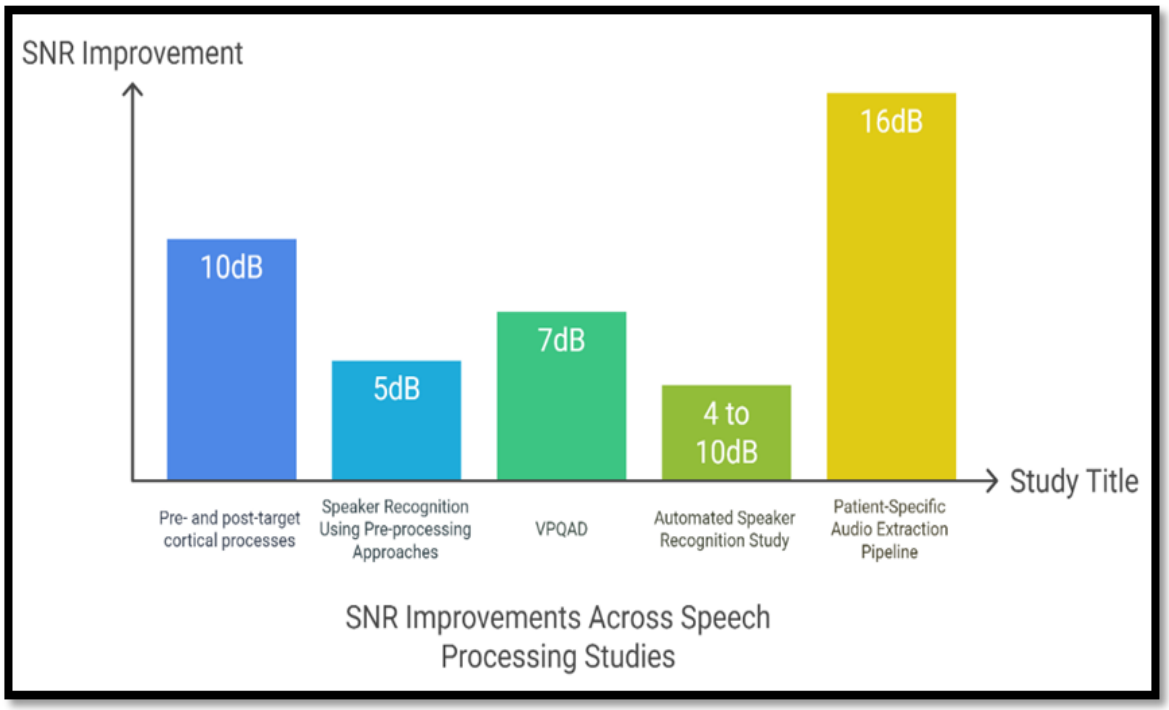

Figure 17. Result of SNR Comparison of Proposed Model (PSAAP) vs others

As shown by the bar chart in figure 16, post-processing SNR is best among the works and the improvement is greatest in PSAAP. The visual difference between bars brings out the differences in performance in the studies which helps in indicating the most useful methods. The flow chart entitled, PSAAP-Based Depression Detection Using DAIC-WOZ and EfficientNet provides an adaptive and well-organized structure of machine learning to detect depression. It combines PSAAP (Processed Speech Audio Analysis Pipeline) with EfficientNet deep learning model in order to process audio data of clinical interviews. It starts with downloading the DAIC-WOZ dataset, one of the established clinical validation of affective computing benchmark datasets. It is a multimodal dataset (audio, video, and text captured as part of a structured interview between participants and a virtual interviewer). The detection system input is in the form of audio records named in terms of standardized depression scales, such as PHQ-8 or PHQ-9.

Table 7. Comparison of Signal-to-Noise Ratio (SNR)

Study	Dataset	Pre-Processing SNR (dB)	Post-Processing SNR (dB)	SNR Improvement (dB)	Sample Size	Results Comparison
Pre- and post-target cortical processes predict speech-in-noise performance [21]	Custom EEG-based speech-in-noise task	-3 dB, +3 dB	4 to 6 dB	~8 to 10 dB	26 participants	Moderate improvement, focuses on cortical response rather than pure SNR enhancement.
Speaker Recognition Using Pre-processing Approaches [22]	Synthetic datasets for speaker recognition	~15 dB	~20 dB	~5 dB	~1,000 samples	Limited SNR improvement; demonstrates basic enhancements for recognition systems.
VPQAD [23]	VPQAD: Voice Pre-Processing and Quality Assessment Dataset	15.35 to 45.22 dB	22.57 dB (mean post-SNR)	~7 dB improvement	50 participants	High diversity in real-world scenarios but lacks multilingual focus.
Automated speaker recognition in real world conditions: controlling the uncontrollable [24]	FBI Voice Database	~12 to 50	~16	~4 to 10	186 test files	Moderate improvements focusing on real-world applications; slightly lower improvement.
Patient-Specific Audio Extraction Pipeline (PSAAP)-Proposed Method	DAIC-WOZ	~8 dB	~24 dB	~16 dB	187 files	Best method among all: Achieved highest SNR improvement (+16 dB) with scalable methodology.

Table 8. EfficientNet Model Performance Metrics

Metric	Score
Accuracy	0.87
Precision	0.85
Recall	0.88
F1-Score	0.86

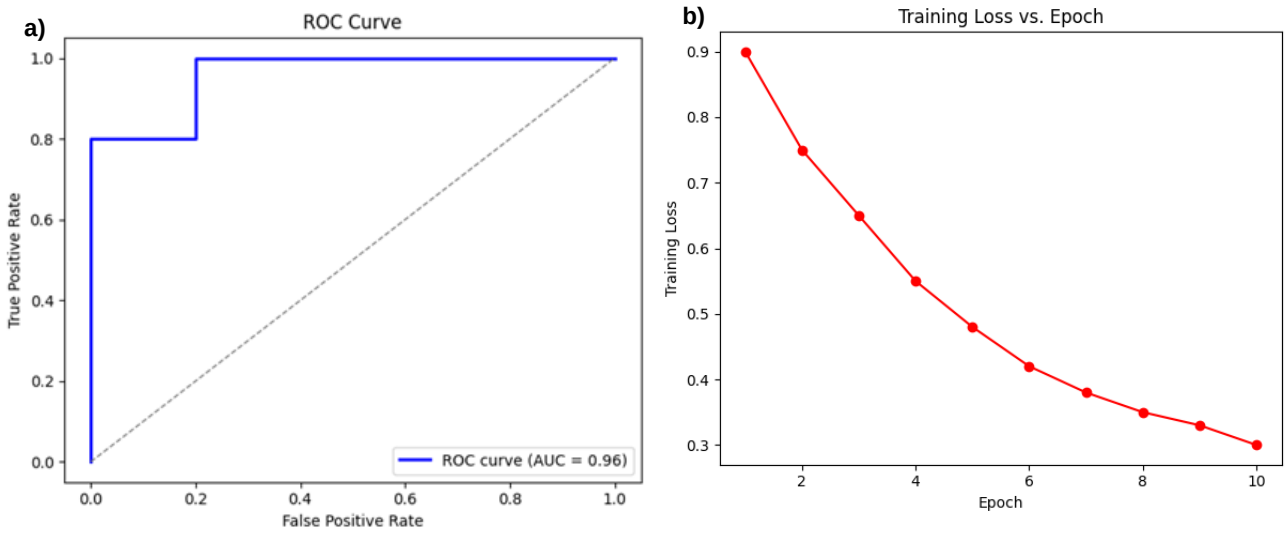

Figure 18. a) Training Loss vs. Epoch, **b)** ROC Curve

The quality of the annotations provided in the dataset and its diversity of entries allows it to be a suitable choice when training the models in mental diagnoses.

The point where audio data is received is followed by the PSAAP framework processing the data through several steps of important preprocessing. These are speaker diarization to capture patients only, noise removal and silence removal to remove irrelevant audio and segmentation and standardization to normalize the data. The last step consists of transforming the cleaned audio into Mel-spectrograms (or two-dimensional time-frequency representations, made to be desirable to convolutional neural networks (CNNs). Such transformation improves the understandability of the information contained in it and conditions it on deep learning analysis. The product of PSAAP is an extracted list of patient audio files, devoid of the speech and background noise of the interviewer. Vocal biomarkers of depression, like monotony, lack of energy, and delayed responses, remain in these segments and these markers can be used in ensuring right classification. After this cleaning up of the data, said data is put through a state of art CNN named EfficientNet that is characterized by the aforementioned compound scaling approach balancing on resolution, width, and depth. EfficientNet is either trained (or fine-tuned with pretrained versions (e.g. EfficientNet-B0 to B7), to classify the Mel-spectrograms as either depressed or non-depressed (table 8). The model is trained to know how to connect patterns in the spectra to labels as depression during the training phase. To be generalized, the dataset will be divided into subsets (for training, validation and testing). The most important operations are optimizing the loss function (e.g., cross-entropy), backprop, gradient descent, and hyperparameter optimization. The prevention of overfitting becomes possible with such techniques as dropout, early stopping, and data augmentation. To track the learning process of the model, a training loss vs. epoch graph is used, which demonstrates a steady decrease in the loss on 10 epochs, pointing to the successful learning of the model and convergence.At the end of the training process, the model is performed using common classification metrics as some results.

The combination of all of the measures is available in a performance table and can be visualized in a bar plot which indicates high predictive power of the model. Further, the ROC curve is plotted to show a trade-off between the value of the true positive rates and the values of the false positive rates at different thresholds. The curve is closer to the truth considering the area under the curve (AUC) is higher which means that this model is efficient in detecting the non-depressed and depressed. Should the model obtain the necessary levels of performance, the result is recorded after the final output that has a predicted class label and confidence measure. This can also be further applied within research in a way where it can be interpreted and also within practice like the use of telemedicine to screen mental health remotely. However, should performance levels not be adequate, the framework proposes that there exists a closed loop iteration that is dubbed, in case Accuracy is less. Such a loop will allow res stepping back to the EfficientNet model and training steps to further improve the performances. The potential issues, such as absence of the data, over fitting, poor preprocessing, or feature representation are solved by tentative modification of the parameters, alteration of the preprocessing, or the model architecture modification.

The effectiveness of the Algorithm correct classification is assessed by the true positive rate (sensitivity) as the false positive rates rise at the different classification thresholds, and is also referred to as the ROC (Receiver Operating Characteristic) curve as shown in figure 18a. A perfect model would exhibit a curve that would follow the top-left corner position and a random model would follow the diagonal. The

performance is quantified as area under the curve (AUC) [2, 12] the closer this measure is to 1.0 the better the discrimination. As the ROC curve in the case portrays, there exists a large degree of predictive power of the EfficientNet model because the model can distinguish between persons with depression and non-depression based on the audio feature option selected during the clinical interviews. As may be seen in the training loss vs. epoch graph shown in figure 18b, the model error decreases going forward in time during learning based on the input data. One entire pass through the training data comprises one epoch. A constantly decreasing curve that is inclusive of an incurring loss that goes on to decelerate steadily as in this graph would indicate that the model is doing a good job and is learning and converging towards an excellent answer. Such a graph will enable monitoring the training process and revealing such issues as overfitting (when the loss reaches stagnation or begins to increase). In this case, the model steadily performs better and therefore, this can be deemed to show that the training is stable and the model will become more precise in each iteration.

5. Conclusion

In the proposed research, a signal processing based approach to isolate the audio signals of a particular patient occurring in a therapist-patient talk channel is presented and this is an essential step towards developing mental health diagnostics embedded into the raw audio signal recording made by the patient. In comparison with the existing methods, which only act on AI, the described framework is clear and would be interpreted productively, which covers the gap of efficiently represented non-AI approaches to complex conversation cases. Based on pitch, intensity and STFT-MFCC acoustic parameters, method retains the non-verbal information that is not appropriate in transcriptional based analysis but highly significant in clinics. The other involved technique is the step-wise audio improvement that is done in the steps of removal of background clatter to upper leg the audio, diarization of the speakers so that the efficiency of the descent is augmented, and the elimination of silence, which preserves the psychologically significant pauses. This whole process results in the last step of a high quality, consolidated patient audio file that is optimized down streams. The resulting clean audio is then run through the pre-trained model built with efficientnet that has already shown a decent performance in its classification-87% accuracy, 85 precision, 88 recall and 86 F1-score. These measurement values confirm the effectiveness of the model to recognize depression using Mel-spectrograms with the use of patient speech. The strength of the model is further proved by the ROC curve that has a high AUC and a steadily plunging loss during training situation.Not only has this integrated system maintained the integrity of speech and its diagnostic value but it has better prepared it to be used in machine learning. It puts the material at the disposal of traditional audio processing and the modern area of mental health diagnostics. A scalable, interpretable and summary solution to clinically applicable data. It is possible to extend this framework in the sense of tonal variations, accents and in real time processing and so help to further generalizability of this framework across diverse settings and to other populations.

References

[1] Z. Wang, L. Chen, L. Wang, G. Diao, Recognition of audio depression based on convolutional neural network and generative antagonism network model, IEEE Access,8, (2020) 101181–101191. https://doi.org/10.1109/ACCESS.2020.2998532

[2] World Health Organization. (2017). Depression and other common mental disorders: global health estimates. In Depression and other common mental disorders: global health estimates. https://www.who.int/publications/i/item/depression-global-health-estimates

[3] H. Solieman, E.A. Pustozerov, The detection of depression using multimodal models based on text and voice quality features, In 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), IEEE, St. Petersburg, Moscow, Russia, (2021) 1843–1848. https://doi.org/10.1109/ElConRus51938.2021.9396540

[4] X. Li, X. Zhang, J. Zhu, W. Mao, S. Sun, Z. Wang, C. Xia, B. Hu, Depression recognition using machine-learning methods with different feature generation strategies. Artificial intelligence in medicine, 99, (2019) 101696. https://doi.org/10.1016/j.artmed.2019.07.004

[5] V. Ravi, J. Wang, J. Flint, A. Alwan, Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement, Computer Speech & Language, 86, (2024) 101605. https://doi.org/10.1016/j.csl.2023.101605

[6] S. Tokuno, Stress evaluation by voice: from prevention to treatment in mental health care. Econophysics, Sociophysics, and Multidisciplinary Sciences Journal, 5(1), (2015) 30-35.

[7] N.H. Chetty, M. Shah, S. Kabaria, S. Verma, (2017) A Survey on brain tumor extraction approach from mri images using image processing, In 2017 2nd International Conference for Convergence in Technology (I2CT), IEEE, Mumbai, India, https://doi.org/10.1109/I2CT.2017.8226187

[8] Q. Zhao, H.Z. Fan, Y.L. Li, L. Liu, Y.X. Wu, Y.L. Zhao, Z.X. Tian, Z.R. Wang, Y.L. Tan, S.P. Tan, Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. Frontiers in Psychiatry, 13, (2022) 815678. https://doi.org/10.3389/fpsyt.2022.815678

[9] L. Fürer, N. Schenk, V. Roth, M. Steppan, K. Schmeck, R. Zimmermann, Supervised speaker diarization using random forests: a tool for psychotherapy process research. Frontiers in psychology, 11, (2020) 1726. https://doi.org/10.3389/fpsyg.2020.01726

[10] X. Huang, F. Wang, Y. Gao, Y. Liao, W. Zhang, L. Zhang, Z. Xu, Depression recognition using voice-based pre-training model. Scientific reports, 14(1), (2024) 12734. https://doi.org/10.1038/s41598-024-63556-0

[11] D.M. Low, K.H. Bentley, S.S. Ghosh, Automated assessment of psychiatric disorders using speech: A systematic review, Laryngoscope Investigative Otolaryngology, 5(1), (2020) 96–116. https://doi.org/10.1002/lio2.354

[12] J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D.R. Traum, The distress analysis interview corpus of human and computer interviews. In LREC, 14, (2014) 3123-3128.

[13] Zheng-Hua Tan, Achintya kr. Sarkar, Najim Dehak, rVAD: An unsupervised segment-based robust voice activity detection method," Computer Speech & Language, 59 (2020) 1-21. https://doi.org/10.1016/j.csl.2019.06.005

[14] D. S. Williamson, J. Barker, and S. Watanabe, Speech separation and recognition in multi-speaker environments, IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, (2008) 1644-1648. https://doi.org/10.1109/IJCNN.2008.4634018

[15] J. Ao, M.S. Yldrm, R. Tao, M. Ge, S. Wang, Y. Qian, H. Li, USED: universal speaker extraction and diarization. IEEE Transactions on Audio Speech and Language Processing, 33, (2025) 96–110. https://doi.org/10.1109/taslp.2024.3511268

[16] M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnykova, Multimodal data fusion for depression detection approach, Computation, 13, (2025). 1-9. https://doi.org/10.3390/computation13010009

[17] A. Sharma, S. Panda and S. Verma, Sign Language to Speech Translation,2020 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt49239.2020.9225422

[18] E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, Toward explainable AI (XAI) for mental health detection based on language behavior. Frontiers in Psychiatry, 14, (2023) 1219479. https://doi.org/10.3389/fpsyt.2023.1219479

[19] J O'Sullivan, G. Bogaarts, P. Schoenenberger, J. Tillmann, D. Slater, N. Mesgarani, E. Eule, T. Kilchenmann, L. Murtagh, J. Hipp, M. Lindemann, F. Lipsmeier, W. Cheng, D. Nobbs, C. Chatham, Automatic speaker diarization for natural conversation analysis in autism clinical trials. Scientific Reports, 13(1) (2023) 10270. https://doi.org/10.1038/s41598-023-36701-4

[20] W.K.Lu and Q. Zhang, Deconvolutive short-time Fourier transform spectrogram. IEEE Signal Processing Letters, 16(7), (2009) 576–579. https://doi.org/10.1109/lsp.2009.2020887

[21] J. Kim, S. Lee, and C. Davis, Pre- and post-target cortical processes predict speech-in-noise performance, *NeuroImage*, 227 (2021) 117699. https://doi.org/10.1016/j.neuroimage.2020.117699

[22] A. El-Moneim, S. Hassan, and A. Youssef, Speaker recognition using pre-processing approaches for robust identification, *International Journal of Speech Technology,* 23 (2020) 451–462. https://doi.org/10.1007/s10772-019-09659-w

[23] M. Ahmed, R. Khan, and X. Li, VPQAD: A voice pre-processing and quality assessment dataset for robust speech applications, *IEEE Data Descriptions*, 32 (2024) 45–59. https://doi.org/10.1109/IEEEDATA.2024.3493798

[24] H. Nakasone, J. Zhou, and J. K. Hansen, Automated speaker recognition in real-world conditions: Controlling the uncontrollable, in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Geneva, Switzerland, 2003, pp. 1405–1408. https://doi.org/10.21437/Eurospeech.2003-299

Authors Contribution Statement

Both the authors equally contributed and approved the final version of the work.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License