DOI 10.15622/ia.23.6.7

J. JACOB, K.S. KANNAN

ENHANCED MACHINE LEARNING FRAMEWORK FOR AUTONOMOUS DEPRESSION DETECTION USING MODWAVE CEPSTRAL FUSION AND STOCHASTIC EMBEDDING

Jacob J., Kannan K.S. **Enhanced Machine Learning Framework for Autonomous Depression Detection Using Modwave Cepstral Fusion and Stochastic Embedding.**

Abstract. Depression is a prevalent mental illness that requires autonomous detection systems due to its complexity. Existing machine learning techniques face challenges such as background noise sensitivity, slow adaptation speed, and imbalanced data. To address these limitations, this study proposes a novel ModWave Cepstral Fusion and Stochastic Embedding Framework for depression prediction. Then, the Gain Modulated Wavelet Technique removes background noise and normalises audio signals. Difficulties with generalisation, which results in a lack of interpretability, hinder extracting relevant characteristics from speech. To address these issues, an Auto Cepstral Fusion extracts relevant features from speech, capturing temporal and spectral characteristics caused by background voice. Feature selection becomes imperative when choosing relevant features for classification. Selecting irrelevant features can result in overfitting, the curse of dimensionality, and less robustness to noise. Hence, the Principal Stochastic Embedding technique handles high-dimensional data, minimising noise influence and dimensionality. Furthermore, the XGBoost classifier differentiates between depressed and non-depressed individuals. As a result, the proposed method uses the DAIC-WOZ dataset from USC for detecting depressions, achieving an accuracy of 97.02%, precision of 97.02%, recall of 97.02%, F1-score of 97.02%, RMSE of 2.00, and MAE of 0.9, making it a promising tool for autonomous depression detection.

Keywords: depression detection, machine learning, ModWave Cepstral Fusion, background noise, XGBoost classifier, DAIC-WOZ dataset, autonomous detection system, accuracy.

1. Introduction. Depression is the most significant reason for non-fatal health loss. In 2017, there were 322 million individuals worldwide who suffered from depression, according to the World Health Organization [1]. Depression can cause self-harm in addition to having a severe negative influence on one's family, career, and educational performance. Depression in adolescence is linked to mood disorders and severe mental illness in later life [2]. While depression is most commonly identified in those in their 30s and 40s, it can also be seen in older adults, youngsters, and those under stress in their relationships and academics [3], also frequently producing results comparable to those of major depression, minor depression, however less severe, adds significantly to the economic and social burden [4, 5]. As the most common mental illness in the world, MDD affects approximately 300 million individuals and is associated with significant financial burden and impairment [6, 7]. In Brazil, depression is the sixth most common health problem, with a lifetime prevalence of up to 16.8% and a prevalence rate of 5.8%. Depression symptoms include low mood, irritation, anhedonia, fatigue, psychomotor slowness, cognitive impairment, and disturbances in internal systems. Early identification of depression symptoms, such as modified speech patterns, may facilitate immediate action and help avert the development of suicidal thoughts and poor social function [8].

A vital component of healthcare is mental health assessment, which enables early intervention and individualised therapy for people in psychological distress. Clinical examinations and subjective self-reporting are the foundations of conventional evaluation approaches, and they can be costly, time-consuming, or biased. However, new technological developments, especially in machine learning (ML), have opened the door for creative methods of mental health evaluation [9]. Using voice recordings in conjunction with ML techniques is one such potential method for identifying and tracking mental health issues. Tone, pitch, rhythm, articulation, and other aspects of voice carry much information that can reveal underlying emotional states and cognitive processes. Studies have demonstrated that people suffering from mental health conditions such as sadness, anxiety, and schizophrenia display unique patterns in their speech characteristics [10].

By analysing these fine-grained audio characteristics from speech recordings, ML techniques may detect patterns linked to particular mental health issues. These algorithms can acquire the ability to discriminate between normal and abnormal speech patterns with a high degree of accuracy by training on many annotated voice samples [11]. Furthermore, as time passes, ML algorithms can adjust and improve, constantly enhancing their prediction power. There are various benefits to combining ML and voice recording in mental health assessments. It offers a scalable, affordable, and non-invasive way to test people who could be at risk of mental health issues [12]. It also makes it possible to continuously track how well patients are doing and how they are responding to treatment, which makes tailored treatments easier to implement and enhances clinical results.

Common symptoms of depression include depressed emotions, loss of interest, mental slowness, and other symptoms. It is challenging to diagnose and has a protracted therapy cycle, a high incidence rate, and a sluggish onset [13, 14]. Psychotherapy and medication therapy are the primary forms of treatment. However, the diagnosis of depression has several deficiencies. First of all, depression is a prevalent mental illness, but many individuals avoid getting active therapy because they are embarrassed to admit they have it. Second, the widely utilised instrument for subjectively diagnosing depression is the Diagnostic and Statistical Manual of Mental

Disorders (DSM-5) [15, 16]. Misdiagnoses and missed diagnoses result from this. Thirdly, patients frequently lack the expertise required for self-assessment, and large-scale, low-cost depression screening instruments are absent [17]. As an outcome, many patients are unaware of their condition, which restricts their options for treatment. Finding an objective technique for quick screening and early warning for depression is therefore essential.

Several studies have shown an association between depression and an individual's behaviour. Research has shown that voice recordings of people can be a beneficial tool for characterising mental health and can offer significant insights into people's mental health. Furthermore, it has been proposed that precise results for depression prediction can be achieved by refining a combination of features. Since Mel-Frequency Cepstral Coefficients (MFCC) are trustworthy and effective even at low dimensionalities, they are the most often employed feature for audio signal processing. Several studies have demonstrated the effectiveness of an algorithm developed via ML techniques in detecting depression in voice samples. This study offers the following:

– With an autonomous system for detecting depression, the proposed ModWave Cepstral Fusion and Stochastic Embedding Framework addresses the complexity of conventional clinical diagnosis methods and offers an effective solution to the increasing number of depression cases worldwide.

– The proposed strategy, which introduces the Gain Modulated Wavelet Method, improves the quality of pre-processed data by efficiently removing background noise from audio recordings, normalising amplitude levels, and capturing both low and high-frequency information.

– The Auto Cepstral Fusion Feature extraction method is utilised to extract relevant features, thereby minimising the impact of noise, improving the robustness of the model, and capturing the temporal and spectral characteristics essential for depression prediction.

– Moreover, the Principal Stochastic Embedding method reduces dimensionality, minimises noise impact, and manages high-dimensional data, enhancing feature selection and classification precision.

– In addition to effectively predicting depression, the proposed approach thoroughly studies performance parameters like accuracy, precision, ROC, F1 score, recall, and sensitivity, offering valuable data for model comparison and evaluation.

The article will discuss recent studies on depression detection in machine learning, proposed methods and their explanations, the outcomes of the proposed work, and future directions with references.

2. Related Structure. Distinctive speech patterns such as lower articulation rate, pauses, slower speaking, lesser intensity, and unusual voice quality can be used to diagnose depression. Pitch, intensity, rhythm, speed, jitter, shimmer, energy distribution, and cepstral characteristics are examples of speech's prosodic, phonetic, and spectral aspects that must be taken into account to identify fluctuations in emotional state. Because jitter is sensitive to abrupt changes in speech, it is essential for identifying mood states. Cepstral coefficients – in particular, MFCC – have been well-researched for vocal analysis-based depression diagnoses that are well-suited for identifying depression speech.

The study by the authors in study [18] aimed to develop an ML tool for diagnosing depressive disorders. They used vocal feature extraction algorithms and ML classification techniques such as MLP, polynomial kernel SVM, polynomial kernel, normalised SVM PUK kernel, and random forest (RF) to extract vocal acoustic features from recordings. The results showed the tool's viability for cost-effective and non-invasive recognition and screening of MDDs, demonstrating its potential in diagnosing and screening these disorders. However, this technique lacks interpretability and also contains inconsistencies in voice, leading to misclassification of depressive disorders.

Paper [19] proposed a unique attention-based deep neural network that enables the merging of several modalities. This network is used to regress the depression level. This network has been trained using acoustic, text, and visual modalities. The regression process relies primarily on verbal input, which validates the therapist's experience. It can be challenging to combine text, graphics, and audio to estimate depression levels since integrating and synchronising multiple data sources is complicated and may require specialised technological knowledge and resources.

Study [20] extracted voice data features using Python programming and stored them in CSV files. For modelling, a database of 1479 voice feature samples was created. Utilising algorithmic selection and 10-fold cross-validation, a decision tree screening framework for depression was developed. Enhanced accuracy in forecasting was attained by the approach, enabling patients with depression to get early warning and care. It shows that clinical depression may be quickly identified and diagnosed using speech data. Depending on the complexity of this model, there is a risk of overfitting and limited generalizability.

An investigation on the creation of a supplementary tool for detecting depressive illnesses was carried out by the authors in paper [21], whereas 33 participants – 22 with a history of MDD and 11 healthy controls – were used to test automated classification algorithms and extract voice acoustic characteristics. ML approaches and an approach for extracting vocal features were applied to the recordings. According to outcomes, random tree models with 100 trees outperformed other models in terms of categorisation, pointing to a non-invasive, low-cost technique for severe depressive illness detection and screening.

In study [22] suggested utilising multiple regression to predict the risk of depression using the context-DNN model. The expertise required to forecast circumstances and surroundings impacting depression while considering context information makes up the context of the suggested context-DNN. Every context data about depression predictor variables enters the DNN as an input, and each variable is used to predict the depression output of the DNN. Regression analysis was utilised to forecast the risk of depression for DNN connections to predict the possible context that may influence that risk. Due to their high learning capacity, DNNs may overfit, mainly when working with noisy or limited datasets.

To automatically identify depressed individuals on social media and provide an explanation for the model prediction, the authors in [23] suggested explainable Multi-Aspect Depression Detection with Hierarchical Attention Network (MDHAN). They've considered user posts that had been enhanced with extra Twitter functionality. Specifically, the author computes the relevance of each tweet and word, encodes user posts using two levels of attention mechanisms applied at the tweet and word levels, and extracts semantic sequence features from user timelines (posts). The hierarchical attention approach was designed to identify patterns that provide interpretable outcomes. However, this model may have problems comprehending complicated models and raising privacy issues because the data it uses is sensitive.

This cross-sectional, descriptive-analytical study involved 205 pregnant Iranian patients under the care of Tabriz health centres. Cluster sampling was the sampling technique employed by the authors in paper [24]. Pregnant women completed the online Depression, Anxiety and Stress Scale-21 (DASS-21) and the sociodemographic characteristics questionnaire as part of the data-gathering process. The general linear model was employed to ascertain the components that were predictive of stress, anxiety, and depression. If a sample is not randomly selected or consists exclusively of people from a particular demographic, the study may be biased towards selection.

As a result, there were many restrictions on the method for identifying depression-related disorders. It must be interpretable and encounters consistent voice data, which could result in incorrect classifications. It takes specialised technological knowledge and resources to integrate and synchronise text, video, and audio to determine depression levels. The existing works could have limited generalizability and run the risk of overfitting, especially with noisy or small datasets. Although it provides a low-cost, non-invasive method for detecting and screening severe mental disorders, there were several significant obstacles, including the model's complexity, potential biases in sample selection, and privacy issues because the data was sensitive.

3. Proposed Methodologies. ML algorithms have shown potential in detecting voice signal depression, potentially transforming mental disorder diagnosis. However, challenges remain, such as background noise sensitivity, limited adaptation speed, and low signal clarity due to class imbalances. It is challenging to reliably distinguish depression detection from speech signals due to these issues. Despite ongoing efforts to improve precision and accuracy, ML efficacy in mental health assessment remains limited by issues like interpretability, robustness, computational complexity, generalisation issues, overfitting, and dimensionality reduction. This study presents a novel ML paradigm for mental health depression detection, aiming to advance mental health diagnostics by enabling more precise and scalable detection. A rising number of individuals worldwide suffer from depression, a severe mental illness that affects people of any age. Traditional methods for diagnosing depression through mental health evaluations are complex and require machine learning techniques. However, limitations such as background noise sensitivity, less adaptation speed, and imbalanced data can impair the accuracy of existing machine learning systems. This study proposes a novel ML framework, ModWave Cepstral Fusion and Stochastic Embedding Framework, to predict depression. To overcome these challenges, a Gain Modulated Wavelet Technique is employed to remove background noise from audio recordings, capturing low- and high-frequency information. The next step is feature extraction, which reduces noise impact and improves model robustness. The Auto Cepstral Fusion Feature extraction technique is introduced to capture temporal and spectral characteristics caused by background voice. Feature selection is crucial for classification, as selecting irrelevant features can lead to overfitting, the curse of dimensionality, less robustness to noise, and low interpretability. The Principal Stochastic Embedding technique handles high-dimensional data, minimises noise influence, and enhances model performance. Classification is performed using the XGBoost classifier to determine if a person is depressed. Figure 1 shows the proposed workflow diagram comprising pre-processing, feature extraction, feature selection, and classification procedures.

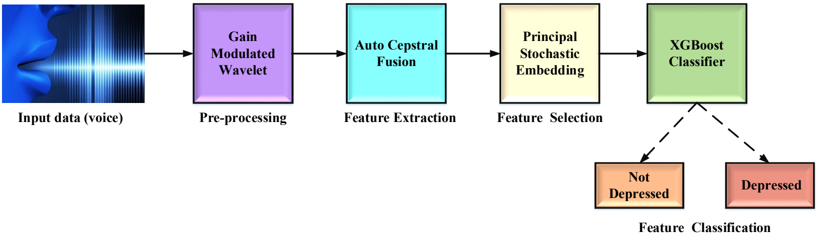


Fig. 1. Proposed Workflow diagram

3.1. Pre-processing based on Gain Modulated Wavelet Technique. Gain Modulated Wavelet (GMW) Technique removes background noise from audio recordings, capturing both low- and high-frequency information in voice signals and normalising audio signals to ensure consistent amplitude levels across recordings. Pre-processing of audio signals typically involves several steps aimed at enhancing the signal's quality or extracting useful information from it. To remove noise from audio signals, the GMW technique strengthens the adaptive gain model and discrete wavelet transform methods. Pre-processing audio signals with adaptive gain control and discrete wavelet transform can be effective for several functions, including audio denoising and capturing information at low and high frequencies.

Decompose the audio signal $x(n)$ into wavelet coefficients using discrete wavelet transform (DWT), which is a powerful tool for time-frequency analysis. The decomposition of the audio signal into wavelet coefficient using DWT is determined in equation (1):

$$x(n) = \sum_{k=0}^{N-1} c_{J,k\,\phi\,J,k}(n) + \sum_{j=1}^{J} \sum_{k=0}^{N/2^{j}-1} d_{j,k\,\psi\,j,k}(n), \qquad (1)$$

where $c_{J,k}$ are the approximation coefficients at scale $J, d_{j,k}$ are the detail coefficients at scale j, and $\phi\,J,k$ and $\psi\,j,k$ are the scaling and wavelet functions, correspondingly.

Estimate the noise level σ_j in each detail coefficient sub-band employing robust methods, including median absolute deviation (MAD) or local variance estimation. The estimation of noise level is measured in equation (2):

$$\sigma_j = \mathrm{MAD}(d_{j,k}) / \, 0.6745, \qquad (2)$$

where σ_j represents the estimated noise level in the j^{th} detail coefficient sub-band. $MAD(d_{j,k})$ refers to the mean absolute deviation of wavelet coefficients in j^{th} sub-band. 0.6745 is a constant scaling factor used to scale MAD to estimate standard deviation for normally distributed data. Combining everything, the formula determines the noise level (σ_j). The $MAD(d_{j,k})$ of the wavelet coefficients in a given sub-band is obtained by dividing it by a scaling factor (0.6745).

Apply adaptive gain control to the detail coefficients using the estimated noise levels. The gain factor g_j for each sub-band is calculated in equation (3):

$$\mathrm{g_j} = \frac{\sigma_{\mathrm{target}}}{\sigma_{\mathrm{j}}}, \tag{3}$$

where σ_{target} is the target noise level. Normalise the detail coefficients by multiplying them with the respective gain factors g_j. The normalised coefficients are determined in equation (4):

$$\mathrm{d'_{j,k}} = \mathrm{g_j} . \mathrm{d_{j,k}}, \tag{4}$$

where $d'_{j,k}$ are the modified detail coefficients after gaining control. Reconstruct the denoised signal $x'(n)$ by applying the inverse discrete wavelet transform to the modified coefficients. The reconstruction of the denoised signal is calculated in equation (5):

$$\mathrm{x'(n)} = \sum_{\mathrm{k=0}}^{\mathrm{N-1}} \mathrm{c'}_{\mathrm{J,k}\,\phi\,\mathrm{J,k}}(\mathrm{n}) + \sum_{\mathrm{j=1}}^{\mathrm{J}} \sum_{\mathrm{k=0}}^{\mathrm{N/2^j-1}} \mathrm{d'}_{\mathrm{j,k}\,\psi\,\mathrm{j,k}}(\mathrm{n}). \tag{5}$$

Adaptive gain control and discrete wavelet transform are combined in this hybrid method to efficiently reduce noise in the audio stream while maintaining significant signal characteristics. The raw data is normalised to eliminate background noise from audio recordings, capturing both low- and high-frequency information present in voice signals and normalising audio signals to ensure consistent amplitude levels across recordings using the combined power of adaptive gain control and discrete wavelet transform. The next stage is to extract features from pre-processed data. The dataset was balanced using the SMOTE approach, which comes after the noise removal procedure. The Synthetic Minority Over-sampling Technique (SMOTE) is a machine learning algorithm that addresses class imbalance, where a minority class is underrepresented in a dataset, leading to biased models. SMOTE balances class distribution, making the model more robust and less biased towards the majority class. Although its effectiveness depends on the dataset and problem, it is a valuable tool for handling imbalanced datasets by generating synthetic samples for the minority class, improving machine learning model performance, especially in situations where the minority class is of significant interest. The process of SMOTE algorithm could be divided into several steps:

1. To create additional samples, choose one minority sample and generate it as x_i.

2. Determines k closest neighbors x_i. Make a random selection from them and mark it with x_j.

3. To create a new sample of x_{new}. Equation (6) uses θ, a random number between 0 and 1.

$$x_{new} = x_i + \theta * \left(x_j - x_i\right). \tag{6}$$

4. Repeat steps 2 and 3 a total of round $(N/100)$ times to create round $(N/100)$ minority samples.

5. Apply the aforementioned process to every minority sample $(i = 1,2,\ldots,T)$.

In imbalance situations, the SMOTE method produces minority samples to improve classifier performance and balance datasets. All samples are treated equally, though, thus it might miss samples that could be mistakenly labelled. Proposed algorithms for over-sampling aim to increase the accuracy of minority samples by giving greater weights to samples that are prone to misclassification. In unbalanced datasets, this method guarantees that minority samples are given greater weight than majority samples.

3.2. Feature extraction based on Auto Cepstral Fusion technique. In audio signal processing, extracting relevant features from pre-processed data is crucial for various applications. However, difficulties arise with generalisation and interpretation due to the raw nature of voice signals, leading to poor performance in feature extraction. A technique called Auto Cepstral Fusion feature extraction is introduced to address these challenges. This method combines Autocorrelation and Mel Frequency Cepstral Coefficients (MFCCs) to enhance feature extraction capabilities. Autocorrelation and MFCCs are combined to capture spectral and temporal information from audio signals, making the technique versatile and practical for various audio analysis tasks. The process begins with pre-processed audio signals. MFCC extraction is applied to capture the spectral envelope, while autocorrelation extracts the audio's pitch period and harmonic structure. Combining these techniques allows for a comprehensive understanding of audio signals, improving interpretability and performance in feature extraction tasks.

3.2.1. Mel Frequency Cepstral Coefficients (MFCCs). Feature extraction is the technique of considering a stationary speech segment that is small enough while identifying and computing a collection of features for every short time frame of the input speech signals to provide meaningful modelling. Since the MFCC method's computation depends on short-time power, features are extracted in this study utilising the mel-frequency cepstral coefficient. The spectrum obtained from the vocal cords of humans further maps the known fluctuation of critical bandwidth frequencies of the human ear using two filters to capture the essential components of speech: a logarithmic filter at high frequencies above 1 kHz and a linear filter at low frequencies below 1 kHz. Figure 2 illustrates the MFCC feature extraction process. The MFCC includes some extraction process which follows.

Pre-emphasis. It needs to go through a filter to make up for the high-frequency part muted through the human sound-generating process. The high pass filter is applied to the voice signal in equation (7):

$$\mathrm{x_1(n) = x(n) - \alpha * x(n-1)}, \tag{7}$$

where $x_1(n)$ represents the output signal, $x(n)$ and $x(n-1)$ signifies present and past signal individually. The value α lies between 0.9 to 1.

Frame Blocking. The continuous speech signal is split into N sample-sized frames, with N-M samples overlapping and M samples (M<N) separating neighbouring frames. This procedure keeps going till the signal is divided into smaller frames.

Windowing: the windowing process involves tapering the signal to zero at the beginning and end of every frame to reduce spectral distortion. After multiplying the signal $x(n)$ by a window $w(n)$ at time n, the extracted signal is obtained by equation (8):

$$\mathrm{y_2(n) = x(n) * w(n),\ \ 0 \leq n \leq N-1}, \tag{8}$$

where N is the number of samples in every frame. Since the Hamming window sinks the sidelobe level of window transfer while reducing the frequency resolution of spectral analysis, it is used in this case; the spectral analysis for reducing the frequency resolution is determined in equation (9):

$$\mathrm{w(n) = 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right],\ \ 0 \leq n \leq N-1}. \tag{9}$$

Fast Fourier Transform. Transfers N frequency domain samples to the time domain. The Discrete Fourier Transform (DFT), which depends on a collection of N Samples (yn), is developed using the widely used FFT approach. The estimation of the DFT process is determined in equation (10):

$$Y_n = \sum_{n=0}^{N-1} y_n e^{-j2\pi kn/N}, \ k = 0,1,2 \dots \dots, N-1. \tag{10}$$

A spectrum or periodogram is a concept used to describe the outcome of the FFT process.

Mel-frequency wrapping. Mel frequency depends primarily on research on how humans perceive frequency. All frequency bands exhibit varying degrees of sensitivity in human hearing. It becomes less responsive to increased frequencies over 1000 Hz. Mel-frequency, defined as linear frequency spacing below 1 kHz, is the voice signal. The estimation of the Mel-frequency wrapping process is measured in equation (11):

$$\mathrm{Mel}(f) = 2595 * \log_{10}(1 + f/700). \tag{11}$$

Cepstrum. In this final step of the MFCC procedure, the log mel spectrum is transformed into the time domain. Since DCT's results include significant quantities of energy, DCT typically conducts this conversion. The DCT output is expressed as MFCC and is represented in equation (12):

$$C[n] = \sum_{n=0}^{N-1} \log \left| \sum_{n=0}^{N-1} x(n) \exp\left(\frac{-j2\pi kn}{N}\right) \right| \exp\left(\frac{j2\pi kn}{N}\right), \tag{12}$$

where n=0, 1,2,……N-1. $C[n]$ means MFCC, and twelve cepstral coefficients are retrieved from every frame, where n is the number of coefficients (n=12).

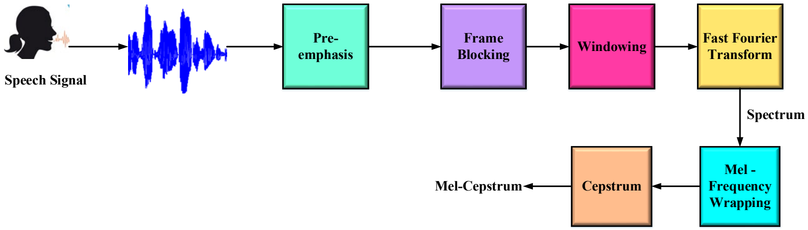


Fig. 2. MFCC feature extraction process

3.2.2. Auto Correlation. The autocorrelation function compares the similarity between the delayed and original signals to determine how self-similar a signal is in the temporal domain. A strong positive association is indicated by an autocorrelation value of +1, a negative association by -1, and no association by 0. Because the signal has a perfect correlation with itself, the autocorrelation at lag zero is always 1. Autocorrelation is instrumental in capturing periodic and repetitive patterns in speech signals. It's computed by correlating the signal with itself at various time lags, revealing crucial speech characteristics like formants and pitch. Calculate the autocorrelation function for each speech signal frame to better understand its self-similarity. The autocorrelation function is defined in equation (13):

$$\mathrm{R(k)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{x(n).x(n+k)}, \tag{13}$$

where N is the frame length, $x(n)$ represents the signal at the time index n, k is the lag at which autocorrelation is computed, and $R(k)$ is the autocorrelation value at lag k. The autocorrelation function captures the self-similarity of the signal, highlighting periodic components. Autocorrelation values may vary depending on the amplitude and energy of the signal. Normalisation helps make the feature more robust and invariant to changes in amplitude. Each autocorrelation value is divided by the autocorrelation at lag 0 to normalise it. The autocorrelation process for normalising the value is measured in equation (14):

$$\mathrm{R'(k)} = \frac{\mathrm{R(k)}}{\mathrm{R(0)}}. \tag{14}$$

From the computed autocorrelation function, you can extract various features that are useful for speech recognition and periodicity of the signal, such as:

– Pitch period: the pitch period of the signal is frequently correlated with the lag corresponding to the first peak following lag 0.

– Harmonic Structure: the autocorrelation function's regularly spaced peaks can indicate that the signal has harmonic components.

– Envelope Information: the signal envelope can be obtained from the decay rate of the autocorrelation values. In audio and voice processing applications, autocorrelation facilitates the extraction of significant signal features, making tasks like pitch estimation, harmonic analysis, and envelope identification easier. While the autocorrelation features offer more details about the temporal and periodic patterns, the MFCC features capture the speech's spectrum qualities. Auto Cepstral Fusion, a Feature extraction technique, extracts relevant features, reducing the impact of noise, improving the robustness of the model, and aiming to capture temporal and spectral characteristics caused by background voice. The proposed strategy provides feature selection to minimise dimensionality based on the particular application and requirements. This is described in more detail in the steps that follow and can aid in identifying the most valuable features while reducing computing complexity.

3.3. Feature Selection based on the Principal Stochastic Embedding technique. Selecting the most pertinent characteristics is the next step after feature extraction. The principal stochastic embedding technique, which supports t-distributed Stochastic Neighbor Embedding (t-SNE) along with Principal Component Analysis (PCA), is used to carry out this procedure. This combination allows for capturing global and local structures identified in voice signal data, making it possible to visualise complex relationships more thoroughly. PCA is good at capturing global structures, which helps it discover broad patterns and trends within the data. Still, t-SNE focuses on maintaining local structures, which allows it to capture complex interactions among neighbouring data points. When these two approaches work together, complicated relationships in the data can be shown more effectively, leading to a more informed feature selection technique.

3.3.1. t-SNE algorithm. The essential elements strongly associated with the target characteristic are selected by applying dimensional reduction techniques. Significant and highly representative features are collected to achieve high accuracy. Decreasing the amount of variables in a dataset is known as dimensional reduction. A proposed technique for reducing the dimensionality of nonlinear data is to drop it from a high-dimensional space into a low-dimensional one using the t-SNE technique. The method focuses on the variance of neighbourhood points and data inclusion in a low space, producing random, unconfirmed probability. It assigns comparable traits with greater probability and dissimilar characteristics to lower probability when distributing pairs of X_i and X_j. The pairwise similarity in the high-dimensional data space is determined in equation (15), and the data points representation by t-SNE in a low-dimensional space is demonstrated in equation (16). Equation (17) illustrates how the technique iteratively operates the same probability distribution across a smaller space to show data points in a low-dimensional space and lower the Kulback-Leibler (KL) variance. The probability distribution with low KL variance is determined in equation (17):

$$P(X_i/X_j) = \frac{S(X_i\ ,\ X_j)}{\sum_{m\neq i}^{N} S(X_i,\ X_m)}, \tag{15}$$

$$Q(Y_i/Y_j) = \frac{S(Y_i,\ \ Y_j)}{\sum_{m\neq i}^{N}\ S\ (Y_i,\ \ Y_m)}, \tag{16}$$

$$KL = \sum_i \sum_j\ P(X_i,\qquad X_j) \log \frac{P(X_i/X_j)}{Q(Y_i/Y_j)}, \tag{17}$$

where $P(X_i/X_j)$ high-dimensional data is space and X_i/X_j are pairs in the P space; Y_i/Y_j is low-dimensional data space, and Y_i/Y_j are pairs in the Q space (Figure 3).

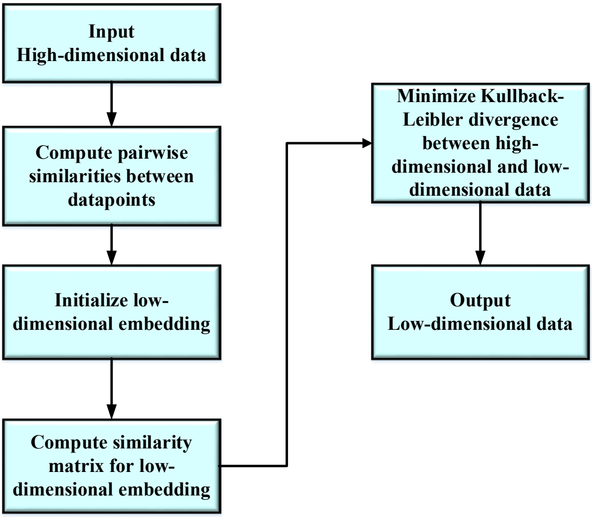

Input
High-dimensional data
Compute pairwise similarities between datapoints
Initialize low-dimensional embedding
Compute similarity matrix for low-dimensional embedding
Minimize Kullback-Leibler divergence between high-dimensional and low-dimensional data
Output
Low-dimensional data

Fig. 3. Flowchart for t-SNE algorithm

3.3.2. PCA Algorithm. Principal components of PCA, an unsupervised statistical technique, are linearly uncorrelated when correlated features are converted. Normalising the dataset uses mathematical ideas like variance, covariance, eigenvalues, and eigenvectors. Correlation is the connection between two characteristics, whereas dimensions are the amount of features in the collection. To ensure there are high-variance features, the method divides the individual value by the standard deviation of all features. The Z covariance matrix includes the variance among the two feature pairs, and eigenvectors represent high-variance information axes. In the P matrix, the technique places eigenvalues and eigenvectors in descending order. The Z covariance matrix is multiplied by the P matrix to generate new features. Important and pertinent characteristics are kept, while less important ones are eliminated to produce a new dataset. The dimensionality reduction feature selection technique aids in reducing features while preserving the most significant amount of relevant information. PCA identifies the principle components that primarily explain the variation in the data, whereas t-SNE produces a low-dimensional representation that preserves the local structure. The hybrid strategy lowers the risk of overfitting and improves model generalisation by combining both techniques to achieve more effective dimensionality reduction. This optimised feature selection process produces a subset of characteristics that are very discriminative and informative of the underlying structure in the data. Enhanced visualisation, complete data representation, improved dimensionality reduction, optimal feature selection, and increased performance in machine learning tasks are just a few benefits of the hybrid feature selection approach utilising principle stochastic embedding. Figure 4 shows the flowchart for the PCA algorithm. The next step involves a classification process, which classifies depressed patients.

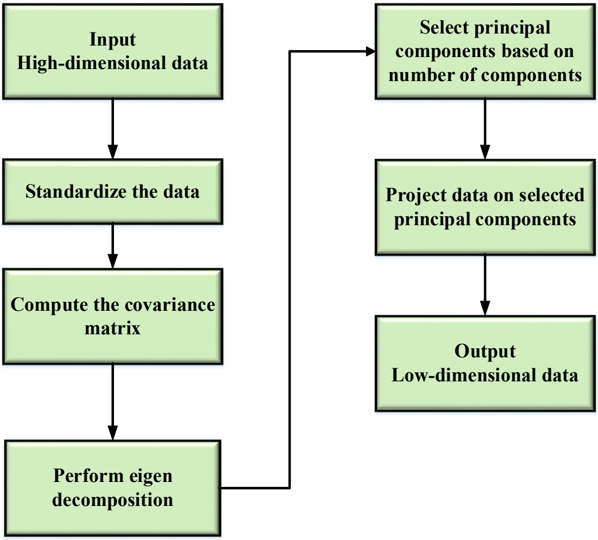

Input
High-dimensional data
Standardize the data
Compute the covariance matrix
Perform eigen decomposition
Select principal components based on number of components
Project data on selected principal components
Output
Low-dimensional data

Fig. 4. Flowchart for PCA

3.4. XGBoost Classifier. Using the XGBoost classifier, also known as the Extremely Gradient Boosted Decision Tree, can effectively segment data into smaller subsets, leading to remarkable accuracy across many Natural Language Processing (NLP) applications and ML models. This classifier boasts several advantages, including scalability, parallelizability, and swift execution times, making it a preferred choice in various settings. Moreover, it is a regularised model in which formalisation helps to prevent overfitting, thereby enhancing performance compared to other algorithms. Utilising the XGBoost classifier, the proposed technique demonstrates superior performance in identifying depressed individuals compared to existing approaches. Through its expertise in handling intricate relationships within data and its robustness against overfitting, XGBoost ensures high accuracy and reliability in predicting outcomes, thereby offering a promising avenue for advancing depression detection. The XGBoost workflow schematic is shown in Figure 5. The blue-coloured zone represents the training and testing data. The boxes inside the dashed lines represent the testing and training procedures, where GBM stands for gradient boosting machine and T is for tree. The results obtained from XGBoost from the dashed box are displayed in the two oval boxes on the right.

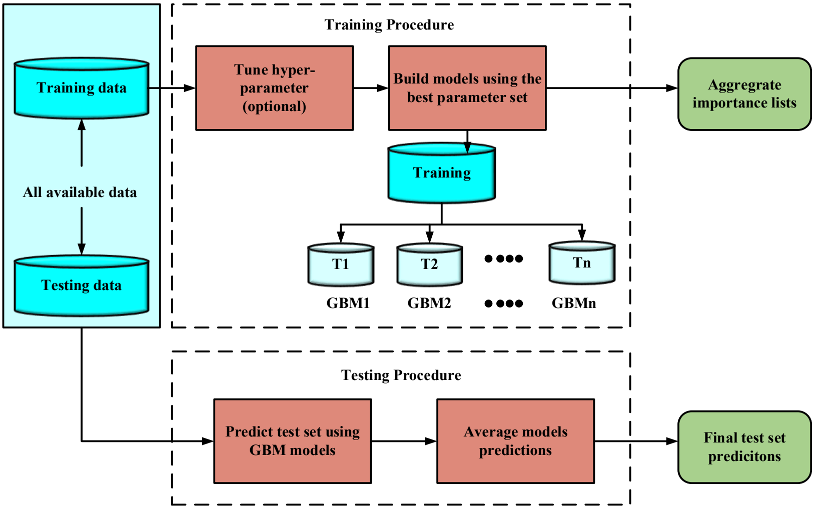

Training data
All available data
Testing data
Training Procedure
Tune hyper-parameter (optional)
Build models using the best parameter set
Training
T1
T2
●●●●
Tn
GBM1
GBM2
●●●●
GBMn
Aggregrate importance lists
Testing Procedure
Predict test set using GBM models
Average models predictions
Final test set predicitons

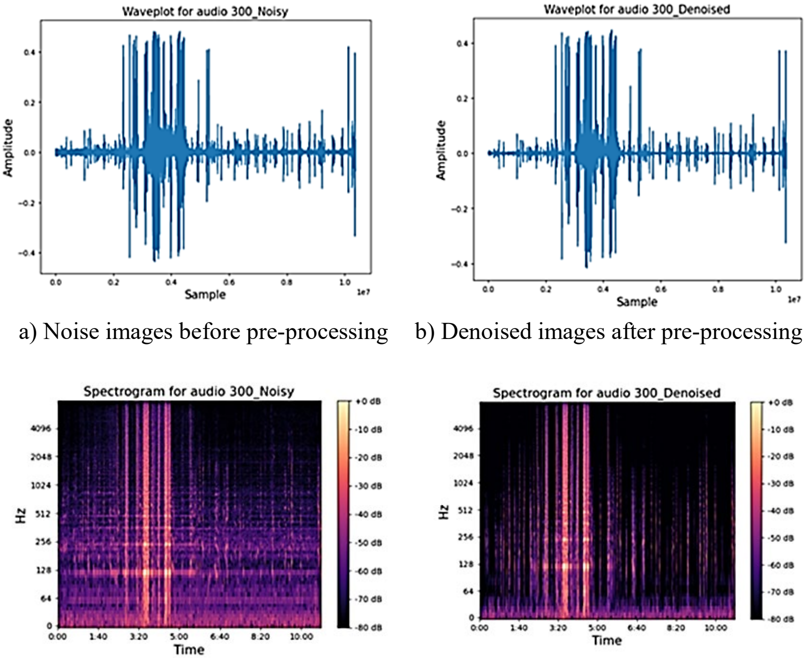
Fig. 5. XGBoost classifier

4. Results and Discussion. The Distress Analysis Interview Corpus: Wizard of Oz (DAIC-WOZ) dataset is employed in this research [25]. Based on feature vectors, data was chosen and split into two categories:

80% for training and 20% for testing. The dataset comprises text, audio, and video exchanges among individuals and an automated interviewer. The interview questions are based on the physical manifestations of depression. Two versions of this dataset have been released up to this point; the expanded DAIC-WOZ is utilised in this study. There are 189 sessions in the sample (102 men and 87 women), of which 133 are not depressed, and 56 are depressed. The dataset is split into 80% and 20% for training and testing. The outcomes achieved are presented in the subsections that follow. The Python tool, Windows 7 (64-bit) OS, and Intel Premium CPU with 8GB RAM are used to carry out this proposed work. In conclusion, it compared previous approaches and the proposed system. This section discusses how well the approach we propose works for identifying depressed patients.

4.1. Performance analysis. The amplitude of the audio signal is shown in the waveplot graph. Unequal or fluctuating amplitude in a noisy audio waveplot indicates the existence of background noise. The noisy waveplot implies that the audio signal has been affected by unwanted noise, which might affect the clarity of the audio. Figure 6(a) shows that the x-axis depicts the sample, and the y-axis is the audio signal's amplitude. The audio signal amplitude following noise reduction processing is seen in the denoised waveplot: a less amplitude fluctuating, smoother waveplot than the noisy version. The denoised waveplot shows that the noise reduction procedure has successfully eliminated or reduced background noise, producing a more precise and cleaner audio stream. Figure 6(b) shows that the x-axis depicts the sample, and the y-axis is the audio signal's amplitude.

A spectrogram shows the audio signal's frequency content with time. The presence of noise is shown in spectrograms of noisy audio as extra energy in different frequency bands, which frequently take the form of irregular patterns or streaks. A noisy spectrogram may mask or distort the properties of the underlying signal by highlighting spectral contamination carried on by background noise. Figure 6(c) shows that the y-axis indicates hz, and the x-axis represents time. The frequency content of the audio signal is seen on the denoised spectrogram following noise reduction. Reduced energy in background noise-corresponding frequency regions improves the visibility of signal characteristics and produces more apparent spectral patterns. A denoised spectrogram shows how noise reduction may improve the signal-to-noise ratio, which makes it easier to identify and analyse relevant audio properties. Figure 6(d) shows that the y-axis indicates Hz, and the x-axis represents time in seconds. With the wave plot showing changes in amplitude and the spectrogram indicating changes in frequency content, both offer useful information on how noise reduction affects audio signals. These graphics help evaluate how well noise reduction methods work and how they affect the audio signal's overall quality.

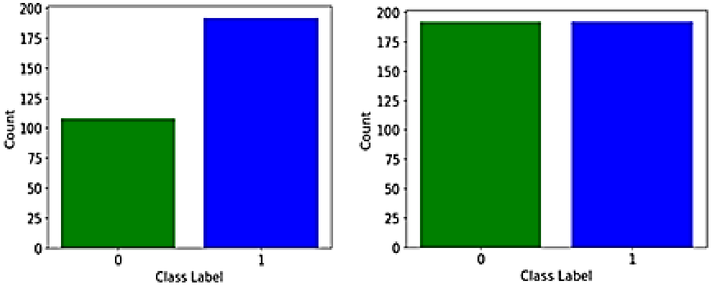


a) Noise images before pre-processing b) Denoised images after pre-processing

c) Noisy spectrogram audio d) Denoised spectrogram audio

Fig. 6. Wave plot and Spectrogram visualisation

4.2. Imbalanced Datasets. Figure 7 below shows the dataset modelling (before and after balancing the dataset). There are two types of class labels: 0 is represented as depressed, and 1 is represented as not depressed. The dataset utilised in this study comprises individuals categorised as depressed and not depressed, but it exhibits an inherent imbalance and is contaminated with background noise. Consequently, the pre-processing methodology outlined in this study addresses these challenges. Following pre-processing, the dataset achieves a balance between the depressed and not-depressed categories while also effectively eliminating background noise. Furthermore, this pre-processing method ensures the capture of low and high-level features within the dataset. Here, we used the SMOTE technique to balance the dataset, which was performed after the noise removal process.



a) Imbalanced dataset b) Balanced dataset

Fig. 7. Dataset modelling

4.3. Confusion matrix. One kind of performance metric employed in ML and classification to evaluate a model's ability to identify depression is a confusion matrix. It offers a summary of the variations among predicted labels and the actual ground truth labels so that the model's accuracy may be evaluated. The four groups in the confusion matrix are TP, FP, FN, and TN, where 0 denotes people who are depressed, and 1 represents those who are not depressed. Several performance measures, including accuracy, precision, recall, and F1-score, may be computed using the confusion matrix to assess how well the framework identifies depression. Figure 8 shows the confusion matrix for the proposed work.

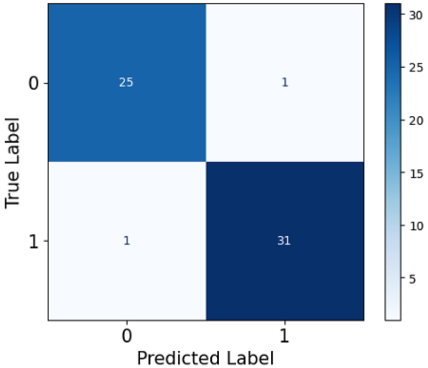


Fig. 8. Confusion matrix

4.4. ROC curve. A graphical depiction of a binary classification model's accuracy across various threshold values is called a Receiver Operating Characteristic (ROC) curve. At various threshold values, it shows the true positive rate (TPR) versus the false positive rate (FPR). The ROC curve, shown in Figure 9, depicts the false positive rate (specificity) on the x-axis and the true positive rate (sensitivity) on the y-axis. An area under the curve (AUC) of 0.97 on an ROC curve for detecting depression denotes good discriminating power in differentiating between those who are depressed and those who are not. A high likelihood of ranking a randomly selected sad person higher than a randomly chosen non-depressed person is indicated by an AUC of 0.97. This suggests that the model makes few prediction errors, achieving high sensitivity while maintaining low false positive rates across various threshold settings.

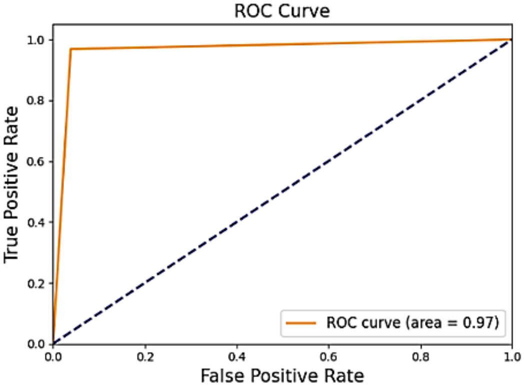


Fig. 9. ROC curve

4.5. Comparative analysis. The depression prediction framework was evaluated with several performance metrics: root mean square error (RMSE), mean absolute error (MAE), accuracy, precision, recall and F1-score. The performance metrics are compared with several existing works such as Deep Convolutional Neural Network-Deep Neural Network ((DCNN-DNN) [26], Deep Convolutional Generative Adversarial Network (DCGAN) [27], Transformer Encoder + Convolutional Neural Network (TE+ CNN) [28], Bidirectional-Long Short term Memory + Attention (Bi-LSTM + Attention) [29], Graph Convolutional Neural Network (GCNN) [30], Convolutional Neural Network (CNN) [31], Gated Recurrent Unit (GRU) [31], Bimodal Attention-GRU (BiAtt-GRU) [31], Two-dimensional CNN-LSTM (2D-CNN-LSTM) [32], Decision Tree (DT) [32], Deep AudioNet (DAN) [32], Transformer-CNN-CNN (TCC) – Softmax [32], Unimodal Ensemble (UE) [33], Multimodal + Selective dropout + Transfer Learning (MM + SD + TL) [33], Multimodal + Selective dropout-Normalization-Attention + Transfer Learning + Spectral-Normalized Neural Gaussian Process (MM+ SD-Norm-Att + TL + SNGP) [33].

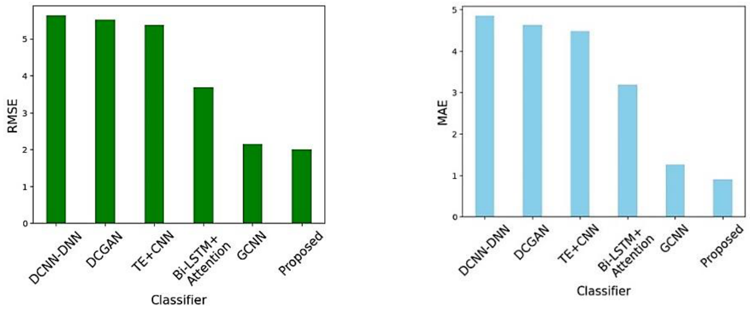
The Root Mean Square Error (RMSE) is a frequently employed metric for assessing a model's or prediction's accuracy. The definition of this term is the square root of the average of the squared differences between the actual and predicted values. Mathematically, RMSE is represented in equation (18):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{18}$$

The accuracy of a model's predictions is gauged by the Mean Absolute Error (MAE) metric. The mean of the absolute differences between the actual and predicted values is how it is defined. Without considering direction, the mean absolute error (MAE) quantifies the average magnitude of mistakes in a set of predictions. A lower MAE value denotes a more precise model. The MAE is determined in equation (19):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{19}$$

where $\hat{y}_i$, y_i and n display the actual severity score, the predicted score generated from the model, and the amount of test data corresponding to it. By using equations (18) and (19), the RMSE and MAE values are calculated. The proposed method performs significantly fewer errors than other existing techniques. Figure 10 compares RMSE and MAE graphs for the proposed model with several existing works. Table 1 shows the comparison values for the RMSE and MAE metrics.



a) b)

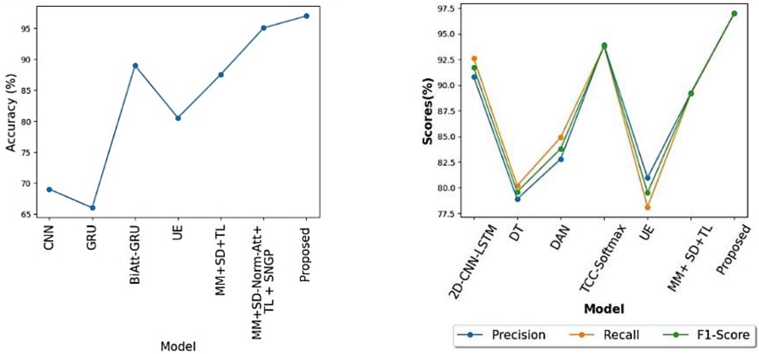
Fig. 10. Comparison graph for: a) RMSE; and b) MAE

Table 1. Comparison of error metrics

Method	RMSE	MAE
DCNN-DNN [26]	5.63	4.85
DCGAN [27]	5.52	4.63
TE+ CNN [28]	5.37	4.48
Bi-LSTM + Attention [29]	4.76	3.61
GCNN [30]	2.15	1.25
Proposed	2.00	0.90

The efficiency of the suggested strategy is demonstrated by comparing approaches for predicting depressive illnesses based on their Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). With an MAE of 1.25 and the lowest RMSE of 2.15, the GCNN demonstrates excellent prediction accuracy. The suggested approach, on the other hand, outperforms the others in terms of predicting the severity of depression, with an even lower RMSE of 2.00 and MAE of 0.90. While models such as the Bi-LSTM + Attention demonstrate competitive performance (RMSE of 4.76 and MAE of 3.61), they are less effective than the proposed technique. The Transformer Encoder + CNN and DCGAN approaches show slightly higher prediction errors, with corresponding MAE values of 4.48 and 4.63 and RMSE values of 5.37 and 5.52, respectively. The suggested approach stands out for its precision and accuracy in estimating the severity of depressive disorder, providing encouraging developments in this area of study.

Evaluation metrics. Accuracy, precision, recall, and F1 score were used as performance indicators for this study. These metrics finally demonstrate the proposed technique's performance reliability. Figure 11 below displays the comparison graph for the proposed work.



a) b)

Fig. 11. Comparison graph for performance metrics: a) Accuracy, b) Precision, Recall and F1-score

The work proposed provides all of the current methods, and Tables 2 and 3 display the comparative analysis of the performance metrics for the proposed model with several existing works.

Table 2. Comparison of Accuracy metrics

Model	Accuracy (%)
CNN [31]	69.00
GRU [31]	66.00
BiAtt-GRU [31]	89.00
UE [33]	80.54
MM + SD + TL [33]	87.55
MM + SD-Norm-Att + TL + SNGP [33]	95.07
Proposed	97.02

The suggested model outperforms multiple existing models in detecting depression-related conditions, with the maximum accuracy of 97.02%. By contrast, the accuracy of the GRU model is 66%, the CNN model is 69%, and the BiAttention-GRU model is 89%. While the Multimodal + SD + transfer learning model achieves 87.55% accuracy, the Unimodal Ensemble model only manages 80.54%. 95.07% is achieved by the Multimodal + SD-Norm-Att + transfer learning + SNGP model, demonstrating a notable improvement with the suggested approach.

Table 3. Comparison of Precision, Recall and F1-score

Model	Precision (%)	Recall (%)	F1-score (%)
2D-CNN-LSTM [32]	90.80	92.60	91.70
DT [32]	78.90	80.20	79.60
DAN [32]	82.80	84.90	83.80
TCC-Softmax [32]	93.90	93.80	93.80
UE [33]	80.96	78.13	79.52
MM + SD + TL [33]	89.26	89.17	89.21
Proposed	97.02	97.02	97.02

A high degree of accuracy and consistency in its predictions is indicated by the suggested model's precision, recall, and F1-score, which are all at 97.02%, suggesting outstanding performance in diagnosing depression. It performs better than a number of other models, including the TCC-Softmax model, which attains a precision, recall, and F1-score of 93.8%, and the 2D-CNN-LSTM, which has a precision of 90.8%, recall of 92.6%, and F1-score of 91.7%. Some models perform better than others, such as the Unimodal Ensemble and Deep AudioNet, which scored 79.52%

and 83.8%, respectively, on the F1 score. Furthermore, the F1-score for the Multimodal + SD + transfer learning model is 89.21%.

This research demonstrates the superiority of the novel method in terms of depression detection, with a low MAE of 0.9 and an exceptional accuracy of 97.02%. In terms of precision and forecast accuracy, this strategy outperforms earlier ones. The study also emphasises the significance of accuracy and F1-score, standard metrics in evaluation processes, in determining the model's efficiency. The proposed method is a potential growth in effective depression detection methods as it outperforms currently available methods and produces reliable findings regarding the accuracy and F1 score.

5. Conclusion. In conclusion, the high incidence of depression necessitates the development of autonomous detection systems, given the complexities associated with traditional clinical diagnosis methods. Existing ML techniques for depression detection encounter challenges such as sensitivity to background noise, slow adaptation speed, and imbalanced data, which can compromise accuracy. To overcome these limitations, this study introduces a novel ModWave Cepstral Fusion and Stochastic Embedding Framework for depression prediction. They address issues like background noise in audio signals and low amplitude levels during pre-processing by employing the Gain Modulated Wavelet Technique. This technique removes background noise while capturing low and high-frequency information in voice signals, subsequently normalising the audio signals. Difficulties in generalisation and lack of interpretability pose obstacles to extracting relevant characteristics from speech. To tackle these challenges, an Auto Cepstral Fusion extraction technique was proposed to extract pertinent features, aiming to capture both temporal and spectral characteristics caused by background voice. Moreover, feature selection is crucial to ensure robust classification. To address this, the Principal Stochastic Embedding technique handles high-dimensional data, reduces the influence of noise, and minimises dimensionality. Utilising the XGBoost classifier, the proposed method distinguishes between depressed and non-depressed individuals using the DAIC-WOZ Datasets from USC. The proposed approach achieved a remarkable accuracy of 97.02% and a low MAE of 0.9, positioning it as a promising tool for autonomous depression detection. This proposed model provided enhanced accuracy by effectively integrating multiple data modalities. Developing advanced machine learning methods presents interesting chances to improve depression detection systems. In particular, deep learning has demonstrated great promise for identifying complicated patterns and characteristics in large, complex data sets. Future studies might examine how deep learning architectures can be applied to assess multimodal data more effectively and accurately identify minor signs of depression. By utilising wearable technology, smartphone apps, and other digital platforms, these systems might continually monitor people's physiological and behavioral signals and offer therapies and notifications when early indications of depression are identified. These kinds of systems can lessen the harmful effects of depression and enhance people's general well-being by providing early detection and access to mental health services.

References

1. Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). World Health Organization. 2017. 22 p.
2. Uddin M.Z., Dysthe K.K., Følstad A., Brandtzaeg P.B. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications. 2022. vol. 34(1). pp. 721–744.
3. Jacobson N.C., Chung Y.J. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors. 2020. vol. 20(12). DOI: 10.3390/s20123572.
4. Ormel J., Kessler R.C., Schoevers R. Depression: More treatment but no drop in prevalence: how effective is treatment? And can we do better? Current opinion in psychiatry. 2019. vol. 32(4). pp. 348–354.
5. Culpepper L. Understanding the burden of depression. The Journal of Clinical Psychiatry. 2011. vol. 72(6). DOI: 10.4088/JCP.10126tx1c.
6. Sadock B.J., Sadock V.A., Ruiz P. Compêndio de Psiquiatria: Ciência do Comportamento e Psiquiatria Clínica. Artmed Editora. 2016. 1490 p.
7. Mundt J.C., Vogel A.P., Feltner D.E., Lenderking W.R. Vocal acoustic biomarkers of depression severity and treatment response. Biological psychiatry. 2012. vol. 72(7). pp. 580–587.
8. Hashim N.W., Wilkes M., Salomon R., Meggs J., France D.J. Evaluation of voice acoustics as predictors of clinical depression scores. Journal of Voice. 2017. vol. 31(2). DOI: 10.1016/j.jvoice.2016.06.006.
9. Khoo L.S., Lim M.K., Chong, C.Y., McNaney R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors. 2024. vol. 24(2). DOI: 10.3390/s24020348.
10. Low D.M., Bentley K.H., Ghosh S.S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope investigative otolaryngology. 2020. vol. 5(1). pp. 96–116.
11. Asci F., Costantini G., Di Leo P., Zampogna A., Ruoppolo G., Berardelli A., Saggio G., Suppa A. Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender. Sensors. 2020. vol. 20(18). DOI: org/10.3390/s20185022.
12. Chen Z.S., Galatzer-Levy I.R., Bigio B., Nasca C., Zhang Y. Modern views of machine learning for precision psychiatry. Patterns. 2022. vol. 3(11).
13. Jiang H., Hu B., Liu Z., Wang G., Zhang L., Li X., Kang H. Detecting depression using an ensemble logistic regression model based on multiple speech features. Computational and mathematical methods in medicine. 2018. vol. 1. DOI: 10.1155/2018/6508319.

14. Na K.S., Cho S.E., Geem Z.W., Kim Y.K. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. Neuroscience Letters. 2020. vol. 721. DOI: 10.1016/j.neulet.2020.134804.
15. Hochman E., Feldman B., Weizman A., Krivoy A., Gur S., Barzilay E., Gabay H., Levy J., Levinkron O., Lawrence G. Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study. Depression and anxiety. 2021. vol. 38(4). pp. 400–411.
16. Narziev N., Goh H., Toshnazarov K., Lee S.A., Chung K.M., Noh Y. STDD: Short-term depression detection with passive sensing. Sensors. 2020. vol. 20(5). DOI: 10.3390/s20051396.
17. Ware S., Yue C., Morillo R., Lu J., Shang C., Kamath J., Bamis A., Bi J., Russell A., Wang B. Large-scale automatic depression screening using meta-data from wifi infrastructure. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018. vol. 2(4). pp. 1–27.
18. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning. medRxiv. 2020. DOI: 10.1101/2020.06.23.20138651.
19. Qureshi S.A., Hasanuzzaman M., Saha S., Dias G. The Verbal and Non Verbal Signals of Depression--Combining Acoustics, Text and Visuals for Estimating Depression Level. 2019. arXiv preprint arXiv:1904.07656.
20. Chen X., Pan Z. A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health. International Journal of Environmental Research and Public Health. 2021. vol. 18(12). DOI: 10.3390/ijerph18126441.
21. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning – an exploratory study. Research on Biomedical Engineering. 2021. vol. 37. pp. 53–64.
22. Baek J.W., Chung K. Context deep neural network model for predicting depression risk using multiple regression. IEEE Access. 2020. vol. 8. pp. 18171–18181.
23. Zogan H., Razzak I., Wang X., Jameel S., Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web. 2022. vol. 25(1). pp. 281–304.
24. Effati-Daryani F., Zarei S., Mohammadi A., Hemmati E., Ghasemi Yngyknd S., Mirghafourvand M. Depression, stress, anxiety and their predictors in Iranian pregnant women during the outbreak of COVID-19. BMC psychology. 2020. vol. 8. pp. 1–10.
25. Gratch J., Artstein R., Lucas G.M., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo S., Morency L.-P. The distress analysis interview corpus of human and computer interviews. LREC. 2014. pp. 3123–3128.
26. Yang L., Jiang D., Xia X., Pei E., Oveneke M.C., Sahli H. Multimodal measurement of depression using deep learning models. Proceedings of the 7th annual workshop on audio/visual emotion challenge. 2017. pp. 53–59.
27. Yang L., Jiang D., Sahli H. Feature augmenting networks for improving depression severity estimation from speech signals. IEEE Access. 2020. vol. 8. pp. 24033–24045.
28. Lu J., Liu B., Lian Z., Cai C., Tao J., Zhao Z. Prediction of Depression Severity Based on Transformer Encoder and CNN Model. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE. 2022. pp. 339–343.
29. Fang M., Peng S., Liang Y., Hung C.C., Liu S. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control. 2023. vol. 82. DOI: 10.1016/j.bspc.2022.104561.

30. Ishimaru M., Okada Y., Uchiyama R., Horiguchi R., Toyoshima I. A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network. Diagnostics. 2023. vol. 13(4). DOI: 10.3390/diagnostics13040727.

31. Cao Y., Hao Y., Li B., Xue J. Depression prediction based on BiAttention-GRU. Journal of Ambient Intelligence and Humanized Computing. 2022. vol. 13(11). pp. 5269–5277.

32. Yin F., Du J., Xu X., Zhao L. Depression detection in speech using transformer and parallel convolutional neural networks. Electronics. 2023. vol. 12(2). DOI: 10.3390/electronics12020328.

33. Ahmed S., Yousuf M.A., Monowar M.M., Hamid M.A., Alassafi M. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. IEEE Access. 2023. vol. 11. DOI: 10.1109/ACCESS.2023.3315243.

Jacob Jithin — Research scholar, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education. Research interests: computer science, engineering, machine learning, voice analysis, cyber security. The number of publications — 7. jithu7771@gmail.com; Srivilliputhur, Tamil Nadu, 626126, Krishnankoil, India; office phone: +91(0474)257-7958.

Kannan K.S. — Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education. Research interests: machine learning, deep learning. The number of publications — 20. k.s.kannan@klu.ac.in; Srivilliputhur, Tamil Nadu, 626126, Krishnankoil, India; office phone: +91(04563)289-042.

УДК 004 **DOI** 10.15622/ia.23.6.7

Д. Джейкоб, К. Каннан

УСОВЕРШЕНСТВОВАННАЯ СИСТЕМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОНОМНОГО ОБНАРУЖЕНИЯ ДЕПРЕССИИ С ИСПОЛЬЗОВАНИЕМ МОДУЛИРОВАННОГО ВЕЙВЛЕТ-КЕПСТРАЛЬНОГО СЛИЯНИЯ И СТОХАСТИЧЕСКОГО ВСТРАИВАНИЯ

Джейкоб Д., Каннан К. **Усовершенствованная система машинного обучения для автономного обнаружения депрессии с использованием модулированного вейвлет-кепстрального слияния и стохастического встраивания.**

Аннотация. Депрессия – это распространенное психическое заболевание, требующее систем автоматического обнаружения из-за своей сложности. Существующие методы машинного обучения сталкиваются с проблемами, такими как чувствительность к фоновому шуму, медленная скорость адаптации и несбалансированные данные. Для устранения этих ограничений в этом исследовании предлагается новая структура модулированного вейвлет-кепстрального слияния и стохастическая структура встраивания для прогнозирования депрессии. Затем техника модулированных волновых функций удаляет фоновый шум и нормализует аудиосигналы. Трудности с обобщением, которые приводят к отсутствию интерпретируемости, затрудняют извлечение соответствующих характеристик речи. Для решения этих проблем используется автоматическое кепстральное слияние, которое извлекает соответствующие характеристики речи, захватывая временные и спектральные характеристики, вызванные фоновым голосом. Выбор признаков становится важным, когда выбираются релевантные признаки для классификации. Выбор нерелевантных признаков может привести к переобучению, нарушению размерности и меньшей устойчивости к шуму. Поэтому метод стохастической иммерсии справляется с высокоразмерными данными, минимизируя влияние шума и размерности. Кроме того, классификатор XGBoost отличает людей с депрессией и людей без депрессии. В результате предложенный метод использует набор данных DAIC-WOZ Университета Южной Калифорнии для обнаружения депрессий, достигая точности 97,02%, прецизионности 97,02%, полноты 97,02%, оценки F1 97,02%, среднеквадратической ошибки 2,00 и средней абсолютной ошибки 0,9, делая его многообещающим инструментом для автономного обнаружения депрессии.

Ключевые слова: обнаружение депрессии, машинное обучение, ModWave Cepstral Fusion, фоновый шум, классификатор XGBoost, набор данных DAIC-WOZ, автономная система обнаружения, точность.

Литература

1. Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). World Health Organization. 2017. 22 p.
2. Uddin M.Z., Dysthe K.K., Følstad A., Brandtzaeg P.B. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications. 2022. vol. 34(1). pp. 721–744.
3. Jacobson N.C., Chung Y.J. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors. 2020. vol. 20(12). DOI: 10.3390/s20123572.

4. Ormel J., Kessler R.C., Schoevers R. Depression: More treatment but no drop in prevalence: how effective is treatment? And can we do better? Current opinion in psychiatry. 2019. vol. 32(4). pp. 348–354.
5. Culpepper L. Understanding the burden of depression. The Journal of Clinical Psychiatry. 2011. vol. 72(6). DOI: 10.4088/JCP.10126tx1c.
6. Sadock B.J., Sadock V.A., Ruiz P. Compêndio de Psiquiatria: Ciência do Comportamento e Psiquiatria Clínica. Artmed Editora. 2016. 1490 p.
7. Mundt J.C., Vogel A.P., Feltner D.E., Lenderking W.R. Vocal acoustic biomarkers of depression severity and treatment response. Biological psychiatry. 2012. vol. 72(7). pp. 580–587.
8. Hashim N.W., Wilkes M., Salomon R., Meggs J., France D.J. Evaluation of voice acoustics as predictors of clinical depression scores. Journal of Voice. 2017. vol. 31(2). DOI: 10.1016/j.jvoice.2016.06.006.
9. Khoo L.S., Lim M.K., Chong, C.Y., McNaney R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors. 2024. vol. 24(2). DOI: 10.3390/s24020348.
10. Low D.M., Bentley K.H., Ghosh S.S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope investigative otolaryngology. 2020. vol. 5(1). pp. 96–116.
11. Asci F., Costantini G., Di Leo P., Zampogna A., Ruoppolo G., Berardelli A., Saggio G., Suppa A. Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender. Sensors. 2020. vol. 20(18). DOI: org/10.3390/s20185022.
12. Chen Z.S., Galatzer-Levy I.R., Bigio B., Nasca C., Zhang Y. Modern views of machine learning for precision psychiatry. Patterns. 2022. vol. 3(11).
13. Jiang H., Hu B., Liu Z., Wang G., Zhang L., Li X., Kang H. Detecting depression using an ensemble logistic regression model based on multiple speech features. Computational and mathematical methods in medicine. 2018. vol. 1. DOI: 10.1155/2018/6508319.
14. Na K.S., Cho S.E., Geem Z.W., Kim Y.K. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. Neuroscience Letters. 2020. vol. 721. DOI: 10.1016/j.neulet.2020.134804.
15. Hochman E., Feldman B., Weizman A., Krivoy A., Gur S., Barzilay E., Gabay H., Levy J., Levinkron O., Lawrence G. Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study. Depression and anxiety. 2021. vol. 38(4). pp. 400–411.
16. Narziev N., Goh H., Toshnazarov K., Lee S.A., Chung K.M., Noh Y. STDD: Short-term depression detection with passive sensing. Sensors. 2020. vol. 20(5). DOI: 10.3390/s20051396.
17. Ware S., Yue C., Morillo R., Lu J., Shang C., Kamath J., Bamis A., Bi J., Russell A., Wang B. Large-scale automatic depression screening using meta-data from wifi infrastructure. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018. vol. 2(4). pp. 1–27.
18. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning. medRxiv. 2020. DOI: 10.1101/2020.06.23.20138651.
19. Qureshi S.A., Hasanuzzaman M., Saha S., Dias G. The Verbal and Non Verbal Signals of Depression--Combining Acoustics, Text and Visuals for Estimating Depression Level. 2019. arXiv preprint arXiv:1904.07656.
20. Chen X., Pan Z. A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health. International Journal of

Environmental Research and Public Health. 2021. vol. 18(12). DOI: 10.3390/ijerph18126441.

21. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning – an exploratory study. Research on Biomedical Engineering. 2021. vol. 37. pp. 53–64.

22. Baek J.W., Chung K. Context deep neural network model for predicting depression risk using multiple regression. IEEE Access. 2020. vol. 8. pp. 18171–18181.

23. Zogan H., Razzak I., Wang X., Jameel S., Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web. 2022. vol. 25(1). pp. 281–304.

24. Effati-Daryani F., Zarei S., Mohammadi A., Hemmati E., Ghasemi Yngyknd S., Mirghafourvand M. Depression, stress, anxiety and their predictors in Iranian pregnant women during the outbreak of COVID-19. BMC psychology. 2020. vol. 8. pp. 1–10.

25. Gratch J., Artstein R., Lucas G.M., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo S., Morency L.-P. The distress analysis interview corpus of human and computer interviews. LREC. 2014. pp. 3123–3128.

26. Yang L., Jiang D., Xia X., Pei E., Oveneke M.C., Sahli H. Multimodal measurement of depression using deep learning models. Proceedings of the 7th annual workshop on audio/visual emotion challenge. 2017. pp. 53–59.

27. Yang L., Jiang D., Sahli H. Feature augmenting networks for improving depression severity estimation from speech signals. IEEE Access. 2020. vol. 8. pp. 24033–24045.

28. Lu J., Liu B., Lian Z., Cai C., Tao J., Zhao Z. Prediction of Depression Severity Based on Transformer Encoder and CNN Model. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE. 2022. pp. 339–343.

29. Fang M., Peng S., Liang Y., Hung C.C., Liu S. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control. 2023. vol. 82. DOI: 10.1016/j.bspc.2022.104561.

30. Ishimaru M., Okada Y., Uchiyama R., Horiguchi R., Toyoshima I. A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network. Diagnostics. 2023. vol. 13(4). DOI: 10.3390/diagnostics13040727.

31. Cao Y., Hao Y., Li B., Xue J. Depression prediction based on BiAttention-GRU. Journal of Ambient Intelligence and Humanized Computing. 2022. vol. 13(11). pp. 5269–5277.

32. Yin F., Du J., Xu X., Zhao L. Depression detection in speech using transformer and parallel convolutional neural networks. Electronics. 2023. vol. 12(2). DOI: 10.3390/electronics12020328.

33. Ahmed S., Yousuf M.A., Monowar M.M., Hamid M.A., Alassafi M. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. IEEE Access. 2023. vol. 11. DOI: 10.1109/ACCESS.2023.3315243.

Джейкоб Джитин — научный сотрудник, факультет информатики и инженерии, Каласалингамская академия исследований и образования. Область научных интересов: компьютерные науки, инженерия, машинное обучение, анализ голоса, кибербезопасность. Число научных публикаций — 7. jithu7771@gmail.com; Шривиллипутур, Тамил Наду, 626126, Кришнанкойл, Индия; р.т.: +91(0474)257-7958.

Каннан К.С. — профессор, факультет информатики и инженерии, Каласалингамская академия исследований и образования. Область научных интересов: машинное обучение, глубокое обучение. Число научных публикаций — 20. k.s.kannan@klu.ac.in; Шривиллипутур, Тамил Наду, 626126, Кришнанкойл, Индия; р.т.: +91(04563)289-042.