A Stacking-based Ensemble Framework for Automatic Depression Detection using Audio Signals

Suresh Mamidisetti[1]*, A. Mallikarjuna Reddy[2]
Research Scholar, Department of CSE, Anurag University, Telangana, India[1]
Associate Professor, Department of Artificial Intelligence (AI), Anurag University, Telangana, India[2]

Abstract*—Mental illnesses are severe obstacles for the global welfare. Depression is a psychological disorder which causes problems to the individual and also to his/her dependents. Machine learning based methods using audio signals can differentiate patterns between healthy and depressive subjects. These methods can assist health care professionals to detect the depression. Literature in depression detection, based on audio signals, used only single classifier, lacks to take advantage of diverse classifiers. The current work combines predictive capabilities of diverse classifiers using stacking method to detect depression. Audio clips are reordered while a predefined paragraph is being read out, for acoustic analysis of speech. The dataset is created which has extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features, that are extracted using openSMILE toolkit. The normalized feature vectors are given as input to multiple classifiers to give an intermediate prediction. These predictions are combined using a meta classifier to form a final outcome. K-Nearest Neighbours (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Trees (DT) classifiers are utilised on the normalized feature vector for intermediate predictions and Logistic Regression (LR) is used as meta classifier to predict final outcome. Our proposed method of using diverse classifiers achieved significant accuracy of 79.1%, precision of 83.3%, recall of 76.9% and F1-score of 80% on our dataset. Results are discussed while using stacking method on our dataset, then compared with various baseline methods also while applying on a publicly available bench marking dataset. Our results showed that combining predictive capability of multiple diverse classifiers helps in depression detection.

Keywords*—Health care; depression detection; acoustic features; speech elicitation; feature selection; openSMILE; ensemble methods

I. Introduction

Mental health disorders are key impediments to the global health agenda's progress (World Health Organisation, 2020). According to WHO data, about 280 million people suffered from depression in 2019 including 23 million children [1]. The COVID-19 epidemic has become a worldwide health disaster. The epidemic has resulted in various damages viz. massive employment loss, lowered earnings, halted children's education, and hindered economic progress. During COVID-19 pandemic, several lockdowns were imposed which made people to stay indoors, for quite a long period. Measures like stay-at-home orders, isolation and quarantine further resulted in emotional, financial risks and aroused several mental health disorders, mainly depression. Depression is one of the world's most accepted serious health issues. Depression is the root cause for triggering other mental health disorders, if not attended properly. Even it leads to psychosomatic disorders, in which actual physiological symptoms will surface, on diagnosis. Thus the detection of depression early and its treatment plays vital role in maintaining health. The study focuses on detection of depression, with more accuracy, with audio signals, using speech elicitation from the subjects.

Generally, there are two methods for depression detection. These are 1) based on standard questionnaires and 2) with the interaction of health care professionals. These methods have their own disadvantages. The method using standard questionnaires needs proper attention while administering them. Health care professionals like Psychiatrists, Psychologists, and Counselors involve upon their availability and expertise in their domain. With the rapid growth and advancement in Information Technology sector, there is an increasing interest in the researchers to develop Machine Learning (ML) models in health care domain for disease diagnosis. ML models learn the meaningful patterns from the individual biomarkers [2], [3]. The automatic depression detection method using audio is the 'Gold standard' among machine learning researchers. Such a method can be utilised as a pre-assessment test for a depression suspect (also named as client) at home. This can benefit users by saving time and cutting down expenses and travel costs which help in the speedy and efficient diagnosis. Subsequently, it results in timely medical care which can be given to identify clients, now named as patients [4], [5].

The studies of audio-based depression detection were started in the late 1980s. Speech has several qualities that make it a desirable candidate for integration into an automated assessment system, according to Scherer et al. [6]. It offers reasonably priced, remote, non-intrusive measurement capabilities. Due to the complexity of speech production, it acts as a sensitive output system, where even slight changes in physiology and cognition can produce audible changes in sound. It is conceivable that the acoustic quality of speech may be impacted in a quantifiable and objectively measurable way given the anticipated effects of depression on both cognitive and physiological aspects that influence speech control.

Studies in developing machine learning models to detect depressed and control subjects support that a person's verbal communication can indicate depression [7], [8]. During interaction between the clinicians and the subject (client), health care professionals rely on the subject's acoustic cues to assess their level of depression. Speech-based depression methods suggest that there exists a relationship between

acoustic speech and depression. Typically, the clinician observes acoustic speech (how they speak) as a main parameter [9]. So ideally, the automatic audio-based depression methods have to consider clinician's acoustic observations in the assessment of depression.

More over the use of audio-based approaches for diagnosing depression has the advantage of protecting individuals' privacy and identification. Visual-based approaches, on the other hand, cannot provide the same level of privacy and may reveal the person's identity known throughout the assessment process. The literature on depression detection methods demonstrates a strong preference for multimodal approaches [10], with audio-based methods receiving lesser attention. Rather than exploring on unimodal systems, researchers have largely concentrated on integrating many modalities. These multimodal systems incur costs for data collecting and computing time, which might impair processing speed. The existing studies also lack the larger dataset, and relatively lesser number of studies were reported using ensemble methods. The current study investigated on the stacking method of ensemble method which is not utilised in depression detection but utilised in other applications of medical diagnosis problems [11], [12], [13].

To overcome these issues we developed a method to create a dataset of voice samples. We have extracted numerous features by using state-of-the-art feature extraction tool kit. We have investigated with two ensemble techniques. Then, we conducted experiments with four developed models that classify depressed and healthy subjects. The present work investigates the acoustic components of speech, to design a comprehensive speech-based depression diagnosis method. An end-to-end machine learning solution was proposed to classify subjects' depressed and non-depressed classes. We believe that such a method can be more effective in the diagnosis of depression.

Main contributions of the current work are:

- To develop a ML model that will classify a subject as being healthy or depressed.
- To create a Dataset of voice samples using speech elicitation technique where subjects were instructed to provide voice samples in two different scenarios: i) while participants are reading out a phonetically balanced short story called as “The north and the south wind” ii) while participants are giving an open form of talk on the choice of their interest.
- To extract baseline features in acoustic speech recognition called as eGeMAPS feature set using state-of-the-art feature extraction toolkit called as openSMILE.
- The ensemble method called stacking was experimented and results were analysed on the created dataset. The performance is compared with the baseline methods and evaluated on the publicly available DAIC-WOZ dataset. Their results are reported.

II. Related Work

Speech plays a crucial role in detecting depression for the following reasons: First, the acoustic component of speech is often influenced by the subject's mental state. Second, the clinician observes acoustic speech characteristics to analyse the condition of the individual. Several studies have found distinguishable acoustic characteristics between non-depressed and depressed individuals. Third, ease of recording and the availability of open source tools for extracting acoustic features such as openSMILE, COVEREP,etc. [14], [15]. For instance, studies employing the acoustic tier of speech indicated that depressed subjects show variations in measures like pitch, intensity, energy, etc., compared with non-depressed speakers [16]. Studies on non-voiced aspects of speech like pauses, hesitations reveal that they do have an association with the depression.

The field of mental health care has seen promising applications for audio-based depression detection techniques. These methods examine audio recordings, such as speech patterns and emotional signals, to find probable symptoms of depression by utilising machine learning techniques. An inexpensive and non-intrusive method of screening people for depression symptoms is, automated diagnosis by audio analysis, which enables early intervention and support. Sentiment analysis is another application in speech, which enables the detection of unfavourable emotions and depressive symptoms, offering insightful information about a person's mental condition. Additionally, depression-related emotional patterns can be recognised using speech emotion recognition applications, allowing for a more thorough understanding of a person's emotional health [17].

Andrew et al. in [18] examined the potential gender bias in automated depression detection systems based on audio features. They analysed on a publicly available dataset called DAIC dataset of clinical interviews among participants with and without depression. They found that the classification accuracy is significantly higher for female participants compared to male participants, which is partly due to differences in the acoustic characteristics of male and female speech. Their work suggests the importance of considering gender bias in the development and evaluation of automated depression detection systems. Speech is used in prediction of not only depression but also in other diseases like dementia, stress disorders, parkinson’s disease etc. [19], [20].

Espinola et al. [21] recorded audio samples during the interaction of subjects (11-control and 22-depressed) with the psychiatrist. The samples were processed with the audacity audio software to eliminate the interviewer's voice signals and other noises. These samples were provided as input for an open-source vocal feature extraction tool called GNU Octave, to extract statistical features. Further, the Weka tool was used for analyzing the classification results using different ML classifiers. Among all classifiers that were experimented, random forest classifier achieved 87.5% accuracy. Their study concluded that the acoustic tier of voice samples is promising for depression detection.

Wegina et al. in [22] conducted a study on 144 volunteers, out of which 54 were diagnosed with depression and 90 as healthy subjects. Audio samples were recorded while

participants responded to a personal interview adapted by vocal screening protocol [23]. Subsequently, voiced responses of participants were also recorded while they were responding to the self-report questionnaires called Beck Depression Inventory (BDI) [24] and Self Reporting Questionnaire (SRQ) - 20 questionnaires [25]. The vocal samples were processed by the PRAAT tool for feature extraction. Acoustic parameters of speech, such as pitch, jitter shimmer, etc., were extracted using PRAAT. Then statistical features like median, mean and standard deviation, etc., were extracted from the acoustic parameters. A multiple linear regression model was used to demonstrate that acoustic parameters of speech were discriminative indicators for depression.

Thati et al. in [26] collected audio data through a mobile application that recorded participants' speech during a task-based depression assessment. They extracted various audio features, including pitch, loudness, and spectral features, and used machine learning algorithms to classify the audio recordings as either depressed or not. Their study found that audio features were effective in distinguishing between non-depressed and depressed participants, achieving an accuracy of 86.3%. Thati et al. in [27] extended their work with different fusion strategies to detect depression. In their work, they combined audio, video, text and also smart phone usage records to build a multimodal feature vector. Among the multimodal feature vectors audio features showed more correlation than the other modalities. Their work suggested that audio features are highly correlated with the depressive behaviour and resulted in efficient depression diagnosis.

Naulegari Janardha et al. in [28] proposed a novel method to enhance the prediction accuracy of depression using speech's acoustic features. Their study employed the Fisher score-based feature selection to choose the most informative features and dynamic ensemble selection to enhance the classification performance. They conducted experiments on a dataset of speech recordings from both depressed and non-depressed individuals. Their findings demonstrate that the proposed approach significantly outperforms existing methods, achieving an accuracy of 82.1%.

Brain sumali et al. in [29] used acoustic features to detect depression and dementia. Authors showed the statistical significance between the acoustic features and depression. In their work, they experimented with feature selection method called as Least Absolute Shrinkage and Selection Operation (LASSO) and SVM classifier with linear kernel to predict depression. They achieved high performance in terms of accuracy, sensitivity and specificity with age matched depressive subject predictions on both training and testing phases. Sara sardari *et al.* in [30] proposed deep learning based convolutional neural network based autoencoder method to detect depression. They showed the association of audio features and depressive behaviour with the help of most discriminative features. Authors used DAIC dataset for extracting hand crafted features then use deep learning based methods to detect healthy and depressive subjects. Their results suggested that audio features are helpful to improve the accuracy of the deep learning method with at least 7% of F-measure performance parameter.

Xiaolin mino et al. in [31] experimented with fusion of feature combining the higher-order spectral characteristics and standard speech features retrieved with classification weights. To validate the suggested features, traditional machine learning models such as support vector machine and k-nearest neighbour algorithms were used, along with convolutional neural network. The accuracies of speech-related feature extraction utilising the collaborative voice analysis repository, higher-order spectral analysis, and their fusion features were 63.15%, 68.42%, and 73.68%, respectively, using the support vector machine technique. Similarly, the equivalent accuracies using the k-nearest neighbour classification technique were 68.18%, 72.73%, and 77.27%, respectively. For the same characteristics, the convolutional neural network model resulted in accuracies of 70%, 77%, and 85%. These findings emphasise that fusion feature's achieved higher accuracy, which can improve the precision of depression recognition when using both classical machine learning and deep learning models.

Balijeet kaur et al. in [32] investigated on wide range of spectral, temporal, and spectro-temporal characteristics which were extracted from speech bio signals of both healthy and depressed subjects. Their work presented a two-stage technique that uses the Quantum-based Whale Optimisation Algorithm (QWOA) to determine the most significant and non-redundant speech variables for effective depression identification. The suggested strategy is tested against three proven univariate filtering approaches and four well-known evolutionary algorithms for feature selection on the DAIC-WOZ dataset. Their findings showed that the suggested model outperforms all univariate filter approaches and evolutionary algorithms, displaying greater performance while requiring lesser computer complexity than existing wrapper-based evolutionary methods.

Despite significant advancements in audio-based depression detection, there are few notable research gaps that warrant attention. Firstly, the majority of existing studies discussed above focus on using traditional machine learning techniques for feature extraction and classification. Secondly, most research works have been conducted on relatively small and homogenous datasets, limiting the robustness of the proposed methods and third, the ethical implications of using audio-based depression detection methods, includes privacy concerns and potential biases in the data, need to be addressed to ensure responsible and ethical deployment of such technologies in real-world settings. The current work aims at addressing these research gaps that will contribute to the development of more efficient, reliable, and ethically sound audio-based depression detection systems.

III. METHODOLOGY

Subjects are carefully chosen to represent the whole population and acoustic speech dataset is created using speech elicitation technique. The overview of the proposed work is represented in Fig. 1. The details are discussed in the following subsections.

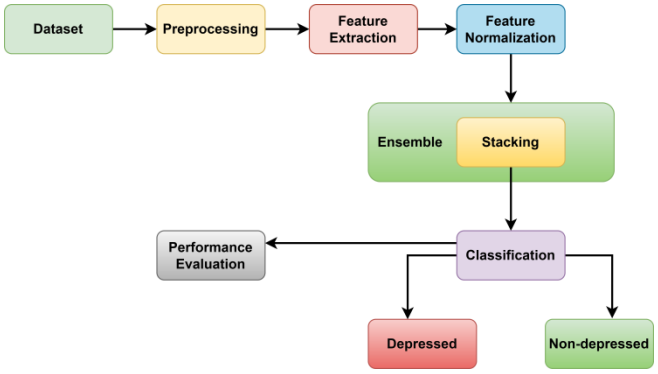

Fig. 1. Overview of the proposed work.

A. Dataset Construction

Students for participating in the dataset of the current work are recruited at Government polytechnic, Hyderabad, India. Out of the 225 students (here onwards called clients) who have accepted the participation in the experiment, 219 were actually turned up to read and understand English language.

Before initiating the dataset construction, all the clients (students) were explained the purpose and procedure of the study. Clients gave their written consents before they involve in dataset construction process. Approvals for conducting the study from the Institute were taken to conduct our study.

A recording room with good lighting condition is setup with a laptop on the table and a chair is placed one metre away from the laptop. Students were given prior information on their time slot to attend the study. Each student has been asked to give his voice samples in two phases. In the first phase, student is asked to read a short tale of "The north wind and south wind" which is shown on the monitor. This fable is taken because it has all the phonetically balanced sounds to stimulate the acoustic characteristics of the speech. In the second phase, they were asked to give an open form of speech of the topic of their interest in the choices that are displayed in the list. List has topics related to their goal in the life, dish they like the most, role model, worst situation in their life, etc. A total of approximately 11 hours of voice samples, from 219 participants have been recorded.

B. Data Preprocessing

The recorded voice samples are used in this pre-processing step. Here, the voices where other than the participant is found, they are cropped manually. The resultant voice samples are denoised using SOX software[1]. It is a cross-platform audio editing tool. SOX has a command-line interface, and is built using standard C language. The denoised voice samples are obtained at the end of this stage.

C. Feature Extraction

The denoised voice samples are used in this feature extraction step. Here, the denoised voice records are given as input for openSMILE toolkit for feature extraction. This tool is a state-of- the-art feature extraction toolkit. The eGeMAPS features are extracted in the current study. The next sections go into more detail about eGeMAPS and the openSMILE toolkit.

[1]https://sox.sourceforge.net/

1) Open SMILE: The open-source Speech and Music Interpretation by Large-space Extraction (OpenSMILE) [33] is an open-source toolkit for audio and speech processing developed by the Audio Communication Group at the Technical University of Munich. It is designed for feature extraction from speech signals and can be used for a wide range of tasks, such as speaker identification, speech recognition, and emotion recognition in speech. It provides a rich set of audio features that can be used for analysis, including low-level signal features such as spectral and cepstral coefficients, as well as higher-level features such as prosodic features, which capture characteristics such as pitch and speaking rate.

Open SMILE is written in C++, and includes a command-line interface for feature extraction, as well as a C++ API for integration with other software. It can run on Windows, Linux, and macOS operating systems. OpenSMILE is licensed under the GNU Lesser General Public License, which allows for free use and modification of the software. Overall, openSMILE is a powerful and flexible toolkit for audio and speech processing that has gained popularity in both academic and industrial settings. Its open-source nature and broad feature set make it a valuable resource for researchers and developers working in the field of speech processing. One of the key features of openSMILE is its modularity. The toolkit consists of a core engine that provides basic audio processing and feature extraction functions, as well as a set of modules that can be added to extend the functionality of the system. These modules include support for various audio formats, filtering and pre-processing, segmentation, and classification.

2) eGeMAPS Features: eGeMAPS [34] (extended Geneva Minimalistic Acoustic Parameter Set) is a widely-used feature set for speech analysis. It includes a comprehensive set of 88 high-level and low-level acoustic features that capture a wide range of properties of speech signals, such as voice quality, prosody, and spectral characteristics. The eGeMAPS feature set was designed to be a compact and efficient feature set for speech analysis, while still providing a rich set of information about speech signals. It includes low-level features such as spectral and cepstral coefficients, as well as higher-level features such as prosodic features, which capture characteristics such as pitch, speaking rate, and voice quality.

The eGeMAPS features contain 25 Low Level Descriptors (LLD) of three parameter groups: Frequency related parameters, Energy related parameters and Spectral related parameters. Frequency related parameters have 8 LLDs. They are pitch, jitter and formants frequency 1-3 and their bandwidths of voice samples. Energy related parameters include 3 LLDs, namely loudness, shimmer, and Harmonics-to-noise ratio (HNR) components of voice signals. Spectral parameter consists of 14 LLDs. They are alpha ratio, Hammarberg Index, Spectral Slope of 0-500 Hz and 500-1500 Hz, Formant 1, 2, and 3 relative energy, Harmonic difference between first harmonic (H1) and second harmonic (H2) energy, Harmonic difference H1 and third formant range (A3), MFCC 1-4 and spectral flux.

Now for these 25 LLDs arithmetic mean and standard deviation normalized by arithmetic mean functional are computed to form 50 components. Then for pitch and loudness, 8 different functional values are calculated. The functionals are as follows: 20^{th}, 50^{th} and 80^{th} percentile and their ranges, mean and standard deviation of falling components signals. A total of 58 components are computed. Now arithmetic mean of first four components of spectral parameters is computed. 66 components are computed in total.

Six temporal features are computed. They are: rate of loudness peaks, mean and standard deviation of voiced samples, mean and standard deviation of devoiced samples, syllable rate. Resultant is 72 features. Now arithmetic mean of spectral flux in non-voiced parts, mean and standard deviation of spectral flux and Mel-Frequency Cepstral Coefficients (MFCC) 1-4, the LLDs of spectral flux and MFCC 1-4 values in total 16 components. Total 88 feature vectors of eGeMAPS are formed.

3) Size of features and ground truth: The eGeMAPS features are extracted for the voice records to form acoustic characteristics of speech biomarkers. The purposes of extracting this feature vector are listed as follows: i) eGeMAPS features are published as baseline features since 2016 in AVEC depression detection challenge [35]. ii) lot of works carried out in depression detection using speech showed reliability of eGeMAPS features [28], [34]. iii) eGeMAPS features are used in similar works of healthcare and also in emotion recognition in the literature of speech based detection methods [36].

For each participant 88 features are formed. By the end of this step, 219 X 88 size matrix is formed. Our dataset has 89^{th} column as ground label. Ground truth is given by a Psychologist. Hence our dataset is clinically validated by the Psychologist forming more reliable and coherent dataset. Thus dataset contains 219 X 89 size matrix.

A Psychologist was recruited to label the subjects. Post audio collection, Psychologist gave 17-item questionnaire called as Hamilton Rating Scale for Depression (HRSD) [37]. Subjects were asked to fill the HRSD form. Then Psychologist validated the label as not depressed and mild depressed. In our experiment we could not find any cases of moderate, severe cases of depression. Out of 219 subjects 137 are labelled as non-depressed and 82 as depressed.

Our justification of choice of size of the dataset is based on a comprehensive related work section presented in the current study on depression detection using machine learning and audio-based approaches. As per the knowledge of the authors, cited studies that utilised sample sizes range from a few tens to hundred participants, thereby establishing a strength on context for our sample size choice. The selected sample size of 219 clients exceeds these cited studies ensuring that our study is adequately powered to detect meaningful effects and relationships between audio-based features and depression. While we acknowledge that larger sample sizes could potentially improve the generalizability of our findings, we believe that the current dataset size is appropriate for the initial validation of our proposed framework.

D. Feature Normalization

Feature normalization is a common preprocessing step in machine learning that involves scaling the input features to have a similar scale and distribution. The purpose of normalization is to improve the performance and stability of machine learning models by reducing the influence of features with large values and preventing some features from dominating others [38].

In the current study, each group of parameters has different values and ranges. Hence to normalize them into a standard set of values we used Min-max scaling mechanism. All the experiments were conducted using these normalized values in this study. In this technique, values are scaled to a fixed range. We used 0 to 1 range for normalizing the values.

E. Ensemble

The ensemble method in machine learning combines several classifiers predictive results into an optimal result [39]. The current work used stacking technique of the ensemble method to derive the optimal predictions. In stacking, multiple heterogeneous classifiers are used to form an intermediate result. These intermediate results are provided as input for the meta classifier to predict the final outcome. Stacking aims to improve the meta classifier's performance.

The mathematical equation for a stacking ensemble can be represented as follows:

Consider D = {(x1, y1), (x2, y2), ..., (xn, yn)} be the training dataset, where xi is the i-th instance and yi is the corresponding target value.

Let M = {M1, M2, ..., Mk} be a set of base models, where Mi is a machine learning model that maps an instance x to a predicted target value Mi(x).

Let f be the meta-model that takes the predictions of the base models as inputs and outputs a final prediction, i.e., f(M1(x), M2(x), ..., Mk(x)).

The stacking ensemble can be trained in two stages:

1) Base model training: Each base model Mi is trained on the training dataset D to predict the target values.

2) Meta-model training: The predictions of the base models on the training dataset D are used to train the meta-model f. Specifically, a new dataset D' is created by replacing the target values in D with the predictions of the base models, i.e., D' = {(M1(x1), M2(x1), ..., Mk(x1), y1), (M1(x2), M2(x2), ..., Mk(x2), y2), ..., (M1(xn), M2(xn), ..., Mk(xn), yn)}. The meta-model f is then trained on D' to predict the final target values.

The mathematical equation for the stacking ensemble can then be written as in (1):

$$f\big(M1(x), M2(x), \dots, Mk(x)\big) = g(w1 * M1(x) + w2 * M2(x) + \cdots + wk * Mk(x)) + b \quad (1)$$

where w1, w2, ..., wk are the weights assigned to the predictions of the base models, and b is the bias term. These

weights and bias term are learned during the meta-model training stage.

The block diagram of the stacking ensemble method is given in the following Fig. 2 and its explanation is as follows:

1) Input Data: It consists of training and test data, with the input data (training data) split into training and validation sets. The base models are trained using the training data. The validation data is used to assess the performance of the base models and choose the best-performing models. Once the meta-model is trained, it is used to make predictions on the test data. The test data is a distinct dataset from the training and validation sets that the model has never seen before.

2) Base models: The base models are the individual models that will be trained on the input data. These models can be of different types, such as Decision trees, SVMs, or Neural networks. Each model produces its own output or prediction for the given input.

3) Predictions: The trained base models produce outputs or predictions for each input in the validation set.

4) Meta-classifier: The meta-model is a model that combines the outputs of the base models to make a final prediction. It takes the outputs of the base models as input features and learns to combine them in an optimal way. The meta-model can be a simple linear model, such as Logistic regression, or a more complex model, such as a Neural network.

5) Final prediction: The meta-model produces a final prediction for each input in the test data. This prediction is based on the outputs of the base models and the learned combination of these outputs.

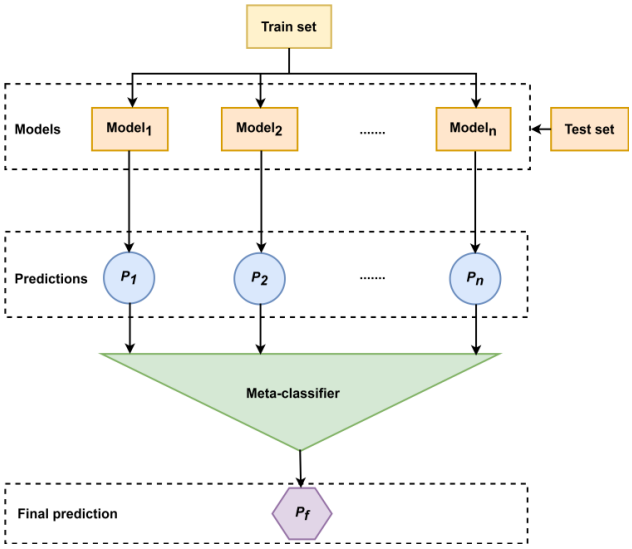

Fig. 2. Block diagram of the stacking ensemble method.

Algorithm for the proposed ensemble method:

The proposed ensemble method algorithm is a 7-step process for classifying acoustic data using a combination of base classifiers and a meta-classifier. Here is a brief explanation of each step as follows:

Step 1: Obtain the acoustic dataset which is of the size 219 X 89 matrix.
Step 2: Perform feature normalization using Min-max scaling technique.
Step 3: Divide the normalized feature vector to training set(80%) and test set(20%) without any overlap.
Step 4: Implement stacking with the base classifiers for training set obtained from step 3.
Step 5: Construct a meta classifier with the input from the outcomes of the base classifiers as features.
Step 6: Now using the test set, form intermediate outputs with the same classifiers. These intermediate outcomes are provided as input for the trained meta classifier to form final prediction.
Step 7: These final predictions are compared with the ground truths to check the performance of the ensemble method.

F. Performance Metrics

In the current work, the ensemble method's performance is evaluated utilising a variety of performance metrics. The performance metrics accuracy, precision, recall and F1-score are measured to show the effectiveness of the model.

The model's overall performance is evaluated using a confusion matrix. It contains the actual values and test predictions. The following Table I gives the details of the confusion matrix.

Confusion matrix contains four terms: True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN). TP and TN are correct predictions made by the model. TP and TN are correctly classified as compared with the actual ground truths. FN and FP are incorrect predictions made by the model. TN and TP are correct predictions where as FN and FP are the incorrect predictions. For any model, TP and TN must be high and FP and FN must be as low as possible. Then performance of the model is higher. The following Table II gives more details of the performance metrics used in the current study.

TABLE I. CONFUSION MATRIX

		Actual Values	
		Positive	**Negative**
Predicted Values	**Positive**	TP	FP
	Negative	FN	TN

TABLE II. PERFORMANCE METRICS UTILISED IN THIS WORK

Performance parameter	Description
Accuracy	It measures proportion of correctly classified instances in our dataset out of all the instances. $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$ (2)
Precision	It measures the accuracy of positive predictions by calculating the proportion of true positives among predicted positives. $\text{Precision} = \frac{TP}{TP + FP}$ (3)
Recall	It measures the ability of the model to identify positive instances by calculating the proportion of true positives among actual positives. $\text{Recall} = \frac{TP}{TP + FN}$ (4)
F1-score	It provides a balance between precision and recall by calculating the harmonic mean of the two metrics. $\text{F1-score} = \frac{2\,(\text{Precision X Recall})}{\text{Precision} + \text{Recall}}$ (5)

These metrics are essential for assessing the performance of the classification model in the context of depressed and non-depressed classification. Accuracy is a measure of the proportion of cases overall that were correctly identified as being either depressed or not. A high accuracy rating indicates that the model is adept at correctly classifying both depressed and non-depressed people, making it a useful indicator of the classification model's overall effectiveness. On the other hand, when dealing with depression detection, precision becomes especially important. It gauges how well positive predictions pan out, which in this instance translates to how well depressed people are identified. A high precision means the model has a low false positive rate and successfully recognises depressed people without mistakenly classifying healthy people as depressed.

Recall is extremely important since it measures how well the model can identify depressed people among all real-world depressed instances. A high recall means that the model successfully identified the majority of the depressed individuals in the dataset and has a low false negative rate. The F1-score offers a useful compromise between recall and precision. It provides a thorough assessment of the model's effectiveness in the classification of depression by taking into account both metrics. As a result, a classification model with high accuracy, precision, recall, and F1-score performs well in accurately identifying people with depression while minimising misclassifications, offering crucial help for those in need of mental health care and therapies.

IV. Results

This section discusses the results obtained in the current study. Jupiter python notebook is used for all the experiments conducted in the current study. The results are discussed in four forms: First, the effectiveness of the proposed model is evaluated in terms of various performance metrics. In order to show the performance gain of the proposed model, the performance of our model is then compared to the baseline models on our dataset. Third, the proposed model is applied to the benchmarking dataset to show the reliability of the study. In the end, the proposed model is compared with state-of-the-art baseline techniques on benchmarking dataset.

A. Performance Evaluation of the Proposed Model

In stacking protocol, first, the choice of the base classifier is made among all the available classifiers significantly important to acquire the performance improvement. In the present study, K-Nearest Neighbours (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Trees (DT) classifiers are selected as base classifiers. This choice is because of the following reasons: 1) they are easy to implement; 2) achieves higher accuracy when the size of the dataset is relatively small; and 3) conducted rigorous experiments and both performed relatively better on trial and error basis of all the classifiers. Thus when both these classifiers are combined, results in the performance enhancement.

Second, meta classifier uses the results of the diverse predictions obtained with the base classifiers. This will be helpful to avoid misclassifications of the model because feature size is reduced when compared with the original dataset size. In our study, the feature size is brought down to 2 from 88. This reduction helps in meta classifier to reduce mis-classifications and results in performance optimisation. Logistic Regression (LR) classifier is trained and tested as the meta classifier in our current work. Table III presents summary of the parameters used in the current study.

TABLE III. Summary of the Parameters of the Proposed Ensemble Method

Parameter	Tuned Parameter
Base Classifier-1	K-NN
Base Classifier-2	NB
Base Classifier-3	SVM
Base Classifier-4	DT
Meta Classifier	LR
Hyperparameters	Default values

Hence the proposed work integrates the predictive capabilities of five classifiers namely K-NN, Naïve Bayes, SVM, Decision Trees and Logistic Regression classifier, to obtain optimal results. Table IV shows the performance metrics obtained on our dataset using a stacking ensemble model. The four metrics evaluated are accuracy, precision, recall, and F1-score. This implies that the proposed framework can be utilized for clinical applications. These findings suggest that the proposed framework can be applied to the real-world scenarios. The hypothesis of the study is that the proposed framework holds as a valuable tool for aiding healthcare professionals in early diagnosis and intervention for depression. Comparable results obtained using accuracy, precision, recall, and F1-score values by our approach demonstrate its capability to precisely detect the depression in individuals.

TABLE IV. Performance Metrics Values of Proposed Stacking Method on Our Dataset

Performance metric	Value obtained on our dataset (in %)
Accuracy	79.1
Precision	83.3
Recall	76.9
F1-score	80.0

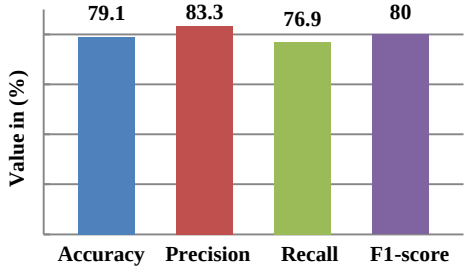

Fig. 3. Depression detection performance of proposed stacking ensemble model.

As presented in Fig. 3, our method of using diverse classifiers achieved accuracy of 79.1%, precision of 83.3%, recall of 76.9% and F1-score of 80.0%. These scores show that the performance of the proposed method is reliable. These results suggest that the model performed well on the dataset,

with high precision and F1-score values. However, it's important to note that the evaluation is based on a single dataset and may not generalize well to other datasets. Therefore, further testing and validation are necessary to confirm the effectiveness and robustness of the model. The following investigation is carried out to prove the reliability of the study.

B. Performance Comparisons of the Proposed Model with the Baseline Models

To show the performance gain of the proposed method we compared the proposed model's performance with the baseline methods. The baseline methods of the current work are the base classifiers used in the proposed model to build the meta-classifier. These methods are chosen because following this method of comparison gives deep understanding of the amount of performance gained with the stacking method over the base classifier utilisation. The feature set used in the current work is also used for training (80%) and testing (20%) base classifiers namely K-NN, Naïve Bayes, SVM, and Decision Trees classifiers. Table V contains details about comparison of the proposed stacking approach with the baseline methods.

Table V shows that the proposed method has outperformed the baseline methods in terms of all performance metrics. It contains a comparison of the performance metrics achieved by our proposed approach using ensemble and on using base classifiers such as K-NN, Naïve Bayes, SVM, and Decision Trees on our dataset. The reason behind choosing these classifiers is its vast use in the literature of the study (presented in the related work section). Our proposed approach outperformed the baseline methods in terms of Accuracy (79.1% vs. KNN: 62.5%, NB: 62.5%, SVM: 75.2%, DT: 76.8%) and Precision (83.3% vs. KNN: 61.5%, NB: 58.3%, SVM: 82.1%, DT: 80.7%), indicating its ability to correctly classify instances and minimize false positive predictions using our model. Although Recall values were similar between two baseline methods (our approach: 76.9% vs. KNN: 66.6%, NB: 63.4%, SVM: 77.2%, DT: 76.8%), our proposed approach achieved a higher F1-score (80.0%) compared to the baseline methods (KNN: 64.0%, NB: 60.8%, SVM: 76.3%, DT: 78.2%), signifying a better balance between Precision and Recall. These results suggest that the effectiveness of our proposed approach in accurately detecting depression and highlight its potential as a robust and reliable tool for classification of depression.

TABLE V. PERFORMANCE COMPARISON OF THE PROPOSED STACKING METHOD WITH THE BASELINE MODELS USING OUR DATASET

S.No	Method	Performance metrics			
		Accuracy	Precision	Recall	F1-score
1	K-NN	62.5	61.5	66.6	64.0
2	NB	62.5	58.3	63.4	60.8
3	SVM	75.2	82.1	77.2	76.3
4	DT	76.8	80.7	76.8	78.2
5	**Our proposed Method-Stacking approach**	79.1	83.3	76.9	80.0

C. Performance Comparison of the Proposed Model using Benchmarking Dataset

Now, the generality of the proposed model is tested. Our approach not only worked for our dataset but also can be applied to the other datasets. We applied our stacking method on the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database [40] which is publicly available benchmarking dataset in depression detection. This corpus contains audio and visual recordings of interviews between a real-life interviewee and a computer-generated interviewer who either acts as a supportive listener or a diagnostician. The participants were instructed to discuss few topics, including personal experiences and emotional states, allowing for the elicitation of genuine emotional expressions. Researchers who are interested in developing and validating depression detection models use the DAIC-WOZ database as a benchmark dataset. Additionally, the public dataset allows for comparisons and cross-validation of different methods, facilitating the advancement of automated depression detection technologies for real-world applications.

To apply our proposed stacking method, the audio clips of the subjects are used to generate the eGeMAPS feature vectors. Then after standardising the dataset stacking framework is applied.

TABLE VI. PERFORMANCE METRICS VALUES OF PROPOSED STACKING METHOD ON A DAIC-WOZ DATASET

Performance metric	Value obtained on our dataset (in %)
Accuracy	78.0
Precision	71.7
Recall	90.6
F1-score	79.4

Table VI shows the details of the performance of the proposed stacking method on the DAIC-WOZ dataset. Our method achieved comparable performance when applied on the benchmarking dataset. Among all the performance metrics recall gave a performance of 90.6% using the proposed method. From the Table VI it can be inferred that the proposed method can be applied to any dataset with similar audio cues.

D. Performance Comparison with State-of-the-Art Baseline Methods using Benchmarking Dataset

Table VII presents a comparison of performance metrics obtained by our proposed ensemble method on the public accessible benchmarking DAIC-WOZ dataset. We compared our results with two state-of-the-art methods: (1) Thati et al. in [27] used multimodal fusion approach by Late fusion using secondary SVM, and (2) Janardhan et al. in [28] used Fisher score-based feature selection method for classification of depressed or non-depressed individuals.

As presented in Fig. 4, our proposed ensemble method outperformed both existing state-of-the-art methods across all performance metrics. Specifically, our method acquired an accuracy of 78.0%, surpassing the late fusion using secondary SVM (74.0%) and Fisher score-based feature selection mechanism (74.0%). The precision of our method was 71.7%, significantly higher than the late fusion (52.0%) and also Fisher

score-based (49.0%) approaches. Also, our ensemble method demonstrated a remarkable recall of 90.6%, outperforming the late fusion (58.0%) and Fisher score-based (50.0%) methods in correctly identifying the depressed subjects. Lastly, F1-score of our proposed method was 79.4%, exceeding both the late fusion (73.0%) and Fisher score-based (74.0%) methods. These results clearly demonstrate the superiority of our ensemble method for depression classification on the DAIC-WOZ dataset, showcasing its effectiveness in achieving higher accuracy, precision, recall, and F1-score compared to the state-of-the-art methods mentioned.

TABLE VII. PERFORMANCE METRICS COMPARISON WITH THE STATE-OF-THE-ART BASELINE METHODS ON DAIC-WOZ DATASET

Performance metric	Our Proposed stacking method (in %)	Late fusion using secondary SVM	Fisher score-based feature selection
Accuracy	78.0	74.0	74.0
Precision	71.7	52.0	49.0
Recall	90.6	58.0	50.0
F1-score	79.4	73.0	74.0

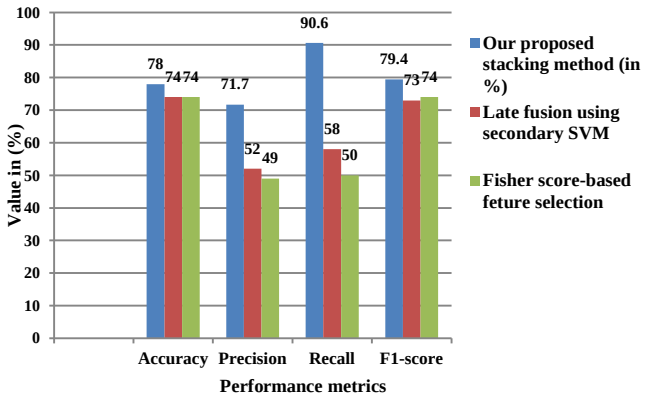

Fig. 4. Performance of proposed stacking method compared with existing Models on a DAIC-WOZ dataset.

V. CONCLUSION AND FUTURE WORK

Mental illness, manifestation of depression and anxiety, is a serious impediment to global well-being. Depression is an illness that can cause financial, social, family, emotional losses for individuals and leads to suicide causing sorrow to their families. A machine learning-based approach using audio can use patterns which are different in healthy and depressed subjects. These methods help healthcare professionals detect depression. The literature on depression detection through audio-based used only single classifier, but not the multiple classifiers. The present work combines the predictive patterns of different classifiers to detect depression using a stacked approach. Audio clips are recorded as subjects read predefined passages and free-form speech. Then eGeMAPS features are extracted using the openSMILE toolkit. The normalized feature vectors are trained with multiple classifiers. Then meta-classifier is also trained to form final predictions. On this audio dataset, our proposed method, which combines a variety of classifiers, produced impressive results with an accuracy of 79.1%, precision of 83.3%, recall of 76.9%, and an F1-score of 80.0%. Our method demonstrated a remarkable performance improvement in terms of accuracy when compared to single classifier technique. Identical performance improvements are seen across various evaluation parameters. Our method's validity on the DAIC-WOZ dataset further confirms its dependability and potency to generalize on other clinical applications. Notably, our method achieved a recall of 90.0%, demonstrating its potential as a global approach for depression detection and demonstrating its ability to generalize and be applicable to other datasets with similar audio cues.

Our future works aim to investigate the implementation of regression-based methods to identify not only the presence of depression but also the severity or levels of depression in individuals. The importance of expanding the dataset size by increasing the number of samples is recognised. Also research on linguistic characteristics and language pragmatics of speech, seeking to extract and utilize the valuable information beyond acoustic features for more accurate depression detection.

ACKNOWLEDGMENT

This project was supported by students of Government Polytechnic, Hyderabad, India, and the authors would like to thank them for their help.

REFERENCES

[1] WHO, "Mental disorders Fact sheets,". 2022. https://www.who.int/news-room/fact-sheets/detail/mental-disorders (accessed Jan. 05, 2023).

[2] N. K. Iyortsuun, S. H. Kim, M. Jhon, H. J. Yang, and S. Pant, "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis," Healthcare, vol. 11, no. 3, pp. 1–27, Jan. 2023, doi: 10.3390/healthcare11030285.

[3] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, "Leveraging bio-maximum inverse rank method for iris and palm recognition", International Journal of Biometrics, Vol. 14, No. 3/4, pp. 421-438, doi: 10.1504/IJBM.2022.124681.

[4] V. Ravi, J. Wang, J. Flint, and A. Alwan, "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement," in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Sept. 2022, pp. 3338–3342, doi: 10.21437/Interspeech.2022-10798.

[5] A. Mallikarjuna Reddy et al., "An Efficient Multilevel Thresholding Scheme for Heart Image Segmentation Using a Hybrid Generalized Adversarial Network", Journal of Sensors, pp. 1-11, Nov. 2022. doi: 10.1155/2022/4093658.

[6] K. R. Scherer Justus, "Vocal Affect Expression: A Review and a Model for Future Research," Psychological Bulletin, vol. 99, no. 2, pp. 143–165, 1986, doi: 10.1037/0033-2909.99.2.143.

[7] G. A. Miller, "Language and Communication,". MacGraw-Hill, 1973.

[8] S. Newman and V. G. Mather, "Analysis of spoken language of patients with affective disorders," American Journal of Psychiatry, vol. 94, no. 4, pp. 913–942, Jan. 1938, doi: 10.1176/ajp.94.4.913.

[9] S. Schoicket, R.A. MacKinnon, and R. Michels, "The Psychiatric Interview in Clinical Practice,". The Family Coordinator, vol. 23, no. 2, pp. 216-217, Apr. 1974, doi: 10.2307/581746.

[10] R. Kuttala, R. Subramanian, and V. R. M. Oruganti, "Multimodal Hierarchical CNN Feature Fusion for Stress Detection," IEEE Access, vol. 11, pp. 6867 - 6878, Jan. 2023, doi: 10.1109/ACCESS.2023.3237545.

[11] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," Journal of Imaging, vol. 6, no. 6, pp. 1-13, May. 2020, doi: 10.3390/jimaging6060039.

[12] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced

feature subset," Computer Methods and Programs in Biomedicine, vol. 198, pp. 1057-1070, Jan. 2021, doi: 10.1016/j.cmpb.2020.105770.

[13] Sudeepthi Govathoti et al., "Data Augmentation Techniques on Chilly Plants to Classify Healthy and Bacterial Blight Disease Leaves", International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 13, No. 6, Jun. 2022, pp. 131-139, doi: 10.14569/IJACSA.2022.0130618.

[14] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Sept. 2018, pp. 1716–1720, doi: 10.21437/Interspeech.2018-2522.

[15] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May. 2014, pp. 960–964, doi: 10.1109/ICASSP.2014.6853739.

[16] A. Esposito, G. Raimo, M. Maldonato, C. Vogel, M. Conson, and G. Cordasco, "Behavioral Sentiment Analysis of Depressive States," in 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Sept. 2020, pp. 209–214, doi: 10.1109/CogInfoCom50765.2020.9237856.

[17] A. Shatte, D. Hutchinson, and S. Teague, "Machine learning in mental health: a scoping review of methods and applications," Psychological Medicine, vol. 49, no. 9, pp. 1426-1448, Feb. 2019, doi: 10.1017/S0033291719000151.

[18] A. Bailey and M. D. Plumbley, "Gender Bias in Depression Detection Using Audio Features," in 29th European Signal Processing Conference (EUSIPCO), Aug. 2021, pp. 596–600, doi: 10.23919/EUSIPCO54536.2021.9615933.

[19] Y. Ozkanca, M. Göksu Öztürk, M. N. Ekmekci, D. C. Atkins, C. Demiroglu, and R. Hosseini Ghomi, "Depression Screening from Voice Samples of Patients Affected by Parkinson's Disease," Digital Biomarkers, vol. 3, no. 2, pp. 72–82, Jun. 2019, doi: 10.1159/000500354.

[20] D. Mizuguchi et al., "Novel Screening Tool Using Voice Features Derived from Simple, Language-independent Phrases to Detect Mild Cognitive Impairment and Dementia," pp. 1–10, May 2023, doi: 10.21203/rs.3.rs-2906887/v1.

[21] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study," Research on Biomedical Engineering, vol. 37, no. 1, pp. 53–64, Mar. 2021, doi: 10.1007/s42600-020-00100-9.

[22] W. J. Silva, L. Lopes, M. K. C. Galdino, and A. A. Almeida, "Voice Acoustic Parameters as Predictors of Depression," Journal of Voice, Aug. 2021, doi: 10.1016/j.jvoice.2021.06.018.

[23] A. A. F. de Almeida, L. R. Fernandes, E. H. M. Azevedo, R. S. de A. Pinheiro, and L. W. Lopes, "Characteristics of voice and personality of patients with vocal fold immobility," Codas, vol. 27, no. 2, pp. 178–185, Apr. 2015, doi: 10.1590/2317-1782/20152014144.

[24] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An Inventory for Measuring Depression," Archives Of General Psychiatry, vol. 4, no. 6, pp. 561–571, Jun. 1961, doi: 10.1001/archpsyc.1961.01710120031004.

[25] T. W. Harding et al., "Mental disorders in primary health care: a study of their frequency and diagnosis in four developing countries," Psychological Medicine, vol. 10, no. 2, pp. 231–241, May 1980, doi: 10.1017/S0033291700043993.

[26] R. P. Thati, A. S. Dhadwal, P. Kumar, and P. Sainaba, "A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms", Multimedia Tools and Applications, vol. 82, no. 4, pp. 4787-4820, Apr. 2022, doi: 10.1007/s11042-022-12315-2.

[27] R. P. Thati, A. S. Dhadwal, P. Kumar, and P. Sainaba, "Multimodal Depression Detection: Using Fusion Strategies with Smart Phone Usage and Audio-visual Behavior," International Journal on Artificial Intelligence Tools, vol. 32, no. 2, Apr. 2023, doi: 10.1142/s0218213023400080.

[28] N. Janardhan and N. Kumaresh, "Improving Depression Prediction Accuracy Using Fisher Score-Based Feature Selection and Dynamic Ensemble Selection Approach Based on Acoustic Features of Speech," Traitement du Signal, vol. 39, no. 1, pp. 87–107, Feb. 2022, doi: 10.18280/ts.390109.

[29] B. Sumali et al., "Speech Quality Feature Analysis for Classification of Depression and Dementia Patients," Sensors (Switzerland), vol. 20, no. 12, pp. 1–17, Jun. 2020, doi: 10.3390/s20123599.

[30] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using Convolutional Autoencoder," Expert Systems with Applications, vol. 189, pp. 1160-1176, Mar. 2022, doi: 10.1016/j.eswa.2021.116076.

[31] X. Miao, Y. Li, M. Wen, Y. Liu, I. N. Julian, and H. Guo, "Fusing features of speech for depression classification based on higher-order spectral analysis," Speech Communication, vol. 143, pp. 46–56, 2022, Sept. 2022, doi: 110.1016/j.specom.2022.07.006.

[32] B. Kaur, S. Rathi, and R. K. Agrawal, "Enhanced depression detection from speech using Quantum Whale Optimization Algorithm for feature selection," Computers in Biology and Medicine, vol. 150, pp. 106122, Sept. 2022, doi: 10.1016/j.compbiomed.2022.106122.

[33] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on Multimedia, Oct. 2010, pp. 1459–1462, doi: 10.1145/1873951.1874246.

[34] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, June 2016, doi: 10.1109/TAFFC.2015.2457417.

[35] M. Valstar et al., "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Oct. 2016, pp. 3–10, doi: 10.1145/2988257.2988258.

[36] B. Stasak, "An investigation of acoustic, linguistic, and affect based methods for speech depression assessment," Ph.D. dissertation, School of Elec. Eng. & Tele comms., Univ., New South Wales, Syd., Aus., 2018.

[37] J. Endicott, J. Cohen, J. Nee, J. Fleiss, and S. Sarantakos, "Hamilton Depression Rating Scale.," Archives Of General Psychiatry, vol. 38, no. 1, pp. 98-103, Jan. 1981, doi: 10.1001/archpsyc.1981.01780260100011.

[38] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in Proceedings of the 12th Annual Conference of the International Speech and Communication Association (INTERSPEECH), Aug. 2011, pp. 2997–3000.

[39] V. NavyaSree et al., "Predicting the Risk Factor of Kidney Disease using Meta Classifiers," in IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Oct. 2022, doi: 10.1109/MysuruCon55714.2022.9972392.

[40] J. Gratch et al., "The Distress Analysis Interview Corpus of human and computer interviews," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 2014, pp. 3123–3128.