How to Cite:
Janardhan, N. ., & Kumaresh, N. (2022). Major depressive disorder detection in polytechnic college students: A minimal acoustic-features-based solution with reading and free form-speech recording tasks. *International Journal of Health Sciences*, *6*(S1), 7492–7518.
https://doi.org/10.53730/ijhs.v6nS1.6596

Major depressive disorder detection in polytechnic college students: A minimal acoustic-features-based solution with reading and free form-speech recording tasks

Naulegari Janardhan
Department of Computer Science, School of Mathematics and Computer Sciences, Central University of Tamil Nadu, Thiruvarur, India
Corresponding author: janardhan17@students.cutn.ac.in

Nandhini Kumaresh
Department of Computer Science, School of Mathematics and Computer Sciences, Central University of Tamil Nadu, Thiruvarur, India

***Abstract*---**Over 322 million people worldwide are affected by depression, the leading cause of disability. Major depressive disorder (MDD) is the most frequent mental disorder and contributes significantly to the global disease burden. Current depression diagnoses, on the other hand, are beset by issues such as patient denial, clinical experience, and self-report bias. Early detection of depression can assist in lessening or even eliminating its detrimental effects. To help the conventional diagnostic approaches in psychiatry, there has been much research into automated depression prediction in recent years. Automated depression identification based on machine learning techniques can help disorder analysts diagnose depression more effectively. The literature on depression has proven that the acoustic features of a depressed individual vary from a normal individual. This paper aims to identify the minimal acoustic features that can be used to detect depression accurately and propose a majority voting classifier for detecting depression (MVCDD). MVCDD is designed with the base classifiers, including Random Forest Classifier (RFC), Support Vector Classifier (SVC), and Logistic Regression (LR). The experiments are performed on the data collected from the students of Government Polytechnic, Masabtank, Hyderabad, India. Results show that the proposed MVCDD model using the feature selection technique of Recursive Feature Elimination with Support Vector Machine (SVM-RFE) given the accuracy, specificity, and balanced accuracy of 81%, 94%, and 77%, respectively. These results comparatively outperform the existing state-of-the-art models

for depression detection. To evaluate the work 10-fold cross-validation technique is used.

***Keywords*---**Acoustic features, Depression, Feature selection, Majority voting, SVM-RFE.

Introduction

Since the last decade, the number of people affected by depression has risen. Depression, according to the World Health Organization (WHO), is a common mental condition that affects 322 million people worldwide [1]. Major Depressive Disorder (MDD) is a medical disorder that involves emotion and mood, as well as neurovegetative activities (such as hunger and sleep difficulties), cognition (such as improper guilt and feelings of worthlessness), and psychomotor activity (such as agitation or retardation) [2][3][4].

Machine learning techniques are being explored fast to provide technical assistance to psychiatrists in diagnosing depression. There are good publications on the most prevalent machine learning approaches for brain imaging categorization and prediction. A review of research found that magnetic resonance imaging data was utilized to differentiate MDD from other mood disorders and look for treatment outcome predictors for individual patients [5][6]. Using magnetic resonance imaging (MRI) and deep learning algorithms for diagnosis, neuroscientists can predict a positive rate of 81% for autism [7].

The depressed person tends to differ in speaking from a normal person; therefore, it can be an effective biomarker to identify depression [8][9]. In the last five years, there have been several attempts to diagnose depression using speech signals worldwide [10][11].

Feature Selection (FS) plays a crucial role in any predictive model of machine learning to select the significant features in predicting the target [12][13]. FS increases the model's performance by eliminating the less-important features and reducing the time complexity [14]. FS also helps avoid the curse of dimensionality [15][16]. Numerous features were extracted in previous studies to enhance the classification effect. Still, it is uncertain what acoustic features are optimal for detecting depression [17]. In spite of the continuing research, accurate diagnosis of depression using a limited number of speech features remains difficult. This work extracted several acoustic features from the speech recordings of the students of Government Polytechnic, Masabtank, Hyderabad, India, to ensure diversity in the feature subspaces. To overcome the issue of higher dimensionality, the highly correlated features among themselves are discarded. The SVM-RFE FS technique has been employed for the selection of essential features. Later, the classification performance of those features is examined against different machine learning models.

Further, a majority voting classifier model for depression detection (MVCDD) was proposed to achieve better classification performance. The classifiers used for this task are Random Forest Classifier (RFC), Support Vector Classifier (SVC), and

Logistic Regression (LR). These classifiers are chosen in this work as they are extensively used in the healthcare domain for depression detection [18][19][20][21][17].

The remainder of this paper is structured as Section 2 provides the related work; Section 3 describes the Participants used to experiment with the proposed work. Section 4 explains the proposed methodology and subsections, including feature extraction, and FS Section 5 describes how the classification is carried on to identify whether a person is depressed or not. Section 6 furnishes the experimentation outcomes and a discussion on that, and the last section offers the conclusion of the proposed work.

Literature survey

A recent study has indicated that using voice as an investigative and monitoring tool for depression holds promise [8][17][22][23]. As the human speech production system is complicated, acoustic changes in speech can be caused by even slight cognitive or physiological changes. Listeners can detect various distinct features in a depressed voice, according to [24], who conducted an initial assessment of patients with major depression.

Many speech characteristics have been investigated to identify depression. The work [25] established that the duration of the speech pause for 60% of depressed participants was extensively linked with their Hamilton Depression Rating Scale (HAM-D) score. Research [26] conducted a sequence of hearing experiments utilizing signals made of F0 contours to examine if analysts, none of whom knew anything about psychology, could discover if a particular contour is of a speaker who was diagnosed as depressed or had completed therapy claiming accuracy of 80 percent.

A previous study [27] found considerable variations in the duration of speech pauses between the spontaneous speech of their depressed group and the control group. Previous studies have shown that F0 is affected by variations in a person's agitation and anxiety [27], personality characteristics [28], and underlying mood [29]. However, few studies [30][31] have identified no significant relationship between F0 and depression.

The work in [32] examined vocal prosody and discovered that differences in fundamental frequency and delay of responses to interviewer questions could differentiate people with moderate/severe depression from those who were not. Research work of [33][34] worked with classification systems for spectral, glottal, and prosodic features.

Some studies, such as [35], discovered that variations in the rate of speech are more noticeable at the phoneme level and that employing phone-duration and phone-specific measures rather than a global speech rate yielded more significant associations between speech rate and depression severity.

Scherer's research [36][37][38] found a strong link between voice quality characteristics and depression levels. The works mentioned in [39][40][10]

identified depression using low-level descriptors (LLDs) and statistical features. Research [17][41][42][43][44] looked at the mel-frequency cepstrum coefficients (MFCCs) and found that the depression categorization recognition performance was statistically significant.

An investigation [31] discovered that depressed people have higher levels of energy in the glottal spectra. The work [33] employed fundamental frequency and speak-switch time as input components to a logistic regression classifier on 57 depressive patients. When it came to identifying participants who reacted or did not respond to a depression diagnosis, they got an accuracy rate of 79%.

Explorative study [45] studied the speech of adolescent boys and girls who had regular conversations with their parents (68 participants had depression, 71 participants were controls). They divided their acoustic characteristics into five categories: spectral, cepstral, prosodic, glottal, and Teager Energy operators (TEO), where the TEO is a non-linear energy operator. They obtained better results with sex-dependent models than with sex-independent models. Within a group of 20 depressive individuals, an investigation [46] discovered no significant differences between fundamental, prosodic aspects of single-dimensional and clinically identified subgroups. According to the authors, both the complexities of speech and depression necessitate a multivariate approach to diagnosing depression through speech features.

In [34], the authors have investigated 30 English-speaking participants and provided an ensemble classification method based on GMM classifiers which incorporated glottal and prosodic characteristics. They received a 74 percent classifier result.

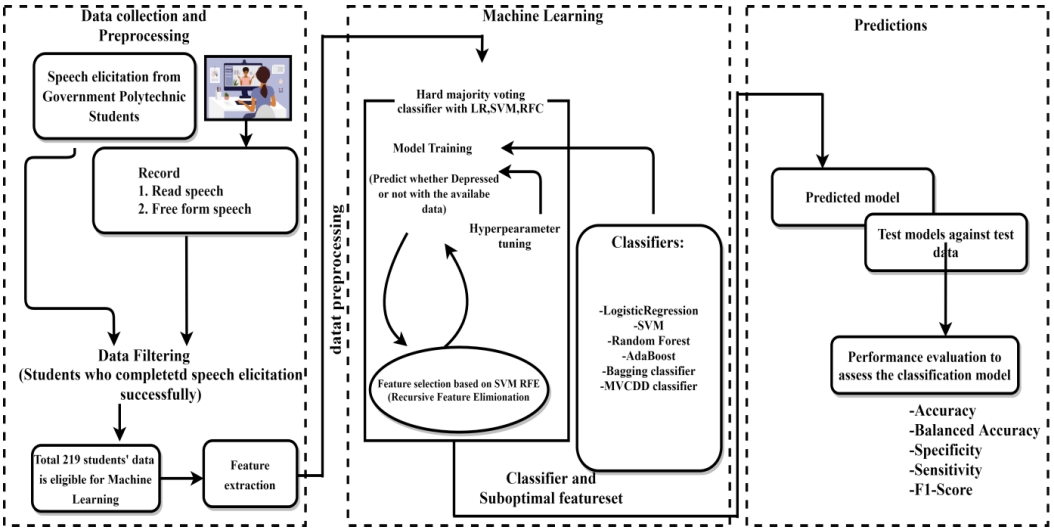

Figure 1. The complete framework of the proposed work

To identify depression from spontaneous speech, [39] selected three classifiers and a variety of feature sets. They discovered that loudness and intensity

variables were the most discriminatory, and pitch and formant features would be more relevant for intra-individual comparisons. They gathered 30 healthy people and 30 people with depression who could communicate in English. They concluded that GMM and SVM performed well in the classification.

Research [11] evaluated 35 people with Hamilton Depression Scale (HAMD) scores of less than seven or more than seventeen. They observed that SVM showed better performance than GMM for diagnosing depression severity, using related dynamics and the first three formant trajectories as characteristics. Another work [10] investigated 170 people and developed an SVM-based computational technique. They recorded a 75.96% of accuracy rate.

Therefore, upon the kin observation of the literature, it is found that there is a gap in research in achieving improved classification accuracy with a relatively small feature subspace. Therefore the current work focuses on tackling this point by attaining 81% classification accuracy only with the best ten features of the EMOBASE feature subspace with the proposed MVCDD model, utilizing the best available machine learning concepts.

Proposed method

The complete framework of the proposed work is presented in Figure 1. Various features are extracted from the speech recordings as part of the data collection. Data preprocessing is performed before moving on to the FS. Finally, classification models are applied to the features that have been chosen. The following subsections briefly explain these steps.

Participants

For this study, the data is collected from the students of Government Polytechnic College, Masabtank, Hyderabad, India. The dataset consists of the recordings of 219 subjects (110 female) with a mean age of 18.5 years; the total duration of all the recordings is 29.2 hours. The students were explained the purpose of the data collection and taken their signatures on the written informed consent before starting the process. Initially, each student was informed to fill out the PHQ-8 questionnaire. In major clinical investigations, the eight-item Patient Health Questionnaire depression scale (PHQ-8) has been proven as a viable diagnostic and severity assessment for depressive disorders. The total score (out of 24) of each student is calculated, and the labeling is done as “Majorly Depressed” if the score is >=10. Filling out the PHQ-8 questionnaire and labeling as “Majorly Depressed” or “Normal” is carried out in a psychiatrist’s presence.

Recording set-up:

Once the labeling was completed, the students were asked to enter the recording laboratory one after the other. The recording laboratory is set up in such a way that it is a closed room with adequate lighting conditions and a negotiable amount of noise. In order to reduce the channel-related effects and provide consistency throughout the data collection, a similar microphone and the recording equipment are used for all the students. A wearable microphone is

supplied to avoid recording errors caused by the slumped body posture, increased agitation, and psychomotor retardation.

Speech elicitation:

As part of the speech elicitation, the students were asked to read a phonetically balanced paragraph called "The North wind and the Sun" from Aesop's Fables. It is frequently employed in acoustic analysis in international multilingual clinical studies. The International Phonetic Association (IPA) recommends it for eliciting all phonemic contrasts that occur in English when testing global users or regional usage.

Read speech may not always accurately reflect the acoustic impacts of depression. Using free speech eliminates practice effects and may enable a broader range of induced emotional effects, such as asking participants to narrate situations that have elicited strong emotions. Many AVEC 2014 participants observed that free form speech performed better than reading speech in terms of predictor performance. The students were given a choice to record a free form speech, including the topics like their favorite holiday spot, whether advanced technology is either a boon or bane, a role model for their life, etc. A mixture of reading and free form speech is collected in this work, as each has its own set of benefits.

Feature Extraction

The openSMILE tool is used to extract the acoustic features from the speech recordings. The openSMILE (Munich open Speech and Music Interpretation by Enormous Space Extraction) technology was created to extract large audio feature spaces in real-time [48]. It is open-source software that provides different configuration files (*.config) that can be executed against the speech recordings depending on the required features. This paper uses eGeMAPS, EMOBASE, and Is09-13 config files. These features were chosen for their ability to index affective physiological changes in voice production, their relevance in previous research, and their theoretical significance. These three sets of features include prosodic, cepstral, and emotional characteristics of a speech.

Prosody is concerned with linguistic functions like intonation, stress, and rhythm, which are not individual phonetic segments but rather properties of syllables and larger units of an expression. The cepstrum is a metaphor used in homomorphic signal processing to transform convolutionally combined signals (such as a source and filter) into cepstra sums for linear separation. The power cepstrum, in particular, is often used as a function vector for portraying the human voice and musical signals. Emotion in speech can be thought of as a two-part communication system: the speaker's expression or representation of the emotion (the encoding) and the auditory signals (e.g., sound intensity) that reflect the perceived or expected emotion.

eGeMAPS features

It stands for the extended Geneva Minimalistic Acoustic Parameter Set. This feature set was inspired by an interdisciplinary gathering of voice and speech scientists in Geneva and improved at the Technische Universitat München (TUM) [49]. All 18 Low-level descriptors (LLD) are given functionals of arithmetic mean and coefficient of variation, generating 36 parameters. Loudness and pitch were calculated using the 20th, 50th, and 80th percentiles, the range of the 20th to 80th percentiles, and the standard deviation and mean of the slope of the rising/falling signal sections. As a result, there were a total of 52 parameters.

A total of 56 parameters are presented, including the arithmetic mean of the Alpha Ratio, the Hammerberg Index, and the spectral slopes from 0–500 Hz and 500–1500 Hz overall unvoiced segments. The number of continuous voiced regions per second, the rate of loudness peaks, the mean length and standard deviation of continuously voiced areas, the mean length and standard deviation of unvoiced regions, and the mean length and standard deviation of unvoiced regions were all included as temporal features. An extended minimalistic set is presented, which includes seven extra LLDs relating to MFCCs, spectrum flux, and formants bandwidth in addition to the 18 LLDs of the minimalistic set. In total, the eGeMAPS contains 88 parameters. Some of the prominent features and their descriptions are given in Table 1.

Table 1
Description of eGeMAPS features

Name of feature	Meaning
Pitch	logarithmic F0 on a semitone frequency scale
Jitter	individual F0 period length deviations
Formant 1,2,3 frequency	the centre frequency of the first, second, and third formant
Formant 1	the bandwidth of the first formant
Shimmer	the difference between consecutive F0 periods' peak amplitudes
Loudness	the estimate of perceived signal intensity from an auditory spectrum
Harmonics-to-Noise Ratio	the relationship between harmonic component energy and noise-like component energy
Alpha Ratio	ratio of total energy between 50 and 1000 Hz and 1 to 5 kHz
Hammerberg Index	the ratio of the 0-2 kHz region's strongest energy peak to the 2-5 kHz region's strongest peak.
Spectral Slope 0-500 Hz and 500-1500 Hz	within the two bands, the linear regression slope of the logarithmic power spectrum.
Formant 1,2,3 relative energy	the ratio of the energy of the spectral harmonic peak at the first, second, and third formants' center frequencies to the energy of the spectral harmonic peak at F0
Harmonic difference H1-H2	the ratio of the first F0 harmonic's (H1) energy to the second F0 harmonic's (H2) energy (H2)

Harmonic difference	the energy of the first F0 harmonic (H1) to the energy of
H1-A3	the third formant range's highest harmonic (A3)
MFCC 1-4	Mel-Frequency Cepstral Coefficients 1-4
Spectral flux	the difference between two consecutive frames' spectrum
Formant 2-3 bandwidth	Formant 1-3 parameters have been added for completeness.

Is09-13 features

The INTERSPEECH 2009 Emotion Challenge feature set is defined in this config file. There are 384 statistical functionals applied to LLD contours in this study [50]. The features are stored in .arff format (for WEKA), which allows adding additional instances to a file that already exists (This is meant for batch processing, in which the openSMILE is called repeatedly to extract the features from numerous files and save them to a single feature file). The suffix _sma added to the names of the low-level descriptors denotes that they were smoothed using a moving average filter with a window length of three. The suffix '_de' attached to the '_sma' suffix indicates that the current feature is a smoothed low-level descriptor's 1st order delta coefficient (differential). Because of the MFCCs' performance in speech, based on previous research [51] and the importance of prosodic features [52][53], the 16 LLDs are made up of 12 cepstral descriptors (MFCCs) and four prosodic descriptors [53]. The names of the 16 LLDs, the 12 functionals, and how they are found in the Arff file are given in Table2.

EMOBASE features

For emotion recognition, a total of 988 auditory features can be retrieved. The following LLDs are part of the feature set defined by emobase.conf: Pitch(F0), Probability of Voicing, F0 envelope, 8 Line Spectral Frequencies(LSFs), Zero-Crossing Rate, Intensity, Loudness, 12 MFCCs, 8 Line Spectral Frequencies(LSFs), Probability of Voicing, Zero-Crossing Rate, Pitch(F0), F0 envelope. The LLD and delta coefficients are subjected to the following functionals: max./min. value and relative location within the input, arithmetic mean, range, two linear regression coefficients and linear and quadratic error, skewness, kurtosis, quartile 1-3, standard deviation, and three inter-quartile ranges. There are 988 features in total [54]. The essential features of EMOBASE are listed in Table 3.

Table 2
Description of Is09-13 features

Name of feature	Meaning
pcm_RMSenergy	The energy of the root-mean-square signal frame
mfcc	MFCCs 1-12
pcm-zcr	Zero-crossing rate of time signal (frame-based)
voiceProb	voicing probability calculated from the Auto Correlation Function
F0	the fundamental frequency calculated from the cepstrum

max	the contour's highest value
min	the contour's lowest value
range	max minus min
maxPos	the largest value's absolute position (in frames)
minPos	the smallest value's absolute position (in frames)
amean	the contour's arithmetic
linregc1	the contour's slope (m) when approximated using a linear approximation
linregc2	the contour's linear approximation's offset (t)
stddev	the standard deviation of the contour's values
skewness	3rd order moment: spectral tilt of the long-term average spectrum
kurtosis	4th order moment: peakedness of the long-term average spectrum

Table 3
Description of EMOBASE features

Name of feature	Meaning
pcm loudness	the loudness as the normalized intensity raised to a power of 0.3
mfcc	15 mel-frequency cepstral coefficients
logMelFreqBand	log power of 8 mel-frequency bands distributed between 0 & 8 kHz
lspFreq	8 Line Spectrum Pair frequencies computed from 8 Linear Prediction Coefficients
F_0 final	the smoothed fundamental frequency contour
$F0_{nenv}$	the envelope of the fundamental frequency contour
jitterLocal	the local frame-to-frame jitter (pitch period length deviations)
jitterDDP	The differential frame-to-frame Jitter (the `Jitter of the Jitter')
shimmerLocal	the local (frame-to-frame) Shimmer
voicingFinal	the final fundamental frequency candidate's voicing probability
quartile1	1^{st} quartile: 25 percentile
quartile2	2^{nd} quartile: 50 percentile
quartile3	3^{rd} quartile: 75 percentile
iqr1-2	quartile2-quartile1: The inter-quartile range
iqr2-3	quartile3-quartile2
iqr1-3	quartile3-quartile1

The summary of the features is shown in Table 4, where the total number of features extracted is 1460. The feature subspaces are composed to experiment with different features and the combinations of the features. Hence, the seven unique feature subspaces are composed of three chosen feature sets, as shown in Table 5.

Table 4
Summary of acoustic features extracted for this work

Baseline feature category	# of features
eGeMAPS	88
Is09-13	384
EMOBASE	988
Total	1460

Data Preprocessing

There were no missing values for any feature that was extracted. The normalization technique is used to convert the numeric columns to a standard scale; in our case, as all the columns are of type numeric, the values are transformed to be between 0 and 1. Normalization is performed using (1).

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

All of the features will be transformed to have the properties of a standard normal distribution with a mean (μ) =0 and a standard deviation (σ) = 1 with standardization. The formula for standardization is given in (2).

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

The Weka tool normalizes and standardizes the data [55]. The University of Waikato created WEKA, which stands for "Waikato Environment for Knowledge Analysis," a Java-based open-source Data Mining tool. It's a mix of machine learning algorithms and data preprocessing software. WEKA provides machine learning algorithms that are simple to apply to the dataset.

After the preprocessing is completed, the machine learning models for classification can be applied. The total number of features extracted for this work is 1460, which, when used, machine learning models suffer from high dimensionality. As this work aims to find out the minimal suboptimal feature set that can classify the participants as depressed or not with increased accuracy, FS is vital.

Table 5
Feature subspaces composed of the baseline features

#	Name of Subspace	#	Name of Subspace	#	Name of Subspace
1	eGeMAPS	2	Is09-13	3	EMOBASE
4	eGeMAPS + Is09-13	5	eGeMAPS + EMOBASE	6	Is09-13 + EMOBASE
7	eGeMAPS + Is09-13 + EMOBASE				

Feature selection

When creating a predictive model, the count of input variables is important. The process of reducing the number of input variables is known as FS. The number of input variables should be reduced to minimize the model's computational cost and, in some cases, to increase the model's performance. FS can be made in multiple ways, but three broad categories exist.

SVM-RFE wrapper method

A wrapper method uses the output of a single machine learning algorithm as an evaluation criterion which means the features are supplied to the chosen machine learning algorithm and then added/removed features based on the model's output. Backward Elimination, Forward Selection, Bidirectional Elimination, and Recursive Feature Elimination (RFE) are some of the wrapper methods available. The RFE method eliminates attributes recursively and constructs a model on the remaining ones. The accuracy metric is used to rank the features in order of significance. The RFE method takes two parameters into account: the model to be used and the number of necessary features. It ranks all of the variables, with '1' being the most significant. It also provides support by giving 'True' for important and 'False' for irrelevant features. We need to determine the optimal number of features for the best accuracy. This is accomplished by employing a loop that begins with one function, progresses to the total number of features, and then chooses the one with the highest accuracy. The optimal number of features, in this case, is ten.

Later, ten is given as the number of features to RFE and obtained the final collection of features from the RFE method. When the features from the three subspaces are combined, it gave the feature subspace of 1046 features, out of which the SVM RFE FS chose the best ten features. The SVM-RFE FS selects the best ten features from three baseline feature subspaces and a feature subspace with the three combined feature subspaces, as shown in tables from Table 6 to Table 9.

Assume the training samples$(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$; $x_i \in R^n$; $y_i \in [-1, +1]$ with a target label y_i
The equation in (3) obtains the SVC discrimination function.

$$f(x) = w^T x + b \tag{3}$$

where x is the input sample, b is a bias, and w^T represents the weight vector calculated by (4).

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{4}$$

where, α_i is the Lagrange multiplier or the support vector.

Table 6
EMOBASE features extracted using SVM-RFE FS

S.No.	Feature
1	mfcc_sma[4]_max
2	mfcc_sma[10]_linregc1
3	lspFreq_sma[2]_max
4	lspFreq_sma[5]_linregerrQ
5	mfcc_sma_de[8]_minPos
6	mfcc_sma_de[9]_kurtosis
7	mfcc_sma_de[11]_max
8	lspFreq_sma_de[3]_quartile2
9	lspFreq_sma_de[4]_amean
10	F0_sma_de_irq2-3

Table 7
Is09-13 features selected using SVM-RFE FS

S.No.	Feature
1	pcm_fftMag_mfcc_sma_de[9]_amean
2	pcm_fftMag_mfcc_sma_de[4]_max
3	pcm_fftMag_mfcc_sma[11]_linregc2
4	pcm_fftMag_mfcc_sma[9]_maxPos
5	pcm_fftMag_mfcc_sma[7]_kurtosis
6	pcm_fftMag_mfcc_sma[6]_min
7	pcm_fftMag_mfcc_sma[5]_kurtosis
8	pcm_fftMag_mfcc_sma[4]_linregc1
9	pcm_fftMag_mfcc_sma[4]_amean
10	pcm_fftMag_mfcc_sma[2]_linregc2

Table 8
eGeMAPS features selected using SVM-RFE FS

S.No.	Feature
1	slopeUV0-500_sma3nz_amean
2	spectralFluxV_sma3nz_stddevNorm
3	hammarbergindexV_sma3nz_amean
4	F3amplitudeLogRelF0_sma3nz_stddevNorm
5	F3amplitudeLogRelF0_sma3nz_amean
6	F3bandwidth_sma3nz_stddevNorm
7	F1frequency_sma3nz_stddevNorm
8	F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope
9	F0semitoneFrom27.5Hz-sma3nz_meanFallingSlope
10	F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope

Table 9
Features selected from the combined feature subspace eGeMAPS+Is09-13+EMOBASE using SVM-RFE FS

S.No.	Feature
1	mfcc_sma_de[6]_maxPos
2	voiceProb_sma_quartile1
3	mfcc_sma[5]_iqr1-2
4	pcm_fftMag_mfcc_sma_de[3]_linregc2
5	F0_sma-linregc1
6	pcm_fftMag_mfcc_sma[9]_maxPos
7	pcm_fftMag_mfcc_sma[6]_skewness
8	pcm_fftMag_mfcc_sma[6]_min
9	pcm_fftMag_mfcc_sma[4]_linregc1
10	pcm_fftMag_mfcc_sma[2]_min

Classification

The aimed problem is a binary classification to classify the participant as depressed or non-depressed based on the chosen feature subspace. To find the suboptimal features and improve the classification results, every feature subspace is tested against the three selected classifiers and the proposed MVCDD ensemble classifier. The proposed algorithm is given as Algorithm 1.

Random Forest Classifier (RFC)

It is a supervised machine learning algorithm that can solve classification and regression problems. It consists of a large number of decision trees that are built on the different subspaces of the dataset. Internally, it follows the technique of ensemble learning to solve complex problems. Rather than depending on the prediction of a single tree, it considers the predictions from every binary tree built and follows majority voting to give the final prediction. The significance of the nodes is calculated for each decision tree using the Gini index, which is calculated by (5).

Algorithm 1: Majority Voting Classifier for Depression Detection (MVCDD)
Input: from students' speech recordings, training speech recording $\{s_1,s_2,\ldots,s_n\}$ and their labels $\{l_1,l_2,\ldots,l_n\}$ and testing speech features $\{t_1,t_2,\ldots,t_m\}$.
Output: Whether a depressed participant or a healthy control, with the labels of $\{t_1,t_2,\ldots,t_m\}$.
//process of training
Begin
Step 1: extract eGeMAPS, IS09-13, and EMOBASE features of every recorded speech from $\{s_1,s_2,\ldots,s_n\}$, and figure the speech statistics as 88 features of eGeMAPS, 384 features of Is09-13 , and 988 features of EMOBASE,
Step 2: feature subspaces are created as $\{X^{(1)},X^{(2)},\ldots X^{(7)}\}$ where $\{X(k)=x1^{(k)}, x2^{(k)},\ldots,xn^{(k)}\}$. Feature subspaces are formed as eGeMAPS, Is09-13, EMOBASE, eGeMAPS+Is09-13, eGeMAPS+EMOBASE, Is09-13+EMOBASE, eGeMAPS+Is09-13+EMOBASE.
For a=1 to 7
Step 3: reduce feature dimensionality for $X(k)$ using Pearson correlation.
Step 4: perform SVM Recursive Feature Elimination (RFE) FS until the suboptimal number of features with better accuracy occurs to get the classifier model for training.
End
// process of Testing
Begin
Step 1: extract eGeMAPS, IS09-13, and EMOBASE features of every recorded speech from $\{t_1,t_2,\ldots,t_m\}$ and figure out the feature statistics as Step 1.
Step 2: feature subspaces are created as $\{D^{(1)},D^{(2)},\ldots,D^{(7)}\}$, where $\{D(k)=d_1^{(k)},d_2^{(k)},\ldots,d_n^{(k)}\}$. Subspaces are formed similar to the Step 2.
For a=1 to 7
Step 3: reduce feature dimensionality for $D(k)$ using Pearson correlation.
Step 4: perform SVM Recursive Feature Elimination (RFE) FS until the suboptimal number of features with better accuracy is given for testing the model.

Step 5: depending on the classifier model for training, calculate the probabilities of whether the samples being tested belong to a depressed participant $\{p_1,p_2,\ldots, p_m\}$ or a healthy participant $\{q_1,q_2,\ldots,q_m\}$, then output the category label depending on greater probability.
End

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \quad (5)$$

where P denotes the class's relative frequency, and c indicates the number of classes.
Entropy can also be employed in a decision tree to branch the nodes. The entropy can be determined using (6).

$$Entropy = \sum_{i=1}^{c} -P_i * \log_2(P_i) \quad (6)$$

where P denotes the class's relative frequency, and c represents the number of classes.

Support Vector Classifier (SVC)

SVC is a widely used machine learning algorithm that works in various classification tasks. Using the distance margin or the distance between two support vectors, it builds a hyperplane to partition the dataset into several classes. Different kernel functions produce the optimal hyperplane, which demands transferring the data to higher dimensions in most situations. The three kernel functions used with SVC are radial basis functions (RBF), polynomial, and linear kernel. The procedure for SVC is as presented below.
Transform the original dataset into a higher-dimensional space using the Gaussian radial bias function kernel, as shown in (7).

$$K(X_i, Y_j) = e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}} \quad (7)$$

The decision function is calculated as follows:
At first, construct the separating hyperplane using (8).

$$S = x + \sum_{j=1}^{l} x_i y_i \quad (8)$$

Where x is a scalar, y is the attribute value, l is the number of attributes, and S is a separating hyperplane.

The data point is above the hyperplane S, if $S > 0$, and lies below the hyperplane S, if $S < 0$.
Weights are adjusted to yield the hyperplane defining the sides of the margin.

$$M_1 = x_0 + \sum_{j=1}^{l} x_i y_i \geq 1 \text{ for } S_j = +1 \quad (9)$$

$$M_2 = x_0 + \sum_{j=1}^{l} x_i y_i \leq 1 \text{ for } \quad S_j = -1 \quad (10)$$

The support vectors are found if a data point satisfies (11).

$$S_j = \left(x_0 + \sum_{j=1}^{l} x_j y_j\right) \geq 1 \; \forall j \quad (11)$$

The maximum margin is calculated as $\frac{2}{\|x\|}$

where Euclidian norm$\|x\| = \sqrt{\sum_{j=1}^{l} x_i^2}$ (12)

The equation for S_j is translated using the Lagrangian formulation and resolved using Karush Tucker conditions, commonly known as first-order derivative tests, to get the hyperplane with the most significant margin. The decision boundary obtained as a result is shown (13).

$$Dec(Y^T) = \sum_{j=1}^{l} S_j b_j Y^j Y^T + x_0 \quad (13)$$

where b, x_0 are the numeric parameters that are acquired from SVC optimization.
Y^T is a testing sample, S_j is a class label of j^{th} sample.
If $Dec(Y^T) > 0$ is considered the sample as positive; otherwise, it is regarded as a negative sample. At last, it predicts the class labels depending on the decision boundary.

Logistic Regression (LR)

It's a classification method that employs a sigmoid function to describe a dichotomous dependent variable with a value between 0 and 1. In LR, the sigmoid function is often an S-shaped curve that emits one when the value is less than 0.5 and zero otherwise. The input to a sigmoid function is computed using the linear regression function. The gradient descent approach is used to approximate the parameters of a linear function, such as weight and bias, using the cost function. The outline of the LR principle is explained below.
Calculate the logistic regression function as

$$z = \beta_0 + \sum_{j=1}^{l} \beta_j y_j \quad (14)$$

Where l is the number of attributes,β_0, β_j are scalar, weight vector, and y_j represents the data sample.
Compute the Predictive probabilities with (15).

$$p_{(\beta_0,\beta_j)}(z) = p_\theta(z) = \frac{1}{1+e^{-z}} \quad (15)$$

Using the cross-entropy, compute the cost function with (16).

$$K(\beta_0, \beta_j) = K(\theta) = \frac{1}{L}\sum_{j=1}^{l} \left[m_j log\left(p_\theta(y_j)\right) + (1 - m_j) log\left(1 - p_\theta(y_j)\right)\right] \quad (16)$$

Update the bias and weights using the gradient descent method.

$$\beta_j = \beta_j - \alpha d\beta_j \quad (17)$$

$$\beta_0 = \beta_0 - \alpha d\beta_0 \quad (18)$$

The estimated values $\beta_0$and β_j are used for the prediction of the test data.
Then, compute the linear equation for the test data as (19).

$$z = \beta_0 + \sum_{j=1}^{l} \beta_j y_j \quad (19)$$

Now, get the probabilities using the sigmoid function

$$p_{(\beta_0,\beta_j)}(z) = p_\theta(z) = \frac{1}{1+e^{-z}} \quad (20)$$

Finally, using the decision boundaries as a guide, translate the probabilities into class labels (21).

$$\widehat{lab} = \begin{Bmatrix} 1; & p_\theta(y) \geq 0.5 \\ 0; & p_\theta(y) < 0.5 \end{Bmatrix} \quad (21)$$

where $\widehat{lab}$ is the predicted label or the class.

Majority voting classifier

This work uses hard/majority voting to sum up the predictions for each class label and predict the final class label with the most votes. In hard majority voting, it predicts the class label $\hat{y}$ using majority voting of each classifierC_j:

$$\hat{y} = mode\left\{C_{1_{(x)}}, C_{2_{(x)}}, \dots, C_{m_{(x)}}\right\} \quad (22)$$

Assuming that three classifiers considered for this work classify a training sample as below:

Classifier 1 predicts label as 1
Classifier 2 predicts label as 0
Classifier 3 predicts label as 1

The final predicted label using the majority voting classifier will be:

$$\hat{y} = mode\{1,0,1\} = 1 \quad (23)$$

Using a majority vote classifier would classify the participant with “label 1”, which is “Depressed”. To evaluate the classification performance of the proposed method, balanced accuracy (Bal. Accu.), as the dataset is imbalanced, accuracy (Accu.), specificity (Spec.), precision (Prec.), sensitivity (Sens.), and F1-score metrics are used. The performance metrics are defined below, denoting true positive, true negative, false positive, and false negative as TP, TN, FP, and FN.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (24)$$

$$specificity = \frac{TN}{TN+FP} \quad (25)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (26)$$

$$balanced\ accuracy = \frac{sensitivity+specificity}{2} \quad (27)$$

$$F1-score = \frac{2*TP}{2*TP+FP+FN} \quad (28)$$

where True Positives (TP): A participant is “depressed”, and the prediction is “depressed” as well.

True Negatives (TN): A participant is “non-depressed”, and the prediction is “non-depressed” as well.

False Positives (FP): A participant is ‘non-depressed,’ but the classifier predicts that he or she is depressed (this is referred to as a “Type I mistake”).

False Negatives (FN): A participant is genuinely depressed, although the classifier predicted that he or she was not (this is known as a “Type II mistake”).

Results and Discussion

Initially, the filter-based FS is performed using Pearson correlation to eliminate the dimensionality curse. The features with 80% of correlation have been removed, and checked the performance metrics. Features chosen based on the Pearson correlation method did not give satisfactory classification results. It has given the highest accuracy of 70% with RFC on eGeMAPS+EMOBASE features. Therefore the SVM-RFE technique is used to select more significant features.

When the best ten features chosen using the SVM RFE FS method are supplied to the classifiers, they give the classification results as shown in Tables, from Table 10 to Table 12. These results are acquired using default parameters, without any hyperparameter tuning, of all the chosen classifiers.

Table 10
Performance of RFC on the features selected using SVM-RFE FS technique

Feature subspace	Bal. Acc. (%)	Accu. (%)	Spec. (%)	Prec. (%)	Sens. (%)	F1- score (%)
eGeMAPS	50	66	90	40	10	64
IS09-13	57	71	92	53	23	78
EMOBASE	63	76	94	71	33	70
eGeMAPS+IS09-13	60	74	96	68	23	75
eGeMAPS+EMOBASE	59	72	91	59	28	72
IS09-13+EMOBASE	56	70	91	53	21	75
eGeMAPS+IS09-13+EMOBASE	57	71	92	60	22	70

Table 11
Performance of SVC on the features selected using SVM-RFE FS technique

Feature subspace	Bal. Acc. (%)	Accu. (%)	Spec. (%)	Prec. (%)	Sens. (%)	F1- score (%)
eGeMAPS	51	70	99	1	36	72
IS09-13	57	72	96	79	17	75
EMOBASE	63	75	95	75	30	72
eGeMAPS+IS09-13	61	75	96	80	27	81
eGeMAPS+EMOBASE	65	**77**	96	81	34	72
IS09-13+EMOBASE	59	72	92	60	26	72
eGeMAPS+IS09-13+EMOBASE	58	72	93	72	24	75

Table 12
Performance of LR on the features selected using SVM-RFE FS technique

Feature subspace	Bal. Acc. (%)	Accu. (%)	Spec. (%)	Prec. (%)	Sens. (%)	F1- score (%)
eGeMAPS	52	70	99	1	5	72
IS09-13	54	71	98	1	10	72
EMOBASE	57	72	97	1	15	70
eGeMAPS+IS09-13	55	72	99	1	10	78

eGeMAPS+EMOBASE	54	71	98	1	10	70
IS09-13+EMOBASE	61	72	89	59	33	70
eGeMAPS+IS09-13+EMOBASE	53	70	96	75	10	75

SVM-RFE FS technique yielded relatively highest accuracy of 77% and the balanced accuracy of 65% for the feature subspace of eGeMAPS+EMOBASE with the SVC classifier, as depicted in Figure 2.

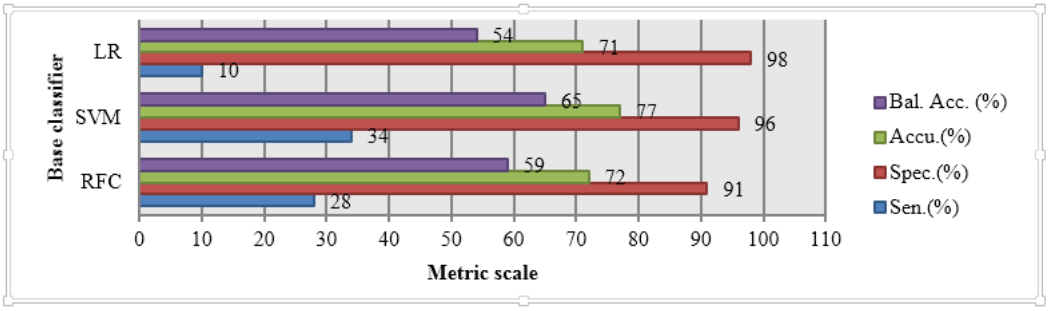

Figure 2 Performance of the base classifiers on the eGeMAPS+EMOBASE feature subspace

The hyperparameter tuning of the chosen classifiers is performed to improve the classification performance. The "GridSearchCV" method is used for tuning each classifier parameter with all the possible values, and the parameters are chosen based on the performance. By looping through predefined hyperparameters, this function assists in fitting the estimator (or model) to the training set. Finally, among the hyperparameters provided, the optimal parameters are chosen. The final hyperparameters for the three selected models are presented in Table 13 to Table 15.

Table 13
RFC parameter values

RFC Parameter	Value
n_estimators	1000
criterion	gini
oob_score	bool
bootstrap	True
random_state	42

Table 14
SVC parameter values

SVC Parameter	Value
kernel	poly
C	1
degree	6

coef0	4.001
gamma	auto

Table 15
LR parameter values

LR Parameter	Value
penalty	l2
dual	FALSE
tol	0.0001
C	51.79474679
solver	liblinear

Cross-validation (cv) is a method of evaluating the classifiers on the limited sample data. The stratified sampling method can deal with the imbalanced classes in the dataset. If the classes are imbalanced, the classification results can be misled or potentially fail in severe imbalanced classes. Hence, stratified sampling is used to avoid this, provided with the Stratified K-fold cv technique. With this technique, the distribution of the class in each split of the data will be enforced to match the distribution in the entire training dataset. This work uses a *10-fold* cv strategy to maintain the ratio between the classes in each fold. The classification results after hyperparameter tuning are tabulated in tables from Table 16 to Table 18.

The graph showing the performance metrics accuracy, specificity, and sensitivity of the hyperparameters tuned classifiers on EMOBASE feature subspace of the best ten features is shown in Figure 3. The SVC classifier gives the highest accuracy of 79%.

The investigational outcomes show that the feature subspace of the ten best features from the EMOBASE feature set gives promising results; hence, the ensemble models are also tested on this feature subspace. AdaBoost and Bagging classifiers are experimented with along with the proposed MVCDD model. The proposed model's performance metrics and the ensemble models are furnished in Table 22.

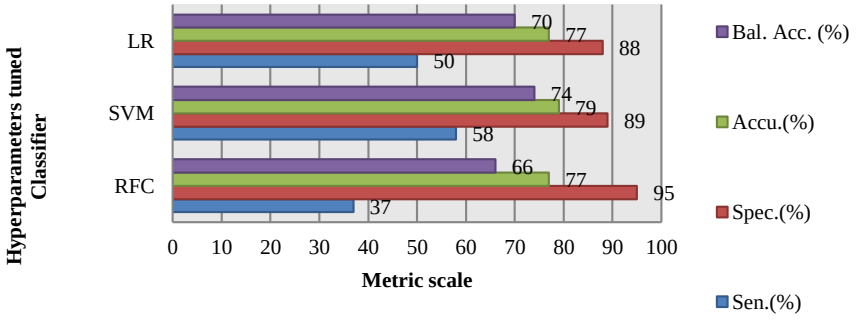

Figure 3 Performance of hyperparameters tuned classifiers on EMOBASE feature subspace

Table 16
Performance of Hyperparameters tuned RFC on the features selected using SVM-RFE FS technique and stratified 10-fold cv

Feature subspace	Bal. Acc. (%)	Accu. (%)	Spec. (%)	Prec. (%)	Sens. (%)	F1- score (%)
eGeMAPS	51	66	89	41	14	65
IS09-13	56	70	92	52	19	78
EMOBASE	66	77	95	77	37	75
eGeMAPS+IS09-13	57	72	94	58	19	75
eGeMAPS+EMOBASE	60	73	93	66	28	75
IS09-13+EMOBASE	55	69	91	50	19	67
eGeMAPS+IS09-13+EMOBASE	58	71	90	64	26	72

Table 17
Performance of Hyperparameters tuned SVC on the features selected using SVM-RFE FS technique and stratified 10-fold cv

Feature subspace	Bal. Acc. (%)	Accu. (%)	Spec. (%)	Prec. (%)	Sens. (%)	F1- score (%)
eGeMAPS	55	65	84	45	26	64
IS09-13	63	68	79	53	47	68
EMOBASE	74	79	89	73	58	79
eGeMAPS+IS09-13	68	74	84	62	52	74
eGeMAPS+EMOBASE	67	70	76	55	58	70
IS09-13+EMOBASE	46	56	76	25	15	56
eGeMAPS+IS09-13+EMOBASE	60	65	74	47	47	65

Table 18
Performance of Hyperparameters tuned LR on the features selected using SVM-RFE FS technique and stratified 10-fold cv

Feature subspace	Bal. Acc. (%)	Accu. (%)	Spec. (%)	Prec. (%)	Sens. (%)	F1-score (%)
eGeMAPS	56	71	93	51	19	14
IS09-13	57	66	81	49	31	10
EMOBASE	54	69	92	20	10	12
eGeMAPS+IS09-13	62	69	79	49	46	14
eGeMAPS+EMOBASE	53	66	85	35	22	10
IS09-13+EMOBASE	53	65	83	37	24	12
eGeMAPS+IS09-13+EMOBASE	58	65	77	47	38	14

Table 19
Recognition performance of ensemble classifiers on EMOBASE features selected using SVM-RFE FS with hyperparameter tuning and stratified 10-fold cv

Classifier	Sen. (%)	Spec. (%)	Accu. (%)	Bal. Accu. (%)	F1-score (%)
Bagging decision tree	64	94	71	60	51
AdaBoost decision tree	60	97	75	65	57
MVCDD	67	94	81	77	70

For each feature subspace, the table displays the results of the classification of each classifier. It is worth noting that each classifier's optimal features differed. These findings show that each feature vector may provide the classifiers with more information. Furthermore, each classifier could not use the same feature subspace that other classifiers had found to be the most effective for better classification results. It implies that classifiers must be built from multiple feature subsets. The experimental results showed that the proposed MVCDD model with EMOBASE feature subspace performed better than other classifiers.

The proposed model exhibits an accuracy of 81%, specificity of 94%, sensitivity of 67%, balanced accuracy of 77%, and an F1-score of 70%, which is comparatively more than the existing models in depression predictive modeling using acoustic aspects of speech. The graph containing the metrics of the ensemble classifiers and the proposed model is shown in Figure 4.

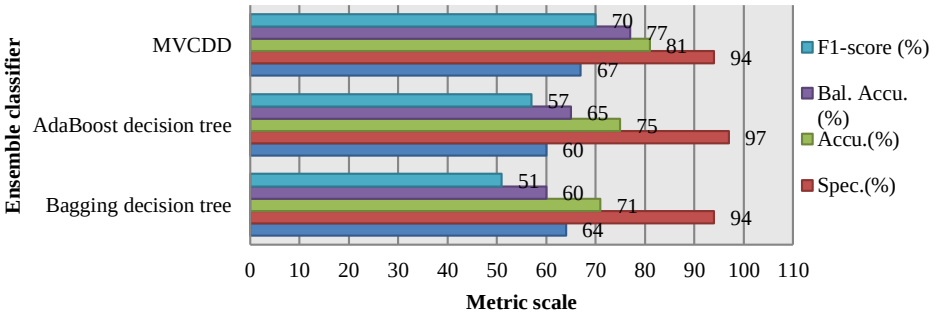

Figure 4. Proposed MVCDD model and other ensemble classifiers' performance on EMOBASE feature subspace with hyperparameter tuning and Stratified 10-fold CV

Conclusion

Depression, the most significant cause of disability globally, affects about 322 million people. MDD is the most common mental illness, accounting for a considerable global disease burden. Depression is such a severe mental disorder, the present diagnosis procedures still necessitate the involvement of a suitably qualified psychiatrist or psychologist, who will use a scale and carefully observe the conversation, depending on the Doctor's experience. The research found that around two-thirds of depression cases were missed from identification [56].

Recently, there has been much research into automated depression prediction to help psychiatrists automate and speed up the entire diagnosis process.

In this paper, the acoustic features of the speech recordings of the students of Government Polytechnic, Masabtank, Hyderabad, India, have been extracted and experimented with two FS methods and different machine learning classification models. The eGeMAPS, Is09-13, and EMOBASE features are extracted from the dataset and formed seven different feature subspaces. To aid the curse of dimensionality Pearson correlation method is used to eliminate the features with high correlation from all the feature subspaces. This work proposed a Majority Voting Classifier for Depression Detection (MVCDD) model. MVCDD has better classification performance when executed on the feature subspace of the ten best EMOBASE features selected by the Support Vector Machine Recursive Feature Elimination (SVM-RFE) method. Classifiers considered for this work are Random Forest Classifier (RFC), Support Vector Classifier (SVC), and Logistic Regression classifier (LR). To deal with the imbalance of the class distribution in the data, the Stratified 10-fold cross-validation technique is used. This work also used GridSearchCV for hyperparameter tuning of the chosen classifiers to improve the results. The ensemble models like Bagging with Decision Tree and AdaBoost with Decision Tree are also experimented with different feature subspaces. Based on the experiments performed, it is evident that the proposed MVCDD model gave better classification results with 81% accuracy and 94% of specificity, which comparatively outperforms the existing state-of-the-art models in this area. In the future, we would like to experiment with multimodal analysis for depression prediction, including audio, visual, and textual modalities.

Acknowledgments
The authors acknowledge the Central University of Tamil Nadu for encouraging the research.

References

[1] M. Marcus, M. T. Yasamy, M. van Ommeren, and Chisholm, “Depression, A Global Public Health Concern,” D., 2012, Accessed: Dec. 19, 2021. [Online]. Available: https://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf.

[2] C. Otte et al., “Major depressive disorder,” Nat. Rev. Dis. Prim., vol. 2, no. Mdd, pp. 1–21, 2016, doi: https://doi.org/10.1038/nrdp.2016.65

[3] K. S. Dobson and M. C. Scherrer, “Major depressive disorder,” Handb. Clin. Interviewing with Adults, pp. 134–152, 2007, doi: http://doi.org/10.4135/9781412982733.n10

[4] A. B. Negrão and P. W. Gold, “Major Depressive Disorder,” Encycl. Stress, vol. 28, pp. 640–645, 2007, http://doi.org/10.1016/B978-012373947-6.00245-2

[5] S. Gao, V. D. Calhoun, and J. Sui, “Machine learning in major depression: From classification to treatment outcome prediction,” CNS Neurosci. Ther., vol. 24, no. 11, pp. 1037–1052, 2018, http://doi.org/10.1111/cns.13048

[6] R. C. Kessler et al., “Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-

reports," Mol. Psychiatry, vol. 21, no. 10, pp. 1366–1371, 2016, http://doi.org/10.1038/mp.2015.198

[7] R. A., T. M., M. F., and C. S., "Neuroimaging-based methods for autism identification: A possible translational application?," Funct. Neurol., vol. 29, no. 4, pp. 231–239, 2014, [Online]. PMID: 25764253; PMCID: PMC4370436.

[8] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," Speech Commun., vol. 71, no. April, pp. 10–49, 2015, http://doi.org/10.1016/j.specom.2015.03.004

[9] K. R. Scherer Justus, "Vocal Affect Expression: A Review and a Model for Future Research," Psychol. Bull., vol. 99, no. 2, pp. 143–165, 1986, PMID: 3515381.

[10] H. Jiang et al., "Investigation of different speech types and emotions for detecting depression using different classifiers," Speech Commun., vol. 90, pp. 39–46, 2017, http://doi.org/10.1016/j.specom.2017.04.001

[11] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 2172–2176, 2013, http://doi.org/10.21437/interspeech.2013-513

[12] K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," Anal. Chim. Acta, vol. 703, no. 2, pp. 152–162, 2011, http://doi.org/10.1016/j.aca.2011.07.027

[13] S. García, J. Luengo, and F. Herrera, "Feature selection," Intell. Syst. Ref. Libr., vol. 72, no. 6, pp. 163–193, 2015, http://doi.org/10.1007/978-3-319-10247-4_7

[14] G. Kiss and K. Vicsi, "Mono- and multilingual depression prediction based on speech processing," Int. J. Speech Technol., vol. 20, no. 4, pp. 919–935, 2017, http://doi.org/10.1007/s10772-017-9455-8

[15] T. Charubin and A. Juś, "Test stand for matteucci effect measurements," Adv. Intell. Syst. Comput., vol. 550, pp. 527–532, 2017, http://doi.org/10.1007/978-3-319-54042-9_52

[16] M. Valstar et al., "AVEC 2013 - The continuous Audio/Visual Emotion and depression recognition challenge," AVEC 2013 - Proc. 3rd ACM Int. Work. Audio/Visual Emot. Chall., pp. 3–10, 2013, http://doi.org/10.1145/2512530.2512533

[17] H. Jiang et al., "Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech Features," Comput. Math. Methods Med., vol. 2018, 2018, http://doi.org/10.1155/2018/6508319

[18] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, and S. Weidt, "Mobile sensing and support for people with depression: A pilot trial in the wild," JMIR mHealth uHealth, vol. 4, no. 3, 2016, http://doi.org/10.2196/mhealth.5960

[19] C. Te Wu, D. G. Dillon, H. C. Hsu, S. Huang, E. Barrick, and Y. H. Liu, "Depression detection using relative EEG power induced by emotionally positive images and a conformal kernel support vector machine," Appl. Sci., vol. 8, no. 8, 2018, http://doi.org/10.3390/app8081244

[20] Z. Wu, L. Lei, J. Dong, J. Hou, and X. Zhang, "Reconfigurable temporal fourier transformation and temporal imaging," J. Light. Technol., vol. 32, no. 23, pp. 4565–4570, 2014, http://doi.org/10.1109/JLT.2014.2361293

[21] M. Deshpande and V. Rao, “Depression detection using emotion artificial intelligence,” Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017, no. Iciss, pp. 858–862, 2018, http://doi.org/10.1109/ISS1.2017.8389299

[22] J. Gratch et al., “The distress analysis interview corpus of human and computer interviews,” Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014, pp. 3123–3128, 2014.

[23] T. C. F. Duarte, H. da S. Lopes, and H. L. M. Campos, “Physical activity, life purpose of community active elderly people: A cross-section study,” Rev. Pesqui. em Fisioter., vol. 10, no. 4, pp. 591–598, 2020, http://doi.org/10.17267/2238-2704rpf.v10i4.3052

[24] Darby J, K, Hollien H: Vocal and Speech Patterns of Depressive Patients. Folia Phoniatr Logop 1977;29:279-291. http://doi.org/10.1159/000264098

[25] H. H. Stassen, S. Kuny, and D. Hell, “The speech analysis approach to determining onset of improvement under antidepressants,” Eur. Neuropsychopharmacol., vol. 8, no. 4, pp. 303–310, 1998, http://doi.org/10.1016/S0924-977X(97)00090-4

[26] A. S. A. Nilsonne, “Differences in Ability of Musicians and Nonmusicians to Judge Emotional State from the Fundamental Frequency of Voice Samples Author (s): Åsa Nilsonne and Johan Sundberg Published by : University of California Press Stable URL : http://www.jstor.org/sta,” vol. 2, no. 4, pp. 507–516, 2016, http://doi.org/10.2307/40285316

[27] M. Alpert, E. R. Pouget, and R. R. Silva, “Reflections of depression in acoustic measures of the patient’s speech,” J. Affect. Disord., vol. 66, no. 1, pp. 59–69, 2001, Article (CrossRef Link)

[28] H. Hollien, “Vocal Indicators of Psychological Stress,” Ann. N. Y. Acad. Sci., vol. 347, no. 1, pp. 47–72, 1980, http://doi.org/10.1111/j.1749-6632.1980.tb21255.x

[29] H. Ellgring and K. R. Scherer, “Vocal indicators of mood change in depression,” J. Nonverbal Behav., vol. 20, no. 2, pp. 83–110, 1996, http://doi.org/10.1007/BF02253071

[30] J. A. R. Jonathan Posner and Bradley S. Peterson, “Vocal Acoustic Biomarkers of Depression Severity and Treatment Response”, NIH Public Access, Bone, vol. 23, no. 1, pp. 1–7, 2008, http://doi.org/10.1016/j.biopsych.2012.03.015

[31] T. F. Quatieri and N. Malyska, “Vocal-source biomarkers for depression: A link to psychomotor activity,” 13th Annu. Conf. Int. Speech Commun. Assoc. 2012, INTERSPEECH 2012, vol. 2, pp. 1058–1061, 2012.

[32] J. F. Cohn et al., “Detecting depression from facial actions and vocal prosody,” Proc. - 2009 3rd Int. Conf. Affect. Comput. Intell. Interact. Work. ACII 2009, 2009, http://doi.org/10.1109/ACII.2009.5349358

[33] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, “Multimodal assessment of depression from behavioral signals,” Handb. Multimodal-Multisensor Interfaces Found. User Model. Common Modality Comb. - Vol. 2, pp. 375–417, 2018, http://doi.org/10.1145/3107990.3108004

[34] K. E. B. Ooi, M. Lech, and N. Brian Allen, “Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system,” Biomed. Signal Process. Control, vol. 14, no. 1, pp. 228–239, 2014, http://doi.org/10.1016/j.bspc.2014.08.006

[35] E. G. Trevino et al., “Effect of irrigants on the survival of human stem cells of the apical papilla in a platelet-rich plasma scaffold in human root tips,” J. Endod., vol. 37, no. 8, pp. 1109–1115, 2011, http://doi.org/10.1016/j.joen.2011.05.013

[36] S. Scherer, G. Stratou, J. Gratch, and L. P. Morency, “Investigating voice quality as a speaker-independent indicator of depression and PTSD,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 847–851, 2013, http://doi.org/10.21437/interspeech.2013-240

[37] S. Scherer et al., “Automatic behavior descriptors for psychological disorder analysis,” 2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013, 2013, http://doi.org/10.1109/FG.2013.6553789

[38] S. Scherer, G. Stratou, and L. P. Morency, “Audiovisual behavior descriptors for depression assessment,” ICMI 2013 - Proc. 2013 ACM Int. Conf. Multimodal Interact., pp. 135–140, 2013, http://doi.org/10.1145/2522848.2522886

[39] S. Alghowinem et al., “A comparative study of different classifiers for detecting depression from spontaneous speech,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 8022–8026, 2013, http://doi.org/10.1109/ICASSP.2013.6639227

[40] M. Valstar et al., “AVEC 2014 - 3D dimensional affect and depression recognition challenge,” AVEC 2014 - Proc. 4th Int. Work. Audio/Visual Emot. Challenge, Work. MM 2014, pp. 3–10, 2014, http://doi.org/10.1145/2661806.2661807

[41] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, “An investigation of depressed speech detection: Features and normalization,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. January, pp. 2997–3000, 2011, http://doi.org/10.21437/interspeech.2011-750

[42] D. Sturim, P. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree, “Automatic detection of depression in speech using Gaussian mixture modeling with factor analysis,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. August, pp. 2981–2984, 2011.

[43] S. Alghowinem, “From joyous to clinically depressed: Mood detection using multimodal analysis of a person’s appearance and speech,” Proc. - 2013 Hum. Assoc. Conf. Affect. Comput. Intell. Interact. ACII 2013, pp. 648–653, 2013, http://doi.org/10.1109/ACII.2013.113

[44] J. Joshi et al., “Multimodal assistive technologies for depression diagnosis and monitoring,” J. Multimodal User Interfaces, vol. 7, no. 3, pp. 217–228, 2013, http://doi.org/10.1007/s12193-013-0123-2

[45] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, “Detection of clinical depression in adolescents’ speech during family interactions,” IEEE Trans. Biomed. Eng., vol. 58, no. 3 PART 1, pp. 574–586, 2011, http://doi.org/10.1109/TBME.2010.2091640

[46] H. H. Stassen, G. Bomben, and E. Gunther, “Speech characteristics in depression,” Psychopathology, vol. 24, no. 2, pp. 88–105, 1991, doi: http://doi.org/10.1159/000284700.

[47] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, “The PHQ-8 as a measure of current depression in the general population,” J. Affect. Disord., vol. 114, no. 1–3, pp. 163–173, 2009, doi: http://doi.org/10.1016/j.jad.2008.06.026

[48] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in openSMILE, the munich open-source multimedia feature extractor,” in MM 2013 - Proceedings of the 2013 ACM Multimedia Conference, 2013, pp. 835–838, http://doi.org/10.1145/2502081.2502224

[49] F. Eyben et al., “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” IEEE Trans. Affect. Comput., vol. 7, no. 2, pp. 190–202, Apr. 2016, http://doi.org/10.1109/TAFFC.2015.2457417.

[50] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 312–315, 2009.

[51] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2, pp. 1–4, 2003, http://doi.org/10.1109/icassp.2003.1202279

[52] A. Batliner, B. Möbius, G. Möhler, A. Schweitzer, and E. Nöth, “Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground,” EUROSPEECH 2001 - Scand. - 7th Eur. Conf. Speech Commun. Technol., pp. 2285–2288, 2001, http://doi.org/10.1007/1-4020-2637-4_3

[53] A. B. M. Nutt, V. W. E. Noth, and J. B. R. Huber, “Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech,” Test, pp. 4–7, 1996.

[54] F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE - The Munich versatile and fast open-source audio feature extractor,” MM’10 - Proc. ACM Multimed. 2010 Int. Conf., pp. 1459–1462, 2010, http://doi.org/10.1145/1873951.1874246

[55] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, “Data mining in bioinformatics using Weka,” Bioinformatics, vol. 20, no. 15, pp. 2479–2481, 2004, http://doi.org/10.1093/bioinformatics/bth261

[56] M. Cepoiu, J. McCusker, M. G. Cole, M. Sewitch, E. Belzile, and A. Ciampi, “Recognition of depression by non-psychiatric physicians - A systematic literature review and meta-analysis,” J. Gen Intern. Med., vol. 23, no. 1, pp. 25–36, 2008, http://doi.org/10.1007/s11606-007-0428-5