Artificial Intelligence Evolution
https://ojs.wiserpub.com/index.php/AIE/

UNIVERSAL WISER PUBLISHER

Research Article

Optimal Combination of Multivariate Filter Feature Selection and Classifier for Speech-Based Depression Detection

Surbhi Sharma[1*], **Anthony J. Bustamante**[2]

[1]School of Computer and Systems Sciences, Jawaharlal Nehru University, Delhi, India
[2]National Institute for Telecommunications Research and Training, National University of Engineering, Peru
Email: surbhisharma9099@gmail.com

Received: 13 September 2021; **Revised:** 17 November 2021; **Accepted:** 17 November 2021

Abstract: In this paper, we have focused to improve the performance of a speech-based uni-modal depression detection system, which is non-invasive, involves low cost and computation time in comparison to multi-modal systems. The performance of a decision system mainly depends on the choice of feature selection method and the classifier. We have investigated the combination of four well-known multivariate filter methods (minimum Redundancy Maximum Relevance, Scatter Ratio, Mahalanobis Distance, Fast Correlation Based feature selection) and four well-known classifiers (k-Nearest Neighbour, Linear Discriminant classifier, Decision Tree, Support Vector Machine) to obtain a minimal set of relevant and non-redundant features to improve the performance. This will speed up the acquisition of features from speech and build the decision system with low cost and complexity. Experimental results on the high and low-level features of recent work on the DAICWOZ dataset demonstrate the superior performance of the combination of Scatter Ratio and LDC as well as that of Mahalanobis Distance and LDC, in comparison to other combinations and existing speech-based depression results, for both gender independent and gender-based studies. Further, these combinations have also outperformed a few multimodal systems. It was noted that low-level features are more discriminatory and provide a better f1 score.

***Keywords*:** speech-based features, multivariate filter feature selection, classification, f1-score

1. Introduction

Depression has emerged as the most challenging health concern of society worldwide. The enormity of the concern is stressed by the alarming rate of suicides worldwide [1]. The pressing need of the hour is to provide a simple, yet highly accurate and easy-to-use decision system, that supports the clinicians and the family. The aim is to aid in the early diagnosis of depression and consistent monitoring of the individual's mental health and to raise an alarm in time so that appropriate measures may be taken.

The research community has suggested many techniques over the past that have aimed to give an objective measure of depression. Many features have been extracted and studied and their relevance established for depression detection. Based on the review of the literature for depression detection, the techniques suggested can be categorized as *uni-modal* and *multi-modal*. Uni-modal techniques are the ones that involve judgement/analysis based on a single modality: only video [2-4], only text [5-7], or only speech [8-10]. Multi-modal techniques, which are a hybrid of two or more

DOI: https://doi.org/10.37256/aie.2220211149

modalities (videos, speech, and text) together [11-13] have been proposed to improve the performance further. These hybrid models take the advantage of individual modalities. The improved results obtained are based on the tedious process of video recording in a controlled environment, that needs the full cooperation of the person involved. In reality, patients are un-cooperative and often in a state of denial. Also, the multi-modal decision systems are more complex and incur more computation time and cost than the uni-modal approach. Hence, there is a need to strengthen the *uni-modal* model, which has the advantage due to its low cost of acquiring data and the low complexity of building the decision model. Among the various modalities, we have focused on speech-based depression detection, which is non-invasive, involves low cost, and does not require the active cooperation of the patient.

After going through various research works of speech-based depression detection systems, it was noted that most of the research work mainly focused only on the feature extraction techniques [11, 14-18]. It is well-known that if an appropriate combination of features selection method and classifier is not used to build the decision system, the performance may degrade. In one of the recent works by Pampouchidou et al. [11], a large set of both low and high-level features from speech has been explored. The importance of individual features was noted by removing each feature or set of features based on the f1 score using a decision tree classifier. However, this wrapper approach is computationally intensive. In literature, filter feature selection methods in combination with a classifier are found to be more successful to find a smaller set of relevant features to improve the performance in many domains such as microarray-based cancer classification, object recognition, and text analysis [19]. This combination of filter feature selection and the classifier is computationally less intensive. But, in speech-based depression detection, filter methods have not been investigated much to improve the performance of the depression detection system, to the best of our knowledge.

Motivated by this, we have investigated combinations of four multivariate filter methods (minimum Redundancy Maximum Relevance [20], Scatter Ratio [21], Mahalanobis Distance [22], Fast Correlation Based feature selection [23], and four well-known classifiers (k-Nearest Neighbour, Linear Discriminant, Tree-based classifier, Support Vector Machine) to obtain a minimal set of relevant and non-redundant features to improve the performance of depression detection system. For our experiments, we have used the high and low-level features [11] individually as well as in combination. This will make the process to acquire the features from speech fast and build the decision system with low cost and complexity. The DAICWOZ [24] repository is used with features based on the COVAREP feature set [25]. Experiments for both gender independent and gender-based features are performed. Experimental results demonstrate the superior performance of the combination of Scatter Ratio and LDC as well as that of Mahalanobis Distance and LDC, in comparison to other combinations and existing speech-based depression results for both gender independent and gender-based studies. Further, these combinations have also performed better than a few multimodal systems.

The next section discusses related work. Section 3 briefly introduces feature selection techniques used in our paper. Experiment set up and results are included in section 5 and conclusion in section 6.

2. Related work

Speech is a piece of natural, non-invasive evidence that has been investigated for depression over the years and has been found indicative in classifying depressed people from the control group. The tenseness of the vocal tract and the vocal cords in depressed people translates to a set of features that are well known and researchers have been investigating them over the years.

France et al. [14] worked with the Fundamental Frequency (F_o), Amplitude Modulation (AM), formants, and Power Spectral Density (PSD) for discriminating control, dysthymic and major depressed persons. Formant frequencies and PSD were successful with 0.94 accuracies to distinguish between the control and depressed female patients and 0.82 between the male patients. Ozdes et al. [16] explored the significance of vocal jitter and spectral slope as indicators of suicidal tendencies. The pairwise classification accuracies among control/depressed, depressed/suicide, control/suicide using jitter was 0.65, 0.60, 0.80 respectively and using spectral slope was 0.90, 0.75, 0.60 respectively. Using both jitter and spectral slope, the accuracies reported were 0.90, 0.75, and 0.85 respectively for the three pairs. The experiments were conducted using a maximum likelihood classifier and no feature selection technique was employed. Sethu et al. [15] evaluated emotions based on pitch, energy slope, and formants for speaker-dependent and speaker-independent studies. Mel Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Coefficients (LPCs) based group delay

performed well in speaker-dependent system, while the first three formants performed well in the speaker-independent system. The Gaussian Mixture Model was used as a classifier and no feature selection technique was used. The study by Scherer et al. [17] involved the analysis of Normalized Amplitude Quotient (NAQ), Quasi-open-Quotient (QoQ), peak slope, and Open Quotient Neural Network (OQ_{NN}). It was observed that these features are independent of gender. 0.75 accuracy was obtained with a Support Vector Machine (SVM) as the classifier. Scherer et al. had also not used any feature selection. Alghowinem et al. [18] evaluated linear features: Fundamental Frequency (F_o), intensity, loudness, voice probability, quality, jitter, shimmer, Harmonics to Noise Ratio (HNR), log energy, root mean square energy, and the non-linear teager energy to understand the difference in voiced/unvoiced and mixed speech on depression detection. They suggested the suitability of Teager Energy Operator (TEO) based features for depression detection using voiced and unvoiced speech. Voicing probability and log energy were found to be more suited for mixed speech. The SVM classifier was used for evaluation.

In 2016, Pampouchidou et al. [11] used a fusion of low-level features and Discrete Cosine Transform (DCT) based features and high-level features. They analyzed two sets of sizes 494 and 1,278 respectively using the DAICWOZ dataset and the COVAREP feature repository. The gender-based results using low-level descriptors are reported in terms of f1 (depressed/non-depressed) as 0.45 (0.85) (Leave-One-Out Cross-Validation (LOOCV)) and 0.59 (0.87) (testing using development set). The DCT based features for gender independent data gave 0.19 (0.71) and 0.47 (0.83) in terms of f1 measure respectively. The decision tree was used as a classifier and the importance of each feature was calculated by removing each feature. In 2017, Pampouchidou et al. [26] evaluated the AVEC dataset [27] using the COVAREP based audio features and reported the precision of 0.948 while using visual OR (ed) with audio gender-based features. Audio alone gave the f1 score of 0.641. A nearest neighbor classifier was used and no feature selection was used for evaluating the performance.

3. Feature selection

Once a set of low-level descriptors and/or the discrete cosine transform coefficients of the acoustic features are obtained to build the decision system for depression detection, it is imperative to reduce the dimensionality of the feature set to handle the curse of dimensionality [28]. Feature selection methods to reduce dimensionality are known to be categorized into two major approaches: Filter and Wrapper [29, 30]. Wrapper methods are computationally intensive due to the repeated training of each candidate subset. On the other hand, filter methods do not involve any learning algorithm to measure the importance of features, hence are simple and computationally less intensive. They are further subdivided into univariate and multivariate methods [29]. Univariate filter methods measure the relevance of the individual features based on the statistical characteristics of the feature and ultimately collect the top-ranked relevant features. Though the univariate method is simple, the set so obtained does not consider the correlation among features and hence suffers from redundancy. Also, features so obtained have a low capacity to discriminate among classes. In literature, multivariate feature selection methods [31] have found success in determining a set of relevant and non-redundant features, which provides better performance. In multivariate feature selection, we have first applied the Fisher's discriminant ratio [21] to select the top 100 relevant features. These top 100 relevant features have been passed to multivariate feature selection methods to select the top 50 non-redundant features.

To the best of our knowledge, multivariate filter feature selection techniques have not been explored in speech-based depression detection systems. This has motivated us to investigate some popular multivariate methods. They overcome the limitations of the univariate filter methods as well as wrapper methods.

3.1 *The ratio of Scatter Matrices (SR)*

Based on the scatter-ness of the feature vectors [21], a trace of the ratio of the scatter matrices is suggested as a popular criterion. Those features are selected that are well separated across classes while being clustered around their class mean. The criterion $J_{SR} = \text{trace}(S_w^{-1}S_b)$ is used where within-class scatter matrix (S_w) and between-class scatter matrix (S_b) are given as:

$$S_w = \frac{1}{N}\sum_{i=1}^{c} P_i \sum_{x \in D_i} (x - \mu_i)(x - \mu_i)^t \quad (1)$$

$$S_b = \sum_{i=1}^{c} (x - \mu_i)(x - \mu_i)^t \quad (2)$$

where P_i, μ_i are the prior probability and mean vector of the *i*th class. N is the total number of data samples.

3.2 *Mahalanobis distance (Maha)*

Given two classes, C_i, i = 1, 2, where μ_i is the mean vector and $\sum$ is the covariance matrix, Mahalanobis distance [22] is given by

$$J_M = (\mu_1 - \mu_2)\sum^{-1}(\mu_1 - \mu_2)^t \quad (3)$$

A higher value of J_M corresponds to more separability between data of two classes.

3.3 *Fast Correlation Based Feature selection (FCBF)*

This is a correlation measure based on the information-theoretic concept of entropy of the random variable. The entropy of variable X is defined as:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) \quad (4)$$

Information gain of X given Y is given by:

$$IG(X|Y) = H(X) - H(X|Y) \quad (5)$$

FCBF is based on the principle of symmetrical uncertainty defined as [23]:

$$SU(X,Y) = \frac{2 * IG(X|Y)}{H(X) + H(Y)} \quad (6)$$

The feature belongs to a set of chosen relevant features if its correlation with the class is more than its correlation with the features selected in the set.

3.4 *minimum Redundancy Maximum Relevance (mRMR)*

mRMR [20] measures the mutual information among two random variables to estimate their relevance. For two variables, x_i and x_j, with marginal probabilities, $p(x_i)$ and $p(x_j)$, and joint probabilities $p(x_i, x_j)$, the mutual information is given by:

$$I(x_i, x_j) = \sum p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \quad (7)$$

The objective is to attain minimum redundancy among selected features and maximum relevance among the features and the class variable, c, is which can be expressed as a single criterion, as either *max*(*Rel-Red*) or *max*(*Rel/Red*), where

$$\text{Rel} = \frac{1}{|s|}\sum_{k \in S} I(c,k) \tag{8}$$

$$\text{Red} = \frac{1}{|s|^2}\sum_{i,j \in S} I(i,j) \tag{9}$$

4. Classifier used

4.1 *K-nearest neighbour*

K-nearest neighbour classifier [32] is a non-parametric classification that classifies the new test sample by comparing it against the training samples with a similarity function. It identifies the K-neighbours of the test sample in the pattern space. The closeness of the test sample is generally calculated with the Euclidean distance function. It requires no learning model to classify a new sample. It is the most simple algorithm with only one parameter K that denotes no of neighbours. K-nearest neighbour becomes computationally intensive as the size of the training sample increases.

4.2 *Linear discriminant classifier*

Linear discriminant classifier [21] assigns a feature vector x to class c with a set of discriminant functions. It generally uses the simplified linear discriminant function as given by

$$H_i(x) = \text{In } p(x \mid c_i) + \text{In } p(c_i)$$

It uses the posterior conditional probability $p(c_i \mid x)$ for the selection of class labels for a given sample x. It is a simple, fast, and portable algorithm.

4.3 *Support Vector Machine (SVM)*

A support vector machine classifier [33] is based on supervised learning that classifies data points by defining a boundary that maximizes the distance between two classes. The equations for the two sides of the separating hyperplane can be written as follows

$$w^t x_i + b \geq 0$$

for all i with $c_i = +1$

and

$$w^t x_i + b \geq 0$$

for all i with $c_i = -1$

The two equations can be merged as

$$c_i (w^t x_i + b) \geq 1 \quad \text{for } i = 1, 2, ..., n$$

Standard SVM is a binary classifier. It is more useful when data is in high dimensions and the number of samples is comparatively small.

4.4 *Decision Tree Classifier (TreeC)*

Decision Tree Classifier [21] is based on non-metric methods in which classification is done in a basic decision tree that proceeds from top to bottom. At each node, a question regarding a particular property of a node is asked which results in the splitting of the node. The terminal nodes or leaf nodes correspond to the category labels. The basic principle underlying tree creation is entropy impurity. Entropy is zero if all the patterns are of the same category and positive if different classes are equally likely. It does not require data to be normalized and scaled before preprocessing. It does not require effort in preprocessing of data.

5. Experiments and results

For our experiment, we have used the three sets of features [11] extracted from the DAICWOZ dataset using the COVAREP toolbox. The first set (set A) comprised of statistical descriptors (refer to Table 1 of [11]) of the following low-level descriptors: F_o (normalized, delta, and delta-delta), NAQ, QOQ, Amplitude difference of the first two harmonics (H1H2), Parabolic Spectral Parameter (PSP), Maximum Dispersion Quotient (MDQ), peak slope, shape parameter of the Liljencrants-Fant glottal model (Rd), Rd confidence measure (Rd_conf), Mel Cepstral Coefficients (MCEP0-24, delta, and delta-delta), Harmonic Model and Phase Distortion Mean (HMPDM 1-24), Harmonic Model and Phase Distortion Deviation (HMPDD 1-12) and 1-3 Formants. The second set (set B) is represented in terms of the DCT coefficients for each low-level descriptor. The third set (Set C) consists of the eight high-level features such as Pause Ratio, Voiced Segment Ratio, Speaking Ratio, Mean laughter Duration, Mean Delay in Response, Mean Duration of Pauses, Maximum Duration of Pauses and the Fraction of Pauses in Overall Time. The experiments were performed on: set A, set B, set C, set AC, set BC, and set ABC on voiced segments only.

Four multivariate methods such as minimum Redundancy Maximum Relevance (mRMR), Scatter Ratio (SR), Mahalanobis distance (Maha), and Fast Correlation Based Feature selection (FCBF) were used to obtain a minimal subset of features that are relevant and non-redundant from each of the six feature sets (set A, set B, set C, set AC, set BC, set ABC). Features so obtained from each set are included incrementally and the decision model is built using a classifier such as KNN [32], LD classifier [21], decision tree [21], and SVM [33]. Both the leave-one-out and hold-out scheme is used to measure the performance of the model which is measured in terms of Classification Error, f1 (Depressed), and f1 (Non-Depressed). The performance of each combination of feature selection and classifier using LOOCV scheme is reported for gender independent, female gender, and male gender in Table 1, Table 2, and Table 3 respectively.

Based on Table 1, the following is observed for gender independent experiments:

i. The minimum classification error of 0.05 is obtained with the combination of SR and LDC with DCT coefficients of low-level descriptor, i.e., Set B.

ii. Among the feature selection methods, the performance of SR and Maha is comparable and better than mRMR and FCBF.

iii. Among the four classifiers, the performance of LDC is best.

iv. The performance with low-level features (Set A and Set B) is better than high-level features (Set C).

Based on Table 2, the following is observed for female gender experiments:

i. The minimum classification error of 0.02 is obtained with the combination of SR and LDC with statistical descriptors of low-level features, i.e., Set A.

ii. Among the feature selection methods, the performance of SR is better than Maha, mRMR, and FCBF for all the sets except for Set C for which Maha gives better performance.

iii. Among the four classifiers, the performance of LDC is best.

iv. The performance with low-level features (Set A and Set B) is better than that with high-level features (Set C).

Based on Table 3, the following is observed for male gender experiments:

i. The minimum classification error of 0.01 is obtained with the combination of Maha and LDC with DCT coefficients of low level descriptor, i.e., Set B.

ii. Among the feature selection methods, the performance of Maha is better than SR, mRMR, and FCBF for all the sets except for Set ABC for which SR gives the better performance.

iii. Among the four classifiers, the performance of LDC is the best.
iv. The performance with low-level features (Set A and Set B) is better than that with high-level features (Set C).

Table 1. Comparative performance of gender independent experiments

		Mahalanobis		Scatter Ratio		mRMR		FCBF	
		Error (#)	F1 D (ND)	Error (#)	F1 D (ND)	Error (#)	F1 D (ND)	Error (#)	F1 D (ND)
Set A	KnnC	0.16 (11)	0.41 (0.9)	0.19 (12)	0.28 (0.89)	0.2 (1)	0 (0.89)	0.2 (1)	0 (0.89)
	LDC	**0.11 (33)**	**0.68 (0.93)**	0.12 (11)	0.65 (0.93)	0.16 (18)	0.52 (0.91)	0.19 (4)	0.27 (0.89)
	TreeC	0.2 (6)	0.46 (0.88)	0.18 (24)	0.55 (0.89)	0.19 (19)	0.55 (0.88)	0.21 (3)	0.38 (0.87)
	SVC	0.14 (20)	0.56 (0.92)	0.14 (33)	0.54 (0.92)	0.16 (28)	0.42 (0.91)	0.2 (1)	0 (0.89)
Set B	KnnC	0.16 (3)	0.3 (0.91)	0.16 (1)	0.3 (0.91)	0.18 (36)	0.29 (0.9)	0.23 (2)	0 (0.87)
	LDC	0.09 (42)	0.78 (0.95)	**0.05 (49)**	**0.88 (0.97)**	0.15 (14)	0.51 (0.91)	0.2 (1)	0 (0.89)
	TreeC	0.23 (24)	0.38 (0.86)	0.21 (1)	0.47 (0.87)	0.18 (9)	0.49 (0.89)	0.24 (1)	0.41 (0.85)
	SVC	0.1 (39)	0.71 (0.94)	0.12 (7)	0.62 (0.93)	0.14 (19)	0.58 (0.91)	0.2 (1)	0 (0.89)
Set C	KnnC	0.18 (2)	0.23 (0.89)	0.18 (2)	0.23 (0.89)	0.20 (6)	0.00 (0.88)	0.20 (1)	0.00 (0.88)
	LDC	**0.20 (1)**	**0.00 (0.88)**	**0.02 (1)**	**0.00 (0.88)**	**0.20 (1)**	**0.00 (0.88)**	**0.20 (1)**	**0.00 (0.88)**
	TreeC	0.26 (5)	0.00 (0.88)	0.26 (5)	0.00 (0.88)	0.25 (4)	0.00 (0.88)	0.28 (1)	0.00 (0.83)
	SVC	0.20 (1)	0.00 (0.88)	0.20 (1)	0.00 (0.88)	0.20 (1)	0.00 (0.88)	0.20 (1)	0.00 (0.88)
Set AC	KnnC	0.16 (11)	0.41 (0.9)	0.16 (11)	0.41 (0.9)	0.2 (1)	0 (0.89)	0.2 (1)	0 (0.89)
	LDC	**0.11 (33)**	**0.68 (0.93)**	**0.11 (33)**	**0.68 (0.93)**	0.16 (21)	0.51 (0.9)	0.19 (2)	0.18 (0.89)
	TreeC	0.2 (6)	0.46 (0.88)	0.2 (6)	0.46 (0.88)	0.17 (9)	0.57 (0.89)	0.29 (2)	0.23 (0.82)
	SVC	0.14 (20)	0.56 (0.92)	0.14 (20)	0.56 (0.92)	0.16 (27)	0.42 (0.91)	0.2 (1)	0 (0.89)
Set BC	KnnC	0.16 (3)	0.3 (0.91)	0.16 (1)	0.3 (0.91)	0.18 (40)	0.29 (0.9)	0.21 (1)	0 (0.88)
	LDC	**0.14 (31)**	**0.65 (0.92)**	**0.14 (31)**	**0.65 (0.92)**	0.16 (26)	0.5 (0.91)	0.2 (1)	0 (0.88)
	TreeC	0.2 (30)	0.52 (0.87)	0.2 (3)	0.44 (0.88)	0.21 (19)	0.43 (0.87)	0.25 (2)	0.31 (0.85)
	SVC	0.18 (15)	0.55 (0.89)	0.17 (22)	0.57 (0.89)	0.16 (32)	0.56 (0.9)	0.2 (1)	0 (0.89)
Set ABC	KnnC	0.16 (2)	0.3 (0.91)	0.16 (1)	0.3 (0.91)	0.16 (13)	0.38 (0.91)	0.21 (1)	0 (0.88)
	LDC	0.12 (50)	0.7 (0.92)	**0.12 (45)**	**0.69 (0.92)**	0.16 (16)	0.54 (0.91)	0.19 (2)	0.13 (0.9)
	TreeC	0.2 (6)	0.48 (0.88)	0.19 (6)	0.49 (0.88)	0.18 (20)	0.58 (0.89)	0.27 (3)	0.27 (0.83)
	SVC	0.15 (4)	0.46 (0.91)	0.14 (26)	0.69 (0.91)	0.13 (49)	0.64 (0.92)	0.19 (1)	0.13 (0.89)

Table 2. Comparative performance of female based experiments

Female		Mahalanobis		Scatter Ratio		mRMR		FCBF	
		Error (#)	F1 D (ND)	Error (#)	F1 D (ND)	Error (#)	F1 D (ND)	Error (#)	F1 D (ND)
Set A	KnnC	0.16 (2)	0.58 (0.9)	0.1 (2)	0.8 (0.94)	0.13 (5)	0.67 (0.92)	0.21 (2)	0.32 (0.88)
	LDC	0.05 (13)	0.9 (0.97)	**0.02 (18)**	**0.97 (0.99)**	0.13 (4)	0.67 (0.92)	0.15 (1)	0.64 (0.91)
	TreeC	0.15 (18)	0.69 (0.91)	0.19 (14)	0.57 (0.88)	0.19 (9)	0.57 (0.88)	0.26 (1)	0.47 (0.83)
	SVC	0.06 (35)	0.86 (0.96)	0.1 (6)	0.8 (0.94)	0.1 (26)	0.77 (0.94)	0.18 (1)	0.59 (0.89)
Set B	KnnC	0.21 (28)	0.43 (0.87)	0.19 (1)	0.4 (0.88)	0.15 (1)	0.71 (0.9)	0.21 (1)	0.43 (0.87)
	LDC	0.26 (1)	0 (0.85)	**0.08 (15)**	**0.84 (0.95)**	0.26 (1)	0 (0.85)	0.24 (1)	0.12 (0.86)
	TreeC	0.34 (2)	0.28 (0.78)	0.24 (6)	0.44 (0.85)	0.18 (4)	0.67 (0.88)	0.24 (3)	0.44 (0.85)
	SVC	0.26 (1)	0 (0.85)	0.13 (15)	0.73 (0.91)	0.24 (47)	0.48 (0.84)	0.26 (1)	0 (0.85)
Set C	KnnC	0.24 (1)	0.28 (0.85)	0.24 (1)	0.28 (0.85)	0.24 (1)	0.28 (0.85)	0.24 (1)	0.28 (0.85)
	LDC	**0.20 (4)**	**0.38 (0.87)**	**0.20 (4)**	**0.38 (0.87)**	0.24 (1)	0.28 (0.85)	0.24 (1)	0.28 (0.85)
	TreeC	0.29 (4)	0.38 (0.87)	0.29 (4)	0.38 (0.87)	0.32 (1)	0.28 (0.85)	0.32 (1)	0.28 (0.85)
	SVC	0.25 (1)	0.28 (0.85)	0.25 (1)	0.28 (0.85)	0.25 (1)	0.28 (0.85)	0.25 (1)	0.28 (0.85)
Set AC	KnnC	0.16 (2)	0.58 (0.9)	0.1 (2)	0.8 (0.94)	0.13 (5)	0.67 (0.92)	0.23 (1)	0.53 (0.85)
	LDC	0.05 (13)	0.9 (0.97)	**0.02 (18)**	**0.97 (0.99)**	0.13 (4)	0.67 (0.92)	0.15 (1)	0.64 (0.91)
	TreeC	0.16 (18)	0.64 (0.9)	0.18 (14)	0.59 (0.89)	0.19 (19)	0.6 (0.87)	0.21 (3)	0.58 (0.86)
	SVC	0.08 (10)	0.81 (0.95)	0.1 (6)	0.8 (0.94)	0.06 (43)	0.86 (0.96)	0.16 (2)	0.64 (0.9)
Set BC	KnnC	0.24 (12)	0.29 (0.85)	0.24 (1)	0.29 (0.85)	0.15 (1)	0.71 (0.9)	0.21 (1)	0.43 (0.87)
	LDC	0.26 (1)	0 (0.85)	**0.1 (4)**	**0.8 (0.94)**	0.23 (17)	0.56 (0.85)	0.24 (1)	0.12 (0.86)
	TreeC	0.31 (25)	0.39 (0.8)	0.24 (2)	0.52 (0.84)	0.18 (4)	0.67 (0.88)	0.24 (3)	0.44 (0.85)
	SVC	0.26 (1)	0 (0.85)	0.16 (3)	0.69 (0.89)	0.19 (27)	0.57 (0.88)	0.26 (1)	0 (0.85)
Set ABC	KnnC	0.24 (17)	0.12 (0.86)	0.1 (2)	0.8 (0.94)	0.11 (7)	0.72 (0.93)	0.23 (1)	0.53 (0.85)
	LDC	0.26 (1)	0 (0.85)	**0.08 (5)**	**0.81 (0.95)**	0.13 (9)	0.67 (0.92)	0.15 (1)	0.64 (0.91)
	TreeC	0.29 (16)	0.44 (0.8)	0.16 (3)	0.67 (0.89)	0.19 (19)	0.6 (0.87)	0.26 (1)	0.47 (0.83)
	SVC	0.26 (1)	0 (0.85)	0.11 (13)	0.74 (0.93)	0.1 (32)	0.77 (0.94)	0.16 (4)	0.64 (0.9)

Table 3. Comparative performance of male based experiments

Male		Mahalanobis		Scatter Ratio		mRMR		FCBF	
		Error (#)	F1 D (ND)	Error (#)	F1 D (ND)	Error (#)	F1 D (ND)	Error (#)	F1 D (ND)
Set A	KnnC	0.06 (2)	0.74 (0.96)	0.06 (1)	0.76 (0.96)	0.06 (9)	0.76 (0.96)	0.06 (1)	0.76 (0.96)
	LDC	**0.04 (1)**	**0.86 (0.98)**	0.04 (22)	0.86 (0.98)	0.06 (4)	0.76 (0.96)	0.08 (2)	0.7 (0.96)
	TreeC	0.06 (6)	0.76 (0.96)	0.08 (5)	0.73 (0.96)	0.06 (5)	0.78 (0.96)	0.1 (1)	0.67 (0.94)
	SVC	0.06 (7)	0.76 (0.96)	0.05 (5)	0.82 (0.97)	0.05 (4)	0.82 (0.97)	0.08 (2)	0.7 (0.96)
Set B	KnnC	0.13 (13)	0.5 (0.93)	0.09 (5)	0.59 (0.95)	0.18 (1)	0.22 (0.9)	0.19 (1)	0 (0.89)
	LDC	**0.01 (42)**	**0.96 (0.99)**	0.01 (44)	0.96 (0.99)	0.12 (14)	0.47 (0.94)	0.15 (1)	0.25 (0.91)
	TreeC	0.18 (24)	0.3 (0.9)	0.18 (14)	0.46 (0.89)	0.12 (11)	0.47 (0.94)	0.19 (2)	0.35 (0.89)
	SVC	0.15 (1)	0 (0.92)	0.15 (1)	0 (0.92)	0.1 (11)	0.56 (0.94)	0.15 (1)	0 (0.92)
Set C	KnnC	0.16 (1)	0.00 (0.90)	0.16 (1)	0.00 (0.90)	0.15 (1)	0.00 (0.90)	0.16 (1)	0.00 (0.90)
	LDC	0.16 (1)	0.00 (0.90)	0.16 (1)	0.00 (0.90)	0.15 (1)	0.00 (0.91)	0.16 (1)	0.00 (0.90)
	TreeC	0.15 (2)	0.00 (0.90)	0.15 (2)	0.80 (0.96)	0.16 (5)	0.00 (0.90)	0.21 (1)	0.00 (0.90)
	SVC	**0.15 (1)**	**0.00 (0.90)**	**0.15 (1)**	**0.00 (0.90)**	**0.15 (1)**	**0.00 (0.91)**	**0.15 (1)**	**0.00 (0.90)**
Set AC	KnnC	0.06 (2)	0.74 (0.96)	0.06 (1)	0.76 (0.96)	0.06 (9)	0.76 (0.96)	0.06 (1)	0.76 (0.96)
	LDC	**0.04 (1)**	**0.86 (0.98)**	0.04 (27)	0.86 (0.98)	0.06 (4)	0.76 (0.96)	0.08 (2)	0.7 (0.96)
	TreeC	0.08 (1)	0.75 (0.95)	0.09 (38)	0.67 (0.95)	0.06 (5)	0.78 (0.96)	0.1 (1)	0.67 (0.94)
	SVC	0.06 (7)	0.76 (0.96)	0.05 (5)	0.82 (0.97)	0.05 (4)	0.82 (0.97)	0.08 (2)	0.7 (0.96)
Set BC	KnnC	0.13 (13)	0.5 (0.93)	0.09 (5)	0.59 (0.95)	0.18 (1)	0.22 (0.9)	0.19 (1)	0 (0.89)
	LDC	**0.01 (42)**	**0.96 (0.99)**	0.01 (44)	0.96 (0.99)	0.12 (14)	0.47 (0.94)	0.15 (1)	0.25 (0.91)
	TreeC	0.18 (24)	0.3 (0.9)	0.18 (14)	0.46 (0.89)	0.12 (11)	0.47 (0.94)	0.19 (2)	0.35 (0.89)
	SVC	0.15 (21)	0 (0.92)	0.15 (1)	0 (0.92)	0.1 (11)	0.56 (0.94)	0.15 (1)	0 (0.92)
Set ABC	KnnC	0.06 (2)	0.74 (0.96)	0.06 (1)	0.76 (0.96)	0.06 (9)	0.76 (0.96)	0.06 (1)	0.76 (0.96)
	LDC	0.03 (38)	0.91 (0.99)	**0.01 (36)**	**0.96 (0.99)**	0.05 (17)	0.82 (0.97)	0.06 (3)	0.76 (0.96)
	TreeC	0.08 (1)	0.75 (0.95)	0.12 (9)	0.61 (0.93)	0.06 (5)	0.78 (0.96)	0.06 (2)	0.76 (0.96)
	SVC	0.06 (10)	0.76 (0.96)	0.06 (4)	0.76 (0.96)	0.05 (4)	0.82 (0.97)	0.06 (3)	0.74 (0.96)

Gender independent and Gender-based experiments were also performed with the hold out method while training on the train set and testing on the development set. A comparison of the performance of the proposed method with existing methods on the DAICWOZ dataset is shown in Table 4. It can be observed that the proposed combination of multivariate feature selection with classifier gives better performance than existing methods for both gender independent and gender-based for LOOCV and hold-out schemes except for one case. It outperformed a few ensemble methods also.

Table 4. Comparison of the proposed method with state of art on DAICWOZ dataset in terms of f1 Depressed (Non-Depressed)

		LOOCV/k-fold	Holdout
GI		**0.88 (0.97) (proposed)**	0.60 (0.93) (proposed)
		0.26 (0.41) [10]	0.52 (0.70) [34]
		0.24 (0.75) [11]	0.46 (0.68) [27]
		0.5 (0.9) [11] (Visual)	0.50 (0.89) [27] (Video)
		0.35 (0.79) [11] (Ensemble)	0.50 (0.89) [27] (Ensemble)
		0.63 [26]	0.57 (mean) [35]
		0.586 [26] (Video)	**0.81 (mean)** [35] **(Ensemble)**
		0.69 [26] (Ensemble)	0.55 (0.79) [36]
		0.77 (0.73) [17]	0.57 (0.89) [37]
			0.63 (0.89) [37] (Video)
			0.63 (0.89) [37] (Ensemble)
GD	Female	**0.97 (0.99) (proposed)**	0.80 (0.96) (proposed)
		0.55 (0.80) [10]	**1.0 (1.0)** [36]
	Male	**0.96 (0.99) (proposed)**	**0.85 (0.96) (proposed)**
		0.49 (0.85) [10]	0.53 (0.71) [36]
	GB	0.59 (0.87) [11]	0.62 (0.91) [11] (Ensemble)
		0.36 (0.83) [11] (Ensemble)	
		0.641 [26]	
		0.58 [26] (Video)	
		0.73 [26] (Ensemble)	

Relevant features selected for depression detection in the proposed work are given in Table 5.

Table 5. Selected features for the best performance of the proposed method

Scheme	GI	Female	Male
LOOCV	Delta F0, Delta Delta F0, HMPDD, HMPDM, MCEP, Peak Slope, Rd	MCEP, Delta MCEP, Delta Delta MCEP, Formant, Peak Slope, HMPDD, HMPDM	MCEP, Delta MCEP, Delta Delta MCEP, HMPDM, Peak Slope
Holdout	MCEP, Delta MCEP, Delta Delta MCEP, HMPDD, HMPDM, Norm F0, Delta Delta F0, MCQ	HMPDD, MCEP	HMPDD

6. The main steps of the algorithm

1. Data Collection.
2. Data Cleaning.
3. Feature Extraction was done from the DAICWOZ Dataset.
4. Fisher Discriminate Ratio (FDR) was employed to select the top 100 features that are relevant to the class label.
5. Feature Selection was done using multivariate analysis.
6. Select the top 50 non-redundant features after multivariate analysis.
7. Features are used incrementally to make a Decision Model for each Classifier.

Both LOOCV and Hold out Scheme was used to analyze the model in terms of the classification error and in terms of the f1 depressed and f1 non-depressed. The low classification error was made the criteria to pick a corresponding number of features along with f1 depressed and f1 non-depressed.

The Combination of SR and LDC, and Maha and LDC provides comparable high performance for all the six sets and for both gender independent and gender-based studies.

Moreover, it was analyzed that the performance with low-level features was better than high-level features.

Step 1 and Step 2 have been performed in constant time.

Step 3 was having the worst time complexity of $N * H$, where N refers to the number of subjects in the dataset and H refers to the number of frames after speech processing of the patients.

Step 4 was having the time complexity of $n * f$, where n refers to the number of samples and f relates to the number of features.

Step 5 was having worst time complexity of $O(f^2)$ where f refers to the number of features and Step 6 is having the time complexity of $O(f \log f)$ because of sorting the features scores in descending order that are achieved after doing multivariate analysis.

In Step 7, the worst time complexity from all the classifiers used is $O(n^2 f + n^3)$ and the model was trained for 10 cross-fold validation.

7. Conclusion

In this paper, we investigated a combination of four well-known multivariate filter feature selection methods (mRMR, Maha, SR, and FCBF) and four classifiers (kNN, LDC, Decision Tree, and SVM) to determine the performance of speech-based depression detection system. Experiments are performed on DAICWOZ dataset using both low and high-level descriptors extracted using the COVAREP toolbox. Results demonstrate that the combination of SR and LDC, and Maha and LDC provide comparable high performance for all the six sets and for both gender independent and gender-based studies. Among all classifiers, the performance of LDC is best on all sets and for both gender independent and gender-based studies. It was also noted that the performance with low-level descriptors is better than high-level descriptors. Further, the performance of the proposed work on speech-based depression detection considerably improves in comparison to existing unimodal and few multimodal depression detection systems. There is a need to develop/explore better combination of feature selection method and classifier, which can provide better or comparable performance than multimodal system to reduce the cost and complexity of the decision system.

Acknowledgement

Author Surbhi Sharma would like to thank my parents and co-author who only helped me in proofreading the paper. The co-author has only one role in proofreading the paper.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

[1] World Health Organization. *Preventing suicide: A global imperative*. World Health Organization; 2014.
[2] Low LSA, Maddage NC, Lech M, Sheeber L, Allen N. Content based clinical depression detection in adolescents. *2009 17th European Signal Processing Conference*. 2009. p. 2362-2366.
[3] Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M. Eye movement analysis for depression detection. *2013 IEEE International Conference on Image Processing*. 2013. p. 4220-4224. Availible from: doi: 10.1109/ICIP.2013.6738869.
[4] Pampouchidou A, Pediaditis M, Maridaki A, Awais M, Vazakopoulou CM, Sfakianakis S, et al. Quantitative comparison of motion history image variants for video-based depression assessment. *EURASIP Journal on Image and Video Processing*. 2017; 64(2017). Avalible from: doi: 0.1186/s13640-017-0212-3.
[5] Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2013. p. 201-213.
[6] Low LSA, Maddage NC, Lech M, Sheeber LB, Allen NB. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*. 2011; 58(3): 574-586.
[7] Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*. 2017; 23(5): 649-685.
[8] Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: Features and normalization. *12th Annual Conference of the International Speech Communication Association*. 2011. Avalible from: doi: 2011.10.21437/Interspeech.2011-750.
[9] Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: A comparison between spontaneous and read speech. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013. p. 7547-7551.
[10] Cummins N, Vlasenko B, Sagha H, Schuller B. Enhancing speech-based depression detection through gender dependent vowel-level formant features. *Proceedings of Conference on Artificial Intelligence in Medicine in Europe*. 2017. p. 209-214.
[11] Pampouchidou A, Pediaditis M, Giannakakis G, Marias K, Simantiraki O, Manonsos D, et al. Depression assessment by fusing high and low level features from Audio, Video, and Text. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016. p. 27-34. Availible from: doi: 10.1145/2988257.2988266.
[12] Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and facial biomarkers of depression based on motor incoordination and timing. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. 2014. p. 65-72.
[13] Zhu Y, Shang Y, Shao Z, Guo G. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*. 2017.
[14] France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*. 2000; 47(7): 829-837.
[15] Sethu V, Ambikairajah E, Epps J. Speaker dependency of spectral features and speech production cues for automatic emotion classification. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009. p. 4693-4696.
[16] Ozdas A, Shiavi RG, Silverman SE, Silverman MK, Wilkes DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*. 2004; 51(9): 1530-1540.
[17] Scherer S, Stratou G, Gratch J, Morency LP. Investigating voice quality as a speaker-independent indicator

of depression and PTSD. In: *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2013. p. 847-851.
[18] Alghowinem S, Goecke R, Wagner M, Epps J, Parker G, Breakspear M. Characterising depressed speech for classification. In: *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2013. p. 2534-2538.
[19] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. *Feature Selection for High-Dimensional Data*. Springer; 2015.
[20] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005; 27(8): 1226-1238.
[21] Duda RO, Hart PE, Stork DG. *Pattern Classification*. John Wiley & Sons; 2012.
[22] Mahalanobis PC. *On the Generalized Distance in Statistics*. 1936.
[23] Yu L, Liu H. Redundancy based feature selection for microarray data. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2004. p. 737-742.
[24] Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. 2014. p. 3123-3128.
[25] Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP-A collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014. p. 960-964.
[26] Pampouchidou A, Simantiraki O, Vazakopoulou CM, Chatzaki C, Pediaditis M, Maridaki A, et al. Facial geometry and speech analysis for depression detection. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017. p. 1433-1436. Availible from: doi: 10.1109/EMBC.2017.8037103.
[27] Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, et al. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016. p. 3-10.
[28] Bellman R. Curse of dimensionality. *Adaptive Control Processes: A Guided Tour*. NJ: Princeton; 1961.
[29] Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003; 3: 1157-1182.
[30] Kohavi R, John GH. Wrappers for feature subset selection. *Artificial intelligence*. 1997; 97(1-2): 273-324.
[31] Devijver P, Kittler J. *Pattern Classification: A Statistical Approach*. 1982.
[32] Cover T, Hart P. Nearest neighbor pattern classification. *IEEE transactions on information theory*. 1967; 13(1): 21-27.
[33] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995; 20(3): 273-297.
[34] Ma X, Yang H, Chen Q, Huang D, Wang Y. Depaudionet: An efficient deep model for audio based depression classification. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016. p. 35-42.
[35] Williamson JR, Godoy E, Cha M, Schwarzentruber A, Khorrami P, Gwon Y, et al. Detecting depression using vocal, facial and semantic communication cues. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016. p. 11-18. Availible from: doi: 10.1145/2988257.2988263.
[36] Vlasenko B, Sagha H, Cummins N, Schuller B. Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. *Proceedings of Interspeech 2017*. 2017. p. 3266-3270. Availible from: doi: 10.21437/Interspeech.2017-887.
[37] Nasir M, Jati A, Shivakumar PG, Nallan Chakravarthula S, Georgiou P. Multimodal and multiresolution depression detection from speech and facial landmark features. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016. p. 43-50.