Comparison of classifiers using robust features for depression detection on Bahasa Malaysia speech

Nik Nur Wahidah Nik Hashim[1], Nadzirah Ahmad Basri[2], Mugahed Al-Ezzi Ahmad Ezzi[1], Nik Mohd Hazrul Nik Hashim[3]

[1]Department of Mechatronics Engineering, Faculty of Engineering, International Islamic University Malaysia, Gombak, Malaysia
[2]Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia, Jalan Hospital, Kuantan, Pahang, Malaysia
[3]Graduate School of Business, Universiti Kebangsaan Malaysia, UKM Bangi, Selangor, Malaysia

Article Info

Article history:

Received May 31, 2021
Revised Nov 3, 2021
Accepted Nov 29, 2021

Keywords:

Classifier
Depression
Features
Robust
Speech

ABSTRACT

Early detection of depression allows rapid intervention and reduce the escalation of the disorder. Conventional method requires patient to seek diagnosis and treatment by visiting a trained clinician. Bio-sensors technology such as automatic depression detection using speech can be used to assist early diagnosis for detecting remotely those who are at risk. In this research, we focus on detecting depression using Bahasa Malaysia language using speech signals that are recorded remotely via subject's personal mobile devices. Speech recordings from a total of 43 depressed subjects and 47 healthy subjects were gathered via online platform with diagnosis validation according to the Malay beck depression inventory II (Malay BDI-II), patient health questionnaire (PHQ-9) and subject's declaration of major depressive disorder (MDD) diagnosis by a trained clinician. Classifier models were compared using time-based and spectrum-based microphone independent feature set with hyperparameter tuning. Random forest performed best for male reading speech with 73% accuracy while support vector machine performed best on both male spontaneous speech and female reading speech with 74% and 73% accuracy, respectively. Automatic depression detection on Bahasa Malaysia language has shown to be promising using machine learning and microphone independent features but larger database is necessary for further validation and improving performance.

Corresponding Author:

Nik Nur Wahidah Nik Hashim
Department of Mechatronics Engineering, Faculty of Engineering, International Islamic University Malaysia
53100, Gombak, Malaysia
Email: niknurwahidah@iium.edu.my

1. INTRODUCTION

Major depressive disorder (MDD) has been reported as one of the psychiatric diagnosis in more than 90% of suicidal cases [1]. Globally, depressive disorder is one of the most common form of mental illness. With the rising number of cases on COVID-19 across the world, an epidemic of depression also emerges. World health organization (WHO) published the results of a survey on the impact of COVID-19 towards mental health [2]. The survey on 130 countries revealed factors that triggers mental health conditions or exacerbating existing ones are due to the death of family members, long period of isolation, loss of income and fear of getting infected. The pandemic is also disrupting the mental health and psychosocial services due to the use of mental health facilities as COVID-19 quarantine or treatment facilities and insufficient number of redeployment of healthcare workers. Approximately 70% of countries have begun to adopt tele-therapy to overcome the in-person service disruption [2]. However, extensive research needs to be done on the early diagnosis of depression using technology to be able to reliably detect remotely those who are at risk.

Conventional methods are generally performed according to a series of questionnaires and rating scales that measure various aspects including thoughts, behaviours and symptoms that are evaluated by a trained clinician or self-reported measurements. The information gathering process is a time-consuming process that requires the trained clinician to perform physical examination to rule out organic conditions and maintain regular interactions with the patient in order to attain accurate assessment. Also, it relies heavily on patients' effort to cooperate in communicating their symptoms and problems, while they are emotionally and psychologically impaired. The reliability of self-reported assessment can also be compromised by patients' overfamiliarity and subjectivity when rating the severity of their symptom [3].

Despite extensive research on automatic depression detection (ADD), we are still far from achieving reliable alternative methods for assisting psychiatric diagnosis. Depression is multifactorial and multi-characteristics with both genetic and environmental factor as the common associated factors in the development of depression, hence the complexity and subjectivity of its diagnosis. Development of an alternative psychological assessment tool is proposed to identify patient's mental condition using speech technology. Subjects with MDD often shows psychomotor retardation which manifests as a slowing or loss of spontaneous movement, reactivity and flattening of emotional expression. Manifestations of psychomotor retardation also include slowed speech and impaired cognitive function [4].

The field of ADD using speech has its own challenges. Each language is unique both in its structure and in the way it reflects the culture of those who speak it. Therefore, in this field of research, there is no one common set of acoustic feature across all languages that can be used as a universal depression detection. Table 1 summarizes information of speech corpus used by previous researches for studying the correlation between speech and psychological impairment. One research has studied the cross-corpus generalization on verbal biomarkers from Australia, USA and Germany database as depression detection [5]. They reported on the need of observations across multiple dataset characteristics in the training part in order to achieve higher accuracy in testing stage. This shows that the model has to learn the language pattern/structure and also learn depression cue patterns from the acoustics parameters in all languages and thus, they are not generalizable.

Researchers have began using deep learning for ADD [6]–[8], but due to the limited availability of large-scale data, most studies in this field still uses machine learning techniques. Deep learning requires a large training data due to the huge number of parameters needed to be tuned by a learning algorithm and computation power for data processing. Previous studies in the field of ADD using machine learning classifiers have used binary logistic regression (LR) [9], [10], gaussian mixture model (GMM) [11]–[14], discriminant analysis [15], [16], support vector machine (SVM) [11], [17], [18], k-nearest neighbour (KNN) [17], gaussian naïve Bayes (GNB) [19] random forest (RF) [20] and decision tree [21]. The common classification method used for detecting depression in speech are SVM and GMM. A study on English language by [14] compared multiple classifiers and concluded that SVM and GMM showed the best classification performance.

This work proposes to investigate the possibility of speech as a biomarker for depression in one of the countries in Southeast Asia, which is Malaysia. The national language in Malaysia is Bahasa Malaysia and thus, we focus to further investigate ADD using Bahasa Malaysia language. This study fills in the research gap on Bahasa Malaysia speech ADD with machine learning classification using microphone independent features. Commonly, only one recording device is used to record the speech to prevent variability in the data. This is the first study that have specifically looked at the effectiveness of using microphone independent features for classification. The hypothesis for this work is that depression can be identified using Bahasa Malaysia speech. Our research question is, can microphone independent acoustic features be deployed to distinguish between depression and healthy speech in Bahasa Malaysia language? This work extends from our recent findings reported in [22] where we explored robust features that are not affected by varying recording devices. In this study, we proceeded with investigating the robust features for depressed and healthy Bahasa Malaysia speech classification. A novel contribution is the analysis of a new database gathered through online platform in Bahasa Malaysia language and the comparison of machine learning classifiers using microphone independent features.

Some limitations should be noted. Gathering speech input through online platform was very challenging due to the difficulty of speaking to a device and instead of a face-to-face conversation. We are also still taunted by the public stigma which involves the discriminatory attitudes to those with mental illness. People are still reluctant to share their data especially when it involves recordings of either audio or video. Male subjects are more reluctant to speak and express themselves compared to women. Thus, we were only able to obtain a small size database for male and an adequate number of samples for female.

Table 1. Summary of depression speech corupus used by previous researchers

1st Published (Name)	Language	Significant acoustic features	Subjects
Vanderbilt II Study France *et al*. (2000) [15]	English (USA)	Fundamental frequency, amplitude modulation	115: 59 Dep, 22 HR, 34 Nat
Moore, Clements, Peifer, and Weisser (2004) [11]	English (USA)	Glottal waveform	33: 15 Dep (6 M, 9 F), 18 Nat (9 M, 9 F)
Mundt, Snyder, Cannizzaro, Chappie, and Geralts (2007) [23]	English (USA)	Formants, pauses, vocalization and speaking rate	35: Dep (15 M, 20 F)
Cohn *et al.* (2009) [12]	English (USA)	Fundamental frequency and respond time	57: Dep (24 M, 34 F)
Oregon Research Institute Low, Maddage, Lech, and Allen (2009) [13]	English (USA)	MFCC	139: 68 Dep (49 F, 19 M) 71 Nat (27 M, 44 F)
Black dog Institute Alghowinem *et al*. (2012) [14]	English (Australia)	Fundamental frequency, energy, intensity, loudness, jitter, shimmer, Harmonic-to-noise-ratio (HNR), voice probability and quality, formants, and MFCCs.	80: 40 Dep, 40 Nat
Shankayi, Vali, Salimi, and Malekshahi (2012) [24]	Persian (Iran)	Pitch, energy, formant, glottal features	56 (Dep, Nat, Rem)
Cincinnati Children's Hospital Medical Centre Scherer, Pestian, and Morency (2013) [25]	English (USA)	Energy, fundamental frequency, peak slope	60: 30 HR, 30 Nat
Audio-Visual Depressive Language Corpus (AViD Corpus) Valstar *et a*l. (2013) [26]	Germany (German)	Open corpus	292 files each containing a range of mix of vocal, free and read speech tasks
Audio-Visual Depressive Language Corpus (AViD Corpus) [27]	Germany (German)	Open corpus	150 files each containing a range of mix of vocal, free and read speech tasks
Distress Analysis Interview Corpus (DAIC-WOZ) [28]	English (USA)	Open corpus	76: Nat (49 M, 27 F) 31: Dep (14 M, 17 F)
University of Nottingham Solomon *et al*. (2015) [29]	English (UK) Local and international students	Energy, pitch, MFCC, articulation rate and total time.	17 (13 F, 4 M): 9 Dep, 8 Nat
Liu *et al.* (2015) [17]	Chinese (China)	N/A	300: 100 HR, 100 Dep, 100 Nat
Kiss *et al.* (2016) [18]	Hungary (Hungarian) And Italy (Italian)	Fundamental frequency, volume dynamics, formants, jitter, shimmer, rate of transient, pauses.	Hungarian: 54: Dep (19 M, 35 F) 73: Nat (29 M, 44 F) Italian: 22: Nat (11), Dep (11)
Azam *et al.* (2016) [16]	Malaysia (Bahasa Malaysia)	MFCC	13: Dep (6 M, 7 F) 20: Nat (10 M, 10 F)
Hashim *et a*l. (2017) [30]	English (USA)	Timing-based features, MFCC and power spectral density	116: Dep (44 M, 72 F)
Pan *et al*. (2019) [9]	Chinese (China-female only)	Intensity, loudness, zero-crossing rate, MFCC, line spectral pair, voicing probability, fundamental frequency.	Database 1: 1132: (584 Dep, 548 Nat) Database 2: 904: (500 Dep, 404 Nat)

Abbreviation: Dep–depressed, HR–high risk suicidal, Nat–natural, Rem–remitted, PTSD–post traumatic stress disorder, M–number of males, F–number of females.

2. DATABASE

This study was conducted during the first wave of COVID-19 in Malaysia. We gathered recordings of depressed and healthy subjects using online platform due to the limitation of visiting the hospitals and having a face-to-face data collection with subjects on site. All procedures performed in studies involving human participants were in accordance with the ethical standards and has been approved by the IIUM research ethical committee (IREC 2019-006). The database was divided based on gender and diagnostic groups of depressed and healthy. Subjects consisted of 43 depressed and 47 healthy were required to sign an informed consent and to be neither under the influence of alcohol, toxicity, nor experiencing respiratory problems. Table 2 summarizes the database demographic characteristics.

Table 2. Summary of database characteristics

Demographic	Subject categorized as depressed	Subject categorized as healthy
characteristics	(n=43)	(n=47)
Gender		
Female	32 (74.42%)	34 (72.34%)
Male	11 (25.58%)	13 (27.66%)
Age (mean □ SD)	28.11±5.71	25.28±7.48
Age (range)		
≤20	1 (2.33%)	3 (6.38%)
21-25	14 (32.56%)	30 (63.83%)
26-30	8 (18.60%)	12 (25.53%)
31-35	12 (27.91%)	1 (2.13%)
>35	8 (18.60%)	1 (2.13%)
Marital status		
Single	29 (67.44%)	44 (93.62%)
Married	11 (25.58%)	3 (6.38%)
Divorced	3 (6.98%)	0
Death of spouse	0	0
Number of children		
none	33 (76.74%)	43 (91.49%)
1-2	5 (11.63%)	3 (6.38%)
3-4	5 (11.63%)	0
5 or more	0	1 (2.13%)
Academic qualification		
Secondary school	4 (9.30%)	1 (2.13%)
Diploma/Degree	19 (44.19%)	2 (4.26%)
Master	17 (39.54%)	39 (82.98%)
PhD	3 (6.98%)	5 (10.64%)
Working status		
Full-time	19 (44.19%)	6 (12.77%)
Part-time	2 (4.65%)	2 (4.26%)
Studying	9 (20.93%)	15 (3.19%)
Housewives	3 (6.98%)	1 (2.13%)
Not working	10 (23.26%)	23 (48.94%)
Working sector		
Government	11 (25.58%)	6 (12.77%)
Private	7 (16.28%)	5 (10.64%)
Self employed	4 (9.30%)	1 (2.13%)
Others	21 (48.84%)	35 (74.47%)
Health level		
Not healthy	8 (18.60%)	0
Moderate	19 (44.19%)	4 (8.51%)
Very healthy	16 (37.21%)	43 (91.49%)

Subjects were asked to fill in an online form that consists of three sections. The first section gathers the demographic information such as age, marital status, number of children, education level, working status, working sector and physical health. Subjects were also asked to report whether they have received a diagnosis of MDD by the hospital's psychiatrist. The second section requires the subject to fill in the Malay BDI-II and PHQ-9.

The final section requires the subject to send a voice note recording through the application called WhatsApp voice note. Two types of recordings were collected which are the spontaneous and reading speech. For spontaneous speech, subjects were asked to answer a question in Bahasa Malaysia about their situation in the past three days which includes any activities or feelings they have experienced. They were also asked about what situation that usually causes them to feel stressed. For reading speech, patients were asked to read a standardized Bahasa Malaysia passage called Cerita Datuk that is commonly used by speech therapists. The recordings received in the voice note were in OGG and MP4 format. Speech files were converted to WAV using the audio.online-convert.com at a sampling rate of 44.1 kHz and 32 bits per sample, with a mono channel. The recordings were normalized prior to the acoustic parameter extraction. Each audio signal was then divided into 20 second segments and acoustic features were extracted for each segments. The total number of 20 second segments for female reading speech, male reading speech and male spontaneous speech are 198, 73, and 156, respectively. We omit the dataset for female spontaneous speech due to poor classification performance. The subjects were categorized as depressed if he/she has a score of BDI-II of 20 or more, a score of PHQ-9 of 16 or more and also has been diagnosed as MDD formally by the hospital's psychiatrist. Healthy subjects were chosen based on their score of minimal range (0-13) for BDI-II and mild range (0-5) for PHQ-9.

3. RESEARCH METHODOLOGY

3.1. Acoustic parameter extraction

Prior to this work, we performed an analysis on robust parameters to identify acoustic parameters that are robust towards multiple mobile recording devices [28]. The reason we conducted the analysis is because of our method of collecting the speech recordings remotely using subject's personal mobile devices with different microphone specifications. Timing-based parameters such as transition parameters (TP), interval length probability density function (ILpdf) and amplitude modulation statistics (AM) were found to be less prone to suffer from microphone variability with coefficient of variance (CoV) less than 30%. Mel-frequency cepstral coefficient (MFCC) parameters were the most affected by microphone variability but power spectral density (PSD) band 1 (0 to 500 Hz) and band 2 (501 to 1000 Hz) were shown to be robust with CoV less than 20%. Therefore, we decided to proceed with the parameters listed in Table 3. All the parameters were extracted using matrix laboratory (MATLAB) software.

Table 3. List of selected robust parameters

Parameter type	List of paramerters
Time-based	t11, t31, t33, t13
	v1, v2, sil1
	avgAM, covAM, varAM, rangeAM, maxAM
Spectrum-based	PSD1, PSD2

3.1.1. Transition parameters (TP)

Each speech signal is divided into a 40-ms frames and labeled as voiced (1), unvoiced (2) and silence (3). Audio speech signal is quasi-periodic for a length of approximately 20-ms to 40-ms. However, due computation time, we use the maximum 40-ms frame. Labels were then concatenated into one sequence of samples (labeled sample sequence). Knowing the future labels in the sequence, we set the states equal to the sequence (not hidden) and estimate the Markov model. The output will be an emission matrix which equals to an identity matrix and a 3-by-3 matrix of transition parameters that describes the probability of being in the current state (row) to the next state (column) within the whole sequence. After the analysis of robust parameter, only four parameters had CoV of less than 30% which are transition probability from voiced-to-voiced (t11), silence-to-voiced (t31), silence-to-silence (t33) and voiced-to-silence (t13).

3.1.2. Interval length probability density function (ILpdf)

The frequency of consecutive 2-ms frames for voiced and silence segments within one labeled sample sequence were extracted and re-distributed into histogram bins of up to five for voiced and ten for silence. Analysis of robust parameters reveled that 2-ms consecutive frames of voiced (v1), 4-ms consecutive frames of voiced (v2) and 2ms of consecutive frames of silence (sil1) were the most robust towards microphone variability with less than 20% CoV. We therefore proceeded with these parameters. The idea of TP and ILpdf parameters are to represent the probability and distribution of pauses, unvoiced sound and voiced sound within a speech signal.

3.1.3. Amplitude modulation (AM)

Each speech signal is squared in order to push half of the energy signal up to higher frequencies and the other half lower towards the mean. The squared signal is then sent through a low-pass filter to eliminate the high frequency energy in speech that are not significant and has low information. To maintain the correct scale, we then take the square root of the signal to reverse the scaling distortion that was due to the squaring in the earlier stage. Finally, the signal is downsampled to reduc e sampling frequency. Five robust statistical measurements were calculated from the envelope which are the average, coefficient of variance, variance, range and maximum.

3.1.4. Power spectral density (PSD)

Speech signal comprised of voiced, unvoiced, and short silence segments that are mixed. A set of third order band-pass filters was applied to each frame segment of the sampled signal. All unvoiced and silence terms were removed and only the voiced terms were collected and concatenated into one new signal. Four normalized equal bands of PSD were obtained using the method of Periodogram and trapezoidal numerical integration for frequencies between 0 to 2000 Hz. The final band was removed due to linear dependent on the other three spectral energy bands. The PSD is calculated on every non-overlapping of 40-ms window frame and averaged over the total number of frames per speech signals.

3.2. Classifiers

Four classifiers were chosen to be reported which are the support vector machine, k-nearest neighbour, random forest, and extreme gradient boosting. These classifiers were determined based on the pairplot distribution on the respective dataset. The pairplot distribution for all dataset categories of both gender reading and spontaneous speech demonstrated overlapping and non-linear behaviour. Thus, from the observation, linear classifiers are not suitable. However, we proceeded to perform training and testing on linear regression and naïve Bayes for verification, and we confirmed our initial observation when the accuracy for these linear classifiers were significantly low and distributed around 0.5.

3.2.1. Support vector machine (SVM)

SVM was chosen as one of the classifiers because of the small number of dataset and it is also one of the most common method of binary classification with discrete target variable. Although SVM is a linear classifier, it can also perform a non-linear classification by changing the kernel model. The idea behind SVM is to find the best boundary or hyperplane that best separates parameters in a two-classes dataset. With N being the number of parameters, the separating hyperplane is an (N-1)-dimensional subspace. SVM will choose a decision hyperplane that maximizes the distance between support vectors (sample data points) to the hyperplane. Three main parameters that can be reading in SVM are kernel, C parameter and gamma.

The kernel functions such as linear, polynomials, radial basis function (RBF) and sigmoid measure the distance of support vectors that are transformed into high-dimensional space to make them more linearly separable when running the machine learning algorithm. The result will then be transferred back to the original input space such as class predictions. However, due to the transformation into higher dimensionality spaces, applying the kernels may cause overfitting. C and gamma can be used for regularization parameters. Gamma acts as the multiplier in the kernel functions where higher gamma value corresponds to a closer fit but higher chance of overfitting. C parameter (penalty/cost) controls the trade-off between increasing the distance between hyperplane to support vectors or minimizing misclassification in the training set. For a small value of C, penalty for misclassification is reduced and thus, higher margin of hyperplane will be chosen, which will result in greater misclassification.

3.2.2. K-nearest neighbour (KNN)

KNN is known to perform better with smaller number of features. KNN classifies a new sample point based on a similarity measure such as Euclidean distance or Manhattan distance. K-value denotes the number k-nearest neighbour’s classes which are the voting class of the new sample data point. The label of the new sample point will be determined based on the most common label closest to most classes in the k sample data points. Two main parameters in KNN are the k-value and the weights. The performance of KNN is mostly affected by the k-value where a smaller k-value might cause the algorithm to be more sensitive towards outlier and a larger k-value might cause the algorithm to include too many sample points. To overcome this issue, a weighted-KNN can be applied. Two options for weight parameter are uniform and distance. Uniform weight indicates that all neighbours get an equally weighted vote, however, distance weight gives a higher weight using a kernel function such as inverse distance function to the nearest k-sample point and less weight to k-sample point that are further than the test sample point.

3.2.3. Random forest (RF)

RF is an ensemble method that trains several decision trees in parallel on multiple subsets of training dataset using different subset of parameters. The final decision combines the decision of each individual trees by means of voting, thus exhibits good generalization. It does not suffer from overfitting because it takes the average of all decisions. However, the process is time-consuming due to having multiple decision trees making predictions for the same input.

The main parameters that are available for tuning are the n-estimator, maximum features, maximum depth and feature importance (criterion). N-estimator is the number of decision trees in the forest. Maximum depth represents the depth of each tree in the forest where more splits correspond to capturing more information about the data. Maximum features determine the number of features to consider when deciding for the best split. The importance of each data feature is calculated using attribute selection indicator called information gain or gini impurity. It captures how much the classifier model’s accuracy decreases when a data feature is dropped. However, for this work, we only used n-estimator and criterion because the initial procedure already performed feature reduction and due to the low number of features, we decided to proceed with a deeper split of trees.

3.2.4. Extreme gradient boosting (XGBoost)

XGBoost uses a gradient boosting framework for problems involving structured or tabular datasets. XGBoost includes boosting where models are built sequentially by minimizing errors from previous models while assigning higher weight (boosting) on the misclassified data points to influence higher performance model. It works on the principle of ensemble where it combines a set of weak learners such as a single split decision tree to improve the model accuracy. The errors in sequential models are minimized using gradient decent algorithm. The output is aggregated along the way instead of at the end of the process.

Multiple parameters can be tuned in XGBoost, however we focused on five parameters which are the learning rate, gamma, maximum depth, minimum child weight and column sample by tree. The learning rate corresponds to how fast the error is corrected sequentially from previous tree to the next which is calculated based on the difference between the two. Gamma reflects the minimal loss reduction that is required to make a split. Maximum depth, minimum child weight and column sample by tree are parameters used in gradient boost that controls depth of tree, minimum sum of weights for all observations and the fractions of columns to be randomly sampled, respectively.

3.3. Optimization

In machine learning, feature selection and hyperparameter tuning are two essential processes to improve the performance of the algorithm. Feature selection removes irrelevant features while increasing the performance of the classifier model and also reduces the number of features in order to reduce computational cost and preventing over-modelling. Hyperparameter tuning is an automatic optimization of machine learning parameters that are specifically performed on specific classifier, feature set and dataset. Different dataset often produces different optimal hyper-parameter settings.

Figure 1 shows the optimization procedure for identifying the best feature set and classifier model using a two-stage optimization process which are the exhaustive feature selection (EFS) and cross-validation grid Search hyperparameter tuning. The EFS algorithm is a wrapper method that uses brute-force evaluation of feature sets. The best feature is selected by optimizing the performance metric such as accuracy score, given a specific classifier. In the first stage, the EFS was executed using each classifier (LR, NB, SVM, KNN, RF and XGBoost) for 200 iterations with each iteration having a different train-test split dataset and using 5-fold cross-validation. The analysis performed in this work uses 70% training and 30% test data (70-30) split. This way, the feature set obtained is expected to be more generalizable. Each iteration outputs a feature set with the best accuracy score. The same feature set with best accuracy score that has a majority of five or more was then selected for validation using its respective classifier.

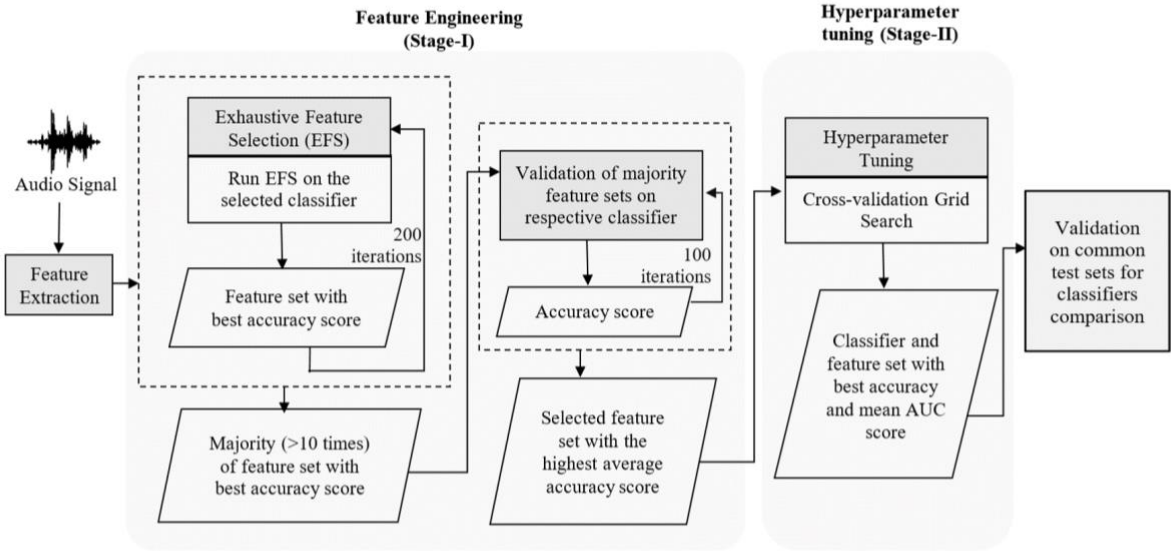

Feature Engineering (Stage-I)
Hyperparameter tuning (Stage-II)
Audio Signal
Feature Extraction
Exhaustive Feature Selection (EFS)
Run EFS on the selected classifier
200 iterations
Feature set with best accuracy score
Majority (>10 times) of feature set with best accuracy score
Validation of majority feature sets on respective classifier
100 iterations
Accuracy score
Selected feature set with the highest average accuracy score
Hyperparameter Tuning
Cross-validation Grid Search
Classifier and feature set with best accuracy and mean AUC score
Validation on common test sets for classifiers comparison

Figure 1. Optimization process using feature engineering and hyperparameter tuning

The performance of the selected feature sets was then validated by executing the classification for 100 iterations using its respective classifier with different test-train set. Each iteration outputs the accuracy score and the classifier model with the feature set that produces a good performance of average accuracy score was selected for hyperparameter tuning. Finally, we tested on one common test set and the classifier with feature set with the best accuracy score and the best mean of area under the receiver operating characteristic (ROC) curve (AUC) was selected. The AUC was calculated for 10 folds cross-validation and the score ranges from 0 to 1 where 1 indicates a 100% correct classification. Different train-test set produce different accuracy scores, thus multiple iterations are required to obtain a suitable and generalizable model. The iterations are chosen based on computation power and time limitation but adequate to generate a normal distribution.

4. RESULTS AND DISCUSSION

4.1. Comparison between spontaneous and reading speech

Based on Table 4, all feature sets consist of at least one time-based feature related to pauses and vocalization which demonstrates the significance of TP in distinguishing between the speech of depressed and healthy. We suspect that recognition of depression from spontaneous speech is cognitively more challenging as compared to reading speech due to the planning required for selecting proper words and controlling the articulatory motor. During data collection, subjects experienced difficulties in coming up with ideas on what to say to fulfil the recording minimum time of one minute. This might also be the effect of non-face-to-face interview. Talking to a device was considered too artificial for them to express information, articulate ideas and sharing emotions as compared to a face-to-face interaction. Thus, increasing the significance of the time-based features as compared to reading a standardized passage. This effect is prominently seen in the male spontaneous speech category where depressed subjects have more difficulty portraying and expressing their thoughts as opposed to healthy male subjects.

The feature values for the female spontaneous speech were predominantly categorized as depressed and thus, was not included in this report. Although the accuracy of the time-based feature classification was high with approximately zero false rate in the depressed group, however, all classifiers that we tried were also classifying majority of healthy speech as depressed, which shows no significant difference between the spontaneous speech of healthy and depressed subjects. This might indicate that Malaysian female, regardless of feeling depressed or not, have a strong inclination towards expressing themselves. When asked about the common condition that may cause them to feel stressful, judging the content of the speech and listening to their tone and rhythm of expression, healthy female subjects tend to express their stressful condition in detail, passionately and a noticeable degree of emotion.

4.2. Comparison of classifiers performance on depressed speech classifications

Table 4 shows the best classifier performance, selected features, and parameter values after performing feature selection and hyperparameter tuning. Classifiers with accuracy and AUC value of less than 0.6 were omitted and not listed in Table 4. The features were selected based on the majority times the feature set produces a high accuracy using EFS as compared to the other sets and produces high accuracy when trained and tested on the respective classifiers. On the other hand, the parameters were selected according to the grid search (GridSearchCV) best score. The test sets are considered balance with depressed to healthy ratio of 5:6 for male reading, 8:7 for female reading, and approximately 7:5 for male spontaneous speech.

When comparing classifiers performance on a balance dataset, appropriate combination of these metric; the f1 score, precision, recall, accuracy, and AUC score should be evaluated. In this work, precision addresses the following question: Of all those who we classified as depressed, how many were actually depressed? while recall addresses the following question: Of all those who are depressed, how many were correctly classified? However, precision and recall are often conflicting. Improving precision usually reduces recall and vice versa. Precision and recall depend on the threshold of the classifier model and the performance of the classifier would vary as the threshold changes.

In this study, the positive refers to the depressed speech and the negative refers to the healthy speech. It is essential to identify depressed subjects correctly and thus, for the model to be considered effective, its true positive rate should be high even at the expense of a few false positive being misclassified. However, a classifier model with a high balance of true positive and true negative would be ideal. Therefore, the classifier with a higher recall than precision would be preferable since it is more desirable to detect as many depressed speeches as possible. On the other hand, f1 score captures the balance between the precision and recall value. However, for this study, it is not necessary to maximize the f1 score considering that we may prefer a higher recall, but it is also important to have an f1 score in the upper range.

Accuracy captures the fractions of correctly assigned positive and negative classes. That means if our problem is highly imbalanced, we get a significantly high accuracy score by simply predicting that all observations belong to the majority class. A high accuracy could indicate a high percentage of true positive but significantly low true negative as demonstrated in the performance of classifiers using female spontaneous speech. Lastly, the ROC curve is more effective in capturing the effect of all thresholds between 0 to 1. The ROC curve plots the true positive rate (sensitivity) with respect to the false positive rate (1-specificity). The area under the ROC curve is captured by AUC score. The higher the AUC score, the better the classifier model at predicting the depressed as depressed and healthy as healthy.

Figures 2 to 4 show the box plot distribution of classification performance for each category of speech signal. Table 5 shows the detailed classifier performance on the selected classifiers based on the analysis of the box plot distribution in Figures 2 to 4. Precision-Recall p-value indicates the significance between mean distribution of precision and recall. Based on a two sample mean t-test with equal variance, we accepted the classifier model with precision higher than recall if the p-value is larger than 0.05 due to the insignificant difference between the two-mean distribution.

Table 4. Selected feature set and classifier model with parameter values

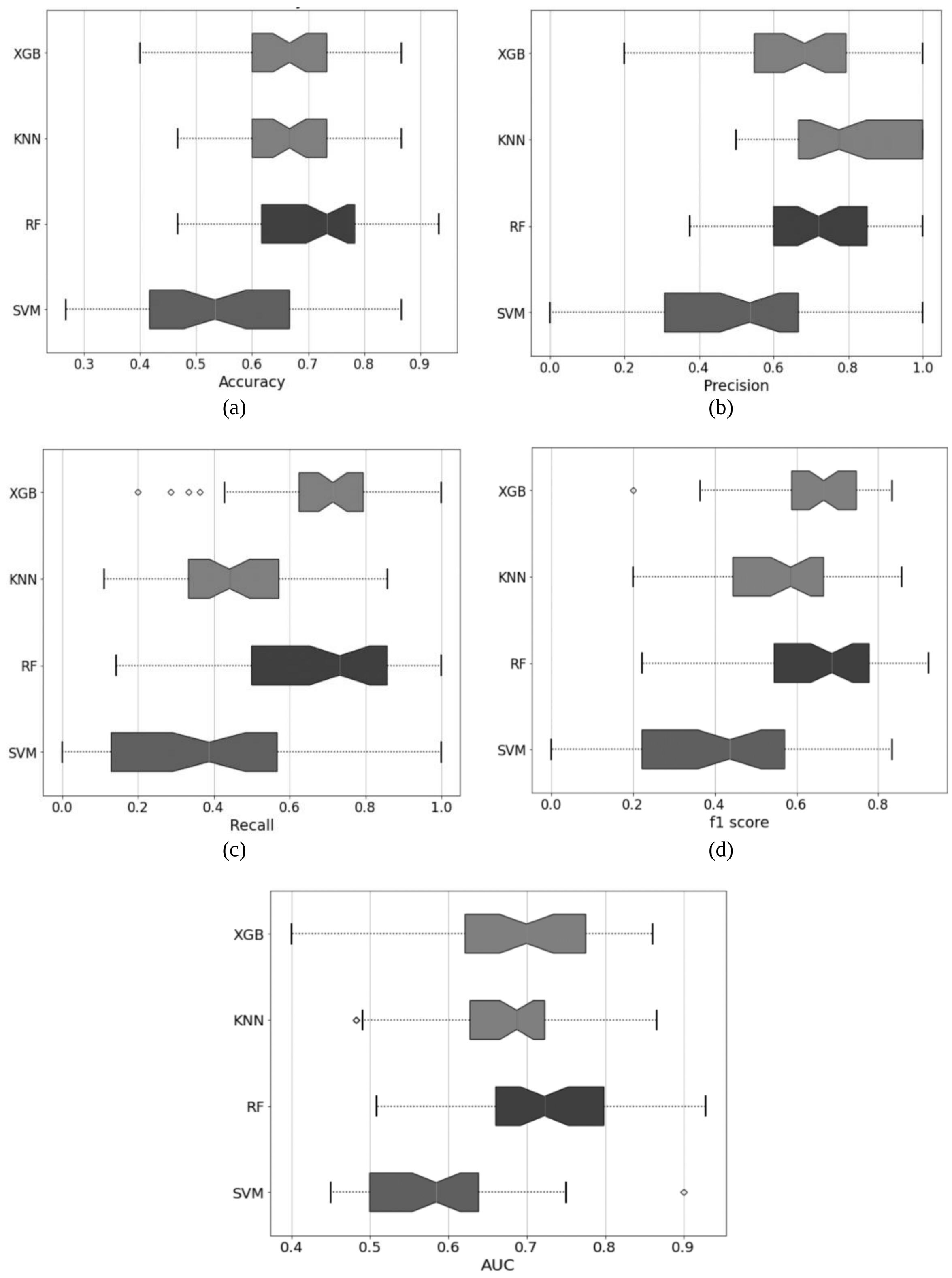
Speech signal type	Classifier	Features	Classifier Parameters	Grid best score (Hyperparameter tuning)
Male Reading Speech (Reading passage)	SVM	t13	C: 100, gamma: 0.3, kernel: RBF	0.6301
	KNN	t33, psd1, psd2, avgam	n:2, weights: uniform	0.6575
	XGBoost	t33, sil1, psd1	min_child_weight: 3, max_depth: 10, learning_rate: 0.2, gamma: 0.2, colsample_bytree: 0.7	0.8103
	RF	t13, psd1, rangeam	criterion: entropy, n_estimator: 90	0.8259
Male Spontaneous Speech (Question and Answer)	SVM	sil1, psd2, covam	C: 50, gamma: 0.8, kernel: RBF	0.7078
	KNN	t11, t31, sil1, avgam	n: 4, weights: distance	0.7468
	XGBoost	t11, t31, sil1	min_child_weight: 7, max_depth: 3, learning_rate: 0.2, gamma: 0.4, colsample_bytree: 0.8	0.7403
	RF	t11, t31, sil1, varam	criterion: gini, n_estimator: 60	0.7273
Female Reading Speech (Reading passage)	SVM	t31, t33, v1	C: 100, gamma: 0.7, kernel: sigmoid	0.7222
	KNN	t13, t33	n: 8, weights: uniform	0.7121
	RF	t11, t33, sil1	Criterion: gini, n_estimator: 160	0.7272

Table 5. Detailed classification performance on the selected classification

Dataset	Classifier	Average Accuracy	Average AUC	Average f1 score	Average Precision	Average Recall	Precision-Recall p-value
Male Reading	RF	0.7000	0.7127	0.6688	0.7223	0.6859	0.3653
	XGB	0.6733	0.6834	0.6571	0.6777	0.6792	-
Male Spontaneous	RF	0.7629	0.7657	0.7811	0.8013	0.7754	0.0836
	XGB	0.7377	0.7366	0.7628	0.7728	0.7679	0.7320
	SVM	0.7368	0.7374	0.7616	0.7711	0.7662	0.7321
Female Reading	SVM	0.7020	0.7075	0.7022	0.7312	0.6952	0.1281
	RF	0.6970	0.6986	0.7094	0.7110	0.7201	-

Figure 2(a) compares the accuracy of the classifiers for male reading speech, followed by comparisons of precision in Figure 2(b), recall in Figure 2(c), f1 score in Figure 2(d), and AUC in Figure 2(e). Based on the accuracy and AUC box plot, RF has the highest value followed by XGB and KNN. Eventhough XGB's scores are slightly lower than 0.7, we include it Table 5 because of its consistency and similar performance to RF. With a larger dataset, we believe that XGB can improve its performance. However, KNN classifier's precision is significantly higher than recall, which can also be seen in the low range of f1 score. Thus, RF and XGB were selected as the best classifier on the male reading speech dataset.

Figure 3(a) compares the accuracy of the classifiers for male spontaneous speech, followed by comparisons of precision in Figure 3(b), recall in Figure 3(c), f1 score in Figure 3(d), and AUC in Figure 3(e). All classifiers show similar performance for accuracy. However, from the precision and recall plots, KNN can be excluded due to its precision distribution being significantly higher than recall (p-value>0.05) and a low f1 score as compared to others. This shows that KNN is better at predicting the healthy speech based on the higher precision value. Therefore, we included RF, XGB and SVM in the list of selected classifiers in Table 5.

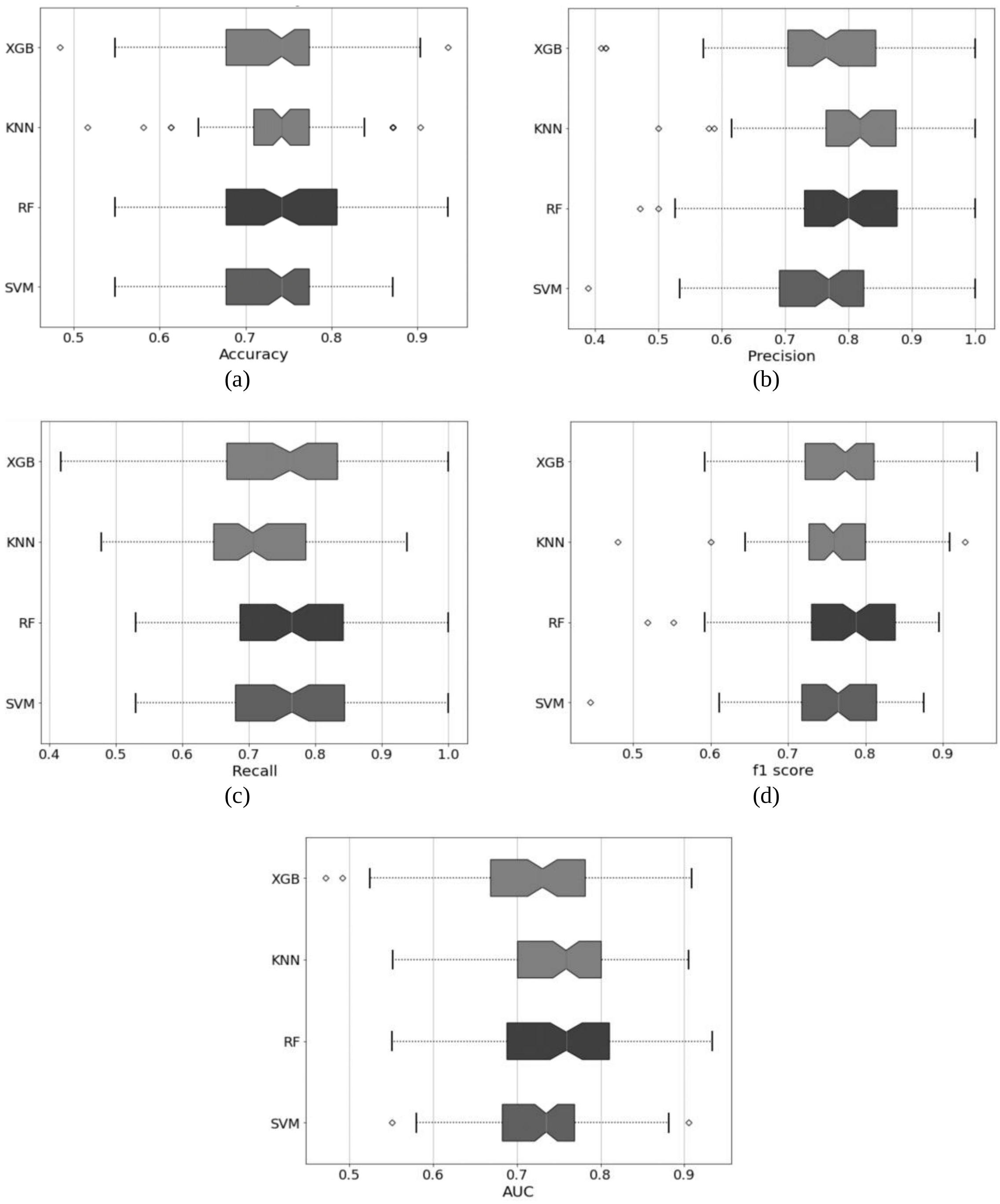


(a) (b)

(c) (d)

(e)

Figure 2. Classifiers performance on male reading speech with respect to (a) accuracy, (b) precision, (c) recall, (d) f1 score, and (e) AUC



(a) (b)

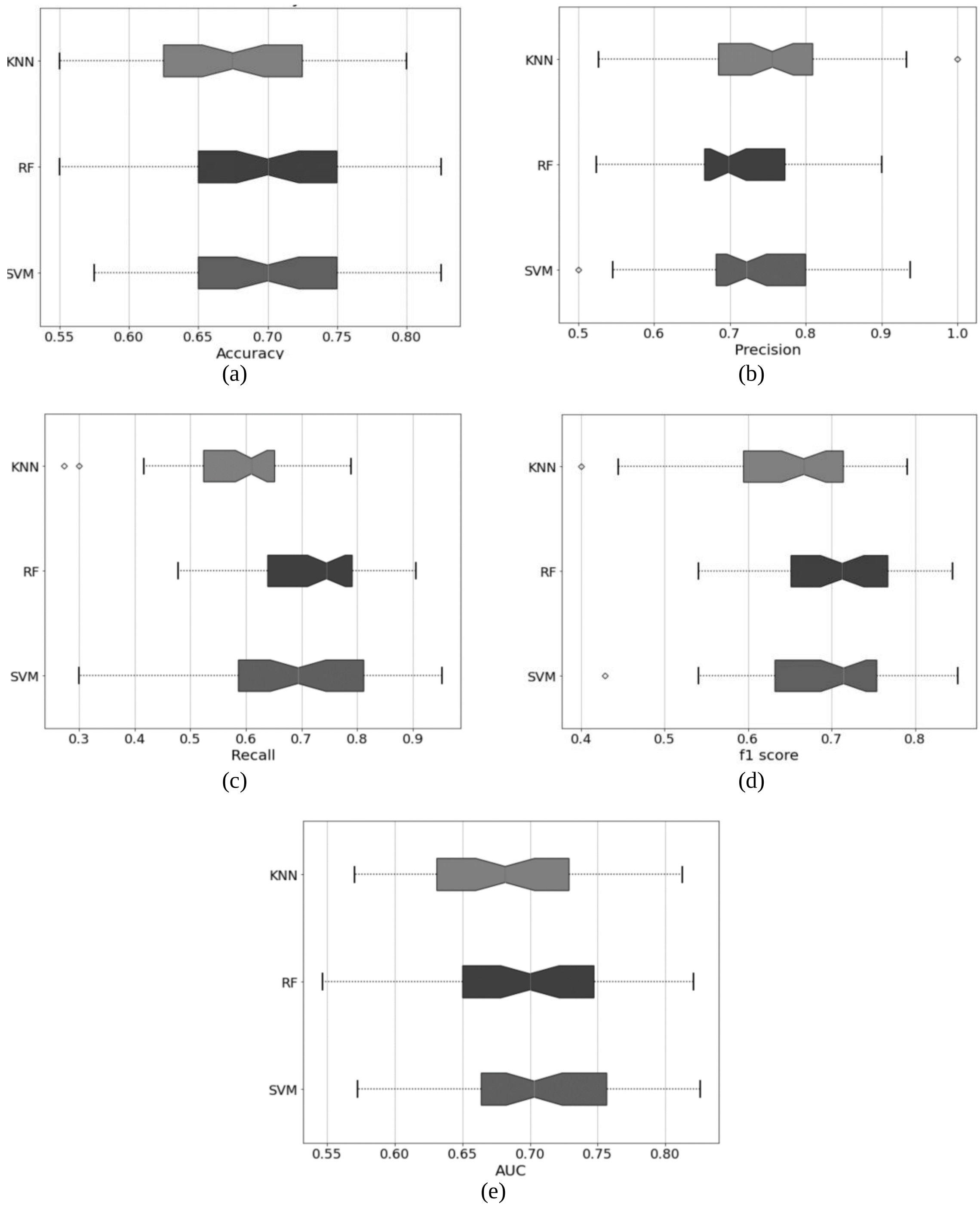
(c) (d)

(e)

Figure 3. Classifiers performance on male spontaneous speech with respect to (a) accuracy, (b) precision, (c) recall, (d) f1 score, and (e) AUC

Figure 4(a) compares the accuracy of the classifiers for female reading speech, followed by comparisons of precision in Figure 4(b), recall in Figure 4(c), f1 score in Figure 4(d), and AUC in Figure 4(e). We excluded KNN based on its low performance on accuracy. KNN also displays low f1 score distribution and a significantly lower recall as compared to precision. Thus, using the robust features for female reading speech, the classification works best with SVM and RF.

Overfitting occurs when the model performs well on the training data but performs poorly on the validation set. To ensure our model was not overfitting, we obtained the training accuracy score and compare it with the validation accuracy. Avoiding overfitting was also considered with using a small number of features by removing redundant features, integrating cross-validation during training and evaluating it over a number of iterations. However, it is best to re-check on the training accuracy and validation accuracy. If the accuracy for training data reaches 1.0, it means the model has memorized the training data instead of learning from the features. Also, a maximum of four features is an optimal dimension for this dataset size following the rule of thumb of one feature per 10 samples. Referring to Figure 1, we noticed that increasing the number of features beyond four does not significantly increases the performance of the classifier during the process in stage-I and it also significantly increases the computational time due to the method of exhaustive feature selection (EFS).



(a)

(b)

(c)

(d)

(e)

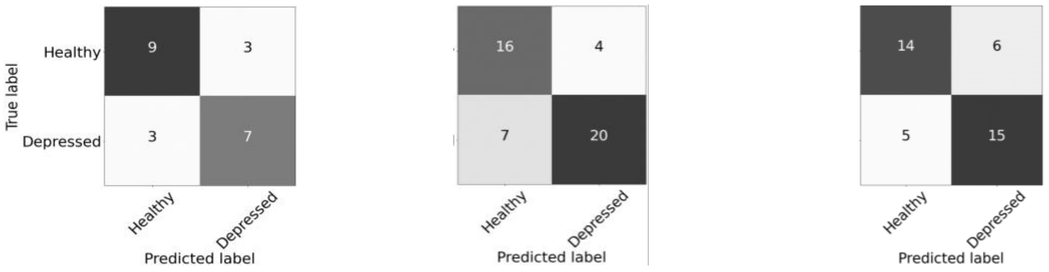
Figure 4. Classifiers performance on female reading speech with respect to (a) accuracy, (b) precision, (c) recall, (d) f1 score and (e) AUC

Table 6 shows the classification report for each category of speech signals. The classifications were analyzed by evaluating the performance on not just the accuracy, but also AUC, precision, recall and f1. Eventhough the performance is within the range of 70 to 80%, they showed consistency in all of classification variables, thus indicating that these are good models for the depression detection using Bahasa Malaysia language. However, with more data and information, we believe that the model can improve.

Table 6. Classification report

Data category	Male Reading			Male Spontaneous			Female Reading		
Classifier	RF (n_estimator=10)			SVM			SVM		
Classification report	Precision	Recall	f1	Precision	Recall	f1	Precision	Recall	f1
Healthy	0.75	0.75	0.75	0.70	0.76	0.73	0.74	0.70	0.72
Depressed	0.70	0.70	0.70	0.79	0.73	0.76	0.71	0.75	0.73
Macro average	0.72	0.72	0.72	0.74	0.75	0.74	0.73	0.72	0.72
Weighted average	0.73	0.73	0.73	0.75	0.74	0.74	0.73	0.72	0.72
Training accuracy	0.9411			0.7383			0.7225		
AUC score	0.7500			0.7949			0.7133		
Validation accuracy	0.7273			0.7447			0.7250		

Using the best classifier from the list in Table 5, we validated and trained the models with the respective feature set and classfier parameters on one selected data set with 70% training and 30% test. Figure 5 shows the Confusion matrix with each explanation Figure 5(a) shows for the male reading speech with random forest classifier, Figure 5(b) for the male spontaneous speech using random forest classifier and Figure 5(c) for female reading speech using support vector machine classifier.



(a) (b) (c)

Figure 5. Confusion matrix for (a) male reading speech with random forest classifier, (b) Male spontaneous speech with random forest classifier, and (c) Female reading speech with support vector machine classifier

For male reading speech, we selected random forest but reduce the number of n_estimator to 10 since there was no significant change on the accuracy. For a tree-based models like RF, we can control the overfitting by tuning the parameters and we have demonstrated the slight difference of the hyperparameter tuning for this dataset. For the male spontaneous speech, RF, XGB and SVM performed equally, and therefore we selected the least complex model, which is SVM. For female reading speech, SVM outperform RF and therefore, we selected SVM as the model for this category.

Based on Table 1, depression identification using speech has been studied with multiple languages and so far, for Bahasa Malaysia, only a small database has been gathered and studied. The work in [16] shows the effectiveness of MFCC on classifying the depression and healthy speech. Based on the literature that shows ADD is dependent on language and the nature of this study's database collected using multiple recordings, we cannot compare our findings with the output reported in [16] because of the dependencies of MFCC on recording devices [28]. However, for a language independent based comparison, a simple classifier such as SVM has shown to be powerful in identifying depression speech [14], [20] which is similar to our findings.

A study using different languages such as English [31] and Japanese [32] have shown that timing-based feature is a promising biomarker for depression speech. Initial studies on pauses and silence found that pause time was longer for depressed patients than in healthy speech. Pause time shows correlation with the questionnaire scores where response time was longer for depressed patients than for healthy patients, and also response time became shorter as questionnaire score decreases [33]. In this work, timing-based features also demonstrate the effectiveness as a feature to distinguish between depressed and healthy speech of Bahasa Malaysia in both male and female.

During the optimization process, there were two issues present pertaining to which procedure to perform first. Selecting the features will have an impact on the optimal hyperparameters and the selection of the hyperparameters will also affect the optimal feature set. This issue can be resolved by performing a unified feature selection and hyperparameter tuning. But, due to the computational time and cost, a unified procedure will be challenging to performed. We proceeded with performing the feature engineering as the first step before proceeding with hyperparameter tuning. Feature engineering is the process of transforming information into an underlying pattern that better represents the overall information but with minimal information, in this case, in the form of acoustic features. This process involves the feature selection. The hypothesis is that the classifier model can give better results if it has found a pattern that is more descriptive of the overall information. In this work, we have seen that the hyperparameter tuning does not significantly improve the performance of the classifier and thus, justifies the separating procedure of stage-I and stage-II. However, we still believe that hyperparameter tuning is crucial as they control the overall behaviour of the machine learning by searching for an optimal combination of hyperparameter that minimizes predefined loss function.

5. CONCLUSION

The classifiers performance demonstrate the effectiveness of robust features for automatic depression detecting using Bahasa Malaysia language. There was no significant difference between the speech of depressed and healthy for the female spontaneous category, but RF classifier performed well for male reading speech and SVM performed well for male spontaneous and female reading speech. Depressive symptoms yield changes in the speech timing of pauses, even in standardize reading. It also shows that recordings through mobile applications or multiple recording devices is also possible, provided with appropriate robust feature sets. It is important to develop a system that can receive input from various microphone specifications such as a mobile app where users will provide speeh input using their personal mobile devices with its own microphone specification. With the output of this study, researchers in the field of ADD are not only limited to gathering data on site with a constant recording device but will also be able to use online platform. A study on a larger collected database will allow for verification on the effectiveness of reading versus spontaneous speech for both genders, and also able to implement ensemble machine learning and deep learning with multiple severity level classification.

ACKNOWLEDGEMENT

This work was supported by funding from the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2018/TK04/UIAM/02/7).

REFERENCES

[1] E. K. Mościcki, “Epidemiology of completed and attempted suicide: toward a framework for prevention,” *Clinical Neuroscience Research*, vol. 1, no. 5, pp. 310–323, Nov. 2001, doi: 10.1016/S1566-2772(01)00032-9.

[2] World Health Organization, *The impact of COVID-19 on mental, neurological and substance use services*. 2020.

[3] L. Baer and M. A. Blais, Eds., *Handbook of clinical rating scales and assessment in psychiatry and mental health*. Totowa, NJ: Humana Press, 2010.

[4] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin, “Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment,” *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 35, no. 2, pp. 395–409, Mar. 2011, doi: 10.1016/j.pnpbp.2010.10.019.

[5] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, “Cross-cultural depression recognition from vocal biomarkers,” in *Interspeech 2016*, Sep. 2016, pp. 1943–1947, doi: 10.21437/Interspeech.2016-1339.

[6] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, “Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated CNNs,” in *Interspeech 2020*, Oct. 2020, pp. 4561–4565, doi: 10.21437/Interspeech.2020-3135.

[7] N. S. Srimadhur and S. Lalitha, “An end-to-end model for detection and assessment of depression levels using speech,” *Procedia Computer Science*, vol. 171, pp. 12–21, 2020, doi: 10.1016/j.procs.2020.04.003.

[8] K. Schultebraucks, V. Yadav, A. Y. Shalev, G. A. Bonanno, and I. R. Galatzer-Levy, “Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood,” *Psychological Medicine*, pp. 1–11, Aug. 2020, doi: 10.1017/S0033291720002718.

[9] W. Pan *et al.*, “Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders,” *Plos One*, vol. 14, no. 6, Art. no. e0218172, Jun. 2019, doi: 10.1371/journal.pone.0218172.

[10] H. Jiang *et al.*, “Detecting depression using an ensemble logistic regression model based on multiple speech features,” *Computational and Mathematical Methods in Medicine*, vol. 2018, pp. 1–9, Sep. 2018, doi: 10.1155/2018/6508319.

[11] E. I. I. Moore, M. Clements, J. Peifer, and L. Weisser, “Comparing objective feature statistics of speech for classifying clinical depression,” in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2004, vol. 3, pp. 17–20, doi: 10.1109/IEMBS.2004.1403079.

[12] J. F. Cohn *et al.*, “Detecting depression from facial actions and vocal prosody,” in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII*, pp. 1–7, 2009.

[13] L.-S. A. Low, N. C. Maddage, M. Lech, and N. Allen, “Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents,” in *2009 8th IEEE International Conference on Cognitive Informatics*, Jun. 2009, pp. 346–350, doi: 10.1109/COGINF.2009.5250714.

[14] S. Alghowinem, R. Göcke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, “From joyous to clinically depressed: mood detection using spontaneous speech,” in *Twenty-Fifth International FLAIRS Conference*, pp. 141–146, 2012.

[15] D. J. France and R. G. Shiavi, “Acoustical properties of speech as indicators of depression and suicidal risk,” *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[16] H. Azam *et al.*, “Classifications of clinical depression detection using acoustic measures in Malay speakers,” in *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, Dec. 2016, pp. 606–610, doi: 10.1109/IECBES.2016.7843521.

[17] Z. Liu *et al.*, “Detection of depression in speech,” in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2015, pp. 743–747, doi: 10.1109/ACII.2015.7344652.

[18] G. Kiss, M. G. Tulics, D. Sztahó, A. Esposito, and K. Vicsi, “Language independent detection possibilities of depression by speech,” in *Smart Innovation, Systems and Technologies*, pp. 103–114, 2016.

[19] D. Shin *et al.*, “Detection of minor and major depression through voice as a biomarker using machine learning,” *Journal of Clinical Medicine*, vol. 10, no. 14, p. 3046, Jul. 2021, doi: 10.3390/jcm10143046.

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

[21] Z. Liu, D. Wang, L. Zhang, and B. Hu, “A novel decision tree for depression recognition in speech,” *Electrical Engineering and Systems Science*, Feb. 2020.

[22] N. N. W. N. Hashim, M. A.-E. A. Ezzi, and M. D. Wilkes, “Mobile microphone robust acoustic feature identification using coefficient of variance,” *International Journal of Speech Technology*, Aug. 2021, doi: 10.1007/s10772-021-09877-1.

[23] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, “Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology,” *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, Jan. 2007, doi: 10.1016/j.jneuroling.2006.04.001.

[24] R. Shankayi, M. Vali, M. Salimi, and M. Malekshahi, “Identifying depressed from healthy cases using speech processing,” in *2012 19th Iranian Conference of Biomedical Engineering, ICBME 2012*, 2012, no. December, pp. 242–245, doi: 10.1109/ICBME.2012.6519689.

[25] S. Scherer, J. Pestian, and L. Morency, “Investigating the speech characteristics of suicidal adolescents,” in *ICASSP 2013-2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 709–713, 2013.

[26] M. Valstar *et al.*, “AVEC 2013,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, Oct. 2013, pp. 3–10, doi: 10.1145/2512530.2512533.

[27] M. Valstar *et al.*, “AVEC 2014,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge-AVEC ’14*, 2014, pp. 3–10, doi: 10.1145/2661806.2661807.

[28] J. Gratch *et al.*, “The distress analysis interview corpus of human and computer interviews,” *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 2014.

[29] C. Solomon, M. F. Valstar, R. K. Morriss, and J. Crowe, “Objective methods for reliable detection of concealed depression,” *Frontiers in ICT*, vol. 2, Art. no. 5, Apr. 2015, doi: 10.3389/fict.2015.00005.

[30] N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs, and D. J. France, “Evaluation of voice acoustics as predictors of clinical depression scores,” *Journal of Voice*, vol. 31, no. 2, pp. 256.e1-256.e6, Mar. 2017, doi: 10.1016/j.jvoice.2016.06.006.

[31] M. W. Hashim, Nik Nur Wahidah Nik and R. S. Jared Meggs, “Analysis of timing pattern of speech as possible indicator for NearTerm suicidal risk and depression in male patients,” in *2012 International Conference on Conference on Signal Processing Systems (ICSPS 2012)*, 2012, Art. no. 58, doi: 10.7763/IPCSIT.2012.V58.2.

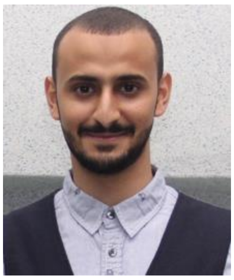
[32] M. Yamamoto *et al.*, “Using speech recognition technology to investigate the association between timing-related speech features and depression severity,” *Plos One*, vol. 15, no. 9, Art. no. e0238726, Sep. 2020, doi: 10.1371/journal.pone.0238726.

[33] P. hardy, R. Jouvent, and D. Widlöcher, “Speech pause time and the retardation rating scale for depression (ERD),” *Journal of Affective Disorders*, vol. 6, no. 1, pp. 123–127, Feb. 1984, doi: 10.1016/0165-0327(84)90014-4.

BIOGRAPHIES OF AUTHORS

Nik Nur Wahidah Nik Hashim (Ph.D.) received her BS, MS, and PhD in Electrical Engineering at Vanderbilt University, Nashville, Tennessee, USA beginning 2005 to 2014. Currently an assistant professor in the Department of Mechatronics, Kuliyyah of Engineering at the International Islamic University Malaysia (IIUM). Her passion in research is studying the field of mental health through speech analysis. Other specific field of research interest include speech processing, machine learning, deep learning, digital signal processing and natural language processing. Email: niknurwahidah@iium.edu.my.

Nadzirah Ahmad Basri (Ph.D.) is currently working in the Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia, Kuantan, Pahang. Nadzirah is a lecturer and a practicing Clinical Psychologist. She carries out psychological assessment and intervention to adults and children of varying mental disorders. She enjoys giving talks on mental health issues and is active working in non-governmental organisations advocating for mental health and academic issues. Email: nadzirahbasri@gmail.com.



Mugahed Al-Ezzi Ahmed Ezzi is an R&D software manager at C&C I Motive Sdn. Bhd. His work focuses specifically on intelligent systems in automotive industries, including machine vision, machine learning, and deep learning. In 2018, he received his bachelor's degree with first-class honors in mechatronics engineering at International Islamic University Malaysia (IIUM). He is currently pursuing his master's degree in mechatronics engineering at IIUM, emphasizing speech analysis and mental health. Email: ezzi.mugahed@gmail.com.

Nik Mohd Hazrul Nik Hashim (Ph.D.) is an Associate Professor of Marketing at the Graduate School of Business, National University of Malaysia. He has more than twenty years of experience as an academician, including a casual teaching stint in Australia. He holds a PhD in Marketing from the University of Melbourne and specializes in Product Innovation. Nik has published in leading international journals, including Journal of Retailing and Consumer Services, Review of Managerial Science, The Service Industries Journal, Journal of Relationship Marketing, International Journal of Emerging Markets, Journal of Financial Services Marketing, European Journal of International Management, and Renewable Energy, and has co-authored a textbook; Product Management and Strategy published by McGraw-Hill. He is keen on knowledge sharing and believes that real projects or practical applications should be embedded in the teaching process. Email: nikhaz@ukm.edu.my.