





Article

Speech Quality Feature Analysis for Classification of Depression and Dementia Patients

Brian Sumali ¹, Yasue Mitsukura ², Kuo-ching Liang ³, Michitaka Yoshimura ³, Momoko Kitazawa ³, Akihiro Takamiya ³, Takanori Fujita ⁴, Masaru Mimura ³ and Taishiro Kishimoto ^{3,*}

¹ Graduate School of Science and Technology, School of Integrated Design Engineering, Keio University, Yokohama 223-8522, Japan; brian.sumali@keio.jp

² Department of System Design Engineering, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan; mitsukura@keio.jp

³ Department of Psychiatry, School of Medicine, Keio University, Tokyo 160-8582, Japan; kcliang@keio.jp (K.-c.L.); y-michitaka@keio.jp (M.Y.); m.kitazawa@keio.jp (M.K.); akihiro.takamiya2017@keio.jp (A.T.); mimura@a7.keio.jp (M.M.)

⁴ Department of Health Policy and Management, School of Medicine, Keio University, Tokyo 160-8582, Japan; tafujita@keio.jp

* Correspondence: tkishimoto@keio.jp

Received: 1 June 2020; Accepted: 23 June 2020; Published: 26 June 2020



Abstract: Loss of cognitive ability is commonly associated with dementia, a broad category of progressive brain diseases. However, major depressive disorder may also cause temporary deterioration of one's cognition known as pseudodementia. Differentiating a true dementia and pseudodementia is still difficult even for an experienced clinician and extensive and careful examinations must be performed. Although mental disorders such as depression and dementia have been studied, there is still no solution for shorter and undemanding pseudodementia screening. This study inspects the distribution and statistical characteristics from both dementia patient and depression patient, and compared them. It is found that some acoustic features were shared in both dementia and depression, albeit their correlation was reversed. Statistical significance was also found when comparing the features. Additionally, the possibility of utilizing machine learning for automatic pseudodementia screening was explored. The machine learning part includes feature selection using LASSO algorithm and support vector machine (SVM) with linear kernel as the predictive model with age-matched symptomatic depression patient and dementia patient as the database. High accuracy, sensitivity, and specificity was obtained in both training session and testing session. The resulting model was also tested against other datasets that were not included and still performs considerably well. These results imply that dementia and depression might be both detected and differentiated based on acoustic features alone. Automated screening is also possible based on the high accuracy of machine learning results.

Keywords: pseudodementia; automated mental health screening; audio features; statistical testing; machine learning

1. Introduction

Dementia is a collective symptoms attributed to loss of recent and remote memory along with difficulty in absorbing new knowledge and trouble in decision making. The most common cause of dementia is Alzheimer's disease which contributes to 60–70% of all dementia cases worldwide. Presently there is no treatment available [1] and recent researches focuses on early detection of dementia signs [2–9] and reducing the risk factors to slow the cognitive decline [10–13].

Preliminary diagnosis of dementia typically performed in a mental hospital by a licensed psychiatrist interviewing and performing tests to the patients [14–16]. Occasionally, diagnosing dementia becomes a complex process, as elderly patients with major depressive disorder often has overlapping symptoms with dementia. To determine whether a patient is truly suffering from dementia, a rigorous test must be performed [17]. A temporary decrease in mental cognition caused by mental disorders is defined as pseudodementia [17–21]. The key difference of pseudodementia is the reversibility of cognitive impairment, in contrast with the progressive nature of dementia. In some cases pseudodementia also serves as biomarker of dementia [21]. Unfortunately, most engineering researches concern only with depression severity or dementia severity [22,23] and almost none focused on pseudodementia.

Features commonly employed for automated mental health screening include facial features (gaze, blink, emotion detection, etc.) [24–26], biosignals (electroencephalogram, heart rate, respiration, etc.) [27–30], and auditory features (intensity, tone, speed of speech, etc.) [23,31]. Although biosignals are the most reliable data source, most of biosignal measurement devices are arduous to equip, limiting their value. In the other hand, facial and acoustic features may be obtained with minimal burden to the patient. As audio feature analysis is comparatively straightforward when compared to facial image analysis, we utilized audio features in this study instead of image features.

The aim of this study was to use an array microphone to record conversations between psychiatrists and depression patients and dementia patients in a clinical setting, and to investigate the differences in acoustic features between the two patient groups and not against healthy volunteers, differing from other conventional studies. Additionally, we are using dataset labelled from licensed psychiatrist to reduce the subjectivity. We revealed the features contributing for pseudodementia screening. In addition, we examined the possibility of utilizing machine learning for automatic pseudodementia screening.

2. Materials and Methods

2.1. Data Acquisition

This study is conducted as a part of Project for Objective Measures using Computational Psychiatry Technology (PROMPT), a research aimed to develop objective, noninvasive, and easy-to-use biomarkers for assessing the severity of depressive and neurocognitive disorders, including dementia. The details of the project may be found in [32].

The PROMPT study was approved by Keio University Hospital Ethics Committee (20160156, 20150427). All participants provided written informed consent. The experiment was conducted on Keio University Hospital and Joint Medical Research Institute. During the interview, the patient and the psychiatrist were seated across a table, as shown in Figure 1.

A single session consists of “free talk” segment followed by “rating” segment. In “free talk”, the psychiatrist conducts a typical clinical interview concerning the patient’s daily life and mood. The length of a “free talk” segment is around 10 min. In the “rating” segment, the patient is interviewed based on a clinical assessment tools related to their mental health history, which may include some tasks such as clock-drawing test and memory test or some personal questions such as their sleep habit and depressive mood in the recent weeks. The duration of “rating” segment typically lasts more than 20 min.

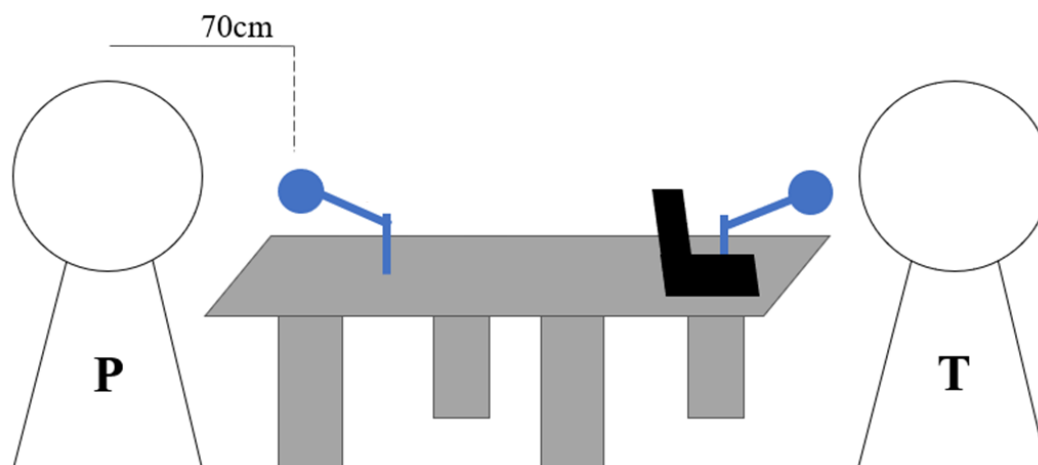


Figure 1. Recording setup during interview session. P is the patient and T is the psychiatrist. There is a distance of approximately 70 cm between the patient’s seat and the recording apparatus.

2.2. Participants

For statistical analysis, the first, and the second parts of machine learning, several datasets were removed from the PROMPT database in consideration of age features and the presence of symptoms. Only datasets which satisfy the following criteria were included:

1. Age between 57 and 84 years-old; 57 is the lowest age for dementia patients and 84 is the highest age for depression patients. The purpose of this criterion was to remove the effect of age which is positively correlated with Dementia.
2. For dementia patients: mini-mental state examination (MMSE) score of 23 or less accompanied with 15-item geriatric depression scale (GDS) score of 5 or less; The purpose of this criterion was to select only patients with dementia symptoms and exclude patients with both symptoms. A person is defined as symptomatic dementia if the MMSE score is 23 or less
3. For depression patients: 17-item Hamilton depression rating scale (HAMD17) of 8 or more. A person is defined to be depressed if one’s score of HAMD17 is 8 or more.
4. The recording session was from “free talk” and the length was at least 10 min long. The purpose of this criterion was to ensure enough information contained within the recordings.

For the third part of machine learning, different criteria were applied to PROMPT database to construct a test set consisting of young depressed and old dementia datasets. Specifically, the criteria were:

1. For dementia patients: mini-mental state examination (MMSE) score of 23 or less accompanied with 15-item geriatric depression scale (GDS) score of 5 or less; The age of the patients should be of 85 years or more.
2. For depression patients: 17-item Hamilton depression rating scale (HAMD17) of 8 or more. The age of the patients should be no more than 56 years.
3. The recording session was from “free talk” and the duration was at least 10 min long.

Each dataset corresponds to a interview session from one subject. In this study, the datasets were considered as independent because (1) time gap between the sessions were long, 2 weeks in the minimum; and (2) the clinical score results may increase or decrease compared to the first visit, especially depression patients. Figure 2 illustrates the dataset filtering for the statistical analysis and machine learning phases.

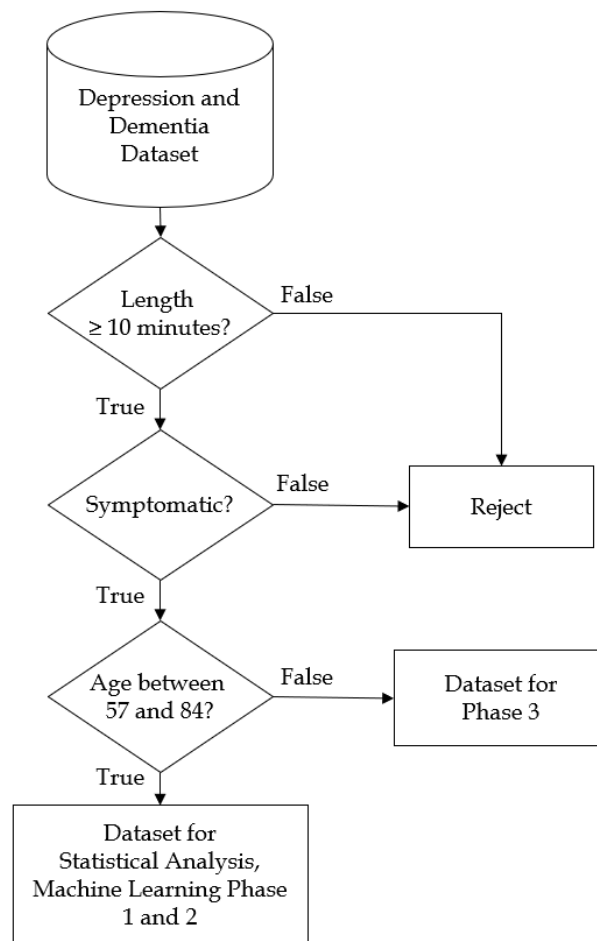


Figure 2. Flowchart of dataset filtration.

2.3. Materials

A vertical array microphone: Classis RM30W (Beyerdynamic GmbH & Co. KG, Heilbronn, Germany) with an internal noise cancellation filter to remove wind and pop noise was utilized to record conversations between patients and psychiatrists. The sampling rate was set to 16 kHz. Feature extraction and analysis was performed utilizing typical processor: Dell G7 7588 with Intel Core i7-8750H@2.20 GHz, 16 GB RAM, manufactured in China. with Windows 10 operating system. All methods were available built-in from software MATLAB 2019b.

Clinical assessment tools utilized were 17-item Hamilton depression rating scale (HAMD) [33], 15-item geriatric depression scale (GDS) [16], Young mania rating scale (YMRS) [34] for depression patients; and mini-mental state exam (MMSE) [14] and clinical dementia rating (CDR) [16] for dementia patients. In this study, HAMD for depression and MMSE for dementia is used as golden standard.

2.4. Audio Signal Analysis

2.4.1. Preprocessing

In some rare cases, the recordings contained some outliers, possibly caused by random errors, and preprocessing of the raw data needs to be conducted. We defined the outliers by using inter-quartile range (IQR). A point in the audio recording is defined to be an outlier if it satisfies one of the following conditions:

1. $X < Q1 - 1.5IQR$
2. $X > Q3 + 1.5IQR$

Here, X is the signal, $Q1$ is the lower (1st) quartile, $Q3$ is the upper (3rd) quartile, and IQR is the inter-quartile range, computed by subtracting $Q1$ from $Q3$. We then apply cubic smoothing spline fitting to the audio signal, without the outliers. The objective of this method is twofold: (1) to interpolate the removed outliers, (2) subtle noise removal.

Additionally, intensity normalization was also performed. This was to ensure that the data is in equal scale to each other and to reduce clipping in audio signals. The normalization was conducted by rescaling the signal such that the maximum absolute value of its amplitude is 0.99. Continuous silence in form of trailing zeroes at front and end of the recordings were also deleted.

2.4.2. Feature Extraction

A subtotal of ten acoustic features were extracted from raw data. They were: Pitch, harmonics-to-noise ratio (HNR), zero-crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCC), Gammatone cepstral coefficients (GTCC), mean frequency, median frequency, signal energy, spectral centroid, and spectral rolloff point, with details in Table 1. These features were chosen as they represent both temporal and spectral features of a signal. Additionally, some of these features relate to closely to speech which is a common biomarker for both depression and dementia [35–37]. These features were computed once in every 10 ms by applying a 10 ms window with no overlap. We then performed feature extraction to the windowed signals. The total count of audio feature is 36, with 14 MFCCs and GTCCs. As we used data with length of at least 10 min, a minimum of 60,000 datapoints were obtained, for each feature. We then computed the mean, median, and standard deviation (SD) of the datapoints and used them for statistical analysis and machine learning, resulting in total feature count of 108.

Table 1. List of features utilized in this study.

Feature	Mathematical Functions and References
Pitch	[38]
Harmonics-to-noise ratio (HNR)	[39]
Zero-Crossing Rate (ZCR)	$ZCR(X) = \frac{1}{2N} \sum_i^N sgn(X_i) - sgn(X_{i-1}) $
Mel-frequency cepstral coefficients (MFCC)	[40]
Gammatone cepstral coefficients (GTCC)	[41]
Mean frequency	Mean of power spectrum from the signal
Median frequency	Median of power spectrum from the signal
Signal energy (E)	$E(X) = \frac{\sigma(X)}{\mu(X)}$
Spectral centroid (c)	$c = \frac{\sum_{i=b_1}^{b_2} f_i s_i}{\sum_{i=b_1}^{b_2} s_i}$ [42]
Spectral rolloff point (r)	$\sum_{i=b_1}^r s_i = \frac{k}{100} \sum_{i=b_1}^{b_2} s_i$ [42]

For ZCR: N , sgn , and X_i denotes the length of signal, signum function extracting the sign of a real number (positive, negative, or zero), and i -th sequence of signal X , respectively. For mean frequency and median frequency: power spectrum from the signal was applied by performing Fourier transform. For signal energy: $E(X)$ is the signal energy of signal X , $\sigma(X)$ denotes the function of standard deviation of signal X and $\mu(X)$ indicates the function of mean of signal X . For spectral centroid: c denotes the spectral centroid, f_i is the frequency in Hertz corresponding to bin i , s_i is the spectral value at bin i , and b_1 and b_2 are the band edges, in bins, over which to calculate the spectral centroid. For spectral rolloff point: r is the spectral rolloff frequency, s_i is the spectral value at bin i , and b_1 and b_2 are the band edges, in bins, over which to calculate the spectral spread.

2.4.3. Statistical Analysis

To investigate the relationship between audio features and clinical symptoms, linear correlations of the acoustic features against the corresponding clinical rating tools were computed. The clinical rating tools were HAMD for depression subjects and MMSE for dementia subjects. In addition, two-tailed t -test were also performed to check statistical significance. The values were adjusted using

Bonferroni correction. Additionally, correlation between age and sex with clinical rating tools were also evaluated for validation purposes.

2.4.4. Machine Learning

Machine learning was performed in three stages: (1) to examine the possibility of automatic pseudodementia diagnosis with unsupervised learning, (2) to examine the possibility of automatic pseudodementia diagnosis with supervised classifier, and (3) to validate its robustness against non age-matched datasets. The unsupervised learning algorithm utilized for the first stage was k-means clustering. The parameters for k-means clustering were $k = 2$ with squared Euclidean distance metric. For stages 2 and 3, the machine learning model utilized was a binary classifier: support vector machine (SVM) with linear kernel, 3rd order polynomial kernel, and radial-basis function (RBF) kernel [43]. The hyperparameters for both linear kernel and polynomial kernel is the cost parameter C while RBF kernel has two hyperparameters: C and gamma. The optimization of hyperparameters was performed using grid search algorithm with values ranging from $\frac{1}{1000}$ to 1000. Linear kernel was chosen as it allows the visualization of feature contributions, as opposed to SVM with nonlinear kernels. For the second phase, the machine learning session was performed using nested 10-fold cross-validation. It is defined as follows:

1. Split the datasets into ten smaller groups, maintaining the ratio of the classes
2. Perform ten-fold cross validation using these datasets.

For each fold:

- (a) Split the training group into ten smaller subgroups.
- (b) Perform another ten-fold cross-validation using these subgroups.

For each inner fold:

- i. Perform LASSO regression [44] and obtain the coefficients.

The LASSO regression solves

$$\min_{\alpha, \beta} \left(\frac{1}{2N} \sum_{i=1}^N (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j| \right)$$

where α is a scalar and β is a vector of coefficients, N is the number of observations, y_i is the response at observation i , x_{ij} is the vector of predictors at observation i , and λ is a nonnegative regularization parameter. High value of λ results in stricter feature selection and in this study, it is computed automatically such that it is the largest possible value for nonnull model. The performance of the model is not considered.

- ii. Mark the features with coefficient of less than 0.01.
- (c) Perform feature selection by removing features with 10 marks obtained from step 2-b-ii.
- (d) Train an SVM model based on features from (c).
3. Compute the average performance and standard deviation of the models.

In the third phase, a SVM model was trained using age-matched subjects and selected features from the second phase. Resulting model's performance is evaluated against the filtered-out subjects: young depression and old dementia subjects. In both cases, the dementia patients were labelled as class 0 (negative) and depression patients were labelled as class 1 (positive). The illustration of the phases are shown in Figure 3.

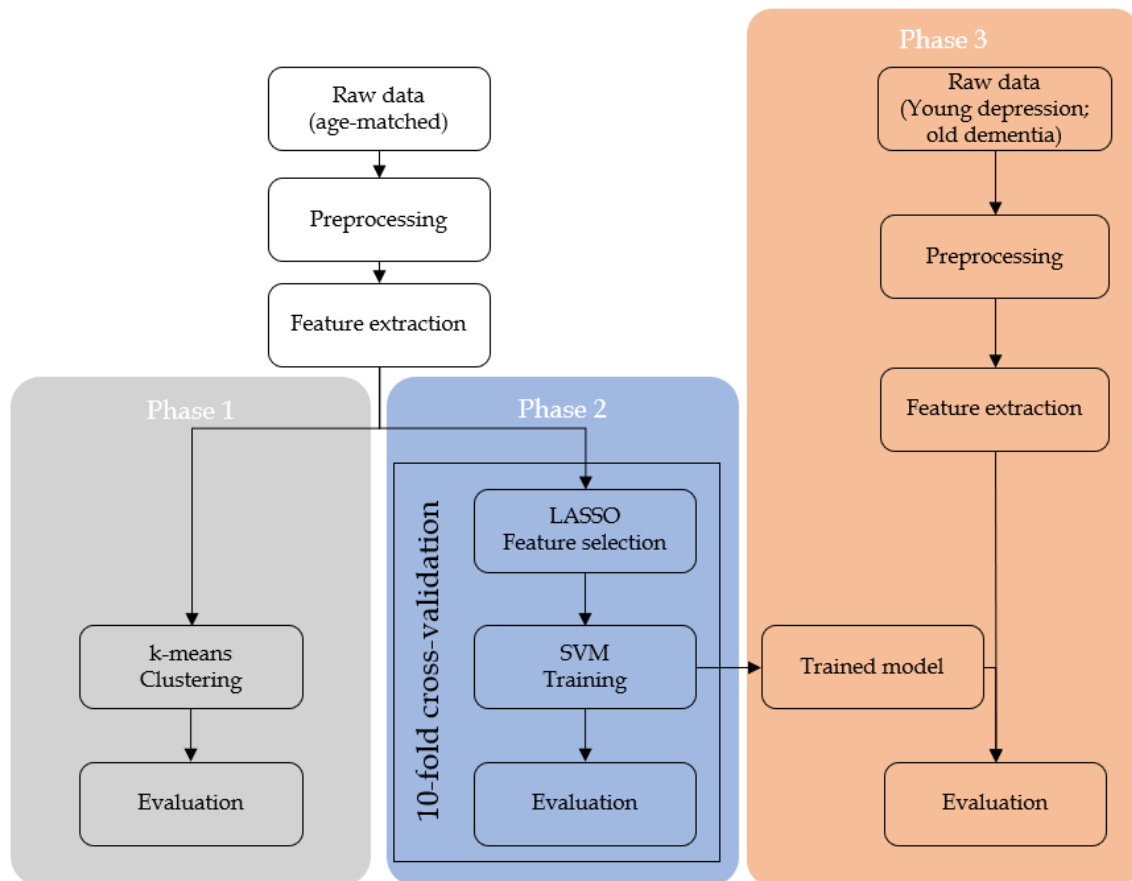


Figure 3. Flowchart of supervised machine learning procedure. The first and second phase used age-matched symptomatic depression and dementia subjects. The first phase consists of unsupervised machine learning clustering while the second phase consists of conventional training and evaluation. The third phase involves of utilizing machine learning model trained from age-matched subjects against non-age matched subjects.

2.4.5. Evaluation Metrics

We utilized eight metrics to evaluate the effectiveness of the machine learning model, all of which are computed based on the ratio of true positive (TP), false positive (FP), true negative (TN), and false negatives (FN). In this study, the class depression was labelled as “positive” and dementia was labelled as “negative”. All of the TP, FP, TN, and FN values were obtained from confusion matrix, as shown on Figure 4. Based on the confusion matrices, the evaluation metrics of observed accuracy, true positive rate (TPR/sensitivity), true negative rate (TNR/specificity), positive predictive value (PPV/precision), negative predictive value (NPV), F1-score, Cohen’s kappa, and Matthew’s correlation coefficient (MCC) can be then computed. The formulas for computing these metrics are described in Table 2. These metrics were conventional evaluation metrics utilized in performance evaluation. Metrics related to inter-rater reliability such as Cohen’s kappa and MCC were included to ensure validity of measurement in cases of imbalanced sample problem.

		Truth	
		Positive (Depression)	Negative (Dementia)
Predicted	Positive (Depression)	True Positive (TP)	False Positive (FP)
	Negative (Dementia)	False Negative (FN)	True Negative (TN)

Figure 4. Confusion matrix and class label utilized in this study.

Table 2. List of evaluation metrics.

Metric	Mathematical Formula
Accuracy (ACC)	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$
True positive rate (TPR)	$TPR = \frac{TP}{TP+FN}$
True negative rate (TNR)	$TNR = \frac{TN}{TN+FP}$
Positive predictive value (PPV)	$PPV = \frac{TP}{TP+FP}$
Negative predictive value (NPV)	$NPV = \frac{TN}{TN+FN}$
F1 score	$F1 = 2 \frac{PPV * TPR}{PPV + TPR}$
Cohen's kappa	$EXP = \frac{(TP+FP)(TP+FN)+(TN+FN)(TN+FP)}{TP+TN+FP+FN^2}$ $Kappa = \frac{ACC-EXP}{1-EXP}$
Matthew's correlation coefficient (MCC)	$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

3. Results

3.1. Demographics

A total of 120 participants (depression $n = 77$, dementia $n = 43$) participated in the study, and 419 datasets (300 of depression and 119 of dementia) were obtained. After age-matching, only 177 datasets (89 of depression and 88 of dementia) from 53 participants (depression $n = 24$, dementia $n = 29$) were qualified for the first and second phase of machine learning. The test dataset for second phase of machine learning consisted of young depression patients and old dementia patients and was used in the third phase of machine learning. There were 242 datasets (211 of depression and 31 of dementia) from 67 patients (depression $n = 53$, dementia $n = 14$). Details of subject demographics were described in Table 3.

Table 3. Subject demographics.

	Demographics	Depression	Dementia
Symptomatic	n (dataset/subject)	300/77	119/43
	age (mean \pm s.d. years)	50.4 \pm 15.1	80.8 \pm 8.3
	sex (female %)	54.5	72.1
Age-matched	n (dataset/subject)	89/24	88/29
	age (mean \pm s.d. years)	67.8 \pm 7.1	77.0 \pm 7.5
	sex (female %)	83.3	72.4
Young depression, Old dementia	n (dataset/subject)	211/53	31/14
	age (mean \pm s.d. years)	42.5 \pm 10.4	88.5 \pm 1.9
	sex (female %)	41.5	71.4

3.2. Statistical Analysis

In this section, the statistical analysis for the extracted features were reported. Pearson's correlation found significant correlations with clinical interview tools in features of GTCCs 1, 3, 12 and MFCCs 1, 3, 4, 7, 12. The average absolute correlation coefficient R was 0.264 and its SD was 0.049. The highest absolute correlation value with statistical significance ($p < 0.05$) was $|R| = 0.346$ for depression and $|R| = 0.400$ for dementia. Features with significant correlation related to depression tend to yield weak to moderate negative Pearson correlation values (average absolute $R \pm SD = 0.289 \pm 0.05$) while features with significant correlation related to dementia tend to yield weak to moderate positive Pearson correlation values (average absolute $R \pm SD = 0.281 \pm 0.06$). The features' distributions were depicted in Figure 5 and their corresponding Pearson correlation values were shown in Table 4. Corrected two-tailed t -test shows significant differences of features in HNR, ZCR, GTCC coefficients 4–14, mean frequencies, median frequencies, MFCC coefficients 4–13, spectral centroid, and spectral rolloff points. No significant difference was found in Pitch and Energy.

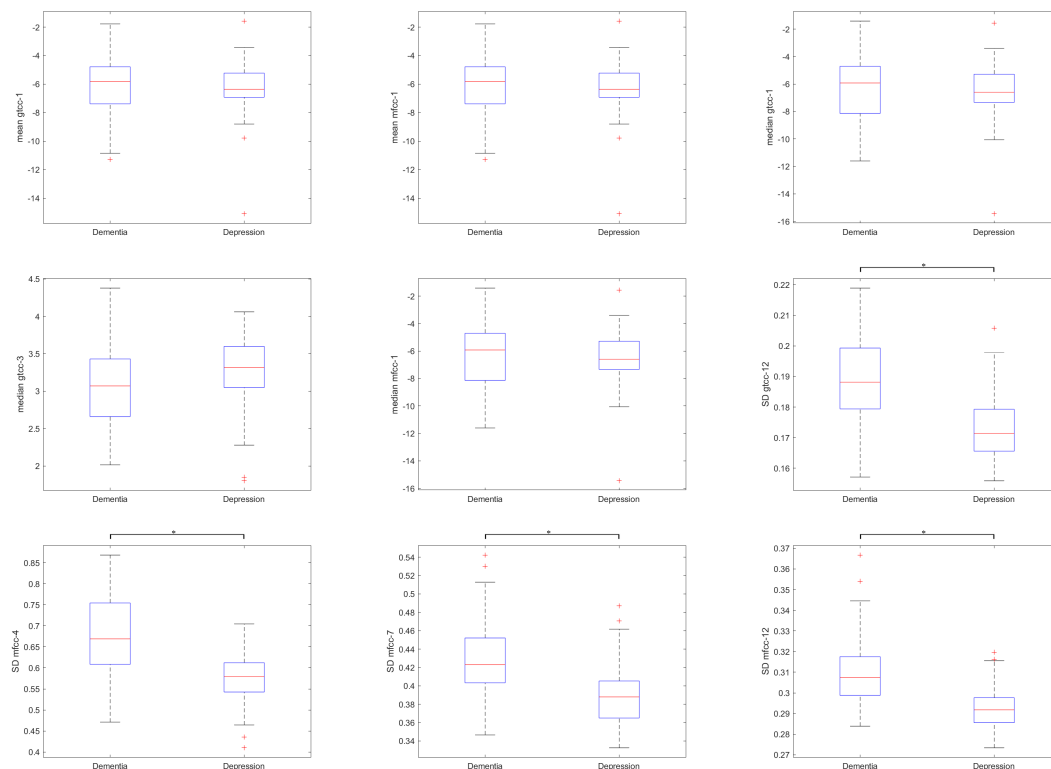


Figure 5. Distribution of features with significant correlation to HAMD and MMSE. * marks the statistically different features between the groups, corrected with Bonferroni correction.

There was no significant correlation was found between sex and clinical assessment tools (depression $R = 0.021$, $p = 0.853$; dementia $R = 0.142$, $p = 0.928$). Age has no significant correlation with depression's clinical assessment tools ($R = 0.097$, $p = 0.403$) but significant, moderate correlation between age and dementia's clinical assessment tools was found ($R = 0.424$, $p = 0.0046$).

Table 4. Features with significant Pearson correlation in both depression and dementia patients.

Feature Description	Pearson's Correlation	
	HAMD Depression	MMSE Dementia
mean GTCC_1	−0.346	0.226
mean MFCC_1	−0.346	0.226
median GTCC_1	−0.325	0.219
median GTCC_3	−0.224	0.230
median MFCC_1	−0.325	0.219
SD GTCC_12	−0.218	0.257
SD MFCC_4	−0.289	0.329
SD MFCC_7	−0.221	0.274
SD MFCC_12	−0.259	0.224

3.3. Machine Learning

In this section, the results of machine learning were presented. The evaluation results from unsupervised learning with kMeans algorithm was shown on Table 5. For the SVM with linear kernels, 26 features were completely rejected in the feature selection, resulting in their removal during creation of the model for second phase. The rejected features were related to pitch, GTCCs 1–3, MFCCs 1–3, signal energy, spectral centroid, and spectral cutoff point. Feature selection in SVM with 3rd order polynomial kernel results in removal of 28 features. The rejected features were related to pitch, GTCCs and MFCCs (1–3, 12–13), signal energy, spectral centroid, and spectral cutoff. LASSO with RBF-SVM similarly rejects 28 features related to low-order (1–4) and high-order (10–13) MFCC and GTCC coefficients, pitch, signal energy, spectral centroid, and spectral cutoff.

Table 5. Phase 1: Unsupervised machine learning result.

Metric	kMeans (%)
Accuracy (ACC)	62.7
True positive rate (TPR)	89.9
True negative rate (TNR)	35.2
Positive predictive value (PPV)	58.3
Negative predictive value (NPV)	77.5
F1 score	70.8
Cohen's kappa	25.2
Matthew's correlation coefficient (MCC)	30.0

Results of feature contributions of the trained linear SVM model was presented in Figure 6 alongside with the list of remaining 82 features. The feature contributions were absolute value of linear SVM coefficients. Machine learning evaluation results for phase 2 were shown on Tables 6–8 and the results for phase were shown on Table 9. Results with and without LASSO algorithm also shown in these tables to confirm effectiveness of feature selection. Here, the label “positive” represents depression patients and “negative” is for dementia patients.

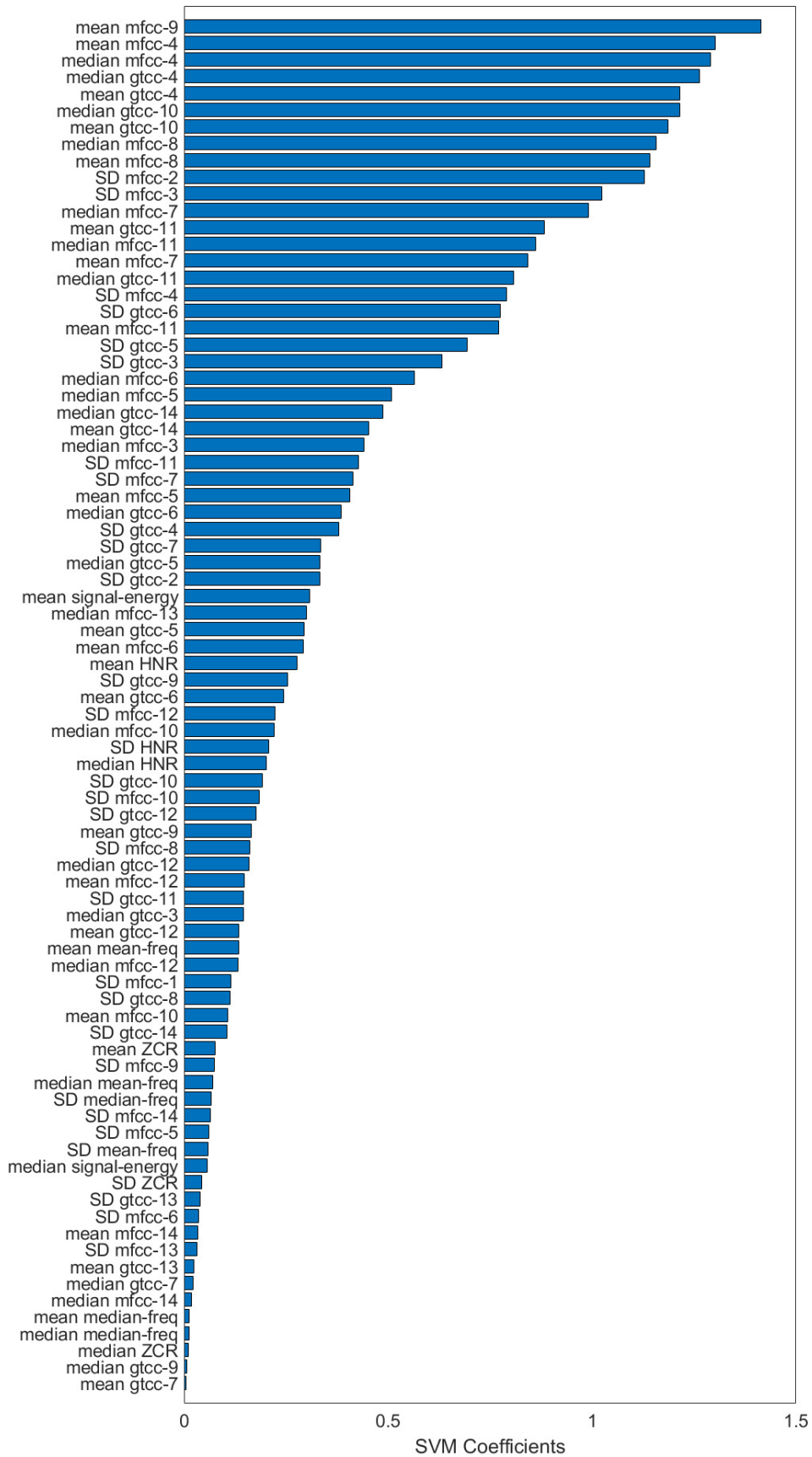


Figure 6. Absolute value of feature contributions of linear SVM with LASSO feature selection, sorted descending.

Table 6. Phase 2: Supervised machine learning result—SVM with linear kernel.

Metrics	Training (Mean \pm SD %)		Testing (Mean \pm SD %)	
	No LASSO	With LASSO	No LASSO	With LASSO
Accuracy (ACC)	90.1 \pm 2.4	95.2 \pm 0.7	84.2 \pm 5.3	93.3 \pm 7.7
True positive rate (TPR)	94.4 \pm 0.9	98.3 \pm 0.9	88.8 \pm 10.5	97.8 \pm 4.7
True negative rate (TNR)	85.7 \pm 4.6	92.6 \pm 1.2	79.6 \pm 11.5	89.4 \pm 13.7
Positive predictive value (PPV)	87.1 \pm 3.5	92.1 \pm 1.2	82.5 \pm 8.8	90.4 \pm 11.7
Negative predictive value (NPV)	93.8 \pm 1.0	98.4 \pm 0.8	88.8 \pm 8.9	98.0 \pm 4.2
F1 score	90.6 \pm 2.0	95.1 \pm 0.7	84.8 \pm 5.5	93.5 \pm 7.2
Cohen's kappa	80.2 \pm 4.7	90.5 \pm 1.4	68.3 \pm 10.5	86.7 \pm 15.0
Matthew's correlation coefficient (MCC)	80.5 \pm 4.4	90.6 \pm 1.4	69.8 \pm 10.3	87.8 \pm 13.5

Table 7. Phase 2: Supervised machine learning result—SVM with 3rd order Polynomial kernel.

Metrics	Training (Mean \pm SD %)		Testing (Mean \pm SD %)	
	No LASSO	With LASSO	No LASSO	With LASSO
Accuracy (ACC)	91.5 \pm 3.1	94.6 \pm 8.1	79.1 \pm 7.6	89.7 \pm 11.4
True positive rate (TPR)	96.4 \pm 2.4	99.1 \pm 1.0	85.3 \pm 10.8	96.7 \pm 5.4
True negative rate (TNR)	86.5 \pm 4.0	90.0 \pm 16.1	72.6 \pm 14.3	83.1 \pm 22.9
Positive predictive value (PPV)	87.9 \pm 3.5	92.3 \pm 9.9	76.9 \pm 8.3	87.6 \pm 13.8
Negative predictive value (NPV)	95.9 \pm 2.7	98.9 \pm 1.2	84.1 \pm 9.9	96.9 \pm 5.0
F1 score	91.9 \pm 2.9	95.3 \pm 6.1	80.3 \pm 6.9	91.1 \pm 8.2
Cohen's kappa	82.9 \pm 6.3	89.2 \pm 16.2	58.0 \pm 15.2	79.7 \pm 21.7
Matthew's correlation coefficient (MCC)	83.3 \pm 6.2	90.1 \pm 13.7	59.4 \pm 14.6	81.8 \pm 17.9

Table 8. Phase 2: Supervised machine learning result—SVM with RBF kernel.

Metrics	Training (Mean \pm SD %)		Testing (Mean \pm SD %)	
	No LASSO	With LASSO	No LASSO	With LASSO
Accuracy (ACC)	90.4 \pm 6.2	95.6 \pm 1.9	75.3 \pm 12.4	88.7 \pm 7.9
True positive rate (TPR)	96.4 \pm 2.9	98.8 \pm 1.0	77.5 \pm 16.6	91.0 \pm 10.3
True negative rate (TNR)	84.3 \pm 10.2	92.4 \pm 3.0	72.9 \pm 17.3	86.1 \pm 13.1
Positive predictive value (PPV)	86.7 \pm 7.9	93.0 \pm 2.6	75.6 \pm 13.8	88.3 \pm 10.4
Negative predictive value (NPV)	95.7 \pm 3.7	98.6 \pm 1.2	77.6 \pm 14.6	91.3 \pm 8.9
F1 score	91.2 \pm 5.4	95.8 \pm 1.7	75.7 \pm 12.5	89.1 \pm 7.9
Cohen's kappa	80.8 \pm 12.3	91.2 \pm 3.7	50.5 \pm 24.8	77.3 \pm 15.9
Matthew's correlation coefficient (MCC)	81.5 \pm 11.7	91.4 \pm 3.6	51.8 \pm 25.0	78.3 \pm 15.4

Table 9. Phase 3: Machine learning result against non-age matched dataset.

Metrics	Linear		Polynomial		RBF	
	All Feats	LASSO	All Feats	LASSO	All Feats	LASSO
Accuracy (ACC)	83.5	82.6	80.2	81.4	65.7	81.0
True positive rate (TPR)	87.7	83.9	82.5	82.9	66.8	82.9
True negative rate (TNR)	54.8	74.2	64.5	71.0	58.1	67.7
Positive predictive value (PPV)	93.0	95.7	94.1	95.1	91.6	94.6
Negative predictive value (NPV)	39.5	40.4	35.1	37.9	20.5	36.8
F1 score	90.2	89.4	87.9	88.6	77.3	88.4
Cohen's kappa	36.5	42.8	34.6	39.3	13.9	37.3
Matthew's correlation coefficient (MCC)	37.2	45.7	37.0	42.2	17.3	39.9

4. Discussion

In the present study, we obtained the audio recordings from clinical interviews of depression and dementia patients. Then, the recordings were filtered according to the analysis criteria. Preprocessing and acoustic feature extraction was then performed to the qualifying datasets. Statistical analysis and machine learning were performed to the acoustic features.

This study has potential limitations. First, although subtle, the recordings were contaminated with the doctor's voice. This naturally reduces the quality of the acoustic features. Next, there is no removal of silence between the dialogues. We hypothesized that long silences correspond to low motivation and therefore useful for predicting depression. Third, we did not consider real-time appliances. We utilized the full length of the recordings for predicting dementia versus depression. Finally, all the experiments were conducted in Japanese hospital, with Japanese doctors, and with Japanese patient. The speech features we extracted might be specific to the Japanese. Needless to say, these limitations imply potential bias in our study and the results of our study must be interpreted with attention to the limitations.

As a result, we found that GTCC coefficients 1, 3, and 12 along with MFCC coefficients 1, 3, 4, 7, 12 showed significant correlation with both clinical assessment tools: HAMD and MMSE, as shown in Table 4. Interestingly, the sign of Pearson's correlation coefficient were different; negative correlation was observed for HAMD and positive correlation was observed for MMSE. This suggests that although the features were important for both depression and dementia, they correlated differently. Another thing to note that the highest absolute correlation value with significance ($p < 0.05$) was 0.346 for HAMD and 0.400 for MMSE, suggesting a weak to moderate correlation between the audio features and clinical rating scores.

The corrected *t*-test between these features in Figure 5 showed statistical differences only in certain features. Interestingly, the standard deviation of a rather high-order MFCC coefficient showed significant difference. Normally, most of the information are represented in the lower order coefficients and their distributions were important for speech analysis. Feature contribution shown in Figure 6 puts these features in the middle of the selected features, and some of the lower-order MFCC features were even removed. This might imply the shared features between dementia and depression did not contribute well for predicting them.

Statistical comparison of acoustic features between two groups found significant differences in both temporal and spectral acoustic features. No significant difference between the two groups can be found in pitch and energy, both in the family of temporal features.

Although the result from unsupervised clustering algorithm was not satisfactory, both the accuracy and inter-rater agreement show that the performance was better than chance, denoting the underlying patterns in the data. In the second part of machine learning, feature selection was performed using LASSO algorithm. Here, both pitch and signal energy features were rejected alongside with other spectral features. Considering that both pitch and signal energy also showed no statistical significance in the *t*-test, it can be inferred that these features do not contribute for classification of depression and dementia. In contrast, GTCCs 4–14 and MFCCs 4–14 had statistically significant difference and were also selected by LASSO algorithm. GTCCs and MFCCs are similar features, related to tones of human speech. Although GFCCs was not developed for speech analysis, both are commonly used for speech recognition systems [45,46]. This finding is consistent with the fact that a person's speech characteristics might be related with their mental health. SVM feature contributions also confirmed that the top contributing features were MFCCs and GTCCs. As the coefficients of the MFCCs and GTCCs are related to the filterbanks utilized when computing them, these coefficients have the benefits of being interpretable [47].

Surprisingly, the best result of the SVM was obtained in SVM with linear kernel, although the the scores were only slightly superior to the nonlinear SVMs. Additionally, the effectiveness of LASSO algorithm for feature selection was evaluated and interesting result was found. For the second phase, all the SVM models benefited from having LASSO feature selection, but for the third phase, nonlinear SVMs seemed to be the most benefited with the feature selection. This might be related by the LASSO algorithm. As LASSO regression is a linear regression with penalty and the feature selection step was basically to discard features that give zero contribution to LASSO regression, linear SVM might be similar to it and was redundant in this case.

Nevertheless, high accuracy and interrater agreement were obtained from the models in both machine learning phases. For comparison, studies [24,25,28,29], and [23] have 87.2%, 81%, 81.23%, 89.71% and 73% as accuracy for predicting depression, respectively. [31] reports 73.6% accuracy for predicting dementia and [30] reports 99.9% TNR and 78.8% TPR. However, most of these studies compared healthy subjects against symptomatic patients, while our study compared patients afflicted with different mental problem. Additionally, most conventional studies measure depression by questionnaire and not with clinical examination, so this cannot be said to be a fair comparison. Low NPV scores and inter-rater during the third phase maybe due to the fact that evaluation in third phase was utilized with heavily imbalanced dataset and with higher number of samples compared to the training phase. These results suggest the possibility of using audio features for automatic pseudodementia screening.

5. Conclusions

We recorded the audio of clinical interview session of depression patients and dementia patients in a clinical setting using an array microphone. Statistical analysis shows significant differences in audio features between depressed patients and dementia patients. A machine learning model was constructed and evaluated; considerable performance was recorded for distinguishing depression patients and dementia patients. Feature contribution analysis reveal features MFCC and GTCC features to be the highest contributing features. The top contributing features were 9th and 4th MFCC features. Based on our findings, we conclude that automated pseudodementia screening with machine learning is feasible.

6. Future Work

Although this study has yielded considerably good results, there are still some rooms for improvements. For example, to eliminate the psychiatrist's voice inside the recordings. Although the microphone was situated against the patient, subtle amount of the psychiatrist's voice also included in the recordings. As such, a specific voice separation algorithm needs to be developed and applied to remove psychiatrist's voice. This will certainly add silent parts in the recordings and the feature extraction methodology needs to be modified; instead of processing audio with 10 ms window, activity-based window might be considered. Additionally, a dynamic cardioid microphone or multichannel array microphone might be beneficial for picking sounds only from the patient's side. In this case, room settings for suppressing reverberation and microphone placement becomes very important.

In conjunction with psychiatrist voice removal, activity-based features might also reveal relevance in aspects we did not consider in this study. Here, we hypothesized that longer silence between answers corresponds with lower patient cognition. We assumed that these silence segments will affect the mean value of the features while minimally affecting the median value and is beneficial for differentiating dementia against depression. However, activity-based or content-based analysis might reveal the difference in features we considered irrelevant in this study, such as signal energy.

Also, this study does not consider patients with overlapping symptoms of depression and dementia. Thus, the next step of this study is to develop a multi-class classifier capable of predicting patients with overlapping symptoms. A regression model trained with clinical assessment tools for both depression and dementia is also a possibility.

In consideration of improving the accuracy, more advanced machine learning techniques such as neural network might be suitable. Although the number of available dataset is relatively small for neural networks, sub-sampling and bootstrapping techniques might help to increase the numbers of dataset. Attention must be paid during the validation such that no data leak may occur. Additionally, feature extraction methods such as the combination of numerous hybrid acoustic features, as listed in [48] might also be beneficial. Nevertheless, the curse of dimensionality should be avoided when handling such numerous predictors.

Additionally, while this study did not consider real-time analysis, shorter audio input length should be considered. In this study we used 10 min recording of the “free talk” session and disregarded the processing time. However, in real case, it is more beneficial if the processing was complete before the patient and psychiatrist started the examination with clinical assessment tools.

Finally, in regards the dataset used for training and testing. All experiments were conducted in a Japanese hospital, with Japanese therapist, and with Japanese patient. Although the audio features relating to mental health are supposed to be independent with the language, there is a need to replicate this research outside of Japan and to evaluate the performance of our model against the publicly available databases. Utilizing other databases also have the benefit of the possibility for fair effectiveness evaluation with our model.

Author Contributions: Conceptualization, T.K.; methodology, B.S., Y.M. and T.K.; software, B.S. and K.-c.L.; formal analysis, B.S. and K.-c.L.; resources, T.F. and T.K.; data curation, M.Y., M.K. and T.K.; writing—original draft preparation, B.S., Y.M. and A.T.; writing—review and editing, B.S., Y.M., A.T. and T.K.; visualization, M.M. and T.K.; supervision, Y.M., M.M. and T.K.; project administration, M.M. and T.K.; funding acquisition, T.K. All authors have read and agreed to the published version of the manuscript.

Funding: As a part of PROMPT project, this research was funded by the Japan Agency for Medical Research and Development (AMED) under Grant Number JP18he1102004. The Grant was awarded on 29 October 2015 and ends on 31 March 2019.

Conflicts of Interest: The authors declare no conflict of interest. The funding source did not participate in the design of this study and did not have any hand in the study’s execution, analyses, or submission of results.

References

- World Alzheimer Report 2019: Attitudes to Dementia. Available online: <https://www.alz.co.uk/research/WorldAlzheimerReport2019.pdf> (accessed on 25 June 2020).
- Yang, H.; Bath, P.A. The Use of Data Mining Methods for the Prediction of Dementia: Evidence From the English Longitudinal Study of Aging. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 345–353. [[CrossRef](#)] [[PubMed](#)]
- Weiner, J.; Frankenberg, C.; Schroder, J.; Schultz, T. Speech Reveals Future Risk of Developing Dementia: Predictive Dementia Screening from Biographic Interviews. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 674–681.
- Shigemizu, D.; Akiyama, S.; Asanomi, Y.; Borevich, K.A.; Sharma, A.; Tsunoda, T.; Matsukuma, K.; Ichikawa, M.; Sudo, H.; Takizawa, S.; et al. Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data. *Commun. Biol.* **2019**, *2*. [[CrossRef](#)] [[PubMed](#)]
- Ju, R.; Hu, C.; Zhou, P.; Li, Q. Early Diagnosis of Alzheimer’s Disease Based on Resting-State Brain Networks and Deep Learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *16*, 244–257. [[CrossRef](#)] [[PubMed](#)]
- Hwang, A.B.; Boes, S.; Nyffeler, T.; Schuepfer, G. Validity of screening instruments for the detection of dementia and mild cognitive impairment in hospital inpatients: A systematic review of diagnostic accuracy studies. *PLoS ONE* **2019**, *14*, e0219569. [[CrossRef](#)]
- Robinson, L.; Tang, E.; Taylor, J.-P. Dementia: Timely diagnosis and early intervention. *BMJ* **2015**, *350*, h3029. [[CrossRef](#)]
- Da Cunha, A.L.V.; de Sousa, L.B.; Mansur, L.L.; Aluisio, S.M. Automatic Proposition Extraction from Dependency Trees: Helping Early Prediction of Alzheimer’s Disease from Narratives. In Proceedings of the 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, Sao Carlos, Brazil, 22–25 June 2015; pp. 127–130.
- Borson, S.; Frank, L.; Bayley, P.J.; Boustani, M.; Dean, M.; Lin, P.-J.; McCarten, J.R.; Morris, J.C.; Salmon, D.P.; Schmitt, F.A.; et al. Improving dementia care: The role of screening and detection of cognitive impairment. *Alzheimers Dement.* **2013**, *9*, 151–159. [[CrossRef](#)]
- Du, Z.; Li, Y.; Li, J.; Zhou, C.; Li, F.; Yang, X. Physical activity can improve cognition in patients with Alzheimer’s disease: A systematic review and meta-analysis of randomized controlled trials. *Clin. Interv. Aging* **2018**, *13*, 1593–1603. [[CrossRef](#)]
- Dominguez, L.J.; Barbagallo, M. Nutritional prevention of cognitive decline and dementia. *Acta Bio Med. Atenei Parmensis* **2018**, *89*, 276–290. [[CrossRef](#)]

12. Geifman, N.; Brinton, R.D.; Kennedy, R.E.; Schneider, L.S.; Butte, A.J. Evidence for benefit of statins to modify cognitive decline and risk in Alzheimer's disease. *Alzheimers Res. Ther.* **2017**, *9*, 1–10. [[CrossRef](#)]
13. Killin, L.O.J.; Starr, J.M.; Shiue, I.J.; Russ, T.C. Environmental risk factors for dementia: A systematic review. *BMC Geriatr.* **2016**, *16*, 175. [[CrossRef](#)]
14. Folstein, M.F.; Folstein, S.E.; McHugh, P.R. "Mini-mental state" a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **1975**, *12*, 189–198. [[CrossRef](#)]
15. Mendes-Santos, L.C.; Mograbi, D.; Spenciere, B.; Charchat-Fichman, H. Specific algorithm method of scoring the Clock Drawing Test applied in cognitively normal elderly. *Dement. Neuropsychol.* **2015**, *9*, 128–135. [[CrossRef](#)] [[PubMed](#)]
16. Yesavage, J.A.; Brink, T.L.; Rose, T.L.; Lum, O.; Huang, V.; Adey, M.; Leirer, V.O. Development and validation of a geriatric depression screening scale: A preliminary report. *J. Psychiatr. Res.* **1982**, *17*, 37–49. [[CrossRef](#)]
17. Wright, P.; Stern, J.; Phelan, M. (Eds.) *Core Psychiatry*, 3rd ed.; Elsevier: Edinburgh, Scotland, 2012; ISBN 978-0-7020-3397-1.
18. Kiloh, L.G. Pseudo-dementia. *Acta Psychiatr. Scand.* **1961**, *37*, 336–351. [[CrossRef](#)]
19. McAllister, W. Overview: Pseudodementia. *Am. J. Psychiatry* **1983**, 528–533.
20. Prakash, R.; Zhao, F.; Daggubati, V.; Giorgetta, C.; Kang, H.; You, L.; Sarkhel, S. Pseudo-dementia: A neuropsychological review. *Ann. Indian Acad. Neurol.* **2014**, *17*. [[CrossRef](#)]
21. American Psychiatric Association (Ed.) *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th ed.; American Psychiatric Association: Washington, DC, USA, 2013; ISBN 978-0-89042-554-1.
22. Jonsson, U.; Bertilsson, G.; Allard, P.; Gyllensvärd, H.; Söderlund, A.; Tham, A.; Andersson, G. Psychological Treatment of Depression in People Aged 65 Years and Over: A Systematic Review of Efficacy, Safety, and Cost-Effectiveness. *PLoS ONE* **2016**, *11*, e0160859. [[CrossRef](#)]
23. Ooi, K.E.B.; Lech, M.; Allen, N.B. Multichannel Weighted Speech Classification System for Prediction of Major Depression in Adolescents. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 497–506. [[CrossRef](#)]
24. Gavrilesco, M.; Vizireanu, N. Predicting Depression, Anxiety, and Stress Levels from Videos Using the Facial Action Coding System. *Sensors* **2019**, *19*, 3693. [[CrossRef](#)]
25. Dadiz, B.G.; Marcos, N. Analysis of Depression Based on Facial Cues on A Captured Motion Picture. In Proceedings of the 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP), Shenzhen, China, 13–15 July 2018; pp. 49–54.
26. Wu, L.; Pu, J.; Allen, J.J.B.; Pauli, P. Recognition of Facial Expressions in Individuals with Elevated Levels of Depressive Symptoms: An Eye-Movement Study. *Depress. Res. Treat.* **2012**, *2012*. [[CrossRef](#)]
27. Nakamura, R.; Mitsukura, Y. Feature Analysis of Electroencephalography in Patients with Depression. In Proceedings of the 2018 IEEE Life Sciences Conference (LSC), Montreal, QC, Canada, 28–30 October 2018; pp. 53–56.
28. Liao, S.-C.; Wu, C.-T.; Huang, H.-C.; Cheng, W.-T.; Liu, Y.-H. Major Depression Detection from EEG Signals Using Kernel Eigen-Filter-Bank Common Spatial Patterns. *Sensors* **2017**, *17*, 1385. [[CrossRef](#)] [[PubMed](#)]
29. Song, H.; Du, W.; Yu, X.; Dong, W.; Quan, W.; Dang, W.; Zhang, H.; Tian, J.; Zhou, T. Automatic depression discrimination on FNIRS by using general linear model and SVM. In Proceedings of the 2014 7th International Conference on Biomedical Engineering and Informatics, Dalian, China, 14–16 October 2014; pp. 278–282.
30. Henderson, G.; Ifeachor, E.; Hudson, N.; Goh, C.; Outram, N.; Wimalaratna, S.; Del Percio, C.; Vecchio, F. Development and assessment of methods for detecting dementia using the human electroencephalogram. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1557–1568. [[CrossRef](#)] [[PubMed](#)]
31. Warnita, T.; Inoue, N.; Shinoda, K. Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 1706–1710.
32. Kishimoto, T.; Takamiya, A.; Liang, K.; Funaki, K.; Fujita, T.; Kitazawa, M.; Yoshimura, M.; Tazawa, Y.; Horigome, T.; Eguchi, Y.; et al. The Project for Objective Measures Using Computational Psychiatry Technology (PROMPT): Rationale, Design, and Methodology. *medRxiv* **2019**. [[CrossRef](#)]
33. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **1960**, *23*, 56–62. [[CrossRef](#)] [[PubMed](#)]
34. Young, R.C.; Biggs, J.T.; Ziegler, V.E.; Meyer, D.A. A Rating Scale for Mania: Reliability, Validity and Sensitivity. *Br. J. Psychiatry* **1978**, *133*, 429–435. [[CrossRef](#)]

35. Mueller, K.D.; Hermann, B.; Mecollari, J.; Turkstra, L.S. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *J. Clin. Exp. Neuropsychol.* **2018**, *40*, 917–939. [[CrossRef](#)]
36. Mundt, J.C.; Vogel, A.P.; Feltner, D.E.; Lenderking, W.R. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biol. Psychiatry* **2012**, *72*, 580–587. [[CrossRef](#)]
37. Darby, J.K.; Hollien, H. Vocal and Speech Patterns of Depressive Patients. *Folia Phoniatr. Logop.* **1977**, *29*, 279–291. [[CrossRef](#)]
38. Gonzalez, S.; Brookes, M. PEFAC—A pitch estimation algorithm robust to high levels of noise. *IEEE Trans. Audio Speech Lang. Process.* **2014**, *22*, 518–530. [[CrossRef](#)]
39. Kim, H.-G.; Moreau, N.; Sikora, T. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*; John Wiley & Sons, Ltd.: Chichester, UK, 2005; ISBN 978-0-470-09336-8.
40. Sahidullah, M.; Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Commun.* **2012**, *54*, 543–565. [[CrossRef](#)]
41. Valero, X.; Alias, F. Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. *IEEE Trans. Multimedia* **2012**, *14*, 1684–1689. [[CrossRef](#)]
42. Peeters, G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO IST Proj. Rep.* **2004**, *54*, 1–25.
43. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [[CrossRef](#)] [[PubMed](#)]
44. Fonti, V.; Belitser, E. Feature Selection using LASSO. *VU Amst. Res. Pap. Bus. Anal.* **2017**, *30*, 1–25.
45. Sukan, N.; Sai Srinivas, N.S.; Kar, N.; Kumar, L.S.; Nath, M.K.; Kanhe, A. Performance Comparison of Different Cepstral Features for Speech Emotion Recognition. In Proceedings of the 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, 5–7 July 2018; pp. 266–271.
46. Adiga, A.; Magimai, M.; Seelamantula, C.S. Gammatone wavelet Cepstral Coefficients for robust speech recognition. In Proceedings of the 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), Xi'an, China, 22–25 October 2013; pp. 1–4.
47. Cheng, O. *Performance Evaluation of Front-end Processing for Speech Recognition Systems*; The University of Auckland: Auckland, New Zealand, 2005; Volume 33.
48. Zvarevashe, K.; Olugbara, O. Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition. *Algorithms* **2020**, *13*, 70. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).