Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(1): 45–65.
DOI: 10.3934/aci.2024004
Received: 20 June 2024
Revised: 19 July 2024
Accepted: 22 July 2024
Published: 29 July 2024

Research article

Classification of dementia from spoken speech using feature selection and the bag of acoustic words model

Marko Niemelä[1,*], Mikaela von Bonsdorff[2,3], Sami Äyrämö[1,4] and Tommi Kärkkäinen[1]

[1] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
[2] Faculty of Sport and Health Sciences and Gerontology Research Center, University of Jyväskylä, Jyväskylä, Finland
[3] Public Health Programme, Folkhälsan Research Center, Helsinki, Finland
[4] Wellbeing Services County of Central Finland, Finland
* **Correspondence:** Email: marko.p.niemela@jyu.fi.

Academic Editor: Pasi Fränti

Abstract: Memory disorders and dementia are a central factor in the decline of functioning and daily activities in older individuals. The workload related to standardized speech tests in clinical settings has led to a growing emphasis on developing automatic machine learning techniques for analyzing naturally spoken speech. This study presented a bag of acoustic words approach for distinguishing dementia patients from control individuals based on audio speech recordings. In this approach, each individual's speech was segmented into voiced periods, and these segments were characterized by acoustic features using the open-source openSMILE library. Word histogram representations were formed from the characterized speech segments of each speaker, which were used for classifying subjects. The formation of word histograms involved a clustering phase where feature vectors were quantized. It is well-known that partitional clustering involves instability in clustering results due to the selection of starting points, which can cause variability in classification outcomes. This study aimed to address instability by utilizing robust K-spatial-medians clustering, efficient K-means++ clustering initialization, and selecting the smallest clustering error from repeated clusterings. Additionally, the study employed feature selection based on the Wilcoxon signed-rank test to achieve computational efficiency in the methods. The results showed that it is possible to achieve a consistent 75% classification accuracy using only twenty-five features, both with the external ADReSS 2020 test data and through leave-one-subject-out cross-validation of the entire dataset. The results rank at the top compared to international research, where the same dataset and only acoustic features have been used to diagnose patients.

Keywords: Alzheimer; classification; spontaneous speech; acoustic features; bag of acoustic words

1. Introduction

Memory disorders impair memory, information processing, reasoning, and other cognitive functions. Cognitive decline is associated with difficulties in finding and understanding words and interruptions in thoughts [1]. Dementia is an advanced form of memory disorder where the deterioration of memory and decline in cognitive abilities impair the patient's ability to perform routine daily activities and hinder social and professional functioning. Consequently, as self-care becomes more difficult, the need for help increases starting from even a mild form of the disorder, and this significantly affects the patient's close ones as well. The most common cause of progressive memory disorders and dementia is Alzheimer's disease [2].

According to the World Health Organization (WHO), every three seconds, someone in the world is diagnosed with a memory disorder [2]. Memory disorders affect over 50 million people worldwide, and the number is expected to triple by the year 2050. The aging of the population is the main driver of the increased prevalence of memory disorders. There is a notable increase in countries such as Japan and many European nations where life expectancy has risen, and birth rates have declined. However, memory disorders impact nearly every country worldwide. In the year 2018, the global costs of memory disorders reached nearly a billion dollars [2].

Currently, there is no curative drug treatment for memory disorders, but early diagnosis, treatment, and rehabilitation can maintain and improve a patient's functional capacity. Studies have shown that individuals with healthier lifestyles among memory disorder patients experience fewer risk factors for worsening memory disorders and can lead a good and fully functional life [3]. The FINGER study was the first in the world to demonstrate that adhering to a comprehensive lifestyle program can improve cognitive functions and prevent the decline of memory functions [4]. Diagnosing memory disorders, especially at an early phase, has been challenging because diagnostics and treatment evaluation require specialized expertise and monitoring. In recent years, there has been increased investment and focus on developing diagnostics and rehabilitation.

Traditionally, cognitive decline has been assessed using conventional pen-and-paper tests, including the Mini-Mental State Examination (MMSE) [5], the Montreal Cognitive Assessment (MoCA) [6], and the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) test, particularly in the early stages of decline [7]. Alongside the paper tests, cognitive impairment related to memory disorders and potential brain injuries has been studied in clinical settings by analyzing patients' speech production conducted by experts. However, this approach has not demonstrated practicality as it is subjective and often challenging to replicate. Consequently, there has been a shift toward developing automatic natural language processing (NLP) machine learning techniques in recent years. These methods have demonstrated utility by being cost-effective, scalable, and providing highly rapid diagnoses [8]. However, the methods are often too complex, demanding significant training data, computationally intensive, or involve too many adjustable training parameters [9, 10].

In machine learning, there has often been a focus on diagnosing Alzheimer's disease or mild cognitive impairment [11]. Speech has been analyzed based on lexical, acoustic, or both lexical and acoustic content. Analyzing the lexical content of speech remains challenging because the content needs to be extracted from the recording either manually, which is error-prone and time-consuming, or by using speech-to-text methods that typically require large speech databases and are not language-independent. Further, lexical features can more easily compromise patient privacy. For instance, a transcript may include information about a patient's inner circle or address [11].

The bag of words approach has been used for processing naturally spoken speech. This approach forms an unordered list from a set of objects, containing a limited number of objects that best describe the set and their occurrences. In the early stage, the model has been used in text categorization [12], such as spam email filtering. The word histogram model has been expanded to include visual words (bag of visual words, BoW), used in image categorization [13], and acoustic words (bag of acoustic words, BoAW) in speech processing [14].

Typically, in a BoaW model, a set of feature vectors is represented as a histogram of quantized vectors. Quantization is performed by clustering the data and associating the feature vectors with the nearest cluster prototypes. The method involves uncertainty caused by the randomness of the clustering result, which this study aims to address by using a robust K-spatial-median clustering method, an efficient K-means++-type clustering initialization, and repeated clustering. In this study, feature selection based on statistical testing aims to reduce the computational complexity of clustering and classification. In addition, the selection of features helps in understanding the most relevant information for the research. Word histogram representations are created for the voiced segments derived from patients' speech recordings, based on the 25 most important acoustic features extracted using the openSMILE library*. The study shows that although the partitioning caused by clustering results in some variance, it is possible to distinguish dementia patients from control patients with an average classification accuracy of 75.2% (*SD:* ±4.0%) on separate test data and with an accuracy of 75% using leave-one-subject-out (LOSO) cross-validation for the full dataset.

2. Related work and reference results

Hernanz-Dominguez et al. [15] conducted a diagnosis based on acoustic features and achieved a classification accuracy of 62.0% using low-level mel-frequency cepstral coefficients features (MFCC 1-13). The features were extracted from 25 millisecond audio segments, and the following statistical measures were computed: averages, kurtoses, skewness, and variances. In [16], memory-impaired patients were diagnosed with an accuracy of 68.0% based on both the frequencies of speech and speech pauses, including averages, variances, minimum and maximum values, and entropies. In [17], classification accuracies ranging from 60.0% to 93.8% were achieved using speech duration, time- and frequency-domain features, and emotional states. However, the dataset was relatively small (40 individuals). Additionally, the control group included relatively young individuals (25% of subjects were 20–60 years old) compared to Alzheimer's patients (all over 60 years old).

In the development of machine learning methods, it is crucial to keep the training and test datasets separate to prevent classification results from being biased. In many studies based on machine learning and natural spoken speech diagnostics, classification methods are not designed to be independent of patients. Results have been published where multiple recordings from the same patients were used to train and test the machine learning method on data obtained from the same patient. In these cases, classification methods have not learned to identify the disease itself but have instead overfit the speech production characteristics of individual patients, rarely generalizing well to new data. Consequently, classification results on the datasets used in experiments may appear better than if applied to a different dataset, as underlined in [18].

*`https://www.audeering.com/research/opensmile/`

In [19], classification was performed based on a subset extracted from the Pitt Corpus audio database[*], consisting of control subjects and dementia patients. The database included a total of 156 individuals, half of whom had a diagnosis of dementia. The dataset was divided into training and testing sets based on individuals' diagnoses, ages, and gender distributions. Noise reduction was applied to the audio recordings, and voiced segments were identified using a signal energy-based threshold for speech. Acoustic features were computed for these segments using the functional emobase, computational paralinguistics challenge (ComPaRe) [20], and Geneva minimalistic acoustic parameter set (eGeMAPS) features from the OpenSMILE library [21]. These included low-level descriptors (LLD) and statistical measures for LLD features [22]. Additionally, statistical measures were selected for multi-resolution cochleagram features (MRCG), as well as features related to pronunciation and speech pauses used in [16]. Correlated features were removed from the feature sets. Results for an individual were based on the modes of classification results for all speaker segments using different feature sets. The best classification accuracy based on acoustic features was achieved with the ComPaRe feature set. For the test dataset, an individual achieved a prediction accuracy of 62.5%, and with the LOSO cross-validation used in the experiments, the overall classification accuracy for the entire dataset was 56.5%.

Syed et al. [23], achieved a classification result of 76.9% using LOSO cross-validation on the training data presented in the ADReSS 2020 challenge[†] [19]. The result was based on acoustic IS10-Paraling features (1582 features sampled from the ComPaRe feature set), a bag of acoustic words (BoAW) model [24] provided by the openXBOW tool[‡], and support vector machine (linear kernel) and logistic regression classifiers. In [18], a similar data division was utilized as in [19], but with a larger number of selected individuals (a total of 164 individuals). The work introduced a new active data representation method based on self-organizing map (SOM) clustering of feature vectors, associating vectors with their nearest clusters, computing durations of voiced segments, and histogram feature extraction. The proposed model methodologically resembles the BoAW model, where clustering is also used to associate feature vectors with their nearest clusters. The best classification result (77.4%) was achieved using non-correlated features from the eGeMAPS feature set (a total of 75 features) and a classifier based on linear discriminant analysis (LDA).

The results from studies [18, 23] are promising. However, the papers do not address the challenges of data clustering, such as the SOM network and K-means clustering used in the BoAW model. In K-means, there are Stirling numbers of the second kind, given by $S(N, K) = \frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i}\binom{K}{i}(i)^{N}$, to partition N observations into K groups which is evidently infeasible for computation [25]. Typically, clustering methods converge to a local minimum of the error function, which is why methods should undergo multiple initializations (often 100 iterations are used [26, 27]), and then the initialization with the smallest clustering error is selected. Multiple initializations do not guarantee an optimal result, but it increases the probability of obtaining a cluster model that effectively describes the internal structure of the data.

[*] `https://dementia.talkbank.org/access/English/Pitt.html`

[†] `https://luzs.gitlab.io/adress/`

[‡] `https://github.com/openXBOW/openXBOW`

3. Data and methods

3.1. Pitt Corpus audio database

The Pitt Corpus audio database is one of the databases provided by DementiaBank*. The data for the database was collected between 1983 and 1988 as part of Alzheimer's research at the University of Pittsburgh [28]. The study included 282 individuals, out of which 101 were healthy control subjects, and 181 were Alzheimer's patients. Participants had to be over 44 years old, have at least seven years of education, no central nervous system abnormalities, and score at least ten out of thirty points on the Mini-Mental State Examination (MMSE) as a preliminary result [5].

The selected participants performed oral tasks, and their performance in everyday tasks was also assessed. Among the participants' oral tasks was a kitchen scene description task designed to measure speech disorders (a mother washing dishes, and children standing on stools stealing pastries). A healthcare professional provided instructions to the patients at the beginning of the image description task, and the patients' responses were recorded with a microphone. For the ADReSS 2020, a random sample of participants in the image description task from the Pitt Corpus database was selected, ensuring that the age and gender distributions of control subjects and dementia patients matched. Furthermore, only one recording was selected from each individual [19]. A total of 78 control subjects and 78 dementia patients were chosen. Approximately 69% of the individuals were selected for the training dataset, and approximately 31% were selected for the test dataset (see Tables 1 and 2). Recordings were acoustically enhanced by removing stationary noise, and audio volume normalization was applied across segmented speech to control for variations caused by recording conditions. It is important to note that in all databases age and gender distributions are not always available. In such cases, it would be good to be able to estimate the ages and genders of the participants (see, for example, the study [29]).

Table 1. ADReSS 2020 training dataset (M=male, F=female, AD=Alzheimer's dementia, MMSE=mini-mental state examination).

	AD			non-AD		
Age	M	F	MMSE (*SD*)	M	F	MMSE (*SD*)
[50, 55)	1	0	30.0 (n/a)	1	0	29.0 (n/a)
[55, 60)	5	4	16.3 (4.9)	5	4	29.0 (1.3)
[60, 65)	3	6	18.3 (6.1)	3	6	29.3 (1.3)
[65, 70)	6	10	16.9 (5.8)	6	10	29.1 (0.9)
[70, 75)	6	8	15.8 (4.5)	6	8	29.1 (0.8)
[75, 80)	3	2	17.2 (5.4)	3	2	28.8 (0.4)
Total	24	30	17.0 (5.5)	24	30	29.1 (1.0)

*https://dementia.talkbank.org/access/

Table 2. ADReSS 2020 test dataset (M=male, F=female, AD=Alzheimer's dementia, MMSE=mini-mental state examination).

	AD			non-AD		
Age	M	F	MMSE (*SD*)	M	F	MMSE (*SD*)
[50, 55)	1	0	23.0 (n.a)	1	0	28.0 (n.a)
[55, 60)	2	2	18.7 (1.0)	2	2	28.5 (1.2)
[60, 65)	1	3	14.7 (3.7)	1	3	28.7 (0.9)
[65, 70)	3	4	23.2 (4.0)	3	4	29.4 (0.7)
[70, 75)	3	3	17.3 (6.9)	3	3	28.0 (2.4)
[75, 80)	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	11	13	19.5 (5.3)	11	13	28.8 (1.5)

3.2. *Data preprocessing*

The audio recordings, recorded in WAVE format with a microphone (duration *Mean*: 55.3 seconds, *SD*: 29.3 seconds), contained a significant amount of noise, which was removed using Adobe Audition's (VERSION 23.6) adaptive noise reduction filter. The adaptive noise reduction filter introduced short noisy segments at the beginning of the recordings, which were manually removed from the recordings. Additionally, extraneous noises not related to participants, such as caregiver speech, overlapping speech, background noise, buzzer sounds, typing sounds, and emergency vehicle sounds, were removed from the recordings. Toward the end, the recordings only contained either participants' speech or silent periods as participants thought about the task. The recordings were resampled[*] from a 44 kHz sampling rate to a 16 kHz sampling rate which accommodates most of the speech frequencies. The resampled audio recordings were normalized to standard audio volumes according to the EBU R128 Standard [30] to minimize the effects of different recording conditions, such as the impact of microphone placement.

3.3. *Speech activity detection*

Segments of voiced speech were identified from the recordings using the open-source Auditok library[†]. The energy threshold for speech detection (65 dB) and the maximum duration of speech segments (10 seconds) were the same as in [19]. Default values of the Auditok library were used as limits for the minimum duration of continuous speech (0.2 second) and the maximum duration of speech pauses (0.3 second). As a result, 2107 audio segments were obtained for the training dataset (duration *Mean*: 1.27 seconds, *SD*: 0.93 second), and 959 audio segments were obtained for the test dataset (duration *Mean*: 1.28 seconds, *SD*: 0.94 second). Finally, the audio segments were normalized to standard audio volumes according to the EBU R128 Standard.

3.4. *Extraction of audio segment features*

The eGeMAPS is a minimal set of features developed for automatic paralinguistic or clinical speech analysis [22]. These features are effective in identifying psychological changes in speech production [21]. They are statistical measures for low-level descriptors (LLD) of audio. The LLD

[*] `https://sourceforge.net/projects/sox/`

[†] `https://github.com/amsehili/auditok`

features include, for example, F0 fundamental frequency (pitch), speech intensity (loudness), spectral flux rate, MFCC features, pitch vibration frequency and amplitude (jitter and shimmer), F1, F2, and F3 formant regions, energy ratio (difference between upper and lower spectrum frequencies, that is, alpha ratio), harmonic differences, harmonic-to-noise ratio (HNR), Hammarberg index, and predicted spectral slope. For the audio segments, the OpenSMILE library (version 3.0.1) was used for forming the LLD and statistical eGeMAPS features (a total of 88 features). The eGeMAPS features are computed from LLDs using, for example, arithmetic mean, standard deviation, coefficient of variation (standard deviation divided by the arithmetic mean), percentiles (25, 50, and 80), and interquartile range 25 to 80. Additionally, the study measured the relative speech pauses before initiating speech production. The speech pause was normalized to the total recording duration and the cumulative speech pause in the recording (a total of two features).

3.5. Feature normalization

Normalization is essential, especially in distance-based applications, because otherwise the features with large scales dominate the distance computations. Unlike the z-score normalization method, min-max normalization is not based on the normal distribution assumption. Min-max normalization to the range $[0, 1]$ corresponds to the following linear transformation:

$$x' = \frac{1}{\max(x) - \min(x)} x + \frac{\min(x)}{\min(x) - \max(x)},$$

where x is the original variable and x' is the normalized variable. The issue with feature normalization lies in the presence of outliers, which distort normalization parameters. In this study, a total of 159 outliers were replaced with non-outlying data, which is a very small number (less than 0.1% of all of the data), ensuring that the test data were scaled accordingly to the range of $[0, 1]$.

3.6. Random forest model with statistical feature selection

Random forest (RF) is a non-linear classification and regression model [31]. Random forest builds a meta-estimator that aims to improve the prediction accuracy of a simple decision tree model and prevent overfitting by dividing the data into multiple decision tree estimators and utilizing the average for regression tasks or the mode for classification tasks. The leaf nodes of the random forest trees determine the values of the response variables given the inputs to the tree. In this study, the Gini index of the random forest classifier was used to describe the average impurity decrease, and this can be computed using the probabilities of two or more classes at each tree node. The more the impurity decreases, the more important the feature is in a classification task. The sum of the importance values of the dataset features, computed with the Gini index, is 100%.

The index is computed as follows:

$$Gini_n = 1 - \sum_{i=1}^{k} p_i^2,$$

where p_i represents the probability that samples in node n belong to class i from k classes. In a binary classification task, the formula reads as: $Gini_n = (p_1 + p_2) - p_1^2 - p_2^2 = p_1(1 - p_1) + p_2(1 - p_2)$.

Search for the best classification accuracy or the simplest model may include the search and selection of only the essential features [32]. In this study, the feature selection process is conducted using RF, as detailed below:

(1) The training dataset is normalized to the $[0, 1]$ scale.

(2) The random forest classification model is trained 100 times on the training dataset. For each iteration, feature importance values are computed based on the Gini index of the RF.

(3) The RF is trained 100 times on the training dataset using permuted response variables. For each iteration, permuted importance values are computed for the features.

(4) The distributions of 100 importance values for each feature and the importance values obtained through permutations are compared using the statistical Wilcoxon signed-rank test. The null hypothesis assumes that the values between the groups come from the same median distribution at a 5% statistical significance level ($p = 0.05$).

(5) The final features are selected as the k highest-ranked features in terms of average importance values that rejected the null hypothesis of the Wilcoxon test.

3.7. Bag of acoustic words model

The word histogram model represents an object as an unordered collection of words. Typically, the content of speech is not directly used in the model. Instead, compact feature vector representations are derived from speech segments, for instance, by analyzing short-term frequency spectra. Following this, a histogram distribution of feature vectors is constructed, which is then utilized in analyzing the speech. The used word histogram method is illustrated in Figure 1. Initially, in the word histogram model, preprocessing, feature extraction, and feature selection are applied to the identified training and test segments as depicted in Sections 3.2–3.6. Subsequently, a limited number of cluster centroids, or cluster prototypes, are formed from the feature vectors of the training data by clustering the data classes separately into a specified number of clusters. K-means, one of the most well-known clustering methods [33], is not robust to outliers because the cluster prototypes are represented by the means of the data sets. The K-spatial-median clustering used in the experiments, combined with an appropriate initialization method, is robust because the prototypes are based on the multidimensional medians of the sets [34, 35]. The initialization of clustering performed with the K-means++ method randomly selects starting points and favors points that are far apart from each other [36]. The clustering result chosen is the one that best minimizes the clustering error based on Euclidean distance:

$$\mathcal{J}(\{\mathbf{c}_k\}) = \sum_{i=1}^{N} \min_{k=1,\dots,K} \|\mathbf{x}_i - \mathbf{c}_k\|_2, \tag{1}$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ denotes the data and $\{\mathbf{c}_k\}_{k=1}^K$ is the set of cluster prototypes that locally minimizes the error function (see Eq (1)). This minimization problem is non-smooth and it is realized using the sequential over-relaxation (SOR) algorithm with $w = 1.5$ as the step size factor [37, 38].

Cluster centers form a collection of prototype vectors (codebook), and the feature vectors of segmented speech segments are quantized by associating the vectors with the nearest prototype vectors (words) in the codebook based on Euclidean distances. The entire content of the speech is described by a histogram of size $2 \times Nc$, where Nc represents the number of clusters for the dementia and control classes. Each position in the histogram represents the occurrence of one codebook word in the speech. After forming the histogram, it is normalized by the total number of words in the histogram. The normalized histograms are used as feature vectors in classification.

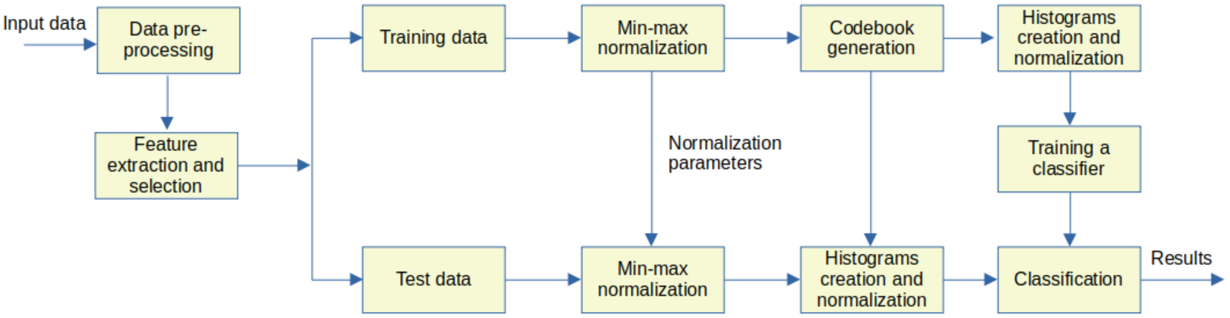

Figure 1. Classification process over the bag of acoustic word model.

3.8. *Comparison of classification methods*

Evaluation of the generalization capability of a dementia classifier is based on its accuracy on a test set. As depicted in Section 3.1, a balanced division into training and test data is provided a priori for the ADReSS 2020 dataset. However, more generally for a set of speech audio samples with labeling into AD and non-AD, our primary interest lies in whether a classifier can reliably identify the class of an unseen, new subject. This capability is measured using LOSO cross-validation. There, in each iteration for N subjects, $N - 1$ cases are used for training the model, and the remaining case is used for testing the model. The validation process is repeated N times, ensuring that each subject is tested exactly once.

Six different methods were compared and used for the binary classification problem of dementia: five-nearest neighbors (5-NN) [39], linear discriminant analysis (LDA) [40], random forest (RF, 50 decision trees and 5 leaf nodes) [31], extreme minimal learning machine (EMLM, with the whole training set as reference points) [41], linear support vector machine (L-SVM, with overfitting parameter $C = 1.0$) [42], and support vector machine with chi-squared kernel (Chi2-SVM, overfitting parameter $C = 0.25$) [42]. Classifier parameters were determined through manual testing. Among the classifiers, 5-NN and L-SVM were based on MATLAB (VERSION R2022B, 64-BIT)[*], LDA and RF were used from the scikit-learn library in Python (VERSION 1.2.2)[†], Chi2-SVM was based on the libsvm library (VERSION 3.32)[‡], and EMLM was based on the source code from GitLab[§].

Because we consider a binary classification problem, the support vector-based technique that maximizes the margin between the two sets is, by construction, a potential approach. Especially, SVM with a chi-squared kernel has demonstrated to be effective in histogram classification [43]. This kernel is based on chi-squared distances, which are incorporated into the kernel function using an extended

[*] https://www.mathworks.com/help/pdf_doc/stats/index.html
[†] https://scikit-learn.org/stable/user_guide.html#user-guide
[‡] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[§] https://gitlab.jyu.fi/hnpai-public/extreme-minimal-learning-machine/

Gaussian kernel:

$$K\left(S_i, S_j\right) = exp\left(\frac{-1}{A} D\left(S_i, S_j\right)\right),$$

where $D(S_i, S_j)$ represents the chi-squared distance between word histograms S_i and S_j, and A denotes a scaling parameter, which is the average of all training histogram chi-squared distances.

The performance of the classification models was evaluated using the ADReSS 2020 test dataset (48 individuals) and by conducting LOSO cross-validation on the entire dataset (156 individuals). The same classifier parameters were utilized in both experiments. Note that only SVMs and RF classifiers included adjustable hyperparameters.

4. Results

4.1. Feature selection

In feature selection (depicted in Section 3.6), a total of 25 features were identified. Among the most important features were the functional features related to MFCC 1-4 coefficients, speech F0 fundamental frequency, spectral slopes predicted by linear regression (from frequency bands 0Hz–500Hz and 500Hz–1.5kHz), and spectral harmonics (two different ratios). Additionally, important features included relative speech pauses (pauseTotalDurationRatio and pauseTotalPausesRatio). The combined sum of the average importance scores for the selected features was 36.7%. Figure 2 illustrates the average importance scores of selected and non-selected features. All features and scores are listed in Appendix A.

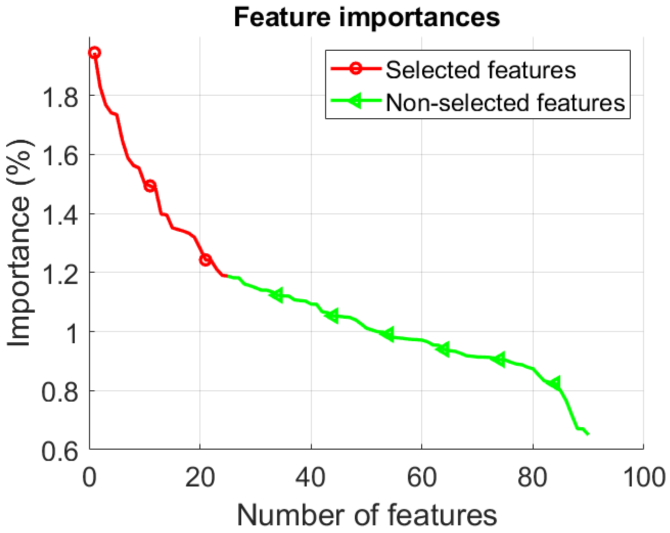

Figure 2. Average feature importances in sorted order for the eGeMAPS feature set and relative speech pauses.

4.2. Classification of test data

The performance of classifiers was measured with the selected 25 features and different-sized codebooks. The cluster prototypes of control subjects and dementia patients were both increased in increments of five from five prototypes to fifty prototypes. Table 3 illustrates the average accuracies and standard deviations over 100 repeated clustering runs (each with 100 clustering reinitializations). From the table, it can be observed that the Chi2-SVM performed the classification task best, with an average classification accuracy of 75.2% (*SD:* ± 4.0%) on the test dataset. The RF model used for feature selection achieved a good classification accuracy of 71.9% (*SD:* ± 4.8%). The best results were obtained with a codebook of 30 prototypes. The variability in classification accuracies can be explained in part by the randomness associated with the classification models and partly by the randomness in the selection of clustering starting points. The variation in 5-NN classification results (2.9%–5.3%) is entirely explained by the randomness of the starting points since the classification model itself does not involve randomness. Table 3 indicates that the performance of the classifiers begins to decline as the codebook size increases beyond 30 prototypes.

Table 3. Average classification accuracies and standard deviations for the ADReSS 2020 test dataset over 100 repetitions of replicated clustering.

	Classifier					
	5-NN	RF	L-SVM	Chi2-SVM	LDA	EMLM
NClust.						
5+5	68.9%	63.7%	64.3%	59.6%	60.6%	57.5%
	(± 2.9%)	(± 4.5%)	(± 1.7%)	(± 2.1%)	(± 2.5%)	(± 2.9%)
10+10	**70.2%**	71.5%	66.9%	70.2%	**67.7%**	66.3%
	(± 5.0%)	(± 4.3%)	(± 3.3%)	(± 3.2%)	**(± 3.9%)**	(± 3.7%)
15+15	66.9%	**71.9%**	**68.4%**	**75.2%**	64.5%	**67.3%**
	(± 5.2%)	**(± 4.8%)**	**(± 4.5%)**	**(± 4.0%)**	(± 5.3%)	**(± 4.6%)**
20+20	64.9%	70.7%	66.3%	73.7%	62.0%	65.3%
	(± 4.5%)	(± 4.7%)	(± 3.9%)	(± 3.8%)	(± 6.0%)	(± 5.3%)
25+25	64.7%	68.5%	64.4%	70.8%	61.0%	63.4%
	(± 4.1%)	(± 4.4%)	(± 3.8%)	(± 4.1%)	(± 5.5%)	(± 5.0%)
30+30	63.7%	67.9%	64.3%	68.8%	59.8%	63.0%
	(± 4.6%)	(± 4.5%)	(± 3.8%)	(± 4.2%)	(± 5.6%)	(± 5.0%)
35+35	63.3%	68.2%	63.7%	68.9%	59.3%	63.6%
	(± 4.6%)	(± 5.4%)	(± 4.5%)	(± 4.4%)	(± 5.9%)	(± 4.9%)
40+40	62.8%	66.5%	63.1%	67.9%	57.2%	61.9%
	(± 4.9%)	(± 5.3%)	(± 4.4%)	(± 4.1%)	(± 6.1%)	(± 4.8%)
45+45	61.8%	65.8%	62.5%	67.4%	55.0%	62.7%
	(± 4.6%)	(± 5.4%)	(± 4.6%)	(± 4.6%)	(± 5.4%)	(± 4.1%)
50+50	62.2%	65.0%	61.9%	66.8%	54.3%	62.6%
	(± 5.3%)	(± 5.8%)	(± 4.8%)	(± 4.6%)	(± 7.2%)	(± 4.9%)

4.3. Performance in LOSO cross-validation

The generalization performance of the classification models was measured by LOSO cross-validation on the dataset of 156 individuals from the ADReSS 2020. The size of the codebook was increased from four cluster prototypes to forty prototypes. The uncertainty associated with cluster error was addressed by repeating LOSO cross-validations a total of 25 times, with each clustering iteration involving 100 reinitializations. The results were obtained by consistently using either the modes or the smallest clustering errors from the 25 binary classification outcomes. Figure 3 shows the classification results for the entire dataset using different classifiers, always choosing the better result between modes and the smallest clustering errors. Note that the results are given separately in Appendix B. Figure 3 illustrates that the best results for most classifiers were achieved with a codebook size of 22 prototypes. The Chi2-SVM classifier performed the classification task best, with a classification accuracy of 72.4%. Additionally, RF and LDA produced classification results of over 70%. Table 4 provides LOSO cross-validation accuracies and F1 scores with a codebook size of 22 prototypes and 75 clustering iterations (each with 100 replicates). The classification results are based on the smallest clustering error outcome. Clearly, the Chi2-SVM classifier emerged as the top performer in LOSO cross-validation with an accuracy of 75.0%. Among the classifiers, RF and EMLM achieved classification accuracies of 71.8% and 71.2%, respectively. Figure 4 shows receiver operating characteristic (ROC) curves for Alzheimer's dementia and non-Alzheimer's dementia groups. The probability values required for forming the ROC curves of the groups were computed from binary classification results over 75 repetitions. The final classes were based on the classification results obtained with the smallest clustering errors. The performance of individual classifiers is represented by the area under the curve (AUC) values, which are provided in the legends of Figure 4. Clearly, control subjects are easier to distinguish from Alzheimer's patients.

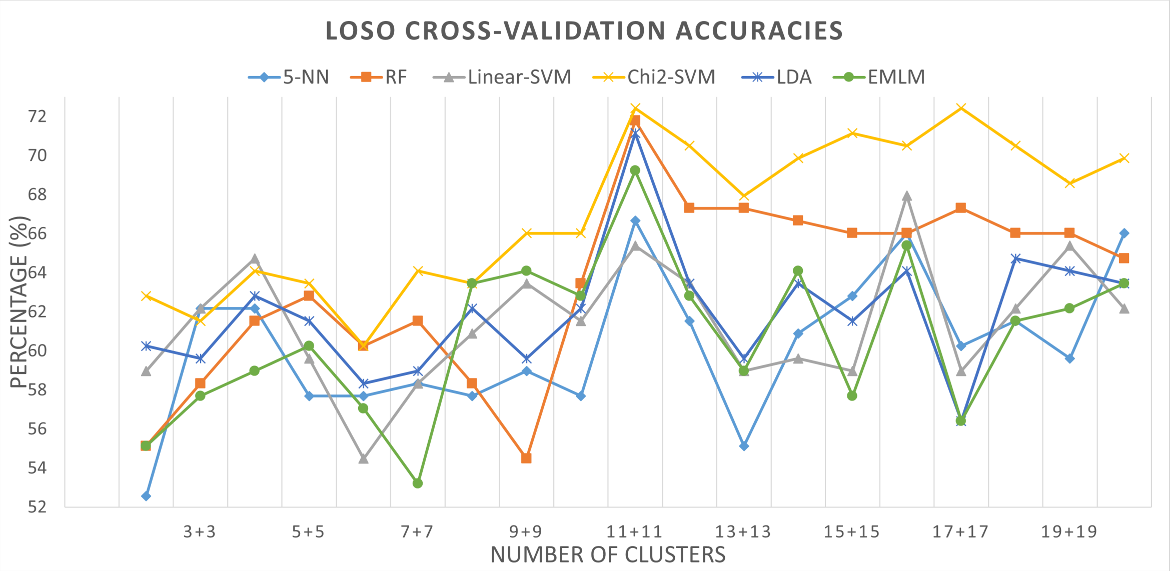

Figure 3. LOSO cross-validation accuracies over 25 repetitions of replicated clustering.

Table 4. LOSO cross-validation results using word histograms with 22 bins and 75 repetitions of replicated clustering (AD=Alzheimer's dementia).

	Classifier					
	5-NN	RF	L-SVM	Chi2-SVM	LDA	EMLM
Accuracy	64.1%	71.8%	66.0%	**75.0%**	69.9%	71.2%
F1 Score (AD)	65.4%	69.9%	62.4%	**72.3%**	70.4%	71.3%
F1 Score (non-AD)	62.7%	73.5%	69.0%	**77.2%**	69.3%	71.0%

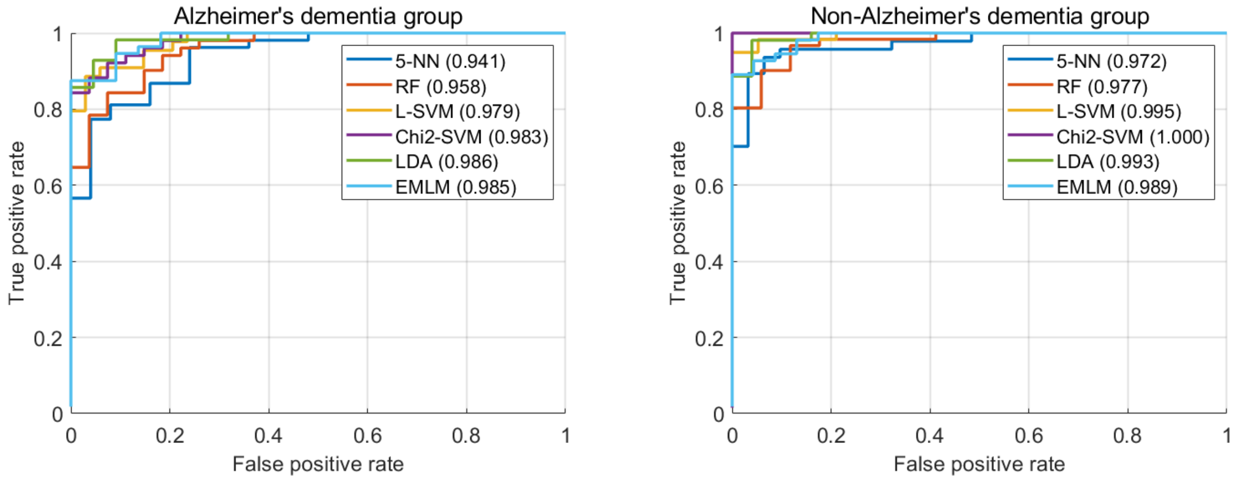

Figure 4. Receiver operating characteristic curves for Alzheimer's dementia and non-Alzheimer's dementia groups, based on histograms with 22 bins and 75 repetitions of replicated clustering.

5. Discussion

In this study, diagnostic analysis related to dementia was conducted on a dataset extracted from the Pitt Corpus audio database, consisting of both control subjects and dementia patients. Statistical features for acoustic speech analysis were extracted from audio segments using the OpenSMILE library, and histograms were constructed based on the BoAW approach. Similar analyses, based on the same dataset and a comparable histogram representation, have been conducted in previous studies, with reported LOSO cross-validation accuracies of 76.9% and 77.4% [18,23]. However, these studies do not provide details on how they accounted for randomness in the clustering process employed for histogram formation.

The study experiments with the ADReSS 2020 test dataset reveal that, through 100 iterations of clustering, the variability in classification outcomes using a deterministic 5-NN classifier typically falls within the range of 4.0% to 5.0%. Therefore, in the experiments, 25 repetitions of LOSO cross-validation were employed when searching for the number of clusters, and 75 repetitions were used in the final results computation. It is important to note that the LOSO cross-validation method itself involves repetitions of grouping equal to the number of subjects to be validated, leading to a multiplication of the final repetitions by the number of subjects (see the beginning of Section 3.8).

In [23], a large set of acoustic features (a total of 1582 features) has been used, leading to computational complexity for the methods. In this study, the computation of average feature importance scores was performed for relative speech pauses and eGeMAPS features (a total of 90 features) using a random forest classifier based on the Gini index. Feature selection was conducted based on the null hypothesis of the Wilcoxon statistical test. Feature selection relying on statistical testing is a straightforward and relatively efficient method for selecting final features [44].

In previous literature, feature selection has been performed for the eGeMAPS feature set in emotion recognition based on speech recordings [45]. The study utilized four different feature selection mechanisms. In the infinite latent feature selection (ILFS) method, feature importance values are based on all possible subsets of features, representing different paths in the feature graph. The ReliefF method uses weight vectors to represent the connections of features to actual class labels. The generalized Fisher score method (Fisher) seeks a set of features that maximizes the lower bound of a credibility function based on the Fisher metric. The active feature selection method (AFS) clusters individual features of the dataset and selects final features based on the distinctiveness of the clusters. In [45], three datasets related to emotion recognition were combined into one set, and the feature selection methods proposed 44–79 eGeMAPS features as the final quantities of features.

Although K-spatial-median clustering is robust, it is important to note that it is computationally more complex than the K-means algorithm, which can become a bottleneck with larger datasets. One may use random swap (RS) clustering, which is computationally more efficient than K-means [46]. In RS clustering, it is possible to estimate the expected number of iterations needed to find the correct clustering. The expected processing time is known to be linearly dependent on the number of data vectors, quadratically dependent on the number of clusters, inversely dependent on the neighborhood size, and logarithmically dependent on the number of successful swaps needed.

In this study, among the key features, statistical features of the MFCC, features related to the fundamental frequency (F0), spectral slopes predicted by linear regression, and harmonic features of spectra were particularly prominent. Especially, spectral slope and harmonic features are known to be associated with cognitive load, that is, the capacity of our working memory to handle information at any given moment [47, 48]. Excessive stress and high cognitive load are known to be harmful to brain structure and function, which is known to increase the risk of cognitive decline and dementia [49].

The fundamental frequency of the speech is known to be associated with Parkinson's disease [50]. Therefore, the fundamental frequency of vowel sounds, along with frequency oscillation, has been investigated in the diagnosis of Parkinson's disease [51]. Schmitt et al. [14] used low-level MFCC features together with signal energy for emotion recognition [14]. However, instead of speech activity detection, features were extracted from unsegmented audio recordings in 25-millisecond durations using a 10-millisecond audio sampling. For low-level features, statistical features (means and standard deviations) were also computed. The final classification results were based on the creation of histograms using the BoAW model. The presented previous research results support the broader scalability of the features and methods used in this study to other speech-identifiable common diseases or different emotions, which could serve as indicators for illnesses or depression [52, 53].

6. Conclusions

In this research, individuals with dementia were identified from healthy controls based on naturally spoken speech. Features related to speech production are known to correlate with the speaker's cognitive ability and changes. The study did not aim to use lexical features derived from challenging speech recognition tasks, such as converting speech to text or analyzing the content of the transcribed text. Instead, spoken speech was categorized based on the acoustic characteristics of the speech. Classifiers and histogram feature extraction based on the BoAW model have demonstrated promise in distinguishing dementia patients from control subjects. The implementation of the BoAW model accounted for the variability in classification results caused by the multiple ways to partition the data by using repeated clusterings, each with multiple reinitializations, as part of the model. The final results were based on either the modes of the classification results or the results with the smallest clustering errors.

The generalization capabilities of the classifiers were tested using the ADReSS dataset, where subjects from the Pitt Corpus database were evenly selected based on age and gender distribution, ensuring that only one audio recording was selected from each individual to avoid overfitting by the classifiers. Furthermore, the classification results were measured with separate test data and by validating the entire dataset using LOSO cross-validation. In the experiments, the classifiers were trained for both tasks using the same classifier hyperparameters. The best accuracies with the test data and the entire dataset were consistent, indicating that overfitting did not occur. Feature selection in this work improved the computational efficiency of the methods. The results were based on only a small number of acoustic features, emphasizing spectral features (slope and harmonics), MFCCs, and statistical characteristics of the fundamental frequency. This minimal feature set provides researchers with opportunities to further develop feature extraction in a more reliable manner. Further, acoustic features are independent of the content of speech. Therefore, the developed diagnostic methods have the potential to be expanded to multiple languages. Overall, the obtained results are promising when fast, cost-effective, and scalable solutions are needed for the rapid diagnosis of cognitive decline or dementia.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The work of the first Author (MN) was supported by the Finnish Cultural Foundation (Grant Number 30231766). The work of the second author (MvB) was supported by the Samfundet Folkhälsan, and the Research Council of Finland (Grant Number 349336).

Conflict of interest

Tommi Kärkkäinen is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. M. W. Bondi, D. P. Salmon, A. W. Kaszniak, The neuropsychology of dementia, In: *Neuropsychological assessment of neuropsychiatric and neuromedical disorders*, Oxford: Oxford University Press, 2009, 159–198.
2. World Health Organization, *Global action plan on the public health response to dementia 2017–2025*, World Health Organization, 2017.
3. R. N. Kalaria, G. E. Maestre, R. Arizaga, R. P. Friedland, D. Galasko, K. Hall, et al., Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors, *Lancet Neurol.*, **7** (2008), 812–826. http://dx.doi.org/10.1016/S1474-4422(08)70169-8
4. T. Ngandu, J. Lehtisalo, A. Solomon, E. Levälahti, S. Ahtiluoto, R. Antikainen, et al., A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial, *Lancet Neurol.*, **385** (2015), 2255–2263. http://dx.doi.org/10.1016/S0140-6736(15)60461-5

5. M. F. Folstein, S. E. Folstein, P. R. McHugh, "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician, *J. Psychiat. Res.*, **12** (1975), 189–198.
6. Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, et al., The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment, *J. Am. Geriatr. Soc.*, **53** (2005), 695–699. http://dx.doi.org/10.1111/j.1532-5415.2005.53221.x
7. A. Heyman, G. Fillenbaum, F. Nash, Consortium to establish a registry for Alzheimer's disease: the CERAD experience, *Neurology*, **49** (1997), 1–26.
8. A. Konig, A. Satt, A. Sorin, R. Hoory, A. Derreumaux, R. David, et al., Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people, *Curr. Alzheimer Res.*, **15** (2018), 120–129. http://dx.doi.org/10.2174/1567205014666170829111942
9. A. Roshanzamir, H. Aghajan, S. M. Soleymani, Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech, *BMC Med. Inform. Decis. Mak.*, **21** (2021), 92. http://dx.doi.org/10.1186/s12911-021-01456-3
10. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, *Proceedings of the 34th International Conference on Machine Learning*, **70** (2017), 1321–1330.
11. S. de la Fuente Garcia, C. W. Ritchie, S. Luz, Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review, *Journal of Alzheimer's Disease*, **78** (2020), 1547–1574. http://dx.doi.org/10.3233/JAD-200888
12. M. F. McTear, Z. Callejas, D. Griol, *The conversational interface: talking to smart devices*, Cham: Springer, 2016. http://dx.doi.org/10.1007/978-3-319-32967-3
13. G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, *ECCV*, **1** (2004), 1–16.
14. M. Schmitt, F. Ringeval, B. Schuller, At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech, *Proceedings of Interspeech*, 2016, 495–499. http://dx.doi.org/10.21437/Interspeech.2016-1124
15. L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, A. Roche-Bergua, Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task, *Alzh. Dement.-DADM*, **10** (2018), 260–268. http://dx.doi.org/10.1016/j.dadm.2018.02.004
16. S. Luz, Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data, *Proceedings of IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, 45–46. http://dx.doi.org/10.1109/CBMS.2017.41
17. K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, et al., On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature, *Cogn. Comput.*, **7** (2015), 44–55. http://dx.doi.org/10.1007/s12559-013-9229-9
18. F. Haider, S. De La Fuente, S. Luz, An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech, *IEEE J.-STSP*, **14** (2020), 272–281. http://dx.doi.org/10.1109/JSTSP.2019.2955022
19. S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, B. Macwhinney, Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge, *Proceedings of Interspeech*, 2020, 2172–2176. http://dx.doi.org/10.21437/Interspeech.2020-2571
20. F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in openSMILE, the munich open-source multimedia feature extractor, *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, 835–838. http://dx.doi.org/10.1145/2502081.2502224

21. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, et al., The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, *IEEE T. Affect. Comput.*, **7** (2016), 190–202. http://dx.doi.org/10.1109/TAFFC.2015.2457417
22. F. Eyben, M. Wöllmer, B. Schuller, OpenSMILE: the munich versatile and fast open-source audio feature extractor, *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, 1459–1462. http://dx.doi.org/10.1145/1873951.1874246
23. M. S. S. Syed, Z. S. Syed, M. Lech, E. Pirogova, Automated screening for Alzheimer's dementia through spontaneous speech, *Proceedings of Interspeech*, 2020, 2222–2226. http://dx.doi.org/10.21437/Interspeech.2020-3158
24. M. Schmitt, B. Schuller, OpenXBOW–Introducing the passau open-source crossmodal bag-of-words toolkit, *J. Mach. Learn. Res.*, **18** (2017), 1–5.
25. M. E. Celebi, H. A. Kingravi, P. A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Syst. Appl.*, **40** (2013), 200–210. http://dx.doi.org/10.1016/j.eswa.2012.07.021
26. J. Hämäläinen, S. Jauhiainen, T. Kärkkäinen, Comparison of internal clustering validation indices for prototype-based clustering, *Algorithms*, **10** (2017), 105. http://dx.doi.org/10.3390/a10030105
27. M. Niemelä, T. Kärkkäinen, Improving clustering and cluster validation with missing data using distance estimation methods, In: *Computational sciences and artificial intelligence in industry*, Cham: Springer, 2022, 123–133. http://dx.doi.org/10.1007/978-3-030-70787-3_9
28. J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, K. L. McGonigle, The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis, *Arch. Neurol.*, **51** (1994), 585–594. http://dx.doi.org/10.1001/archneur.1994.00540180063015
29. K. Hechmi, T. N. Trong, V. Hautamäki, T. Kinnunen, Voxceleb enrichment for age and gender recognition, *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, 687–693. http://dx.doi.org/10.1109/ASRU51503.2021.9688085
30. European Broadcasting Union, *Loudness normalisation and permitted maximum level of audio signals*, EBU Recommendation, 2023.
31. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. http://dx.doi.org/10.1023/A:1010933404324
32. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.*, **3** (2003), 1157–1182.
33. A. K. Jain, Data clustering: 50 years beyond k-means, *Pattern Recogn. Lett.*, **31** (2010), 651–666. http://dx.doi.org/10.1016/j.patrec.2009.09.011
34. S. Äyrämö, *Knowledge mining using robust clustering*, Jyväskylä: University of Jyväskylä Printing, 2006.
35. S. Äyrämö, T. Kärkkäinen, K. Majava, Robust refinement of initial prototypes for partitioning-based clustering algorithms, In: *Recent advances in stochastic modeling and data analysis*, Chania: World Scientific, 2007, 473–482. http://dx.doi.org/10.1142/9789812709691_0056
36. D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, 2007, 1027–1035.
37. T. Kärkkäinen, S. Äyrämö, On computation of spatial median for robust data mining, *Peoceedings of Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems*, 2005, 1–14.

38. M. Niemelä, S. Äyrämö, T. Kärkkäinen, Toolbox for distance estimation and cluster validation on data with missing values, *IEEE Access*, **10** (2022), 352–367. http://dx.doi.org/10.1109/ACCESS.2021.3136435
39. T. Cover, P. Hart, Nearest neighbor pattern classification, *IEEE T. Informa. Theory*, **13** (1967), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964
40. Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, *Biostatistics*, **8** (2007), 86–100. http://dx.doi.org/10.1093/biostatistics/kxj035
41. T. Kärkkäinen, Extreme minimal learning machine: Ridge regression with distance-based basis, *Neurocomputing*, **342** (2019), 33–48. http://dx.doi.org/10.1016/j.neucom.2018.12.078
42. N. Cristianini, J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge: Cambridge university press, 2000. http://dx.doi.org/10.1017/CBO9780511801389
43. J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, *Int. J. Comput. Vision*, **73** (2007), 213–238. http://dx.doi.org/10.1007/s11263-006-9794-4
44. F. Wilcoxon, Individual comparisons by ranking methods, In: *Breakthroughs in statistics*, New York: Springer, 1992, 196–202. http://dx.doi.org/10.1007/978-1-4612-4380-9_16
45. F. Haider, S. Pollak, P. Albert, S. Luz, Emotion recognition in low-resource settings: an evaluation of automatic feature selection methods, *Comput. Speech Lang.*, **65** (2021), 101119. http://dx.doi.org/10.1016/j.csl.2020.101119
46. P. Fränti, Efficiency of random swap clustering, *J. Big Data*, **5** (2018), 13. http://dx.doi.org/10.1186/s40537-018-0122-y
47. T. F. Yap, J. Epps, E. Ambikairajah, E. H. C. Choi, Formant frequencies under cognitive load: effects and classification, *EURASIP J. Adv. Signal Process.*, **2021** (2011), 219253. http://dx.doi.org/10.1155/2011/219253
48. T. F. Yap, J. Epps, E. Ambikairajah, E. H. C. Choi, Voice source features for cognitive load classification, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, 5700–5703. http://dx.doi.org/10.1109/ICASSP.2011.5947654
49. S. B. Scott, J. E. Graham-Engeland, C. G. Engeland, J. M. Smyth, D. M. Almeida, M. J. Katz, et al., The effects of stress on cognitive aging, physiology and emotion (ESCAPE) project, *BMC Psychiatry*, **15** (2015), 146. http://dx.doi.org/10.1186/s12888-015-0497-7
50. D. V. L. Sidtis, W. Hanson, C. Jackson, A. Lanto, D. Kempler, E. J. Metter, Fundamental frequency (f0) measures comparing speech tasks in aphasia and Parkinson disease, *J. Med. Speech-Lang. Pa.*, **12** (2004), 207–213.
51. M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, *Nat. Prec.*, 2008, 1–27. http://dx.doi.org/10.1038/npre.2008.2298.1
52. R. Alshammri, G. Alharbi, E. Alharbi, I. Almubark, Machine learning approaches to identify Parkinson's disease using voice signal features, *Front. Artif. Intell.*, **6** (2023), 1084001. http://dx.doi.org/10.3389/frai.2023.1084001
53. D. Nickson, C. Meyer, L. Walasek, C. Toro, Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review, *BMC Med. Inform. Decis. Mak.*, **23** (2023), 271. http://dx.doi.org/10.1186/s12911-023-02341-x

Appendix A. Average importance scores for eGeMAPS features and relative speech pauses

Table A.1. Average importance scores for the eGeMAPS feature set and relative speech pauses based on the Gini index (the most important 25 features are selected for the study).

Rank	Feature	Importance (%)	Category (appearance)
1	mfcc3V_sma3nz_amean	1.95	(Voiced) MFCC (1/16)
2	logRelF0-H1-A3_sma3nz_amean	1.83	Harmonic ratio (1/4)
3	mfcc3_sma3_amean	1.77	MFCC (2/16)
4	slopeV0-500_sma3nz_amean	1.74	(Voiced) Spectral slope (1/6)
5	F0semitoneFrom27.5Hz_sma3nz_percentile50.0	1.73	F0 fundamental frequency (1/11)
6	F0semitoneFrom27.5Hz_sma3nz_percentile20.0	1.65	F0 fundamental frequency (2/11)
7	loudness_sma3_amean	1.59	F0 fundamental frequency (3/11)
8	slopeUV0-500_sma3nz_amean	1.56	(Unvoiced) Spectral slope (2/6)
9	F3amplitudeLogRelF0_sma3nz_stddevNorm	1.55	Formant energy (1/6)
10	loudness_sma3_percentile20.0	1.5	Loudness (1/10)
11	F0semitoneFrom27.5Hz_sma3nz_percentile80.0	1.49	F0 fundamental frequency (4/11)
12	F0semitoneFrom27.5Hz_sma3nz_amean	1.48	F0 fundamental frequency (5/11)
13	mfcc4_sma3_amean	1.4	MFCC (3/16)
14	logRelF0-H1-H2_sma3nz_amean	1.4	Harmonic ratio (2/4)
15	mfcc2_sma3_amean	1.35	MFCC (4/16)
16	mfcc4V_sma3nz_amean	1.35	(Voiced) MFCC (5/16)
17	F3bandwidth_sma3nz_amean	1.34	Formant bandwidth (1/6)
18	slopeV0-500_sma3nz_stddevNorm	1.33	(Voiced) Spectral slope (3/6)
19	spectralFluxUV_sma3nz_amean	1.32	(Unvoiced) Spectral flux (1/5)
20	pauseDurationRatio	1.28	Relative speech pauses (1/2)
21	mfcc1V_sma3nz_amean	1.24	(Voiced) MFCC (6/16)
22	slopeV500-1500_sma3nz_amean	1.24	(Voiced) Spectral slope (4/6)
23	loudness_sma3_percentile50.0	1.21	Loudness (2/10)
24	pauseTotalPausesRatio	1.19	Relative speech pauses (2/2)
25	mfcc2V_sma3nz_amean	1.19	(Voiced) MFCC (7/16)
26	slopeUV500-1500_sma3nz_amean	1.18	(Unvoiced) Spectral slope (5/6)
27	mfcc2V_sma3nz_stddevNorm	1.18	MFCC (8/16)
28	loudness_sma3_pctlrange0-2	1.16	Loudness (3/10)
29	slopeV500-1500_sma3nz_stddevNorm	1.16	(Voiced) Spectral slope (6/6)
30	loudness_sma3_stddevNorm	1.15	Loudness (4/10)
31	HNRdBACF_sma3nz_amean	1.14	Harmonic noise ratio (1/2)
32	mfcc1_sma3_amean	1.14	MFCC (9/16)
33	mfcc3_sma3_stddevNorm	1.13	MFCC (10/16)
34	loudness_sma3_percentile80.0	1.12	Loudness (5/10)
35	hammarbergIndexV_sma3nz_amean	1.12	(Voiced) Hammerberg index (1/3)
36	shimmerLocaldB_sma3nz_amean	1.12	Shimmer (1/2)
37	alphaRatioUV_sma3nz_amean	1.11	Alpha ratio (1/3)
38	hammarbergIndexUV_sma3nz_amean	1.11	(Unvoiced) Hammerberg index (2/3)
39	logRelF0-H1-H2_sma3nz_stddevNorm	1.1	Harmonic ratio (3/4)
40	mfcc4_sma3_stddevNorm	1.09	MFCC (11/16)
41	F3bandwidth_sma3nz_stddevNorm	1.09	Formant bandwidth (2/6)
42	shimmerLocaldB_sma3nz_stddevNorm	1.07	Shimmer (2/2)
43	spectralFlux_sma3_amean	1.06	Spectral flux (2/5)
44	mfcc4V_sma3nz_stddevNorm	1.05	(Voiced) MFCC (12/16)
45	mfcc1_sma3_stddevNorm	1.05	MFCC (13/16)

Table A.2. Average importance scores for the eGeMAPS feature set and relative speech pauses based on the Gini index (continued).

Rank	Feature	Importance (%)	Category (appearance)
46	mfcc3V_sma3nz_stddevNorm	1.05	(Voiced) MFCC (14/16)
47	equivalentSoundLevel_dBp	1.05	Sound level (1/1)
48	F2bandwidth_sma3nz_amean	1.04	Formant bandwidth (3/6)
49	F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	1.03	F0 fundamental frequency (6/11)
50	F3frequency_sma3nz_amean	1.01	Formant frequency (1/4)
51	F2amplitudeLogRelF0_sma3nz_stddevNorm	1.01	Formant energy (2/6)
52	mfcc1V_sma3nz_stddevNorm	1.00	(Voiced) MFCC (15/16)
53	mfcc2_sma3_stddevNorm	1.00	MFCC (16/16)
54	spectralFlux_sma3_stddevNorm	0.99	Spectral flux (3/5)
55	F1frequency_sma3nz_amean	0.98	Formant frequency (2/4)
56	F1frequency_sma3nz_stddevNorm	0.98	Formant frequency (3/4)
57	HNRdBACF_sma3nz_stddevNorm	0.98	Harmonic noise ratio (2/2)
58	logRelF0-H1-A3_sma3nz_stddevNorm	0.97	Harmonic ratio (4/4)
59	jitterLocal_sma3nz_amean	0.97	Jitter (1/2)
60	F0semitoneFrom27.5Hz_sma3nz_stddevNorm	0.97	F0 fundamental frequency (7/11)
61	F2frequency_sma3nz_amean	0.97	Formant frequency (4/4)
62	jitterLocal_sma3nz_stddevNorm	0.95	Jitter (2/2)
63	loudness_sma3_meanFallingSlope	0.95	Loudness (6/10)
64	F1bandwidth_sma3nz_stddevNorm	0.94	Formant bandwidth (4/6)
65	loudness_sma3_stddevRisingSlope	0.94	Loudness (7/10)
66	MeanVoicedSegmentLengthSec	0.93	Voiced segments (1/3)
67	loudness_sma3_meanRisingSlope	0.93	Loudness (8/10)
68	F1amplitudeLogRelF0_sma3nz_stddevNorm	0.92	Formant energy (3/6)
69	spectralFluxV_sma3nz_amean	0.92	(Voiced) Spectral flux (4/5)
70	alphaRatioV_sma3nz_amean	0.91	(Voiced) Alpha ratio (2/3)
71	F3frequency_sma3nz_stddevNorm	0.91	Formant frequency (1/2)
72	F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	0.91	F0 fundamental frequency (8/11)
73	F1bandwidth_sma3nz_amean	0.91	Formant bandwidth (5/6)
74	F2frequency_sma3nz_stddevNorm	0.91	Formant frequency (2/2)
75	F2bandwidth_sma3nz_stddevNorm	0.9	Formant bandwidth (6/6)
76	hammarbergIndexV_sma3nz_stddevNorm	0.9	Hammerberg index (3/3)
77	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	0.89	F0 fundamental frequency (9/11)
78	alphaRatioV_sma3nz_stddevNorm	0.89	(Voiced) Alpha ratio (3/3)
79	spectralFluxV_sma3nz_stddevNorm	0.88	(Voiced) Spectral flux (5/5)
80	loudnessPeaksPerSec	0.87	Loudness (9/10)
81	VoicedSegmentsPerSec	0.85	Voiced segments (2/3)
82	F3amplitudeLogRelF0_sma3nz_amean	0.83	Formant energy (4/6)
83	MeanUnvoicedSegmentLength	0.83	Unvoiced segments
84	loudness_sma3_stddevFallingSlope	0.82	Loudness (10/10)
85	F2amplitudeLogRelF0_sma3nz_amean	0.8	Formant energy (5/6)
86	F1amplitudeLogRelF0_sma3nz_amean	0.76	Formant energy (6/6)
87	StddevVoicedSegmentLengthSec	0.72	Voiced segments (3/3)
88	F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	0.67	F0 fundamental frequency (10/11)
89	StddevUnvoicedSegmentLength	0.67	Unvoiced segments (1/1)
90	F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	0.65	F0 fundamental frequency (11/11)

Appendix B. LOSO cross-validation results for the ADReSS 2020 audio dataset

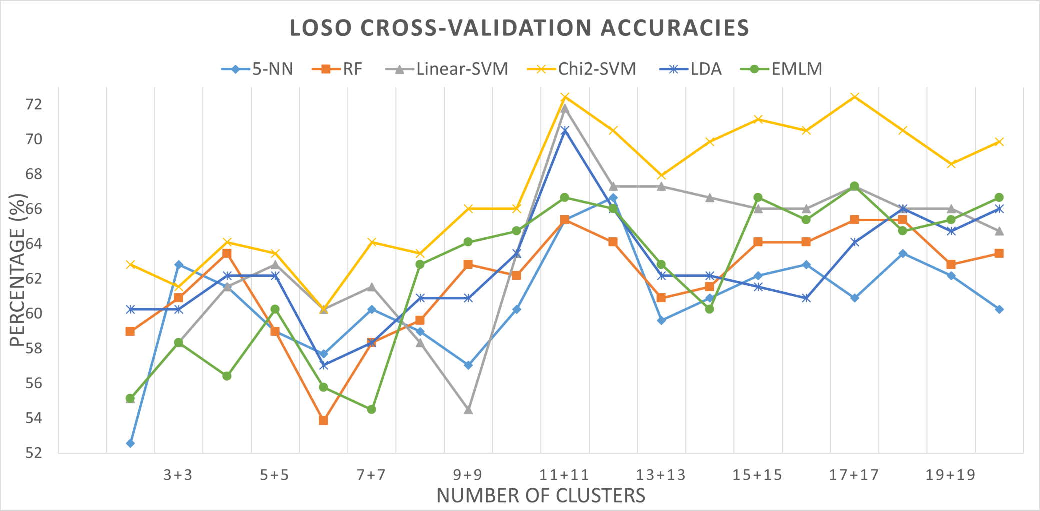

Figure B.1. Mode-based LOSO cross-validation accuracies over 25 repetitions of replicated clustering.

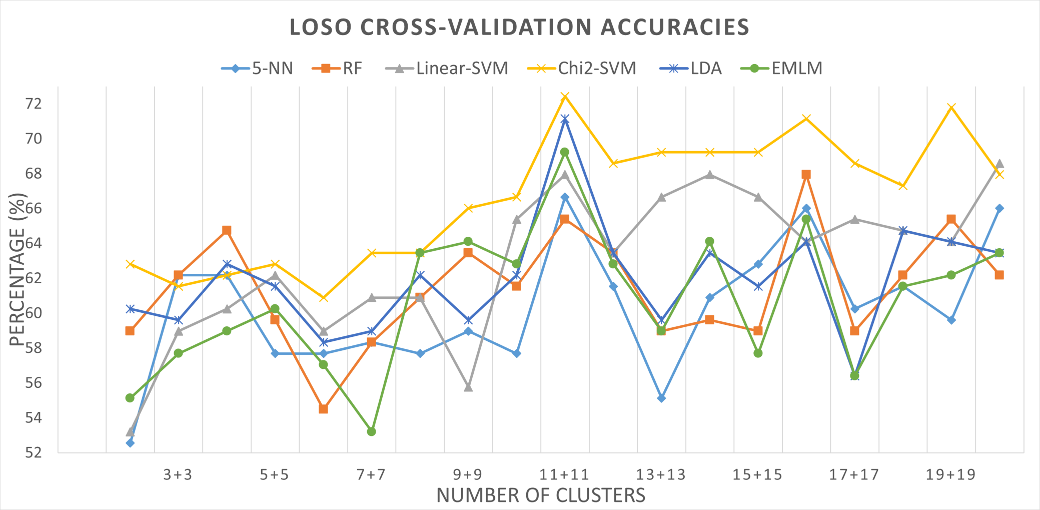

Figure B.2. Minimum clustering error-based LOSO cross-validation accuracies over 25 repetitions of replicated clustering.