Detecting Cognitive Disease from Speech Using Support Vector Machines with Wav2Vec and MFCC Acoustic Features

Bao Hoang, Yijiang Pang, Hiroko Dodge, Jiayu Zhou
ILLIDAN Lab, University of Michigan

January 22, 2025

Abstract

Early detection of cognitive diseases is pivotal in enabling timely and effective treatment interventions. Although brain imaging techniques have achieved remarkable accuracy in disease identification, their widespread adoption is hindered by the complexity of the technology and the associated high costs, rendering them unsuitable for large-scale adult screening. The PREPARE Challenge seeks to address this limitation by developing methodologies that utilize acoustic biomarkers extracted from speech samples to identify cognitive traits. In this work, we propose a framework that integrates advanced acoustic feature extraction techniques, including Wav2Vec and Mel-frequency cepstral coefficients (MFCC), with robust outlier detection strategies to mitigate the impact of noise. These features are subsequently employed in conjunction with Support Vector Machines (SVM), achieving a 10th-place ranking in the challenge. The implementation details and source code are available at https://github.com/illidanlab/PREPARE-Challenge-Acoustic-Track-Code.

1 Introduction

Alzheimer's disease is currently the seventh leading cause of death in the United States and the most common cause of dementia among older adults [KMXA23]. Early detection of Mild Cognitive Impairment (MCI) and Alzheimer's Disease and Related Dementias (ADRD) is essential for effective treatment and intervention. Advanced techniques like brain imaging, including MRI and PET, combined with machine learning algorithms, have demonstrated high efficacy in cognitive disease detection [WGT+18, HPL+24, ZLZ+15]. However, the complexity and cost of these technologies limit their accessibility for large-scale screening, particularly in underserved regions with limited healthcare infrastructure. Digital biomarkers, such as linguistic and acoustic features derived from spontaneous speech, present a promising alternative for cost-effective and scalable Alzheimer's disease detection [HPDZ24, TCDZ22]. The PREPARE Challenge (Pioneering Research for Early Prediction of Alzheimer's and Related Dementias EUREKA Challenge) aimed to identify innovative, advanced, and effective methods for classifying individuals into normal, MCI, and ADRD categories using speech audio data. For this competition, we utilized the pretrained Wav2Vec model to extract audio representation embeddings, complemented by traditional Mel-Frequency Cepstral Coefficients (MFCC) features. We applied robust outlier and noise removal strategies to enhance data quality before using a Support Vector Machine (SVM) for classification. Our method achieved an 10th-place ranking on the PREPARE Challenge leaderboard.

2 Data

The PREPARE Challenge dataset comprises raw audio recordings, each approximately 30 seconds in duration, alongside metadata containing basic demographic attributes such as age and gender. The dataset is annotated with three cognitive class labels: Normal, Mild Cognitive Impairment (MCI), and Alzheimer's Disease and Related Dementias (ADRD). It includes 1,646 audio samples for training and 412 samples for testing. This section provides a detailed analysis of the key statistical characteristics of the dataset.

2.1 Demographics

Tables 1 summarizes the demographic distributions in the training and test datasets. The distributions of sample sizes, gender, and age across the three cognitive labels (Normal, MCI, ADRD) are consistent between the training and test datasets.

Dataset	Cognitive Status	Control	MCI	ADRD
Training	Number of samples	911	217	518
	Gender (% female)	59.276	56.682	57.336
	Age	74.922±8.291	72.498±10.166	76.191±8.477
Test	Number of samples	229	51	132
	Gender (% female)	65.502	56.863	60.606
	Age	75.227±8.512	75.902±9.810	76.280±7.853

Table 1: Demographics of PREPARE Challenge Training&Test Datasets

2.2 Audio Length

Table 2 illustrates the distribution of audio lengths (in seconds) in the training dataset. As shown, the majority of audio files have lengths ranging from 20 to 30 seconds, with 1439 training audio files and 356 testing audio files falling within this range. Notably, no test samples are shorter than 10 seconds. To ensure alignment between the training and test datasets, audio files shorter than 10 seconds in the training set are excluded from further analysis.

Length (seconds)	Number of Train Audio Files	Number of Test Audio Files
0 – 9	18	0
10 – 19	189	56
20 – 30	1439	356

Table 2: Distribution of Audio Length

2.3 Language

The dataset includes audio recordings from speakers of multiple languages such as English, Spanish, Galician, and Chinese. The language distributions across the three cognitive labels (Normal, MCI, ADRD) in the training and test datasets are shown in Tables 3 and 4, respectively. English speakers constitute the majority in both datasets, followed by Spanish, Chinese, and Galician speakers. Notably, all Chinese speakers in the training dataset are labeled as MCI.

Language	Control	MCI	ADRD
English	760	115	456
Spanish	140	61	54
Chinese	0	34	0
Galician	11	7	8

Table 3: Training Dataset Language Distribution

Language	Number of Audio Files
English	324
Spanish	76
Chinese	9
Galician	3

Table 4: Test Dataset Language Dsitribution

Language Detection Model The Whisper model, developed by OpenAI [RKX+22], is a state-of-the-art speech recognition system trained on a vast dataset of audio recordings. Whisper excels at transcribing audio into text and can also detect the language spoken in the audio, which we leverage for the language distribution analysis in this Subsection 2.3. For our implementation, we utilized the pretrained weights of the Whisper large version model, consisting of 1550 million parameters, whose detailed instructions for using the model are available at https://github.com/openai/whisper.

3 Methods

The methodology pipeline employed in this study is designed to effectively process audio data for cognitive classification. It begins with raw audio files, which serve as the input data. These audio recordings are subjected to acoustic feature extraction, where techniques such as Wav2Vec embeddings and Mel-frequency cepstral coefficients (MFCC) are utilized to capture critical speech characteristics. Subsequently, a noise filtering step is implemented to identify and remove samples that exhibit excessive noise or inconsistencies, ensuring a high-quality dataset for model training. The processed features are then fed into a classification model, where a Support Vector Machine (SVM) is employed to learn the cognitive class predictions based on the extracted acoustic patterns. To further enhance the robustness and reliability of the predictions, a logit-smoothing procedure is applied to further mitigate the influence of noise or outliers. Our methodology pipeline is illustrated by Figure 1.

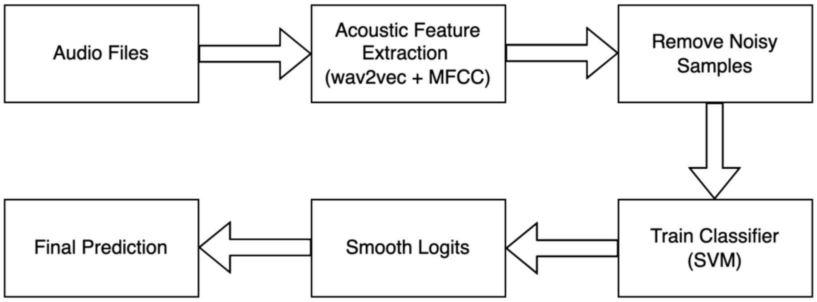

Figure 1: Pipeline of proposed method

3.1 Acoustic Features

The Wav2Vec model [BZMA20] is a self-supervised learning framework designed for deep audio modeling using unlabeled training data. It employs a masked learning strategy similar to that of the BERT language model [DCLT19], enabling the model to learn contextualized audio representations. After pretraining, the model can be fine-tuned on labeled datasets for various downstream tasks, including speech recognition, speaker identification, and emotion classification. For our experiments, we utilized the pretrained weights of the Wav2Vec model from MetaAI, available on Hugging Face at https://huggingface.co/facebook/wav2vec2-large-lv60.

Given a Wav2Vec model f and a raw audio input X, the model generates an embedding $f(X)$ with a shape of $(1024, T)$, where T represents the number of time steps, determined by the length of the audio and the encoder stride. Since different audio inputs X_1 and X_2 generally result in embeddings with different numbers of time steps T_1 and T_2, we standardize the feature dimensions by averaging along the time-step dimension, which produces a fixed-size feature vector of shape 1024.

We also extracted Mel-Frequency Cepstral Coefficients (MFCCs) following the methodology described in [MPSN01]. This process involved computing the first 13 MFCC bands (indices 0–12), along with their corresponding 13 delta coefficients and 13 delta-delta coefficients, which capture the rate of change and acceleration of the MFCCs, respectively. To summarize these features, we applied six descriptive statistical functions—minimum, maximum, mean, standard deviation, skewness, and kurtosis—resulting in a 234-dimensional feature vector for each speech sample. We used the librosa library [BMCRDL+15] to extract the MFCC features.

To summarize, after obtaining the Wav2Vec embedding of shape 1024 and the MFCC features of shape 234, we concatenate them, resulting in final 1258-dimensional acoustic features for each audio file.

3.2 Classifiers

We utilized a wide range of models from the Scikit-learn library [PVG+11], including Gradient Boosting, Random Forest, Decision Tree, Multi-Layer Perceptron (MLP), Logistic Regression, and Support Vector Machine (SVM). To optimize model performance, we employed the Grid Search method to identify the hyperparameters that achieved the highest 5-fold cross-validation accuracy. The optimal hyperparameters for each model, as determined through 5-fold cross-validation, are as follows:

- Decision Tree: `criterion = 'entropy', max_depth = 5, max_features = 'sqrt', min_samples_leaf = 1, min_samples_split = 2`
- Random Forest: `max_depth = 10, min_samples_leaf = 4, min_samples_split = 10, n_estimators = 100`
- MLP: `hidden_layer_sizes = (256, 64)`
- Logistic Regression: `C = 0.01, l1_ratio = 0.1, max_iter = 200, penalty = 'elasticnet', solver = 'saga'`
- Support Vector Machine: `C = 1, degree = 3, gamma = 'scale', kernel = 'rbf'`

Among the evaluated models, the Support Vector Machine achieved the best performance on the competition's test dataset with the optimal hyperparameters: `C = 1, degree = 3, gamma = 'scale', kernel = 'rbf'`.

3.3 Outlier Detection

To enhance the quality of the training data and improve the performance of our models, we identified and removed four categories of noisy or uninformative training samples:

1. **Uninformative Audio Files.** To ensure alignment between the training and test datasets, we excluded audio files from the training set that were shorter than 10 seconds, as described in Section 2.2.
2. **Minority Language.** According to our investigation of the language distribution, specifically, Table 3, we excluded Chinese samples from the training dataset (and directly assigned the MCI label to the nine Chinese samples in the test dataset). Additionally, since the test dataset contains only three Galician samples, we also excluded Galician samples from the training dataset to maintain consistency.
3. **Alexa Speech.** We observed that some audio samples in the competition dataset consisted solely of conversations with Alexa, and these conversations include phrases such as, ”Alexa, what time is it?”, ”Alexa, when is Thanksgiving?”, or ”Alexa, how do you bake chocolate chip cookies?”. With practical validation, we believe these data samples lack cognitive markers and could negatively impact the classifier's ability to learn meaningful cognitive characteristics. In total, using the Whisper model, we identified 85 Alexa-related samples in the training set and 14 in the test set from audio's transcripts. To address this, we excluded these samples from the training dataset.
4. **One-Class SVM Outlier Samples.** The one-class classification (OCC) problem [POP21] focuses on learning a representation that identifies subjects belonging to a specific class among all objects, where the trained model can effectively recognize positively labeled queries during inference. One-class classification is particularly useful for anomaly detection, where it identifies unusual samples that are different from the training set distribution. We implemented a One-Class SVM model [NHR12] using the Scikit-learn library [PVG+11] to filter out outliers from the training data. Using this approach, we excluded 147 audio samples as outliers by leveraging Wav2Vec and MFCC features for feature extraction.

3.4 Smooth Logits prediction

Since the competition's official evaluation metric is multiclass Log Loss, defined as equation (1), which allows us to submit predicted logits for all three labels rather than a single label prediction, we introduced an auxiliary operation to refine the logits for samples with highly confident labels and lowly confident labels.

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_{nm} \log p_{nm} \tag{1}$$

where p_{nm} denotes the predicted logit, while y_{nm} represents the ground truth label for sample n in class m.

If the predicted probability (logit) for any label exceeds 0.85, we assign a value of 1 to that label and set the others to 0. For example, if the model predicts logits for a sample as [0.05, 0.9, 0.05], we modify it to [0, 1, 0], reflecting the model's high confidence that the label for this sample is MCI.

Additionally, for samples with low-probability predictions, we smooth the logits by redistributing the probabilities. If the predicted probability for any label is less than 0.05, we set that label's value to 0 and reassign its probability to the most confident label of the sample. For instance, if the model predicts logits for a sample as [0.04, 0.4, 0.56], we adjust it to [0, 0.4, 0.6].

4 Experimental Results

This section presents an evaluation of the performance of various classifiers, acoustic feature combinations, and noisy sample removal strategies.

4.1 Classifiers

We evaluate the performance of several classifiers, including Decision Tree, MLP, Logistic Regression, Random Forest, and SVM. Table 5 summarizes the 5-fold cross-validation accuracy and test loss achieved by each scikit-learn classifier, optimized using the best hyperparameters specified in subsection 3.2. Among these, the Support Vector Machine (SVM) consistently outperforms the others on both metrics (validation accuracy and test loss), demonstrating its effectiveness in leveraging Wav2Vec + MFCC features for robust classification.

Classifier	Cross-Validation Accuracy (5-Fold)	Test Loss (Multiclass)
Decision Tree	0.605±0.011	1.359
MLP	0.640±0.020	1.584
Logistic Regression	0.655±0.027	0.731
Random Forest	0.662±0.015	0.723
Support Vector Machine	**0.664±0.016**	**0.689**

Table 5: Cross Validation Accuracy and Test Loss of Scikit-learn Classifiers

4.2 Acoustic Features

We evaluate the performance of various acoustic feature types, including Wav2Vec, MFCC, and eGeMAPS, as well as their combinations. The eGeMAPS features are pre-extracted and provided by the PREPARE Challenge organizers, following the procedure outlined in [ESS+16]. Table 7 summarizes the results, presenting the 5-fold cross-validation accuracy and test loss for different combinations of acoustic features, with Support Vector Machines (SVM) as the classifier. From the results, we observe that although Wav2Vec achieves slightly higher accuracy than Wav2Vec + MFCC in the internal validation split (0.667 compared to 0.664), when concatenated with MFCC, Wav2Vec + MFCC outperforms other feature sets in test loss. This highlights the effectiveness of integrating complementary acoustic features to improve the generalization of classification performance. However, the results also highlight the importance of careful feature selection. Naively combining all available features, such as Wav2Vec + MFCC + eGeMAPS, leads to higher test loss compared to Wav2Vec +

MFCC alone, suggesting that excessive feature concatenation can introduce redundancy or noise that negatively impacts model performance.

Features	Cross-Validation Accuracy (5-Fold)	Test Loss (Multiclass)
eGeMAPS	0.590±0.010	1.082
MFCC	0.629±0.019	0.793
Wav2Vec	0.667±0.014	0.712
eGeMAPS + MFCC	0.627±0.019	0.793
Wav2Vec + eGeMAPS	0.662±0.011	0.745
Wav2Vec + MFCC	0.664±0.016	**0.689**
Wav2Vec + MFCC + eGeMAPS	**0.676±0.034**	0.734

Table 6: Cross Validation Accuracy and Test Loss of Different Acoustic Feature Combination

4.3 Noisy Sample Effects

We evaluated the impact of noisy sample removal on the performance of our proposed method by removing different types of noise individually and collectively. As described earlier, the four categories of noise considered are: (1) Uninformative audio files, (2) Minority Language, (3) Alexa samples where participants interact with Alexa, and (4) outlier samples detected by a One-Class SVM. Table 7 summarizes the results, presenting the test loss for different noise removal strategies using an SVM classifier with Wav2Vec and MFCC as acoustic features. The results indicate that removing any individual category of noise improves test loss compared to using the full dataset. Furthermore, removing all noisy samples leads to the best test loss of 0.689, demonstrating a significant reduction in cross-entropy loss. This highlights the importance of carefully identifying and eliminating noisy samples in the training dataset to enhance classifier performance.

Noise Type	Test Loss (Multiclass)
-	0.742
Uninformative audio files	0.722
Minority Language	0.741
Alexa Samples	0.734
One-class SVM Outlier Samples	0.726
All	**0.689**

Table 7: Test Loss of Different Noisy Removal Strategies Combination

5 Conclusion

The PREPARE Challenge provides an unique opportunity for participants to advance methodologies for detecting cognitive diseases through speech. With advancements in deep learning technologies for audio processing, these methods have the potential to play a pivotal role in the early detection of Alzheimer's disease, enabling timely and effective treatment. However, to ensure the effectiveness of models in high-risk domains like healthcare, a thorough analysis of training data is essential to identify and address distributional differences and other data quality issues. Our team implemented a robust outlier detection algorithm alongside careful data analysis, focusing on audio length, speech content, and the language of speakers, to eliminate noisy training samples. We then leveraged powerful pre-trained acoustic models, such as Wav2Vec, and extracted effective acoustic features like MFCCs, combining them with SVM classifiers to predict cognitive conditions. Our approach ranked 10th on the leaderboard, demonstrating the effectiveness of our methodology.

References

[BMCRDL$^+$15] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python, 2015.

[BZMA20] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[ESS$^+$16] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[HPDZ24] Bao Hoang, Yijiang Pang, Hiroko H Dodge, and Jiayu Zhou. Subject harmonization of digital biomarkers: Improved detection of mild cognitive impairment from language markers. *Pac. Symp. Biocomput.*, 29:187–200, 2024.

[HPL$^+$24] Bao Hoang, Yijiang Pang, Siqi Liang, Liang Zhan, Paul M. Thompson, and Jiayu Zhou. Distributed harmonization: Federated clustered batch effect adjustment and generalization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5105–5115, New York, NY, USA, 2024. Association for Computing Machinery.

[KMXA23] Kenneth Kochanek, Sherry L. Murphy, Jiaquan Xu, and Elizabeth Arias. Mortality in the united states, 2022, December 2023.

[MPSN01] S. Molau, M. Pitz, R. Schluter, and H. Ney. Computing mel-frequency cepstral coefficients on the power spectrum, 2001.

[NHR12] Zineb Noumir, Paul Honeine, and Cédue Richard. On simple one-class classification methods, 2012.

[POP21] Pramuditha Perera, Poojan Oza, and Vishal M. Patel. One-class classification: A survey, 2021.

[PVG$^+$11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RKX$^+$22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[TCDZ22] Fengyi Tang, Jun Chen, Hiroko H. Dodge, and Jiayu Zhou. The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment. *Frontiers in Digital Health*, 3, February 2022.

[WGT$^+$18] Qi Wang, Lei Guo, Paul M. Thompson, Jr. Clifford R. Jack, Hiroko Dodge, Liang Zhan, Jiayu Zhou, for the Alzheimer's Disease Neuroimaging Initiative, National Alzheimer's Coordinating Center, and Yudong Zhang. The added value of diffusion-weighted mri-derived structural connectome in evaluating mild cognitive impairment: A multi-cohort validation. *Journal of Alzheimer's Disease*, 64(1):149–169, 2018. PMID: 29865049.

[ZLZ$^+$15] L Zhan, Y Liu, J Zhou, J Ye, and P M Thompson. Boosting classification accuracy of diffusion MRI derived brain networks for the subtypes of mild cognitive impairment using higher order singular value decomposition. *Proc. IEEE Int. Symp. Biomed. Imaging*, 2015:131–135, April 2015.