Crosslinguistic Acoustic Feature-based Dementia Classification using Advanced Learning Architectures

Anna Seo Gyeong Choi[1], Jin-seo Kim[2], Seo-hee Kim[3], Min Seok Back[3], Sunghye Cho[4]

[1] Department of Information Science, Cornell University, Ithaca, NY
[2] School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA
[3] Yonsei Wonju Severance Christian Hospital, Korea
[4] Linguistic Data Consortium, Department of Linguistics, University of Pennsylvania, Philadelphia, PA
sc2359@cornell.edu, jins0904@sas.upenn.edu, fklir@naver.com,
minbaek@yonsei.ac.kr, csunghye@ldc.upenn.edu

Abstract

In this study, we rigorously evaluated eight machine learning and deep learning classifiers for identifying Alzheimer's Disease (AD) patients using crosslinguistic acoustic features automatically extracted from one-minute oral picture descriptions produced by speakers of American English, Korean, and Mandarin Chinese. We employed eGeMAPSv2 and ComParE feature sets on segmented and non-segmented audio data. The Multilayer Perceptron model showed the highest performance, achieving an accuracy of 83.54% and an AUC of 0.8 on the ComParE features extracted from non-segmented picture description data. Our findings suggest that classifiers trained with acoustic features extracted from one-minute picture description data in multiple languages are highly promising as a quick, language-universal, large-scale, remote screening tool for AD. However, the dataset included predominantly English-speaking participants, indicating the need for more balanced multilingual datasets in future research.

Keywords: Alzheimer's Disease, Crosslinguistic approach, Machine learning classification

1. Introduction

Alzheimer's disease (AD) is the most common type of neurodegenerative disease among individuals over 65, affecting 6.7 millions Americans and 50 million people worldwide (Alzheimer's Association, 2023). A recent clinical trial (Sims et al., 2023) of amyloid immunotherapies has showed that patients at an early stage of the disease gained more benefits from the treatment, highlighting the importance of early screening of patients or individuals at risk. However, most diagnostic tools of AD require specialized expertise and equipment and are expensive and/or invasive, making it challenging to implement the tools at scale within diverse communities.

The quest for a cost-effective and scalable early screening tool of AD has led to the rise of speech-based "digital biomarkers" (Hajjar et al., 2023; Robin et al., 2021; Laguarta and Subirana, 2021). While automated techniques to detect cognitive decline using speech have gained much attention among experts in clinical neurology, signal processing, and machine learning, many prior studies have focused on English-speaking patients. This limited scope has resulted in a lack of crosslinguistic and cross-cultural validity and feasibility, and thus health equity. Recently, there has been more attempts to tackle multilingual AD detection, such as a recent Signal Processing Grand Challenge (Luz et al., 2023). This challenge accentuated a critical societal and medical concern, opening research potential for robust, crosslinguistic AD detection. In line with these recent efforts, we trained machine learning classifiers with crosslinguistic datasets to distinguish AD patients from healthy controls (HC). In this study, we only employed acoustic features for training, because acoustic features relied on acoustic signal of speech and could be uniformly extracted across languages. There has been past literature attempting to create classifiers using various linguistic and speech features (Li et al., 2021; Vigo et al., 2022; He et al., 2023), but research only using acoustic features is scarce. We included three languages in the experiment: English, Korean, and Mandarin Chinese. These languages differ in various ways, from writing systems to morphology and syntax to prosody. This extensive linguistic spectrum not only augments the comprehensiveness of our investigation but also ensures the broad utility and applicability of our approach. Also, by employing both conventional and deep-learning machine learning models, we aimed to conduct a comprehensive study for crosslinguistic AD prediction.

2. Methods

2.1. Data Acquisition and Feature Extraction

We employed speech datasets of English and Mandarin from DementiaBank (Lanzi et al., 2023). The English dataset was drawn from the Pitt Corpus (Becker et al., 1994) and the Mandarin dataset

Proceedings of the RaPID-5 @LREC-COLING 2024, pages 95–100

was derived from the Lu Corpus (MacWhinney et al., 2011), both being picture description data. We directly imported the patient grouping of the Pitt corpus from the metadata file that the authors provided, and we followed Li (2019) to determine participants' diagnostic groups in the Lu corpus. Additionally, we incorporated a Korean picture description dataset that our team has collected and fully transcribed. Participants' diagnostic groups in the Korean dataset were determined by an expert clinical neurologist based on published criteria (McKhann et al., 2011). The prosodic systems of these three languages greatly differ in that English has a lexical-stress-based system, whereas Mandarin Chinese is a tone language and Korean is intonational. Therefore, the inclusion of these three languages with diverse phonetic and prosodic characteristics maximizes the crosslinguistic aspect of our study. Since there were not many patients with Mild Cognitive Impairment (MCI) (English=20, Chinese=0, Korean=16), we grouped all patients (either with MCI or AD) as "patients". In terms of participant counts, the datasets include 99 HCs and 192 patients for English, 15 HCs and 33 patients for Mandarin, and 20 HCs and 26 patients for Korean.

For all datasets, we segmented the audio files into utterances based on the timestamps in the transcripts. We excluded interviewers' utterances from the analysis, using the timestamps in the transcripts. We extracted low-level descriptors from segmented and non-segmented data without interviewers' speech and calculated several statistical derivatives (e.g., mean, standard deviation, minimum, maximum) for training. All audio files were configured to be WAV audio files of 44.1 kHz and 16-bit PCM using ffmpeg (Tomar, 2006).

To extract acoustic features from the audio recordings, we employed openSMILE (Eyben et al., 2010), a widely recognized tool for automatic feature extraction in paralinguistic research. Specifically, we utilized eGeMAPS v2 (extended Geneva Minimalistic Acoustic Parameter Set; Eyben et al., 2015) and ComParE (Computational Paralinguistics Challenge; Schuller et al., 2013) feature sets provided by openSMILE. The eGeMAPS v2 and ComParE feature sets were specifically chosen due to their demonstrated performance in previous studies on pathological speech analysis (Valsaraj et al., 2021; Xue et al., 2019; Vats et al., 2021). These feature sets included various acoustic features such as pitch, intensity, voice quality, articulation, and other spectral features, which were essential in distinguishing patients' vocal patterns in our multilingual datasets.

We standardized extracted features using StandardScaler from scikit-learn. Dimensionality was further reduced using Principal Component Analysis (PCA), retaining components that explained 95% of the variance in the data to maintain a balance between data simplification and the retention of crucial information for better performance. Participants speaking different languages were equally distributed to train and test sets to prevent any learning biases.

2.2. Traditional Machine Learning Classifiers

We evaluated the performance of several traditional machine learning classifiers, implementing 10-fold stratified cross-validation for all models for accuracy assessment. The selected array of classifiers, including Random Forest, Support Vector Classifier, and Gradient Boosting, are known for their robustness in handling high-dimensional data and their flexibility in hyperparameter tuning. Each classifier was integrated into a pipeline comprising PCA with a 0.95 variance threshold and the classifier itself. This pipeline was subsequently assessed using 10-fold stratified cross-validation. For each classifier, we computed the mean accuracy and its standard deviation across the 10 folds. Additionally, a grid search was conducted over a range of hyperparameters to identify the optimal parameters that maximized accuracy. The best performance of each classifier was reported after hyperparameter tuning.

2.3. Deep Learning Models

For this study, we employed two distinct deep learning architectures, namely Multi-Layer Perceptrons (MLPs) and Recurrent Neural Networks (RNNs), utilizing the Keras library in Python. Both architectures were tailored to address the heterogeneous nature of acoustic features across the languages under study. The MLP model comprised multiple dense layers and utilized LeakyReLU as the activation function. LeakyReLU was chosen to introduce a small, non-zero gradient for the negative input domain, thereby mitigating the "dying ReLU" problem and allowing the network to learn from the negative input space. Additionally, L2 regularization, BatchNormalization, and AlphaDropout layers were included in the MLP model to ensure generalizability and mitigate overfitting.

In contrast, the RNN model was designed to optimally handle sequences of acoustic features and incorporated L2 regularization, BatchNormalization, and AlphaDropout layers similar to the MLP model. The RNN model employed the sigmoid activation function specifically for the binary classification tasks, facilitating the model's output to be in the range of 0 to 1, thus making it highly interpretable as a probability measure.

We trained multiple instances of each model type

	Acc.	Precision	Recall	F1
LR	56.78	56.42	56.91	56.66
RF	**75.52**	75.76	75.42	75.45
SVC	58.80	59.01	58.70	58.67
GB	66.91	67.06	66.81	66.82
RR	58.07	58.24	57.97	57.99
kNN	59.38	59.44	59.28	59.35
MLP	75.00	74.89	74.93	74.91
RNN	**73.27**	73.21	73.12	73.17

Table 1: Performance metrics in percentage, Non-segmented, eGeMAPSv2. Acc: Accuracy, F1: F1 score.

	Acc.	Precision	Recall	F1
LR	57.22	50.69	57.12	43.14
RF	58.29	56.82	58.19	56.13
SVC	57.46	55.06	57.36	42.41
GB	57.90	55.77	57.80	52.24
RR	57.34	52.07	57.24	43.13
kNN	55.34	54.53	55.24	50.78
MLP	**73.08**	63.19	60.49	61.81
RNN	56.92	56.81	56.78	56.80

Table 2: Performance metrics in percentage, Segmented, eGeMAPSv2. Acc: Accuracy, F1: F1 score.

	Acc.	Precision	Recall	F1
LR	52.71	52.36	52.68	52.42
RF	53.77	52.36	53.67	52.44
SVC	57.81	52.11	57.71	43.61
GB	54.40	53.26	54.30	53.05
RR	52.65	52.29	52.62	52.35
kNN	50.41	48.96	50.31	48.99
MLP	**83.54**	73.68	75.68	74.67
RNN	53.68	53.64	53.62	53.63

Table 3: Performance metrics in percentage, Non-segmented, ComParE. Acc: Accuracy, F1: F1 score.

	Acc.	Precision	Recall	F1
LR	53.98	51.85	53.88	51.70
RF	57.22	55.48	57.12	54.73
SVC	57.50	54.97	57.40	43.63
GB	57.95	55.97	57.85	51.58
RR	57.80	55.69	57.70	46.46
kNN	52.95	52.11	52.85	52.35
MLP	**76.89**	76.54	76.79	76.66
RNN	55.36	55.33	55.32	55.33

Table 4: Performance metrics in percentage, Segmented, ComParE. Acc: Accuracy, F1: F1 score.

independently, and their predictions were subsequently aggregated. The mean of these predictions served as the final prediction for each input sample, thereby enhancing prediction accuracy while diminishing tendencies for overfitting. Further rigor was added to our methodology through the use of stratified 10-fold cross-validation, which ensured the models' robustness and generalizability across unseen, crosslinguistic data. A random hyperparameter search was also conducted to fine-tune each model's parameters, a necessity given the diverse acoustic feature space inherent in crosslinguistic datasets.

3. Results

3.1. Classification results

From the list of multiple machine learning and deep learning classifiers we trained our data on, we report the results from 8 different classifiers: Logistic Regression (LR), Random Forest (RF), Support Vector Classifier (SVC), Gradient Boosting (GB), Ridge Regression (RR), k-Nearest Neighbors (kNN), MLP, and RNN. We report our results on both non-segmented and segmented datasets using eGeMAPS and ComParE feature sets. The comprehensive performance metrics of these classifiers under various configurations are summarized in Tables 1-4.

The MLP classifier trained with non-segmented audio files using the ComParE feature set showed the best performance with an accuracy of 83.54% and an AUC of 0.80 (Table 4). The model correctly identified 190 patients with AD out of 251 and 132 HCs out of 134. Figure 1 shows the Receiver Operating Characteristic (ROC) plot of the best performing model. The optimal threshold for the model is at 0.40, where it attains its best balance of sensitivity and specificity. The specific hyperparameters that we used for this model were a learning rate of 0.1, a dropout rate of 0.6, and a batch size of 32. With these optimal parameters, the model achieved its best precision (73.68%), recall (75.68%), and F1-score (74.67%).

Other models also exhibited relatively good performances under certain configurations. The Random Forest (RF) classifier, for instance, showed great performance with an accuracy of 75.52% on non-segmented data using the eGeMAPSv2 feature set (Table 2). This illustrates the efficacy of the ensemble learning techniques in handling the complexity of crosslinguistic acoustic data. Similarly, the RNN model displayed a high accuracy of 73.27% under the same condition, underscoring the potential of recurrent architectures in screening patients with AD within acoustic features.

3.2. Feature importance

Figure 2 shows 10 features with the highest feature importance values in SHapley Additive exPla-

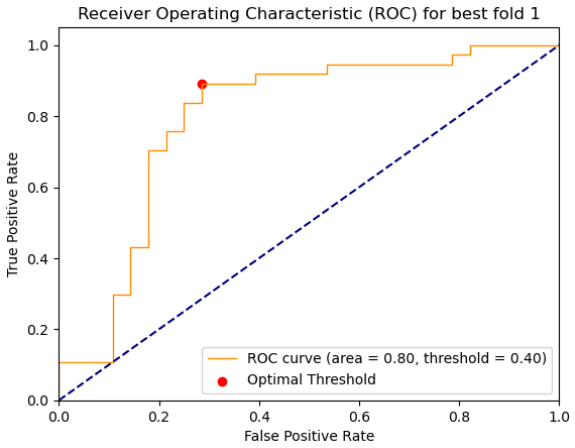

Figure 1: ROC curve illustrating the model's binary classification performance.

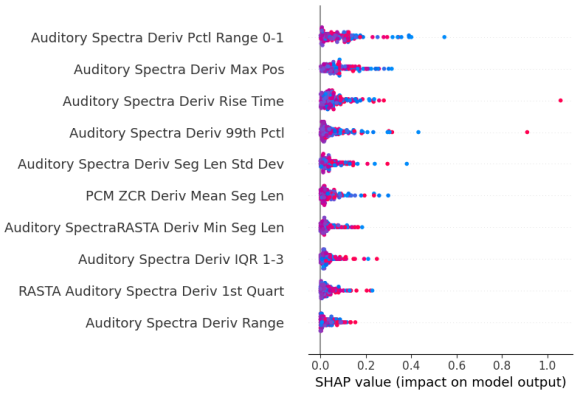

Figure 2: SHAP summary plot showing the influence of various features on model predictions. Absolute values are provided.

nations (SHAP) (Lundberg and Lee, 2017), illustrating the significance and impact of each acoustic feature on the best performing model's predictions. Most features in Figure 2 were spectra-related features, which measured prosodic characteristics of the participants. The only non-spectral feature was the mean values of zero crossing rate from the participants' speech. None of articulation-related features, such as Mel-frequent Cepstral Coefficients, had high importance values in the prediction.

4. Discussion

The MLP classifier trained on the non-segmented audio files using the ComParE feature set showed the best performance in distinguishing patients, showing an accuracy of 83.54% (AUC=0.8). The results suggest that a large-scale screening of AD patients using acoustic features extracted from one-minute picture descriptions in multiple languages is highly promising. Acoustic features that we employed could be automatically and uniformly extracted regardless of languages, which make a large-scale, remote screening of AD possible.

The MLP models generally performed the best with both eGeMAPSv2 and ComParE feature sets and in both segmented and non-segmented conditions, which may suggest that MLP models handle high dimensional features well, such as the acoustic features that we used in this study. Yet, the performance of an MLP model trained on segmented datasets slightly decreased compared to the same model trained on non-segmented datasets, which may suggest possible advantages of employing non-segmented data that retained linguistic nuances. Also, RNNs generally showed worse performance than MLP models in all segmentation and feature set combinations, which may suggest that we need larger datasets for efficient training with deep learning models.

Selected features with high feature importance values mostly included spectral-related features, suggesting that voice timbre and prosody are important features in distinguishing patients with AD from HCs. In contrast, the fact that articulation-related features, such as MFCCs, did not have high feature importance values in these tasks suggest that information on articulation is no longer informative when the dataset includes multiple languages with different phonetic and phonological systems. Future research is needed to confirm this observation. Other future research directions may include the exploration of advanced architectures and a deeper dive into interpretability.

5. Conclusion

In this study, we have rigorously evaluated various machine learning and deep learning classifiers for the binary task of distinguishing AD patients from HCs using acoustic features extracted from crosslinguistic speech data. Acoustic features can be automatically extracted from speech, regardless of languages, which make AD screening in diverse communities using natural speech highly plausible. Our findings contribute to both the methodological advancements and the inclusivity of crosslinguistic machine learning models in the field of AD and speech, benefiting diverse linguistic communities.

While showing promising results, this study has a few limitations in that many participants in the study were English speakers and only three languages were included. Future research will need to have balanced sample sizes for all languages to prevent any learning biases and include more languages to benefit numerous patients speaking non-English languages.

6. Bibliographical References

Alzheimer's Association. 2023. 2023 alzheimer's disease facts and figures. https://shop.alz.org/2023-Alzheimers-Disease-Facts-and-Figures-P1887.aspx. Accessed: 2023-10-18.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Ihab Hajjar, Maureen Okafor, Jinho D Choi, Elliot Moore, Anees Abrol, Vince D Calhoun, and Felicia C Goldstein. 2023. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 15(1):e12393.

Rui He, Kayla Chapin, Jalal Al-Tamimi, Núria Bel, Marta Marquié, Maitee Rosende-Roca, Vanesa Pytel, Juan Pablo Tartari, Montse Alegret, Angela Sanabria, et al. 2023. Automated classification of cognitive decline and probable alzheimer's dementia across multiple speech and language domains. *American Journal of Speech-Language Pathology*, 32(5):2075–2086.

Jordi Laguarta and Brian Subirana. 2021. Longitudinal speech biomarkers for automated alzheimer's detection. *frontiers in Computer Science*, 3:624694.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

Bai Li. 2019. Automatic detection of dementia in mandarin chinese. Master's thesis, University of Toronto.

Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. 2021. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6423–6427. IEEE.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. 2023. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. *arXiv preprint arXiv:2301.05562*.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. 2011. The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269.

Jessica Robin, Mengdan Xu, Liam D Kaufman, and William Simpson. 2021. Using digital speech assessments to detect early signs of cognitive impairment. *Frontiers in digital health*, 3:749758.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

John R Sims, Jennifer A Zimmer, Cynthia D Evans, Ming Lu, Paul Ardayfio, JonDavid Sparks, Alette M Wessels, Sergey Shcherbinin,

Hong Wang, Emel Serap Monkul Nery, et al. 2023. Donanemab in early symptomatic alzheimer disease: the trailblazer-alz 2 randomized clinical trial. *Jama*, 330(6):512–527.

Suramya Tomar. 2006. Converting video formats with ffmpeg. *Linux journal*, 2006(146):10.

Akshay Valsaraj, Ithihas Madala, Nikhil Garg, and Veeky Baths. 2021. Alzheimer's dementia detection using acoustic & linguistic features and pre-trained bert. In *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 171–175. IEEE.

Nayan Anand Vats, Aditya Yadavalli, Krishna Gurugubelli, and Anil Kumar Vuppala. 2021. Acoustic features, bert model and their complementary nature for alzheimer's dementia detection. In *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*, pages 267–272.

Ines Vigo, Luis Coelho, and Sara Reis. 2022. Speech-and language-based classification of alzheimer's disease: a systematic review. *Bioengineering*, 9(1):27.

Wei Xue, Catia Cucchiarini, RWNM van Hout, and Helmer Strik. 2019. Acoustic correlates of speech intelligibility. the usability of the egemaps feature set for atypical speech.