# Identifying Alzheimer's disease through speech: a multilingual approach

# Identificação da doença de Alzheimer através da fala: uma abordagem multilíngue

# Identificación de la enfermedad de Alzheimer a través del habla: un enfoque multilingüe

M.Sc. Guilherme Bernieri[1], Ph.D. Julio Cesar Duarte[1]

1 Military Institute of Engineering – IME, Rio de Janeiro (RJ), Brazil.

Corresponding author: M.Sc. Guilherme Bernieri
E-mail: bernieri@ime.eb.br

**Abstract**

Alzheimer's disease, the leading form of dementia among elderly individuals worldwide, has significant social and economic repercussions. It is characterized by memory loss and changes in language, cognition, and emotions, irreversibly affecting neurons. Early diagnosis is crucial but challenging, as it relies on detailed medical evaluations, cognitive tests, and complex exams that are often expensive and inaccessible, particularly for low-income individuals. In this context, advanced computational techniques, such as machine learning (ML), emerge as promising non-invasive alternatives for the early detection of the disease. This study introduces a multilingual ML-based approach focusing on paralinguistic and emotional speech characteristics as biomarkers for Alzheimer's identification. The experiments yielded results with accuracies reaching 81% for English and 87.50% for Portuguese. Additionally, integrating this methodology with the state-of-the-art model by Haider, Fuente, and Luz(1) resulted in an average accuracy of 81.70%, surpassing their original results.

**Keywords:** Alzheimer's Disease; Automatic Speech Analysis; Machine Learning

**Resumo**

A doença de Alzheimer, principal forma de demência entre os idosos no mundo, tem significativas repercussões sociais e econômicas. É caracterizada pela perda de memória e mudanças na linguagem, cognição e emoções, afetando irreversivelmente os neurônios. O diagnóstico precoce é fundamental, mas desafiador, pois depende de avaliações médicas, testes e exames complexos que, muitas vezes, são inacessíveis para indivíduos de baixa renda. Nesse contexto, técnicas computacionais, como o aprendizado de máquina (AM), surgem como alternativas para a detecção da doença. Este estudo apresenta uma abordagem multilíngue baseada em AM, focando nas características paralinguísticas e emocionais da fala como biomarcadores para a identificação do Alzheimer. Os experimentos produziram resultados com acurácia de 81% para o inglês e 87,50% para o português. Além disso, a integração dessa metodologia com o modelo de Haider, Fuente, e Luz[1] resultou em uma acurácia média de 81,70%, superando os resultados originais dos autores.

**Descritores:** Doença de Alzheimer; Análise Automática da Fala; Aprendizado de Máquina

**Resumen**

La enfermedad de Alzheimer, principal forma de demencia entre las personas mayores en todo el mundo, tiene significativas repercusiones sociales y económicas. Se caracteriza por la pérdida de memoria y cambios en el lenguaje, cognición y emociones. El diagnóstico temprano es desafiante, ya que depende de evaluaciones médicas, pruebas y exámenes complejos que, a menudo, son inaccesibles para las personas de bajos ingresos. Las técnicas computacionales, como el aprendizaje automático (AA), emergen como alternativas para la detección de la enfermedad. Este estudio presenta un enfoque multilingüe basado en AA que se centra en las características paralingüísticas y emocionales del habla como biomarcadores para la identificación del Alzheimer. Los experimentos arrojaron resultados con una exactitud del 81% para inglés y del 87,50% para portugués. Además, la integración de esta metodología con el modelo de Haider, Fuente y Luz[1] resultó en una exactitud promedio del 81,70%, superando sus resultados originales.

**Descriptores:** Enfermedad de Alzheimer; Análisis Automático del Habla; Aprendizaje Automático

## Introduction

The aging of the global population is a phenomenon that is advancing rapidly. Data from the World Health Organization (WHO)[2] indicate that the global population aged 60 years or older is growing at a rate of approximately 3% per year, and it is estimated that, by the year 2050, the total number of elderly people will exceed 2 billion. As the elderly population of the world increases, so does the number of people affected by neurodegenerative diseases, such as Alzheimer's and Parkinson's. Currently, Alzheimer's disease (AD) is the most prevalent neurodegenerative disease among the elderly worldwide, and according to the World Alzheimer Report 2023[3], it is expected that by 2050, about 139 million people will suffer from dementia globally, with the annual costs related to this condition potentially reaching 2 trillion dollars by 2030.

The methodologies used in current studies for identifying AD through speech analysis focus on two main approaches: automatic and semi-automatic. In an automatic approach, features are extracted directly from speech, and a model is trained to perform the classification. In a semi-automatic approach, the manual transcription of speech is included to obtain additional linguistic features[4]. A common limitation of both approaches is the difficulty in generalizing to different languages, as models trained in one language may not be effective in another. This is especially relevant considering the linguistic diversity worldwide. The identification of AD through non-verbal (paralinguistic) approaches has been minimally explored to date. However, acoustic analysis of speech has shown to be capable of assisting in understanding the subtleties that may indicate the presence of the disease[5].

This work aims to contribute to the research area of signal processing applied to health, through a multilingual approach based on ML to identify AD. The study builds upon the research conducted by Bernieri and Duarte[6] where the use of emotion recognition to identify AD was evaluated. Thus, paralinguistic features, automatically extracted from the acoustic signals of spontaneous speech, and emotional characteristics, through the three-factor theory of emotions[7], are now evaluated as biomarkers for the actual classification of AD. It is expected that the results of this work can contribute to the early identification of AD, expanding and improving the diagnostic support tools available to healthcare professionals.

**Related works**

Cai *et al*.[8] explored AD detection using pre-trained language models, graph neural networks (GNN), and acoustic and linguistic features. The authors utilized the Pitt Corpus[9] dataset and data augmentation techniques, with a Generative Pre-trained Transformer (GPT). The results indicated that the combined approach of a pre-trained language model with GNN achieved the best performance, with an accuracy of 85.04%. In terms of individual methods, the text model significantly outperformed the audio model, with an accuracy of 84.60% compared to 77.14%, possibly due to the complexity of audio data. When text and audio were combined, the accuracies were similar to those achieved with text alone, suggesting an excessive dependence on textual information.

Haider, Fuente, and Luz[1] proposed a model to detect AD based only on the analysis of acoustic features of speech. They explored various sets of acoustic features, such as eGeMAPS, emobase, and others. Additionally, a dataset derived from the Pitt Corpus[9], segmented by a voice activity detection system, was used. To model the acoustic information, they employed an active data representation (ADR) that creates a fixed-dimension feature vector. Three disease identification experiments were conducted, with the most effective method being the combination of eGeMAPS, ADR, and linear discriminant analysis (LDA), achieving 77.44% accuracy. An additional strategy. involving the fusion of feature sets, reached 78.70% accuracy by using the decision tree (DT) algorithm and majority voting.

The study conducted by Bernieri and Duarte[6] proposed a methodology for identifying AD by automatically analyzing speech and using emotion recognition as a biomarker of the disease. The authors utilized the Emo-DB[10] dataset to train the emotion classifier and the Pitt Corpus[9] for AD classification. The research implemented a voice activity detector and the support vector machine (SVM) technique for emotion recognition. The disease classification involved extracting emotional characteristics from speech and pattern recognition with a multilayer perceptron (MLP) neural network. The study achieved

72.61% accuracy, with 72.90% precision and a 72.50% recall rate, suggesting that emotional features in speech could potentially serve as valuable indicators for the early diagnosis of AD.

## Materials and Methods

#### Datasets

This study utilized three datasets that are well-known and widely applied in research of this nature, as seen in the works of Haider, Fuente and Luz, Bernieri and Duarte, Cai *et al*., and Aluísio, Cunha and Scarton[1, 6, 8, 11]. The first of these datasets is the Berlin Database of Emotional Speech (Emo-DB)[10], which contains 535 recordings of German speech made by ten actors expressing seven distinct emotions. This was used to train the automatic emotion recognition model. The second dataset, referred to as the Pitt Corpus[9], includes 404 English speech recordings, of which 222 are from patients with AD and 182 from a Control Group. This set was used in training the AD classification model. Finally, the Cinderella[11] dataset, comprising 40 Portuguese speech recordings (20 from participants with Mild AD and 20 from the Control Group), was employed to validate the developed multilingual model.

#### Preprocessing

To ensure the quality and homogeneity of the data, the audio files were preprocessed, removing noises and standardizing the recordings. For this task, an implementation of the Wave-U-Net neural network architecture from the Malaya-Speech[12] library was used, due to its robustness and versatility, with the ability to adjust to various types of noises. The process involved dividing the audio files into 15-second segments, separating the voice from the noise, and, subsequently, recombining the voice segments to restore the original recording without noise. This strategy aimed to facilitate noise removal, reduce computational costs, and maximize efficiency. Due to the nature of the tests applied to the participants, it was necessary to remove the interventions of the examiners from the recordings. In this additional step, through diarization[1] techniques,

[1] Diarization refers to the process of segmenting and attributing audio recordings to specific speakers.

each file was divided into distinct temporal segments, assigning each segment to a specific speaker, and at the end, the speech segments from the examiners were removed.

#### Voice activity detection model

A voice activity detector was used in the methodology to manage the diversity of the audio recordings, which vary in duration and include both speech and silent segments. This detector, based on a pre-trained implementation of the convolutional neural network architecture VGGVox from the Malaya-Speech library, identifies and records the timestamps corresponding to speech and silence within the recordings. VGGVox processes audio spectrograms, transforming them into vector representations used for detecting vocal activity. This network consists of convolutional and fully connected layers, using the ReLU activation function and a softmax layer for classification. The network traverses each recording in small segments for analysis, determining the probability of each segment being voice or silence, and recording their start and end times. Consecutive segments with the same classification are grouped into samples of up to 10 seconds.

#### The three-factor of emotions

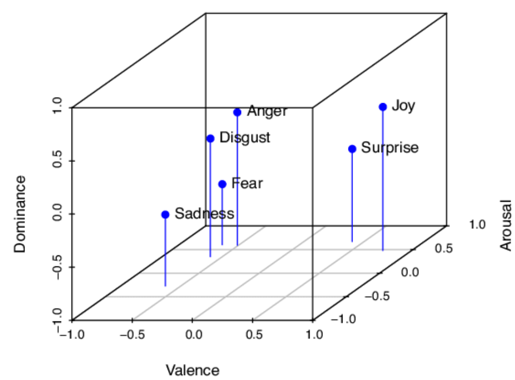
To investigate a combination of acoustic, paralinguistic, and emotional speech characteristics, the main features selected for the model were the dimensional values of valence, arousal, and dominance. The choice of these features was grounded in the three-factor theory of emotions, proposed by Russell and Mehrabian(7). This theory proposes a three-dimensional analysis framework to categorize emotions, assuming that all emotions we experience can be defined and located in a space formed by three basic dimensions:

- Valence refers to the degree of pleasure or displeasure experienced by a person in reaction to an event, object, or situation, with positive emotions such as happiness representing high valence;
- Arousal pertains to the level of activity or stimulation felt, with emotions such as anger associated with high excitement;

- Dominance describes the feeling of control or influence over a situation, with emotions that confer a sense of power, such as confidence, indicating high preeminence.

Thus, any specific emotion can be represented as a point in a three-dimensional space, as illustrated in Figure 1. For instance, the emotion “joy” exhibits high valence, high arousal, and high dominance.

**Figure 1 –** Positions of basic emotions in the emotional space of the valence, arousal, and dominance axis. Adapted from Buechel e Hahn[13].




#### Extraction of speech features

The process of extracting the dimensional features of valence, arousal, and dominance, through the acoustic analysis of speech, is carried out using the implementation of a pre-trained deep learning model, developed by Wagner *et al*.[14]. Based on the wav2vec 2.0 framework, which utilizes the Transformer architecture and self-attention techniques, the model was adjusted to the new features and trained on the MSP-Podcast Corpus emotional dataset. For this purpose, following the final layer of the network, an average pooling operation was applied on the hidden states of the last Transformer layer to synthesize the audio sequence representation. This representation, after being processed by a hidden layer, is fed into the output layer, thus generating the final predictions of the model. The output values vary within a continuous range from 0 to 1 for each dimensional feature.

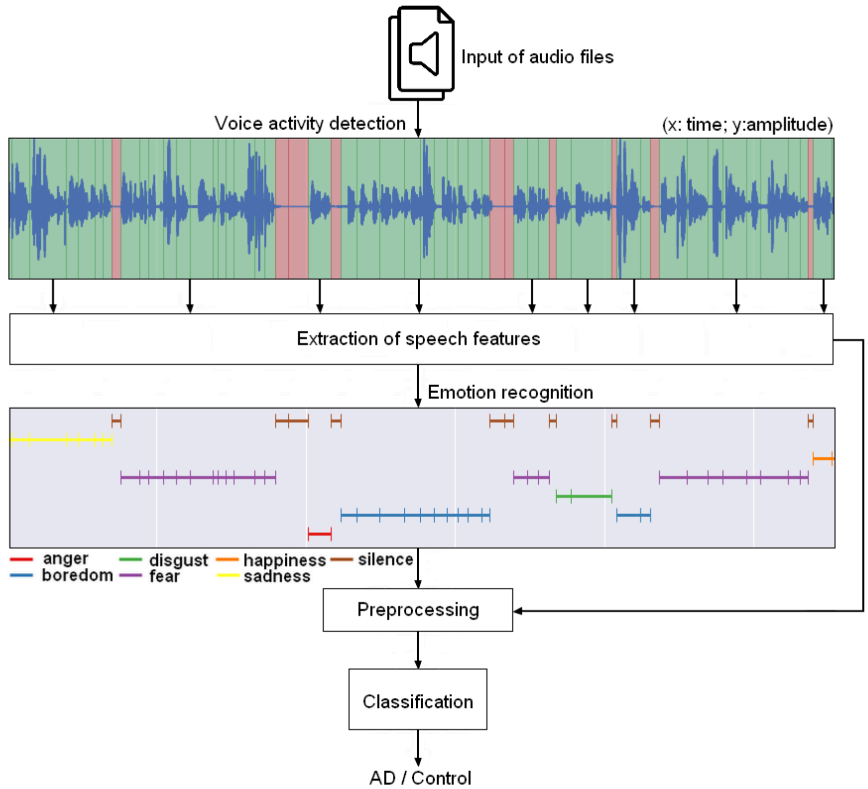
**Emotion recognition model**

Since the model for extracting speech features is adjusted based on the dimensional values of valence, arousal, and dominance, instead of categorical values, it became necessary to train a new model for the classification of emotions. For this purpose, the latent space representation from the final layer of the speech feature extraction model, obtained from the recordings of the Emo-DB dataset, served as input for the emotion recognition model. These latent representations act as a unique “fingerprint” for each emotion. The SVM algorithm was chosen for its robustness against overfitting and efficiency in handling high-dimensional data, making it suitable for detecting the specific patterns associated with each emotion. After the training process, the model was capable of identifying the seven distinct emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral.

**Classification methodology**

The developed classification methodology involves the structured application of the procedures and techniques previously outlined. Additionally, it incorporates an MLP neural network, designed to learn from the paralinguistic patterns and multidimensional values of emotions through speech, as illustrated in the architecture of Figure 2. Initially, the voice activity detection model identifies and records the occurrences of voice and silence activities in each recording. Relying solely on the identified voice activities as a reference, each audio file is analyzed to extract the dimensional features of valence, arousal, and dominance. Subsequently, the latent representations of the extracted features are used as input for the emotion recognition model, which classifies all the emotions identified throughout the speech. Upon performing these tasks, each analyzed recording results in the timestamp data for voice and silent activity segments, along with their corresponding emotions and the dimensional values of valence, arousal, and dominance.

**Figure 2 –** Architecture of the AD classification methodology.




Preparing data for entry into the MLP neural network requires preprocessing, which involves aggregating information from individual segments of the recordings into a single record per sample. As a result, each record includes the age and sex of the patient, the total duration of the recording, the total speech duration, the speech-to-silence ratio, the total number of pauses, their average and maximum durations, the rate of emotional variation, as well as the average values of valence, arousal, and dominance. Additionally, it includes the sum of these values for each individual emotion. Before being forwarded to the classifier, data undergo normalization. Finally, the neural network performs the classification of patients with Alzheimer's, distinguishing them from patients without the disease, by adjusting its synaptic weights.

## Experiments and Results

### Emotion recognition

The emotion recognition model is a fundamental component of the methodology developed in this research. In this context, to evaluate the performance of the model, a "leave one group out" cross-validation was applied to the Emo-DB dataset. In this approach, during each iteration, a group of recordings from a specific actor was withheld as the test set, while the remaining groups were used for training. This helps validate that the model performs well on previously unseen data, thereby enhancing its generalization capability, as can be observed in Table 1.

**Table 1 –** Emotion classification metrics.

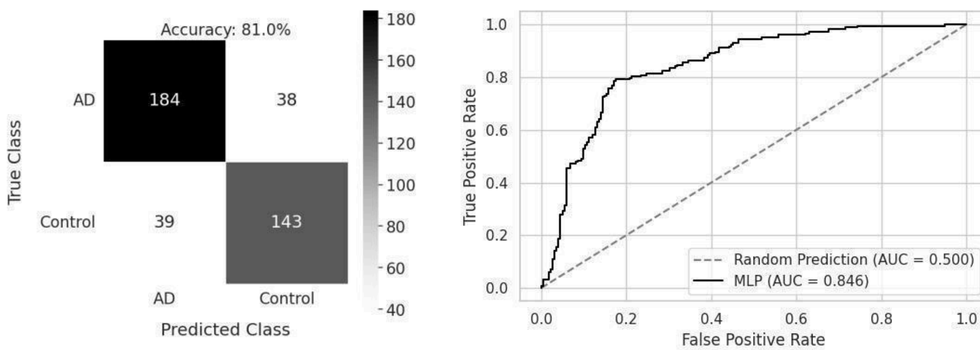
| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Anger | 0.932 | 0.976 | 0.954 | 127 |
| Boredom | 0.928 | 0.951 | 0.939 | 81 |
| Disgust | 0.932 | 0.891 | 0.911 | 46 |
| Fear | 0.912 | 0.899 | 0.905 | 69 |
| Happiness | 0.894 | 0.831 | 0.861 | 71 |
| Neutral | 0.962 | 0.962 | 0.962 | 79 |
| Sadness | 0.968 | 0.968 | 0.968 | 62 |
| Macro avg | 0.932 | 0.925 | 0.929 | 535 |
| Accuracy | | | 0.933 | 535 |

### Alzheimer's disease identification

The efficacy of the classification methodology was validated on the Pitt Corpus dataset through a 10-fold cross-validation. The final architecture of the network, comprising six layers of neurons — including the initial input layer, four hidden layers, and the last output layer — was defined, and its hyperparameters were optimized through the experimentation of a range of settings, based on the number of inputs to the network and definitions outlined in the literature. For the training of the model, the backpropagation algorithm was used. The developed neural network achieved an accuracy of 81.00%, a precision of 80.80%, a recall of 80.70%, and a specificity of 80.80%.

In the healthcare field, a false negative result in disease diagnosis can be more detrimental than a false positive, potentially leading to the neglect of initiating treatment, and causing serious problems. Therefore, classification models in this area should be evaluated for their sensitivity and specificity[15]. Additionally, the analysis of the ROC curve, more specifically, the area under the ROC curve (AUC), can be used to detail the performance of a diagnostic test. A high-quality model typically exhibits an AUC close to 1, indicating a strong measure of separability between classes. The model developed in this study achieved an AUC value of 0.846, as can be observed in Figure 3.

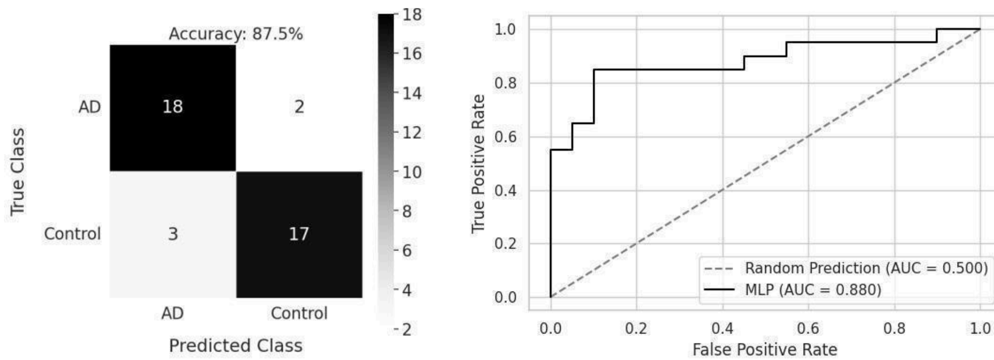
**Figure 3 –** Classifier metrics for the Pitt Corpus dataset.




**Assessment of the multilingual capability of the methodology**

The methodology developed to identify AD through automatic speech analysis is driven by the goal of being language-independent. Therefore, it can be applied in diverse global contexts, covering populations speaking different languages. To evaluate the generalization capability of the methodology in diverse linguistic environments, the Cinderella dataset was used. The preprocessing steps, speech feature extraction, emotion recognition, feature engineering, training, and model evaluation were consistent with those used in the previous experiment, allowing for a fair comparison. The analysis of the performance metrics of the model revealed performance aligned with the results obtained in the experiment using English recordings, achieving an accuracy of 87.50%, a precision of 87.60%, and recall and specificity values of 87.50%. Furthermore, the model achieved an AUC value of 0.846, as illustrated in Figure 4.

**Figure 4 –** Classifier metrics for the Cinderella dataset.




**Integration of the methodology into a state-of-the-art model**

The final stage of the experiments focused on integrating the developed methodology with the related work of Haider, Fuente, and Luz[1], due to its good description for reproduction, aiming to evaluate its potential to enhance another state-of-the-art model. To replicate the study, the best individual identification method developed by the authors was selected, which uses the eGeMAPS parameter set, the ADR technique, and a model derived from LDA. The integration followed three distinct phases: the inclusion of emotional features into the reference model, the modification of its classification algorithm to MLP, and the combination of classifiers using the stacking technique. Despite the comprehensive description of the reference model, some aspects of the technique remained inadequately clarified. Thus, the reproduction of the model, ADR+LDA, achieved an accuracy of 71.30%, while the original model reached 77.40%. Nonetheless, by integrating the emotional features, its accuracy increased to 76.70%, maintaining the same LDA classifier and, after replacing the LDA classification algorithm with MLP, an additional improvement was generated, achieving an accuracy of 79.20%. The most effective method, however, was stacking, which combined the individual classifiers and achieved the highest accuracy, of 81.70%, even outperforming the best value obtained by Haider, Fuente, and Luz[1], as presented in Table 2.

**Table 2 –** Model combination results.

| Model | Accuracy | Precision | Recall | F1-score | Specificity |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| ADR+LDA reported | 0.774 | - | - | - | - |
| ADR+LDA reproduced | 0.713 | 0.710 | 0.708 | 0.708 | 0.708 |
| ADR+Emotions+LDA | 0.767 | 0.765 | 0.764 | 0.765 | 0.764 |
| ADR+Emotions+MLP | 0.792 | 0.790 | 0.791 | 0.790 | 0.791 |
| Stacking | **0.817** | **0.815** | **0.816** | **0.815** | **0.816** |

**Comparison of results with related works**

Table 3 presents a comparative analysis between the results of this research and the related works. In the study by Cai *et al.*[8], multimodal methods utilizing acoustic and linguistic features were investigated, achieving an accuracy of 85.04% using GNN. Haider, Fuente, and Luz[1] developed a paralinguistic model based exclusively on acoustic features, resulting in an accuracy of 78.70% using DT. Bernieri and Duarte[6] proposed the identification of AD through automatic speech analysis and emotion recognition, obtaining an accuracy of 72.61% with MLP. This work implemented a multilingual method that combines paralinguistic and emotional features, achieving the highest accuracy of 87.50% with MLP, standing out for its comprehensive analysis of emotions and provision of a multilingual approach.

**Table 3 –** Comparative table of related works.

| Related work | Main objective | Method | Features | Best result | Emotions | Multilingual |
|---|---|---|---|---|---|---|
| Cai *et al.*[8] | Investigation of multimodal methods for AD detection using acoustic and linguistic features | GNN | Acoustic Linguistic | Accuracy 85.04% | | |
| Haider, Fuente, and Luz[1] | Development of a paralinguistic model for AD detection using acoustic parameters | DT | Acoustic | Accuracy 78.70% | | |
| Bernieri and Duarte[6] | Identification of AD through automated speech analyzing and emotion recognition | MLP | Acoustic | Accuracy 72.61% | ✓ | |
| This work | Implementation of a multilingual approach based on paralinguistic and emotional features for AD identification | MLP | Acoustic | Accuracy 87.50% | ✓ | ✓ |

## Conclusion

The results of the experiments conducted highlighted the effectiveness of the methodology, even when applied to a dataset in another language. Furthermore, the integration of the proposed methodology with a state-of-the-art reference model demonstrated its potential to enhance other models. An additional contribution of this work was the development of a model for emotion recognition that achieved high performance for most emotions. The methodology developed in this research proved to be an important tool to assist in the identification of Alzheimer's. However, it is essential to emphasize that this approach does not constitute a definitive method for the diagnosis of the disease. Instead, it acts as a complementary tool with the potential to raise awareness about the signs of the disease. By assisting healthcare professionals in identifying patients who potentially suffer from the disease, it enables more effective targeting for in-depth diagnostic investigations.

Throughout the execution of this work, promising opportunities have emerged for expanding the scope of the research, like exploring the developed methodology in other neurodegenerative diseases. Another possibility is using more diverse datasets, covering more languages. Moreover, with the ongoing evolution of Large Language Models, an opportunity arises for integrating these models with the methodology of this research since their high attention capacity and ability to process language across multiple languages can enhance the performance and scope of AD detection through speech analysis.

## References

1. Haider F, de la Fuente S, Luz S. An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. IEEE Journal of Selected Topics in Signal Processing. 2020 Feb;14:272-281.

2. World Health Organization. Ageing. World Health Organization (WHO) [Internet]. 2024 [cited 2024 Jan 15]; [about 1 p.]. Available from: https://www.who.int/health-topics/ageing.

3. Long S, Benoist C, Weidner W. World Alzheimer Report 2023: Reducing dementia risk: never too early, never too late. London, England: Alzheimer's Disease International (ADI); 2023.

4. Campbell EL, Mesía RY, Docío-Fernández L, García-Mateo C. Paralinguistic and linguistic fluency features for Alzheimer's disease detection. Computer Speech & Language. 2021. Jul;68:101198.

5. de la Fuente Garcia S, Haider F, Luz S. Cross-corpus Feature Learning between Spontaneous Monologue and Dialogue for Automatic Classification of Alzheimer's Dementia Speech. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020 Jul 20-24; Montreal, Canada. p. 5851-5855.

6. Bernieri G, Duarte JC. Identifying Alzheimer's Disease Through Speech Using Emotion Recognition. Journal of Health Informatics. 2023 Jul 20;15:1-14.

7. Russell JA, Mehrabian A. Evidence for a three-factor theory of emotions. Journal of Research in Personality. 1977 Sep;11:273–294.

8. Cai H, Huang X, Liu Z, Liao W, Dai H, Wu Z, et al. Multimodal Approaches for Alzheimer's Detection Using Patients' Speech and Transcript. Lecture Notes in Computer Science. 2023 Sep 13;395–406.

9. Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. Archives of Neurology. 1994 Jun; 51 (6): 585-594.

10. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Proceedings of Interspeech 2005; 2005 Sep 04-08; Lisbon, Portugal. p. 1517-1520.

11. Aluísio S, Cunha A, Scarton C. Evaluating Progression of Alzheimer's Disease by Regression and Classification Methods in a Narrative Language Test in Portuguese. Computational Processing of the Portuguese Language. 2016;109–114.

12. Husein Z. Malaya, Speech-Toolkit library. Version 1.2.7 [software]. 2020 [cited 2024 Jan 15]. Available from: https://github.com/huseinzol05/malaya-speech.

13. Buechel S, Hahn U. Emotion Analysis as a Regression Problem – Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. European Conference on Artificial Intelligence 2016; 2016. p. 1114-1122.

14. Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Burkhardt F, Eyben F, et al. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022; 1–13.

15. Shreffler J, Huecker MR. Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios. StatPearls. 2023.