Identifying Alzheimer's Disease Through Speech Using Emotion Recognition

Identificação da Doença de Alzheimer Através da Fala Utilizando Reconhecimento de Emoções

Identificación de la Enfermedad de Alzheimer a Través del Habla Mediante el Reconocimiento de Emociones

Guilherme Bernieri[1], Julio Cesar Duarte[2]

1 Master's Student, Military Institute of Engineering – IME, Rio de Janeiro (RJ), Brazil.
2 Professor, Military Institute of Engineering – IME, Rio de Janeiro (RJ), Brazil.

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

Abstract

Alzheimer's disease is the most common neurodegenerative dementia in elderly people in the world and its diagnosis requires a wide medical evaluation, supported by cognitive tests, clinical and imaging exams. Identifying the disease through speech can reduce the cost and time of medical diagnosis. Emotional states are important performance indicators of cognitive processes. Intelligent and non-invasive computational techniques can become relevant support tools for an early medical diagnosis. Therefore, this article addresses the use of emotion recognition through voice as a biomarker to identify the presence of Alzheimer's disease. The proposed method is based on the extraction of emotional features from speech and pattern recognition using neural networks. The results of the experiments reached an accuracy of 72.61%, a precision of 72.90% and a recall of 72.50% through cross-validation of the data.

Keywords: Alzheimer Disease; Automatic Speech Analysis; Machine Learning

Resumo

A doença de Alzheimer é a demência neurodegenerativa mais comum em pessoas idosas no mundo e o seu diagnóstico requer uma ampla avaliação médica, apoiada por testes cognitivos, exames clínicos e de imagem. Identificar a doença através da fala pode reduzir o custo e o tempo do diagnóstico médico. Os estados emocionais são importantes indicadores de desempenho dos processos cognitivos. Técnicas computacionais inteligentes e não invasivas podem se tornar relevantes ferramentas de apoio para um diagnóstico médico precoce. Portanto, este trabalho aborda a utilização do reconhecimento de emoções através da voz como biomarcador para identificação da presença da doença de Alzheimer. O método proposto é baseado na extração das características emocionais da fala e no reconhecimento de padrões utilizando redes neurais. Os resultados dos experimentos alcançaram uma acurácia de 72,61%, uma precisão de 72,90% e uma revocação de 72,50% por intermédio da validação cruzada dos dados.

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

Resumen

La enfermedad de Alzheimer es la demencia neurodegenerativa más común en personas mayores en el mundo y su diagnóstico requiere una amplia evaluación médica, apoyada en pruebas cognitivas, exámenes clínicos y de imagen. Identificar la enfermedad a través del habla puede reducir el tiempo del diagnóstico. Los estados emocionales son importantes indicadores de rendimiento de los procesos cognitivos. Técnicas computacionales inteligentes pueden convertirse en herramientas relevantes para un diagnóstico médico temprano. Este artículo aborda el uso del reconocimiento de emociones a través de la voz como biomarcador para identificar la presencia de la enfermedad de Alzheimer. El método propuesto se basa en la extracción de características emocionales del habla y el reconocimiento de patrones mediante redes neuronales multicapa. Los resultados de los experimentos alcanzaron una exactitud del

72,61 %, una precisión del 72,90 % y una exhaustividad del 72,50 % mediante la validación cruzada de los datos.

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

Introduction

Currently, the world population is aging. Data from the World Health Organization (WHO)[1] indicate that the world population aged 60 years and older is growing at a rate of 3% per year and it is estimated that by 2050 the total number of elderly people aged 60 and older will exceed 2 billion. Keeping up with the worldwide growth of the elderly population, the number of people suffering from neurodegenerative diseases, such as Alzheimer's and Parkinson's diseases, is also increasing. Alzheimer's disease, for example, is currently the most common neurodegenerative disease in elderly people worldwide. According to the World Alzheimer Report 2022[2], in 2050, it is estimated that 139 million people will suffer from dementia worldwide and that the annual expenditure on the disease could reach 2 trillion dollars by 2030. Alzheimer's disease is a debilitating disease that causes damage and affects brain neurons in a progressive, irreversible, and fatal way. The gradual degeneration of nerve cells induces conditions of dementia, involving different circumstances, such as speech impairment[3]. The most common early symptoms are short-term memory loss, mood swings, and difficulty using language. As the disease progresses, the performance of daily activities is affected and, in the final stages of the disease, people are confined to bed and need 24 hour care[4].

Even with positive results in several studies to date, the process for diagnosing Alzheimer's disease requires a broad medical evaluation, supported by cognitive tests, clinical and imaging exams. These exams are important diagnostic tools, however, they may require invasive procedures and are time-consuming and expensive to obtain, especially for low-income patients[5], which may delay the diagnosis of the disease. Although the gradual loss of memory is the main symptom of Alzheimer's disease, problems with speech, communication and changes in emotions[6] also appear as initial symptoms of the disease. Emotions are complex and are among the main characteristics of the human being. The recognition of emotional states can be considered one of the

most important evaluation criteria to measure the performance of cognitive processes[7]. Studies indicate that Alzheimer's disease has its onset silently, up to 20 years before the patient presents observable cognitive symptoms[8].

In this case, the use of intelligent and non-invasive computational techniques to support medical diagnosis can become important tools in the early identification of Alzheimer's disease. These techniques are easy to apply and can be implemented in the medical protocols already used, without prejudice to the daily routine of professionals. Among the non-invasive methods currently being researched, automatic speech analysis provides a powerful, natural, and user-friendly tool capable of providing information easily, quickly, and economically. The automatic recognition of emotions has become the focus of research related to technology and health. The identification of Alzheimer's disease using paralinguistic approaches has been less explored so far, however, acoustic analysis is able to assist in understanding the subtleties of speech that may indicate the presence of the disease[9].

This work aims to contribute to the research area of signal processing for Healthcare, with a multilingual approach based on machine learning for the identification of neurodegenerative diseases, evaluating the use of emotion recognition through speech as a biomarker for classification of Alzheimer's disease. The methodology applied in this study intends to develop a multilayer perceptron neural network model, using voice activity detection and emotion recognition automatic techniques, which can help health professionals in the identification of the disease, reducing the interval between diagnosis and start of medical treatment, as well as providing an analysis of the application of emotional features of speech as an addition to recent machine learning methods.

Related Works

Computational techniques based on the automatic processing of voice signals have found an important place in research applied to the detection of neurodegenerative disorders. With the aim of developing a method to generalize the use of computational tools for the diagnosis of Alzheimer's disease, Campbell *et al.*[10] propose, in their study, the use of multilingual systems with characteristics extracted from voice signals. One of the main objectives of the research was to explore the advantages of using paralinguistic

parameters to develop a multilingual method of fully automatic disease detection. The authors also propose a method using linguistic parameters, although not in a multilingual system, and compare the results obtained. The multilingual approach is based on processing the acoustic waveform of the voice. A multilayer neural network is used as a classifier and has an accuracy rate of 75% for the English language and 77% for the Spanish language. The approach using linguistic and semiautomatic parameters involves the analysis of temporal patterns of silence between words and errors in word production, reaching an accuracy of 88%.

Haider *et al*.[8] intended to identify Alzheimer's disease through emotional characteristics. The authors presented a study of the predictive value of emotional behavior characteristics automatically extracted from spontaneous speech for the detection of Alzheimer's disease. An automatic emotion recognition system, applying a set of acoustic parameters and the Berlin emotion database, was trained using a support vector machine. The classified emotions of each speech segment are represented in a single fixed-dimension vector for the disease classification step. The effectiveness of using emotions for Alzheimer's disease detection was evaluated on a balanced subset of the Pitt Corpus, which is a spontaneous speech database for the study of communication in Alzheimer's disease. The best performing classification model was the random forest, and the experiment was able to detect Alzheimer's disease with an accuracy of 63.42%.

Materials and Methods

Datasets

This study uses two datasets widely used in research of this nature, as can be seen in the related works of Campbell *et al*.[10] and Haider *et al*.[8]. The Berlin Database of Emotional Speech (Emo-DB)[11] was used to train the automatic emotion classifier and the Pitt Corpus[12] was used for the classification of Alzheimer's disease.

The Emo-DB dataset is part of a research project at the Technical University of Berlin and is frequently used in studies on automatic emotion recognition. Ten actors, five men and five women, simulated five short sentences and five long sentences, which can be used in every day's life and which are interpretable in all emotions. The emotions

considered in the research are anger, boredom, disgust, fear, happiness, neutral and sadness. In total, after being evaluated by twenty people in a perception test regarding emotion recognition, 535 recordings in the German language formed the final corpus.

The Pitt Corpus is part of the University of Pittsburgh Alzheimer's Research Program and was collected between 1983 and 1984. To participate in the study, the participants must be over 44 years of age, have at least 7 years of schooling, have no history of nervous system disorders or take neuroleptic medication, and have an initial Mini-Mental State Examination score of 10 or more. The participants were classified into Alzheimer's group (AD), Control group (Control) and Mild Cognitive Impairment group (MCI). The Pitt Corpus data were obtained through the “Cookie Theft” image description activity for the AD group, Control group and MCI group and include recordings and manual transcripts of the activities. For the Alzheimer's group, word fluency, story recall and sentence construction activities were also carried out. For this study, due to the small number of audio samples from the group of patients with Mild Cognitive Impairment, this group was not considered for analysis. Thus, a subset of the Pitt Corpus was used, containing 236 audio recordings from the Alzheimer's group and 242 recordings from the Control group, totaling 478 audio files in English.

Pre-Processing

Since the Pitt Corpus dataset is composed of audio recordings with different qualities and interferences, it was necessary to carry out a pre-processing step, in order to remove background noise and normalize the recordings volume. For this, an implementation of a pre-trained U-Net residual neural network from the Malaya-Speech library[13] was used. The model is tuned on different noises, augmented noises, procedural noises and overlapped noises. Each audio file was divided into 15-second segments, applying noise removal, volume normalization and, then, combined again after the process was complete. The pre-processing of audio files is a delicate process and directly affects the performance of emotion recognition. In this way, the entire dataset used in the experiment must go through the same process.

Emotion Recognition Model

A support vector machine (SVM) with a radial basis function kernel was developed to detect the patterns of each emotion. Although initially designed for binary classification, support vector machines are now also widely used for multiclass problems. The heuristic one-vs-one majority voting ensemble method is applied to divide the multiclass dataset into binary classification problems, where each binary classification model predicts an output and the class that receives the most votes is selected. The extraction of the audio samples characteristics used as inputs in the algorithm is performed by a pre-trained and fine-tuned model for the recognition of the dimensional values of arousal, dominance and valence of emotions, based on the wav2vec 2.0 framework[14]. As the model was adjusted for dimensional and non-categorical values, a fine-tuning is necessary for classifying emotions. Thus, the representation of the model's latent space was used to extract 1024 hidden states for each audio sample. The recordings of the Emo-DB dataset were applied to train the model to classify seven different emotions, namely: anger, boredom, disgust, fear, happiness, neutral and sadness. An exhaustive search algorithm and cross-validation of different combinations of the kernel function and its hyperparameters was used to define the best configuration for the SVM model.

Voice Activity Detection Model

The audio recordings of the “Cookie Theft” image description task from the Pitt Corpus data subset have snippets of silence and speech, and the timing of each audio is variable. Therefore, a voice activity detector was incorporated into the experimentation methodology to extract the timestamps of the voice and silence activities of each recording. The developed voice activity detector was based on a pre-trained multilingual deep learning model, also from the Malaya-Speech library[13]. Each audio sample is divided into small segments of 30 milliseconds, which are analyzed by the model for the probability of being voice or silence, as well as the start and end timestamps of the segments. After that, the segments with equal ratings - voice or silence - are grouped again to form larger samples. So, the extracted data are recorded to be used in the emotion recognition of the speech segments of the audios.

Classification Methodology

In order to identify patients with Alzheimer's and differentiate them from patients without the disease, a multilayer perceptron neural network was built, having, as input, the data referring to the emotions identified in the patients' speeches. Five layers were used in the neural network, one for input, three hidden layers and one layer for output, as shown in Figure 1. The hyperbolic tangent activation function was applied due to its adaptable nonlinear characteristic for classification between two classes. The Adam optimizer was adopted as an alternative to the gradient descent method, as it has a more efficient stochastic optimization and requires only first order gradients, where memory cost is reduced. The neural network was trained using the error backpropagation algorithm, which consists of two processing steps. In the first step, an input is established to the neural network and its signal is transmitted by each layer of the network until a result is obtained by the output layer. In the second step, the error signal is calculated at the network output and transmitted in reverse, adjusting the values of the synaptic weights as the error is back propagated. The hyperparameters from the neural network were also defined using an exhaustive search algorithm and cross-validation of different combinations of hyperparameters.

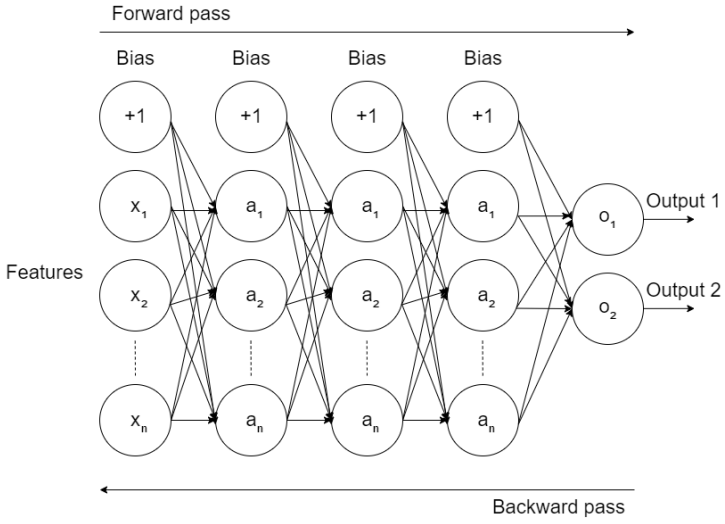

Figure 1 – Neural network architecture

The architecture of the method proposed in this study is presented in Figure 2. The applied methodology makes use of a voice activity detection module to obtain the start

and end timestamps of speech and silence segments detected in the audio samples of the Pitt Corpus subset. From the voice activity timestamps, each audio file is analyzed to perform the extraction of the acoustic characteristics of the identified segments. Then, the automatic emotion recognition module classifies the emotions existing in the patients' speech recordings. Thus, the voice activity time, silence time, number of pauses and total of each emotion recognized in the patients' speech recordings are extracted. To apply the data obtained as input parameters to the neural network, it is necessary to perform the pre-processing of data. Therefore, the categorical labels are transformed into numerical values, and, through a feature scaling algorithm, the parameter values are normalized to the range between zero and one. Finally, based on supervised learning, the neural network is stimulated by the input data and, through an iterative process of adjusting its synaptic weights, learns to classify patients with Alzheimer's disease and differentiate them from patients without the disease.

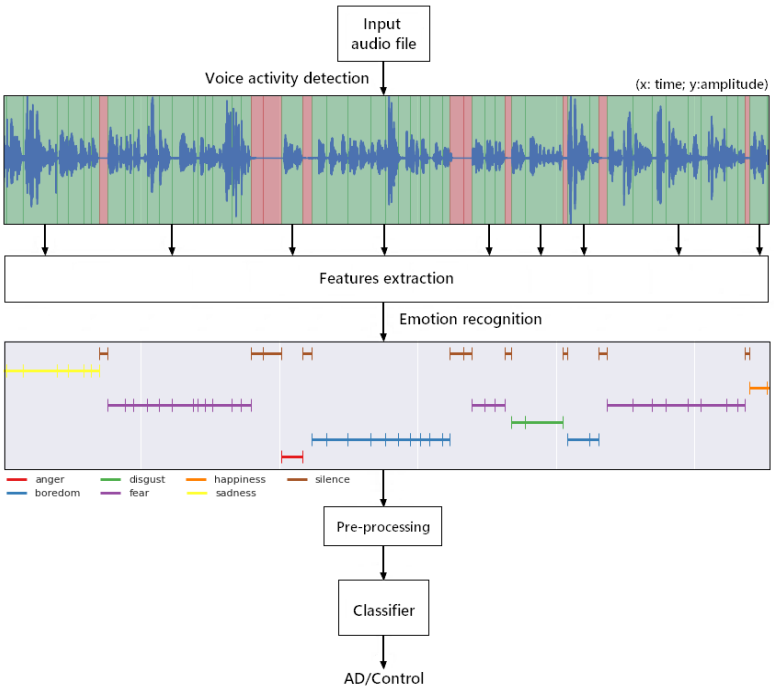

Figure 2 – Experimentation methodology

Results and Discussion

To evaluate the results obtained by the emotion recognition model, labeled audio recordings from the Emo-DB dataset were compared with the classes predicted by the

classifier. As Emo-DB does not define an official training and testing set, a leave-one-group-out cross-validation was applied, where the group of recordings of each actor was validated individually, from the training carried out with the groups' recordings of the other actors. The quantitative metrics for evaluating the model were obtained from the average resulting from the cross-validation of the data. Thus, the emotion recognition model experiment obtained an accuracy of 94.02%, as seen in the confusion matrix in Figure 3. Considering the arithmetic mean of the results obtained by class, the macro precision reached was 94.18% and the macro recall was 93.38%. From these results it is possible to have a view of the performance of the classifier in the recognition of emotions through speech recordings.

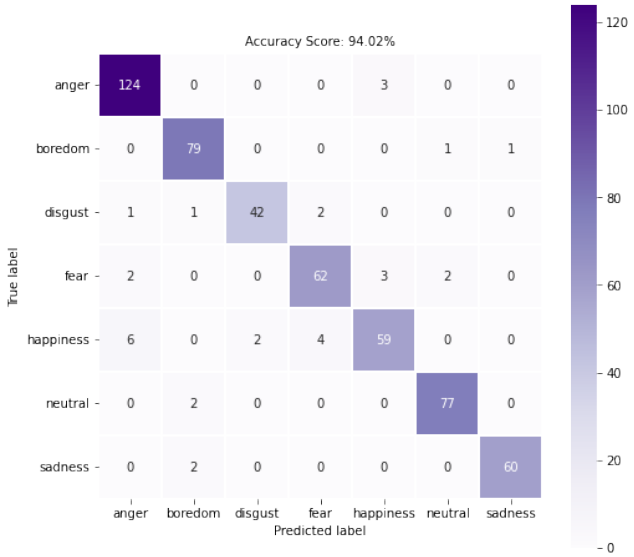

Figure 3 – Emotion recognition model confusion matrix

As input to the neural network developed for the identification of Alzheimer's disease, timestamps of voice activity and emotions recognized in the speech of patients from the Pitt Corpus data subset were used. The evaluation of the classifier results was performed through a 10-fold cross-validation, where the total dataset was divided into ten exclusive subsets of the same size, validating each subset separately, from the training carried out with the other subsets. The neural network developed to classify Alzheimer's disease obtained an accuracy of 72.61%, a macro precision of 72.90% and a macro recall of 72.50%.

In the Healthcare field, a false negative result for the diagnosis of a disease is more critical than a false positive result, as the false negative result can neglect the initiation of the patient's treatment and cause serious problems. Therefore, classification models developed for Healthcare problems must be evaluated by their sensitivity and specificity. The Receiver Operating Characteristic (ROC) curve analysis is a powerful tool for measuring and specifying problems in diagnostic performance in medicine. This analysis allows the study of the variation in sensitivity and specificity for different cut-off values. Also, the Area Under the Curve (AUC) is associated with the discriminating power of a diagnostic test[15]. An excellent model has an AUC close to 1, which means it has a good measure of separability. The model proposed in this article obtained as a result for the AUC the value of 0.764, as can be seen in Figure 4.

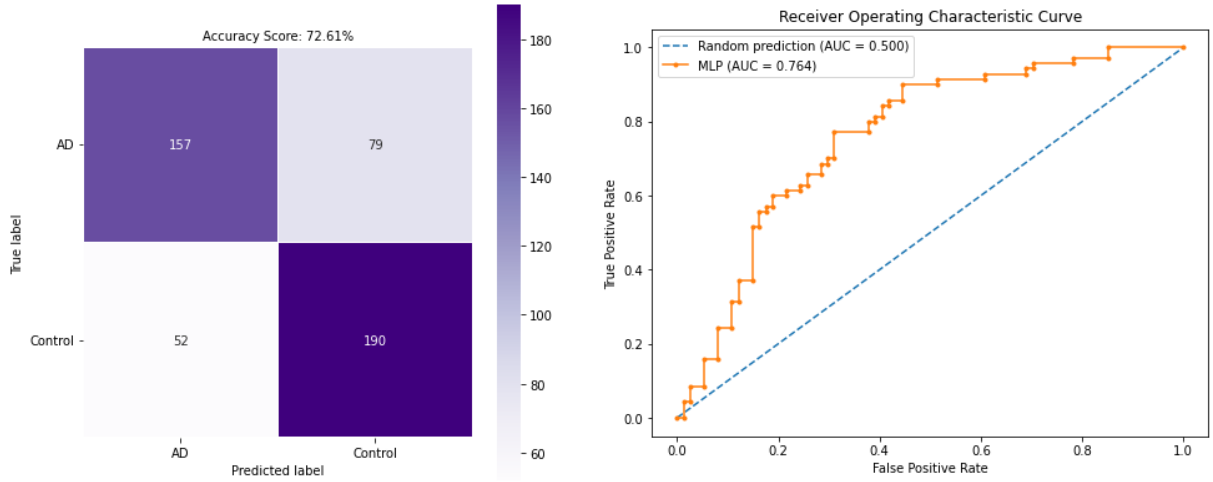

Figure 4 – Metrics of the Alzheimer's disease classifier

The research carried out by Haider *et al.*[8], presents a new approach for the research area by using only the representation of emotions through a fixed-dimension vector to identify Alzheimer's disease. In comparison with the model proposed by the authors, this study differs in that it uses a voice activity detector to delimit the sections where emotions will be classified, as well as mapping the timestamps of recordings to be used, together with emotions, as a parameter for the identification of Alzheimer's disease.

The work developed by Campbell *et al.*[10] proposes the use of multilingual systems based on the processing of the acoustic waveform of the voice, considering the

characteristics of fluency, phonation and voice quality to perform the extraction of acoustic parameters. As the representation of emotions is independent of the spoken language, the analysis of emotional characteristics of speech can be added to this method for optimizing the results obtained by the authors.

Conclusion

The results obtained in this work demonstrate that emotional characteristics represented through speech can be used, regardless of language, as biomarkers for the identification of Alzheimer's disease in patients. Using a multilayer perceptron neural network, it was possible to identify Alzheimer's disease and differentiate patients from the control group of patients with the disease, with an accuracy of 72.61% and AUC value of 0.764.

Machine learning techniques are increasingly present in the medical routine. The development of models that can be integrated into medical practice and that can help health professionals to identify the disease is an open challenge. The proposed model in this study provides new possibilities for the early detection of Alzheimer's disease. Through recording devices, such as mobile or voice assistant microphones, speech can be collected in any medical office in a simple, inexpensive and non-invasive way for the patient. The use of an automated and intelligent test to support medical decisions can reduce time between the diagnosis of the disease and the beginning of the treatment. However, a protocol must be defined to conduce the patient to perform tasks through spontaneous speech, allowing the model to index all subtleties and emotional characteristics present in voice.

In order to resolve the limitations of the research, as future works, the authors intend to verify that the incorporation of emotional characteristics obtained through speech improves the classification methods of Alzheimer's disease. In this way, the combination of the proposed technique with others available in the state of the art will be tested. It is recommended to consider the change in emotional states during speech and a more in-depth temporal analysis. Other acoustic features can also be incorporated into the model to optimize its results. It is also suggested to use other datasets with more samples, especially in other languages, so that the generalization capacity of the proposed model

can be evaluated. Finally, further studies of the validity and reliability of the results obtained through this type of datasets are recommended.

References

1. who.int [Internet]. Ageing - World Health Organization; c2022 [cited 2022 Aug 27]. Available from: https://www.who.int/health-topics/ageing.
2. Gauthier S, Webster C, Servaes S, Morais JA, Rosa-Neto P. World Alzheimer Report 2022: Life after diagnosis: Navigating treatment, care and support. London, England: Alzheimer's Disease International (ADI); 2022.
3. Vizza P, Tradigo G, Mirarchi D, Bossio RB, Lombardo N, Arabia G, et al. Methodologies of speech analysis for neurodegenerative diseases evaluation. International Journal of Medical Informatics. 2019 Feb;122:45–54.
4. Alzheimer's Association. 2022 Alzheimer's Disease Facts and Figures. Chicago: Alzheimer's Association; 2022.
5. Liu L, Zhao S, Chen H, Wang A. A new machine learning method for identifying Alzheimer's disease. Simulation Modelling Practice and Theory. 2020 Feb;99:102023.
6. Sharma P, Sharma A, Fayaz F, Wakode S, Pottoo FH. Biological signatures of Alzheimer Disease. Current Topics in Medicinal Chemistry. 2020 Apr 1;20.
7. Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: A review. Expert Systems with Applications. 2020 Jul 15;150:113213.
8. Haider F, de la Fuente S, Albert P, Luz S. Affective Speech for Alzheimer's Dementia Recognition. In Kokkinakis D, Lundholm Fors K, Themistocleous C, Antonsson M, Eckerström M, editors, LREC: Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID). European Language Resources Association (ELRA). 2020. p. 67-73
9. 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
10. 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.

11. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology; 04-08 de setembro de 2005; Lisboa, Portugal. INTERSPEECH 2005. p. 1517-1520.

12. 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.

13. Husein Z. Malaya, Speech-Toolkit library. Version 1.2.7 [software]. 2020 [cited 2022 Aug 27]. Available from: https://github.com/huseinzol05/malaya-speech.

14. Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Eyben F, Schuller BW. Dawn of the transformer era in speech emotion recognition: closing the valence gap. arXiv:2203.07378v2 [Preprint]. 2022 [cited 2022 Aug 27]: [25 p.]. Available from: https://arxiv.org/abs/2203.07378v2

15. Braga AC, Oliveira P. Diagnostic analysis based on ROC curves: theory and applications in medicine. International Journal of Health Care Quality Assurance. 2003 Jul 1. p. 191-198