Automatic classification of Alzheimer's disease through features extracted from speech recordings

Classificação automática da doença de Alzheimer através de características extraídas de gravações de fala

Clasificación automática de la enfermedad de Alzheimer mediante funciones extraídas de grabaciones de voz

Dieine Estela Bernieri Schiavon[1], Carla Diniz Lopes Becker[2]

1 Master's Student, Federal University of Health Sciences of Porto Alegre – UFCSPA, Porto Alegre (RS), Brazil.
2 Ph.D., Federal University of Health Sciences of Porto Alegre – UFCSPA, DECESA, Porto Alegre (RS), Brazil.

Corresponding author: Dieine Estela Bernieri Schiavon
E-mail: dieineb@ufcspa.edu.br

Abstract

Alzheimer's disease is a progressive neurodegenerative pathology and is among the most common forms of dementia in older people. Changes in memory are common symptoms, and changes in speech and language can be signs of cognitive decline. Intelligent systems have the potential for use as diagnostic support tools. **Objective:** To propose a Convolutional Neural Network model for classifying Alzheimer's disease using features extracted from speech recordings. **Method:** We used speech segments with and without pauses from healthy individuals and those with Alzheimer's disease to extract features and recognize patterns in spectrograms. Model training uses a 5-fold stratified cross-validation method. **Results:** The results showed accuracy, sensitivity, and specificity metrics of 97.37%, 97.04%, and 97.62%, respectively. **Conclusion:** The proposed model presented promising results and could contribute to studying non-invasive biomarkers that detect Alzheimer's disease early.

Keywords: Alzheimer's disease; Deep Learning; Speech Processing

Resumo

A doença de Alzheimer é uma patologia neurodegenerativa progressiva estando entre as formas mais comuns de demência em pessoas idosas. Alterações de memória são sintomas frequentes, e alterações de fala e linguagem podem ser sinais de declínio cognitivo. Os sistemas inteligentes têm potencial para uso como ferramentas de apoio ao diagnóstico. **Objetivo:** Propor um modelo de Rede Neural Convolucional para classificação da doença de Alzheimer utilizando características extraídas de gravações de fala. **Método:** Utilizamos segmentos de fala com e sem pausas de indivíduos saudáveis e com doença de Alzheimer para extrair características e reconhecer padrões em espectrogramas. Para o treinamento do modelo usamos validação cruzada estratificada de 5-*folds*. **Resultados:** Obtivemos métricas de acurácia, sensibilidade e especificidade de 97,37%, 97,04% e 97,62%, respectivamente. **Conclusão:** O modelo proposto apresentou resultados promissores podendo contribuir para o estudo de biomarcadores não invasivos, que detectem precocemente a doença de Alzheimer.

Descritores: Doença de Alzheimer; Aprendizagem Profunda; Processamento de Fala

Resumen

La enfermedad de Alzheimer es una patología neurodegenerativa progresiva y se encuentra entre las formas más comunes de demencia en las personas mayores. Los cambios en la memoria son síntomas comunes y también en el habla y el lenguaje pueden ser signos de deterioro cognitivo. Los sistemas inteligentes tienen potencial como herramientas de apoyo al diagnóstico. **Objetivo:** Proponer un modelo de Red Neuronal Convolucional para clasificar la enfermedad de Alzheimer utilizando características extraídas de grabaciones de habla. **Método:** utilizamos segmentos de habla con y sin pausas de individuos sanos y con enfermedad de Alzheimer para extraer características y

reconocer patrones en espectrogramas. Para entrenar el modelo utilizamos una validación cruzada estratificada de 5-*folds*. **Resultados**: Obtuvimos métricas de precisión, sensibilidad y especificidad del 97,37%, 97,04% y 97,62%, respectivamente. **Conclusión:** El modelo propuesto mostró resultados prometedores y podría contribuir al estudio de biomarcadores no invasivos que detecten tempranamente la enfermedad de Alzheimer.
Descriptores: Enfermedad de Alzheimer; Aprendizaje Profundo; Procesamiento del Habla

Introduction

Alzheimer's disease (AD) and related dementias constitute a significant cause of disability and dependence among older people worldwide and are among the most costly diseases for society. According to the World Health Organization (WHO)[1], in 2019, the estimated global societal cost of dementia was US$1.3 trillion, which is expected to reach US$2.8 trillion by 2030.

Late diagnosis of AD contributes substantially to the cost of treating the disease, mainly due to the high utilization of health services. Early diagnosis can save medical care costs[2] and provide patients and their caregivers with a better quality of life.

Medical research indicates that, although gradual memory loss is the main symptom of AD, language deficits and speech problems are some of the typical clinical symptoms of early cognitive impairment. Initial language characteristics may include difficulty finding words, naming objects, repetition, empty speech without content, vague language, and prolonged pauses.[3]

In recent years, more economical, noninvasive, potentially accurate, and easy-to-collect techniques have been studied to aid in diagnosing AD. These techniques rely on artificial intelligence, specifically deep learning, to classify AD based on linguistic or paralinguistic characteristics extracted from patients' speech.[4,5,6,7,8]

The potential of using speech as a biomarker for AD is based on prospective values, including the ease with which speech can be recorded and tracked over time, the collection being noninvasive, improvements in technologies for speech analysis that have occurred in the last decade driven by advances in artificial intelligence and machine learning, and the fact that speech problems can manifest at different stages of the disease.[8]

This study aims to contribute to research in speech signal processing based on deep learning for the early identification of dementia, evaluating pauses in chained speech for the classification of Alzheimer's disease. The methodology proposed in this study aims to propose a binary convolutional neural network model using speech recordings that can assist healthcare professionals in identifying AD early, enabling improving the patient's quality of life and possibly reducing healthcare costs.

Related Work

Yuan *et al*.[9] presented the pause method to adjust pre-trained models Ernie and Bert (bidirectional encoder representations from transformers). The Ernie model obtained the best accuracy, 89.6%. According to the authors, the proposed method improved accuracy and reduced the variation in model performance. They concluded that the pauses were helpful in the thin adjustment of the pre-trained AD classification models.

The study of Alkenani *et al*.[10] aimed to automatically identify different dementia etiologies, including Alzheimer's disease and mild cognitive impairment, and determine if individuals with initial cognitive decline have language deficits. Some machine learning algorithms in the development of the model are Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The models presented an accuracy between 95 and 98%. The authors concluded that the statistical analysis of lexical-syntax biomarkers showed that linguistic deviations are associated with the neurodegenerative pathologies prodrome, suggesting that language biomarkers can assist in early diagnosis.

Pompili *et al*.[11] intended to provide a more comprehensive characterization of language skills for the automatic identification of narrative description tasks, incorporating pragmatic aspects of speech production. The authors used clustering techniques by dividing data samples into different categories in the space of standards and embeddings, a high-dimensional data conversion process to low-dimension data in the form of a vector in such a way that the two are semantically similar. Classification experiments reached an accuracy of 85.5%. The authors' results confirmed that, by incorporating characteristics that represent aspects of discourse, there is a better characterization of linguistic disabilities in Alzheimer's disease.

Liu *et al*.[12] presented a method for detecting AD using CNN from the acoustic characteristics of spontaneous speech (image description task). The authors used the Data Augmentation Technique, based on minimum modifications in existing data, with the objective of artificially increasing data to deal with sample data scarcity. The accuracy of the authors' model was 82.59% in the DementiaBank database to detect AD using only audio features.

The paper of Gonzalez-Atienza *et al*.[13] used a speech-based binary classifier approach and acoustic and linguistic resources derived from audio recordings and text transcripts. The authors observed that the selection of characteristics with the best results were characteristics based on pauses, incorporations of words, and POS (part-of-speech) categories as nouns and verbs. The model was evaluated separately using two data sets, Pitt Corpus (96%accuracy) and ADReSS (79%accuracy). The authors concluded that from the results found, the approach with linguistic characteristics can distinguish people with Alzheimer's.

Bernieri and Duarte's[14] paper evaluates the use of emotion recognition through speech as a biomarker for classifying Alzheimer's disease. Their proposed method was based on extracting emotional features from speech and pattern recognition using neural network. The results of the experiments reached an accuracy of 72.61%, a precision of 72.90%, and a recall of 72.50%.

The studies covered in this section highlight the potential of CNNs in classifying Alzheimer's disease in various architectures and with different types of input data. Unlike those studies, this work proposes a CNN model with fewer layers and a more economical, noninvasive, and easy-to-collect solution based on features extracted from speech recordings. It can be through recordings made by healthcare professionals in their clinical practice.

Materials and Methods

Datasets

This study uses the Pitt Corpus(15) dataset and Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS).(16) The Pitt Corpus is known to have various linguistic features, cognitive assessments, and clinical information related to dementia, especially Alzheimer's disease. Using Pitt Corpus, we can train and validate our machine learning models on a well-established and widely used dataset. The ADReSS is a dataset derived from the Pitt Corpus.

Pitt Corpus is part of a larger protocol administered by Alzheimer's and Related Dementia Study at Pittsburgh University School of Medicine. A longitudinal study was conducted between 1983 and 1988, with annual collection. Currently, the guardianship is from Dementiabank, which allows free access to researchers through prior email requests, which is the source of this work's data. The dataset has 104 controls, 208 people with possible and likely Alzheimer's disease, and 85 with unknown diagnoses. Pitt Corpus contains data obtained through the following tasks, recorded in MP3 audio and manually transcribed: Cookie Theft, an image presented where the participant must report everything he sees. The Control Group and the dementia group performed this task. The

Verbal fluency and Recall of Stories tasks are only for the dementia group. This study uses only the Cookie Theft task for dementia and control groups. The selected groups contain 236 samples from individuals with AD and 242 samples from individuals without AD. The samples have different durations due to linguistic variations common in free descriptions and also the degree of neurological impairment. In the control group, the variations were from 18 seconds to 145 seconds. In the group with dementia, the duration ranged from 2 seconds to 246 seconds.

The ADReSS is a dataset matched by age and gender to minimize the risk of bias in prediction tasks.[17] Data consist of audio recordings of image descriptions obtained from participants using the Cookie Theft image from the Boston Diagnostic Aphasia Examination (BDAE) [18], wich transcribed and annotated using the CHAT coding system.[19] In the dataset, 4077 audio segments of 10 seconds each are available, already normalized, and treated for noise reduction. Segments are divided into two batches (training and testing). The training set has 1358 speech segments from the control group and 1476 speech segments from individuals with AD, while the test segments contain 1243 distributed in AD and non-AD. The ADReSS dataset was used to increase the data to reduce the possibility of the model's overfitting.

Preprocessing

As the Cookie Theft task requires prior guidance, the initial and final cut of the audios was performed to eliminate the interviewer's speech. The noise was removed using the Noisereduce. [20] Noisereduce is an algorithm in Python that reduces noise in time domain signals such as speech, bioacoustics, and physiological signals. It works by calculating a signal's spectrogram and estimating a noise threshold for each frequency band of that signal/noise. This threshold calculates a mask, which blocks noise below the frequency range threshold.[20]

Normalization was performed at -23 dB LUFS (Loudness Unit Full Scale) using the pyloudnorm[21] algorithm. Pyloudnorm is a Python package software that implements the ITU-R BS.1770 recommendation to measure the perceived intensity of audio signals.[21]

Audios were converted into Mel spectrograms using the librosa[22] library (python package for music and audio analysis). After conversion, the spectrograms were saved in PNG extension for input into the neural network as image format. The images generated in this step were resized to 224 x 224.

To compare the results, we tested the database by removing all pauses from the 478 audios in the Pitt Corpus database, using the librosa library for resampling audios to 22050Hz. The frequency of 30dB was used as a filter, and segments with intensity below this value were removed.

Model Architecture Selection

The proposed model is inspired by EfficientNet-B5 architecture. The EfficientNet Models are based on compound scaling methods.[23] EfficientNet consists of eight architectures based on the Efficient-B0 base model. Variants from EfficientNet-B0 to EfficientNet-B7 have a similar structure; what differentiates them are the size of the input image and the number of parameters that increase with the version. Each has different parameters, from 5.3M to 66M. Version B0 is the base model with fewer parameters and receives an image size of 224 x 224.

Tests were carried out with the EfficientNet-B1 to B7 architectures to choose the architecture. We used the same number of layers as the proposed model, changing only the values of the filters in the convolutional layers and the neurons in the dense layer. For this screening, we used the same image input value (224, 224, 3) for all architectures. We implemented the tests to verify the architecture with the best precision and the highest number of true positives (recall). The screening results are in Table 1.

EfficientNetB5, with its 30.6 million parameters[24], was chosen for our purpose. It emerged as the most suitable option, demonstrating the best sensitivity and accuracy in our architectural pre-tests. The EfficientNet-B5 was used as an architectural basis for the proposed model and the transfer learning task. We added three max_pooling2d layers to reduce the total number of parameters. In addition, we include one dense inner layer with RELU activation functions with 2048 neurons. Finally, a dense output layer contains an output unit for binary classification with a sigmoid activation function. We chose the Adam optimizer, loss binary_crossentropy, and accuracy metric to compile the model.

Table 1 – EfficientNet´s Family – Screening Performance on Validation

Model	Accuracy (%)	Loss	Recall (%)
EfficientNet-B1	87.72	0.2898	49.03
EfficientNet-B2	86.65	0.3000	83.46
EfficientNet-B3	84.38	0.4163	49.04
EfficientNet-B4	90.50	0.3081	54.29
EfficientNet-B5	**90.18**	**0.2755**	**81.50**
EfficientNet-B6	91.11	0.3626	9.76
EfficientNet-B7	91.11	0.4131	68.76

Training and Validation

We split the dataset into two batches: training (80%) 3644 samples for the 5-fold stratified cross-validation (CV) and (20%) 911 for testing. Continuous monitoring of the model's performance on the validation set is conducted to mitigate overfitting. The performance training uses a CV method, with 30 epochs in each fold. The batch size used was 64, and the optimizer Adam's learning rate was set to 0.00001 for all experiments.

Evaluation Metrics

Model evaluation is indicated to measure the algorithm's generalization capacity. It is carried out through performance metrics, which evaluate how well the model correctly classifies the classes. We used the following evaluation metrics: accuracy, precision, sensitivity, specificity, and the ROC (Receiver Operating Characteristics) curve to analyze the performance of the proposed CNN.

Results and Discussion

Results With Pauses

After defining the architecture, cross-validation was performed in several experiments until the model reached a satisfactory result. We employed the repeated validation technique to increase accuracy and sensitivity using the best-performing fold weights (Table 2). After this retraining step, our model achieved a validation accuracy of 91.91% and a sensitivity of 98.03%, a high percentage in identifying individuals with AD using the simple proposed model. The area of the ROC curve was 0.90.

Table 2 – Training accuracy, training loss, validation accuracy, and validation loss for each fold of the proposed model

Fold	Training Acc (%)	Training Loss	Validation Acc (%)	Validation Loss
Fold 1	72.16	0.5099	58.62	0.7650
Fold 2	98.03	0.0643	91.70	0.3190
Fold 3	99.59	0.0059	99.81	0.0142
Fold 4	**99.84**	**0.0022**	**99.74**	**0.0103**
Fold 5	99.76	0.0024	99.73	0.0102
Average	93.88	0.1169	89.92	0.2237

The last step is the use of transfer learning. In the pre-trained network, learning occurs by transfer. Transfer learning is a method of reusing pre-trained model knowledge for another task. It can be used for classification, regression, and clustering problems.[25] We used the pre-trained EfficientNet-B5 network for the transfer learning task. The EfficientNet-B5 model was trained with an image size of 456 x 456 pixels, as this is the default size used by the model. 28,513,527 parameters were used in training, where 172,743 are non-trainable parameters obtained from pre-training with Imagenet and 28,340,784 are trainable parameters. For the image size, we resized it to 224 x 224, and it is not necessary to use the original size (456 x 456). The transfer learning task was divided into two steps: first, 5-fold cross-validation is performed. Then, in step 2, we retrain the model with the best weight (lower validation loss) cross-validation fold with a split holdout and test the model with the reserved test data (911 samples).

The evaluation model on the test data resulted in metrics of accuracy, recall, precision, specificity, and F-score of 97.37%, 97.04%, 97.04%, 97.62%, and 97.04%, respectively, and a ROC curve of 0.97. These results improved by 5.46% (accuracy) after applying the transfer learning technique, except for the recall, which decreased by around 1%. The learning curves are in Figure 1, and the final confusion matrix of the test and the ROC curve are shown in Figure 2.

Figure 1 — Training and Validation Learning Curves

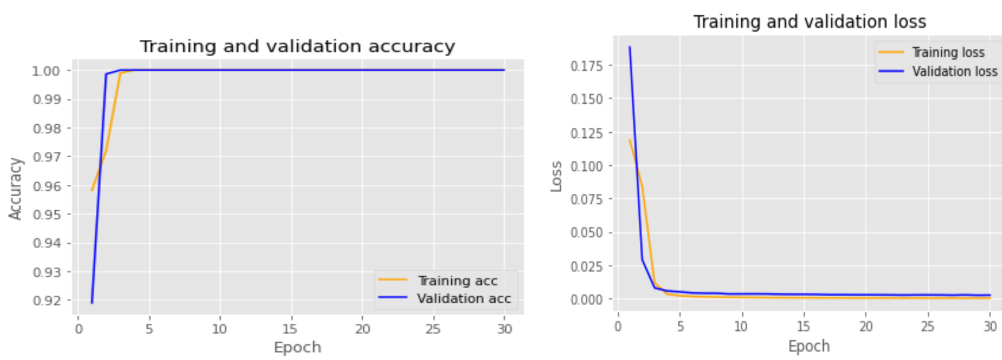
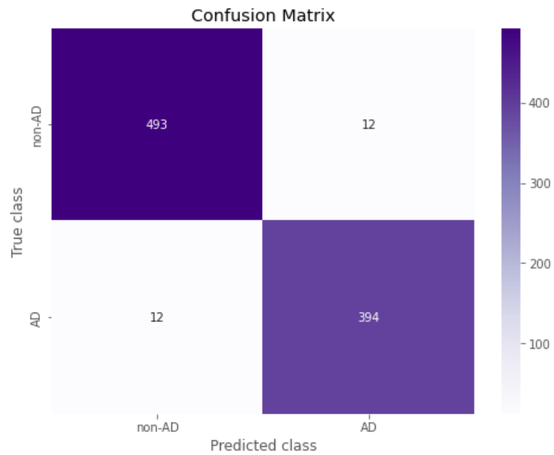
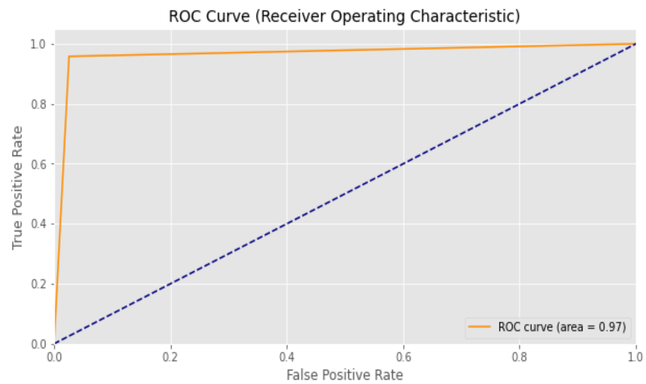


Figure 2 — Confusion Matrix and ROC Curve Metrics of the Alzheimer's disease classifier





Results With Pauses Removal

We used the same databases and methodology described so far to evaluate the model with the Mel spectrograms without pauses.

After the retraining step, validation accuracy reached 90.26%, recall 85.71%, and ROC Curve 0.91. As with pauses data, the transfer learning task was divided into two steps: first, 5-fold cross-validation is performed. Then, in step 2, we retrain the model with the best weight (lower validation loss) cross-validation fold with a split holdout and test the model with the reserved test data.

The evaluation model on the test data without pauses results in metrics of accuracy, recall, precision, specificity, and F-score, 95.39%, 98.03%, 92.13%, 93.27%, and 94.09%, respectively, and a ROC curve of 0.96. After using the transfer learning technique, the metrics demonstrate an increase in the percentages, mainly of accuracy (5.13%) and recall (12.33%).

Discussion

In this part, we compare the proposed model results in two ways: first, we compare the hypothesis raised in the study with similar studies described in the related work section.

Comparing the results of the two hypotheses, the data with pauses were better in practically all metrics. The recall was slightly better in the model without pauses, but data with pauses showed a significantly better result for precision. We agree with the study of Yuan *et al*.[9]; the authors concluded that the pauses helped fine-tune the pre-trained AD classification models, as we did.

The proposed model in this study reached higher evaluation metrics than studies [9,11,12,13,14] using only Convolutional Neural Networks. However, Alkenani *et al*.[10] also show a relevant accuracy of 98% but use the multilayer perceptron.

In our study, we realized that we needed some help because we only had a limited number of samples (478 from the Pitt-Corpus database), so we expanded with the ADReSS database, with an additional 4077 samples. As we did not find references in the literature to using the two concatenated databases, we decided to increase the data in our study with the ADReSS.

In the related works where we compare our results, we find several more conventional machine learning methods, such as Embeddings, Support Vector Machine (SVM), Naive Bayes (NB), or Convolutional Neural Networks (CNN), together with other approaches.

Working with raw audio in a CNN requires much more time to preprocess the data. Standardizing this data type for CNNs, which require standardized information for data entry, may take some time. The alternative of using spectrograms simplified the preprocessing burden, as only a few functions found in currently available Python libraries for machine learning are needed to transform an audio file into a Mel spectrogram. Furthermore, the results of models with CNNs using spectrograms are considerable.

In summary, we consider promising the EfficientNet-B5 architecture for binary classification of Alzheimer's disease and the ADReSS and DementiaBank databases together due to the low number of Pitt Corpus data for use in Convolutional Neural Networks. However, in future work, statistical tests can be considered for a more robust approach in addition to standard performance metrics in machine learning.

For this study, pauses in the connected speech were shown to be more efficient, both in pattern recognition in spectrogram images for the classification of Alzheimer's disease and without Alzheimer's disease.

Conclusion

This article proposes a Convolutional Neural Network to support the diagnosis of Alzheimer's patients. The proposed method implements a CNN architecture and was tested using cross-validation. The authors point out essential limitations regarding the number of data available for analysis. The experience and knowledge acquired with Convolutional Neural Networks state that performance will increase, avoiding overfitting, with a considerable increase in the number of samples available for training. Alternatively, networks pre-trained on large datasets help improve results when the database is small.

The authors encourage further research to consider pauses in connected speech in Alzheimer's patients and the use of a Convolutional Neural Network for classification. The initial results studied here were promising and could be explored in future studies on non-invasive tests that detect Alzheimer's disease early. However, more studies and tests are essential to consolidate as reliable tools for classifying Alzheimer's disease and exploring their integration into healthcare.

References

1. World Health Organization. Dementia [Internet]. World Health Organization. 2023 [cited 2023

Dec 15]. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia.

2. Guo Y, Li C, Roan C, Pakhomov S, Cohen T. Crossing the “Cookie Theft” Corpus Chasm: Applying What BERT Learns From Outside Data to the ADReSS Challenge Dementia Detection Task. Frontiers in Computer Science. 2021 Apr 16.

3. Yang A, Liu C, Wu J, Kou X, Shen R. A review on α-mangostin as a potential multi-targetdirected ligand for Alzheimer’s disease. European Journal of Pharmacology 2021 Ap;897:173950.

4. Mahajan P, Baths V. Acoustic and Language Based Deep Learning Approaches for Alzheimer’s Dementia Detection From Spontaneous Speech. Frontiers in Aging Neuroscience 2021 Feb 5; 13.

5. Haulcy R, Glass J. Classifying Alzheimer’s Disease Using Audio and Text-Based Representations of Speech. Frontiers in Psychology [Internet]. 2021 Jan 15 [cited 2023 Dec 15];11. Available from: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.624137.

6. Chlasta K, Wołk K. Towards Computer-Based Automated Screening of Dementia Through Spontaneous Speech. Frontiers in Psychology [Internet]. 2021 Feb 12 [cited 2023 Dec 15];11. Available from: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.623237.

7. Laguarta J, Subirana B. Longitudinal Speech Biomarkers for Automated Alzheimer’s Detection. Frontiers in Computer Science [Internet]. 2021 Apr 8 [cited 2023 Dec 15];3. Available from: https://www.frontiersin.org/articles/10.3389/fcomp.2021.624694.

8. de la Fuente Garcia S, Ritchie CW, Luz S. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer’s Disease: A Systematic Review. Journal of Alzheimer’s Disease. 2020 Dec 8; 78(4):1547–74.

9. Yuan J, Cai X, Bian Y, Ye Z, Church K. Pauses for Detection of Alzheimer’s Disease. Frontiers in Computer Science [Internet]. 2021 Jan 29 [cited 2023 Dec 15];2. Available from: https://www.frontiersin.org/articles/10.3389/fcomp.2020.624488.

10. Alkenani AH, Li Y, Xu Y, Zhang Q. Predicting Prodromal Dementia Using Linguistic Patterns and Deficits. IEEE Access [Internet]. 2020 [cited 2023 Dec 15]; 8:193856–73. Available from: https://ieeexplore.ieee.org/abstract/document/9218925.

11. Pompili A, Abad A, de Matos DM, Martins IP. Pragmatic Aspects of Discourse Production for the Automatic Identification of Alzheimer’s Disease. IEEE Journal of Selected Topics in Signal Processing [Internet]. 2020 Feb 1;14(2):261–71. Available from: https://ieeexplore.ieee.org/document/8963723.

12. Liu Z, Guo Z, Ling Z, Li Y. Detecting Alzheimer’s Disease from Speech Using Neural Networks with Bottleneck Features and Data Augmentation [Internet]. IEEE Xplore. 2021. p. 7323–7.

Available from: https://ieeexplore.ieee.org/document/9413566.

13. Gónzalez Atienza M, González López JA, Peinado AM. An Automatic System for Dementia Detection using Acoustic and Linguistic Features. digibugugres [Internet]. 2021 Jan 28; Available from: https://digibug.ugr.es/handle/10481/66645.

14. Bernieri G, Duarte JC. Identifying Alzheimer's Disease Through Speech Using Emotion Recognition. Journal of Health Informatics [Internet]. 2023 Jul 20 [cited 2023 Dec 23];15(Especial). Available from: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1093.

15. Becker JT. The Natural History of Alzheimer's Disease. Archives of Neurology. 1994 Jun 1;51(6):585.

16. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. arXiv:200406833 [cs, eess, stat] [Internet]. 2020 Aug 5 [cited 2023 Dec 15]; Available from: https://arxiv.org/abs/2004.06833.

17. Luz S, Haider F, de la Fuente Garcia S, Fromm D, MacWhinney B. Editorial: Alzheimer's Dementia Recognition through Spontaneous Speech. Frontiers in Computer Science [Internet]. 2021 Oct 21 [cited 2023 Dec 12];3. Available from: https://www.frontiersin.org/articles/10.3389/fcomp.2021.780169.

18. Goodglass H, Kaplan E, Weintraub S. BDAE: The Boston Diagnostic Aphasia Examination. Philadelphia, PA: Lippincott Williams & Wilkins; 2001. 2001.

19. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: The ADReSSo Challenge. arXiv:210409356 [cs, eess] [Internet]. 2021 Mar 22 [cited 2023 Dec 23]; Available from: https://arxiv.org/abs/2104.09356.

20. Sainburg T. timsainb/noisereduce: v1.0. zenodoorg [Internet]. 2019 Jun 11 [cited 2023 Dec 23]; Available from: https://zenodo.org/record/3243139.

21. Steinmetz CJ, Reiss J. pyloudnorm: A simple yet flexible loudness meter in Python [Internet]. Audio Engineering Society Convention 150. Audio Engineering Society; 2021 [cited 2023 Dec 23]. Available from: https://csteinmetz1.github.io/pyloudnorm-eval/paper/pyloudnorm_preprint.pdf.

22. McFee B, Metsai A, McVicar M, Balke S, Thomé C, Raffel C, et al. librosa/librosa: 0.9.2 [Internet]. Zenodo. 2022. Available from: https://zenodo.org/record/6759664#.Y6n8x-zMK00.

23. Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [Internet]. proceedings.mlr.press. PMLR; 2019 [cited 2023 Dec 23]. p. 6105–14. Available from: https://proceedings.mlr.press/v97/tan19a.html?ref=jina-ai-gmbh.ghost.io.

24. Team K. Keras documentation: Keras Applications [Internet]. keras.io. [cited 2023 Dec 23].

Available from: https://keras.io/api/applications.

25. Tammina S. Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications Volume 9, Issue 10, ISSN 2250-3153; 2019.