



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

An automatic Alzheimer's disease classifier based on spontaneous spoken English

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Bertini, F., Allevi, D., Lutero, G., Calzà, L., Montesi, D. (2022). An automatic Alzheimer's disease classifier based on spontaneous spoken English. *COMPUTER SPEECH AND LANGUAGE*, 72, 101298-101307 [10.1016/j.csl.2021.101298].

Availability:

This version is available at: <https://hdl.handle.net/11585/833418> since: 2021-09-25

Published:

DOI: <http://doi.org/10.1016/j.csl.2021.101298>




Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

An Automatic Alzheimer’s Disease Classifier Based on Spontaneous Spoken English

Flavio Bertini ^{a,*}, Davide Allevi^a, Gianluca Lutero^a, Laura Calzà ^{b,c},
Danilo Montesi ^a

^a*Department of Computer Science and Engineering, University of Bologna, Mura Anteo
Zamboni 7, Bologna, Italy*




^b*Interdepartmental Centre for Industrial Research in Health Sciences and Technologies,
University of Bologna, Via Tolara di Sopra 41/E, Bologna, Italy*

^c*Department of Pharmacy and Biotechnology, University of Bologna, Via Belmeloro 6,
Bologna, Italy*

Abstract

According to the World Health Organization, the number of people suffering from dementia worldwide will grow to 150 million by mid-century, and Alzheimer’s disease is the most common form of dementia contributing to 60-70% of cases. The problem is compounded by the fact that current pharmacologic treatments are only symptomatic, and therapies are ineffective in slow down or cure the degenerative process. An automatic and standardize classifier for Alzheimer’s disease is thereby extremely important to rapidly respond and deliver as preventive as possible interventions. Speech alterations might be one of the earliest signs of cognitive defect and, recently, the researchers showed that they can be observable well in advance other cognitive deficits become manifest. In this paper, we propose a full automated method able to classify the spontaneous spoken production of the subjects. In particular, we trained an artificial neural network using the spectrogram of the audio signal, which is the visual representation of the speech of the subject. Moreover, to overcome the problem

*Corresponding author

Email addresses: flavio.bertini2@unibo.it (Flavio Bertini ) ,
davide.allevi@studio.unibo.it (Davide Allevi), gianluca.lutero@studio.unibo.it
(Gianluca Lutero), laura.calza@unibo.it (Laura Calzà ) , daniilo.montesi@unibo.it
(Danilo Montesi )

of the large amount of annotated data usually required for training deep learning models, we used a specific data augmentation approach that avoids distorting the original samples. We evaluated the proposed method using the English *Pitt Corpus* from *DementiaBank*. The used dataset consists of 180 subjects: 43 healthy controls and 137 Alzheimer’s disease patients. The proposed method outperformed the other approaches in the literature based on manual and semi-automatic transcription and annotation of speech, improving the classification capability by 5.93%, and obtained good classification results compared to the state-of-the-art neuropsychological screening tests (i.e., the Mini-Mental State Examination and the Activities of Daily Living portion of the Blessed Dementia Rating Scale) exhibiting an accuracy of 93.30% and an F1 score of 88.50%.

Keywords: Alzheimer’s disease, speech analysis, speech classification, data augmentation, autoencoder neural networks

1. Introduction

The rising elderly population is one of the main demographic characteristics of developed countries, and the phenomenon is leading to the emergence of various age-related issues. Dementia is one of the most increasing pathologies
5 and one of the major causes of disability and loss of self-sufficiency among older people. The World Health Organization recognized it as a public health emergency [1], with an estimated 50 million people affected by the disease and nearly 10 million new cases every year worldwide. The economic implications of dementia in terms of direct medical and social care costs are one of the
10 big challenges for health-care systems. The increase of 35% of the costs from 2010 to 2015 has led to an enormous sum that was similar in magnitude to the GDP of countries with medium/large economies, such as the Netherlands [2]. Alzheimer’s disease is one of the most common forms of dementia, which is a chronic syndrome that includes different diseases that affects cognitive abilities
15 to perform everyday activities. Moreover, it impacts family members, caregivers, and society at large. Dementia’s progressive nature is commonly preceded by

deterioration in emotional control, social behaviour, or motivation.

The symptoms linked to dementia can be manifested at different severity levels, and most people undergo a gradual cognitive decline. In particular, 20 Alzheimer’s disease is a neurodegenerative pathology that develops years before clinical manifestations and typically preceded by prodromal stages such as mild cognitive impairment. It is severe enough to be assessed with neuropsychological assessment and to heavily interfere with everyday activities [3]. In [4], the researchers estimated that 70% of early-stage diagnosed dementia subjects 25 progressed to a severe stage with an annual conversion rate from 10% to 15% clinic sample [5]. Even though the researchers have identified several risk factors that affect the likelihood of developing dementia, epidemiological studies have shown that people adopting a better lifestyle have a reduced risk of dementia symptoms [6]. However, no disease-modifying therapies are available for demen- 30 tia, and current treatments are only symptomatic for memory and psychiatric symptoms. Thus, similarly to other pathologies for which there is no cure, such as frailty condition [7], prompt detection is a key challenge to promote early and optimal management of cognitive decline. Furthermore, it has recently become clear that the need for fast and remote digital health assessment tools is 35 of utmost importance during extreme events, such as pandemic diseases, during which the older population is most vulnerable and fragile.

Despite the lack of effective treatments raises both theoretical issues and ethical concerns even about the “detection” [8], the early diagnosis of cognitive decline is a challenging topic. However, the implementation of preventive 40 measures requires to have psychometric tests, with high accuracy, low-cost and suitable for large-scale use. An adequate and timely risk identification might also reduce the economic impact of health spending and the emotional burden for patients and their caregivers. In particular, subjects affected by dementia manifest cognitive alterations in various domains: memory, attention, executive 45 functioning, visuospatial skills, perceptual speed and language also [9].

The diagnosis of prodromal dementia is currently challenging [10] and it was usually done via traditional pen-and-paper screening tests, such as the Mini-

Mental State Examination and the Montreal Cognitive Assessment [11]. However, language has been recently subjected to growing interest, and literature suggests that language impairment is a promising sign to reveal early signs of cognitive decline [12]. In particular, the analysis of spoken production is an in-expensively and ecologically approach to identify alterations related to cognitive functionalities. In literature, several studies obtained good results in cognitive impairment detection using different language features, such as speech errors [13], acoustic features [14], lexical features [15] and a combination of lexical, acoustic, and syntactic features [9]. However, most of the proposed methods require several manual activities, such as transcription, annotation and correction, that results in a prone-to-error, non-standardised and time-consuming approach with the potential loss of useful information reducing the scalability of the screening.

In this paper, we propose an automatic method for classifying potential Alzheimer’s disease patients using their spontaneous spoken English. The method has proven effective results and is based on a specific type of neural networks, that is autoencoder, trained using the visual representation (i.e., the spectrogram) of the audio signal of the subjects. Typically, the autoencoders are used for unsupervised learning of data coding. Firstly, through dimensional reduction (i.e., encoding), the autoencoder learns a representation of the data (i.e., code) and then, in a reconstruction stage (i.e., decoding), it uses the reduced features to generate an outcome as close as possible to the original input. Although it seems that the only purpose of an autoencoder is to copy the input to the output, the encoded representation allows performing different types of tasks, such as dimensionality reduction, image denoising, and anomaly detection to name a few. Among the different types of autoencoders, we used a type of recurrent neural networks, that is *auDeep* [16], whose aim is the unsupervised feature extraction from audio data.

Deep learning models usually require a large amount of annotated data, but unfortunately in some health contexts, and certainly in our study, it is not possible to collect new and large amounts of data. Data augmentation and transfer

learning represent two different approaches to solve this problem. The first in-
80 creases the amount of data adding slightly modified copies of already existing
one, while the latter leverages the information learnt from another dataset/task.
To allow the use of a classifier based on neural networks, we adopted a data aug-
mentation approach to enlarge the size of the input dataset [17]. In particular,
through three different operations, that do not distort the information content
85 useful for Alzheimer’s disease detection, each log mel spectrogram¹ has con-
tributed to increasing the number of inputs. This approach does not require
collecting further input data, and it is computationally cheaper compared to
methods based on audio deformation that require more complex operations on
the audio waveform. Moreover, this makes it possible to reuse small datasets
90 collected more than thirty years ago, such as the one used in this study for the
evaluation: the English *Pitt Corpus* from *DementiaBank* [18].

The strengths of our method include the detection of potential Alzheimer’s dis-
ease subjects, the capability to automatically process the audio files to classify
potential patient, and the possibility to standardize the screening phases avoid-
95 ing manual analysis that may lead to unfair and non-uniform evaluations. In
the last decade, deep learning models have been becoming more and more com-
plex and resource-demanding, and research is orienting towards smaller, faster,
and more efficient models. The proposed model outperforms more complex
deep learning models and requires much less training data, has a short learning
100 time and does not require the fine-tuning of a large number of parameters. In
particular, in comparison to deep learning approaches, it does not require any
pre-trained phase on a large dataset. To the best of our knowledge, the pro-
posed method is the first attempt in the English language to fully automate
processing the audio file without manual or semi-automatic transcription and
105 feature extraction steps. The proposed approach outperforms the other meth-

¹In the log mel format of the spectrogram, the horizontal axis represents the time in linear
scale, the vertical axis represents the frequency in logarithmic scale, and the intensity is
colour-coded.

ods in the literature achieving classification accuracy of 93.30%, which improves the state-of-the-art by 5.93%. The promising results confirm the strength of the linguistic approach and the proposed method allows easy scalability.

The rest of the paper is organized as follows. In Section 2, we review the literature on methods to detect language disorder in Alzheimer’s disease subjects. Section 3 summarises the main characteristics of the English *Pitt Corpus* from *DementiaBank*. We present our automatic method for speech classification for Alzheimer’s disease in Section 4, including the data augmentation approach used to overcome the dataset size problem. In Section 5, we discuss the results of our method, comparing them with state-of-the-art manual and semi-automatic methods based on transcription and annotation of speech. Some concluding remarks are made in Section 6.

2. Related Works

Despite there is an extensive literature that confirms the worth of linguistic features in detecting health-related issues [19], in Section 2.1, we discuss the available studies on dementia classifiers based on speech analysis. Then, we describe various approaches proposed for data augmentation for audio files in Section 2.2.

2.1. Speech Analysis for Dementia Detection

Recent studies have confirmed that language assessments provide an effective, simpler and economic approach [20] to detect earliest signs of cognitive defect and dementia. In particular, automatic speech analysis based on natural language processing, speech recognition, and machine learning techniques can provide objective and fast diagnostic results [21]. The features of the spoken language that may aid in dementia detection can be classified into three different classes: morphological, syntactic and phonological. Researchers have largely investigated morphological features. In [22], the authors used different types of morphological features, such as the number and rate of distinct lemmas,

the number and rate of nouns, verbs, adjectives, pronouns and conjunctions and
135 the number of the first person singular verbs in distinguishing dementia patients
from healthy controls. Where in [23] and [24], the authors found verbs play an
important role especially in a small size corpus, in particular, the frequency of
the verb, the proportion of main clauses with nonfinite or finite verbs, the counts
of nouns, verbs, noun-verb ratio are statistically significant features. Syntac-
140 tic features were explored in [9] and [25] showing that dementia patients tend
to produce shorter and less complex sentences and that syntactic factors may
vary among different patients. Phonological features, such as articulation rate,
speech tempo, hesitation ration, silent pause, have been explored more recently.
In [26], the authors found that voiceless segments produced by patients affected
145 by dementia were highly correlated with fluency. Whereas, a classifier based on
the total duration of 's' phoneme, the pseudo-syllable rate, the average pause
duration, the total count of 'm' phoneme was proposed in [27]. In another study,
the authors showed that dementia patients produce longer vowels during text
reading tasks [28]. Linguistic features are usually used with learning models
150 to facilitate and automate the diagnosis of dementia and researchers explored
different approaches based on support vector machine [29, 30, 31], neural net-
works [32, 33], random forest [34, 35] and Naive Bayes [36] classifiers. Typically,
all these approaches require a combination of different features, a significant
amount of manual processing to extract and clean the features or use large neu-
155 ral networks with many parameters that, consequently, require a lot of labelled
data and training time.

2.2. Audio Data Augmentation

Neural network models usually require a large amount of data for train-
ing, improving the accuracy, and avoiding overfitting. Data augmentation is a
160 technique to increase the training dataset -when extra annotated data is not
available- through slightly modified copies of existing data or newly created
synthetic data. Researchers have explored different techniques for audio data
augmentation. In [37], the authors investigated three distortion methods, that

is vocal tract length distortion, speech rate distortion, and frequency-axis ran-
165 dom distortion to artificially augment training samples. Whereas in [38], the
authors proposed a method to transform the spectrograms using a random lin-
ear warping along the frequency dimension. A different approach involves the
superimposing of a generated noise signal to the original audio [39] or the mixing
of the original audio signal with music and TV/movie audio [40]. Whereas, a
170 method that changes the speed of the audio signal was proposed in [41]. In [42],
the authors used an acoustic room simulator to generate simulated audio data
for speech recognition. The *SpecAugment* method proposed in [17] is simple and
computationally cheap and operates on the log mel spectrogram of the input au-
dio. In particular, the proposed augmentation operations, inspired by computer
175 vision approaches, allow keeping the audio features robust to deformations in
the time direction and partial loss of frequency information and partial loss of
small segments of speech.

3. Pitt Corpus Audio Dataset

We evaluated the proposed method using the *Pitt Corpus*² from *Dementia-*
180 *Bank*. The dataset was gathered longitudinally between 1983 and 1988 as part
of the Alzheimer Research Program at the University of Pittsburgh [18]. The
study initially enrolled 319 participants according to the following eligibility cri-
teria: all the participants were required to be above 44 years old, have at least
seven years of education, have no history of major nervous system disorders,
185 and have an initial Mini-Mental State Examination score above 10. Finally,
the cohort consisted of 282 subjects and the demographic characteristics shown
that the selected patients were older, less well educated, and more likely to be
women than control subjects. In particular, the cohort included 101 healthy
control subjects (HC) and 181 Alzheimer’s disease subjects (AD). An extensive

²The dataset is publicly available and more information to gain access to the dataset can
be found here: <https://dementia.talkbank.org>.

Table 1: Characteristics of the cohort.

	HC subjects	AD subjects
	<i>(n=101)</i>	<i>(n=181)</i>
Age (<i>years</i>)	63.80 ± 8.30	71.40 ± 8.30
Gender (<i>F/M</i>)	57/44	121/59
Education (<i>years</i>)	14.30 ± 2.90	12.10 ± 2.90
MMSE	29.10 ± 1.10	18.40 ± 5.20
ADLB	0.19 ± 0.95	6.97 ± 4.10

190 neuropsychological assessment was conducted on the participants, including verbal tasks, the Mini-Mental State Examination (MMSE) and the Activities of Daily Living portion of the Blessed Dementia Rating Scale (ADLB). Table 1 provides the main characteristics of the cohort. A comprehensive description of the cohort building process can be found in Becker et al. [18].

195 In addition to the standard cognitive evaluation, all subjects were requested to record their spontaneous speech induced by three different tasks: *the description of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination* (the picture shows a familiar domestic scene in a kitchen) [43], *a word fluency task*, *a sentence construction task*, and *a story recall task* using two stories “Uncle Bill and Johnny” and “George and Melanie” [44]. The whole *Pitt Corpus* dataset includes both the audio files and the transcriptions. However, the main problem with using the entire dataset is that the *Pitt Corpus* is highly unbalanced. In particular, only the Cookie-Theft picture description test was administered to all the subjects (i.e., HC subjects and AD subjects),
 200 which means that the classifier may suffer from a learning bias. For instance, a *naïf* classifier may recognize an AD subject by assessing whether the audio does not relate to the description of the Cookie-Theft picture. Thus, we specifically selected a subset of 180 subjects (43 HC subjects and 137 AD subjects)
 205 to avoid possible learning biases due to the characteristics of the dataset, while

210 preserving the statistical characteristics of the original cohort. The length of the resulting audio files varies between approximately 16 seconds to 1 hour. Due to the required input format of the selected deep neural network, the audio files were converted in *.wav* format using the free audio converter tool *fre:ac*³.

4. Alzheimer’s Disease Classifier

215 In this section, we describe the automatic Alzheimer’s disease classifier based on spontaneous spoken English. In particular, we firstly discuss the data augmentation technique adopted. This is important as it will help in understanding the reasons behind the selection of a particular methodology. Then, we present the architecture of the classifier based on a deep recurrent neural network and
220 a multilayer perceptron.

4.1. Data Augmentation Technique

The size of the training dataset has a significant impact on the performance of a deep learning model, however, collecting new and labelled data is very time-consuming or, as in this case, completely impossible. Data augmentation
225 is a state-of-the-art approach to enlarge the training dataset by adding slightly modified copies or newly synthetic data derived from the existing one, for instance, augmentation operations for images include rotation, mirroring, translation, noise overlap, and hue and saturation adjustment.

There are several data augmentation approaches for audio files in the literature,
230 however, because signs of cognitive defect may be reflected in different speech nuances, we excluded techniques that heavily distort the original audio samples introducing undesired artefacts. We were interested in the intrinsic audio features, such as temporal, rhythmic and acoustic features, characterizing Alzheimer’s disease, thus we could not afford to select data augmentation
235 techniques introducing audio artefact in the learning process that could lead to misclassification, such as voice tone alteration, speech rate alteration and speech

³<https://www.freac.org>

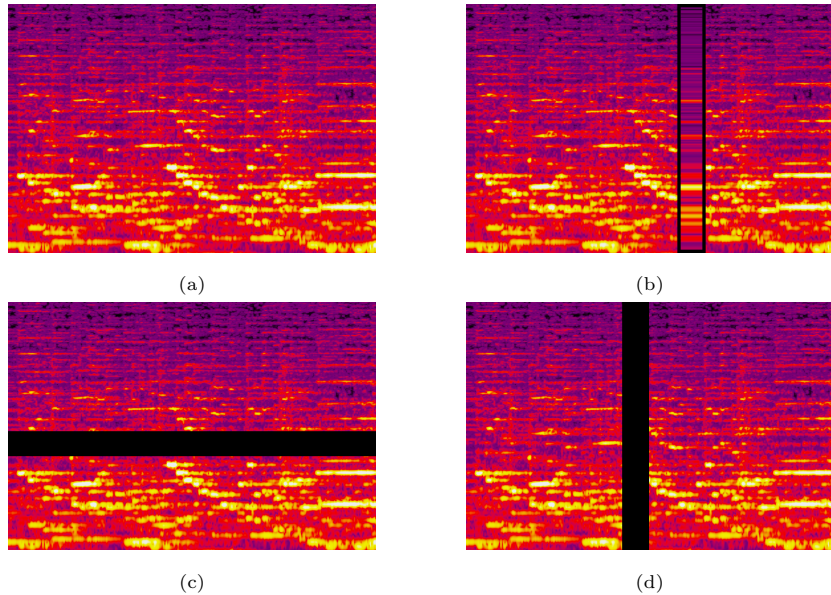


Figure 1: The original log mel spectrogram in (a) and with the time/frequency operations applied (b), (c), and (d). The black box and band show the time warp region in (b), the frequency masking region in (c) and the time masking region in (d).

speed alteration. In practice, we used the *SpecAugment* suite presented in [17] that transforms the log mel spectrogram to increase the number of input data points. In particular, *SpecAugment* consists of the three operations shown in
 240 Figure 1, that is the time warping, that shifts the spectrogram in time in a random direction (Figure 1b); the frequency masking, that masks a random slice of frequencies steps (Figure 1c); and the time masking, that masks a random slice of times steps (Figure 1d). Given the log mel spectrogram as an image where the horizontal and vertical axes represent the time and the frequency,
 245 respectively, the three operations operate as follows:

- *Time warping* operation randomly selects the slice to be warped fixing a point along the horizontal axis within the interval $(W, \tau - W)$, where τ represents the time steps of the spectrogram and W is the time warp parameter. The identified area is randomly warped either to the left or
 250 right by a distance w randomly chosen from a uniform distribution.

- *Frequency masking* operation masks f consecutive frequency channels starting from f_0 . The f number of channels to be masked is randomly chosen from a uniform distribution and f_0 is chosen within the interval $[0, \nu - f)$, where ν is the number of frequency channels.
- 255 • *Time masking* operation masks t consecutive time steps starting from t_0 . The t number of time steps to be masked is randomly chosen from a uniform distribution and t_0 is chosen within the interval $[0, \tau - t)$.

This approach creates new audio samples from the existing ones. In particular, the three operations were applied to each input spectrogram, resulting in a final
 260 training dataset twice the size of the original one. Moreover, it is computationally cheaper compared to methods based on audio deformation that require more complex operations on the audio waveform.

4.2. Autoencoder and Multilayer Perceptron Architecture

The proposed classifier for Alzheimer’s disease combines a deep recurrent
 265 neural network and a multilayer perceptron. In particular, we used *auDeep* [16], that is a specific type of recurrent neural network called autoencoder, to learn efficient audio data coding in an unsupervised way. An autoencoder network aims to reconstruct a given input through two complementary phases, that is encoding and decoding. In other words, an autoencoder aims to copy
 270 the input compressing it into a latent-space representation (encoding) and then reconstruct the output from this representation (decoding). The goal is to reconstruct a copy of the input as accurate as possible and learn the best compact representation of the input. The dimensionality reduction that characterizes the encoding phase produces a code preserving only the most relevant features of
 275 the input. The intuition behind the choice of the autoencoder is that the learnt code, that is a 128 dimensional vector, should best sum up the intrinsic audio features of vocal production of Alzheimer’s disease patients. Thus, we used that code to train a multilayer perceptron able to classify potential Alzheimer’s disease subjects. Figure 2 presents the architecture of the proposed model. In

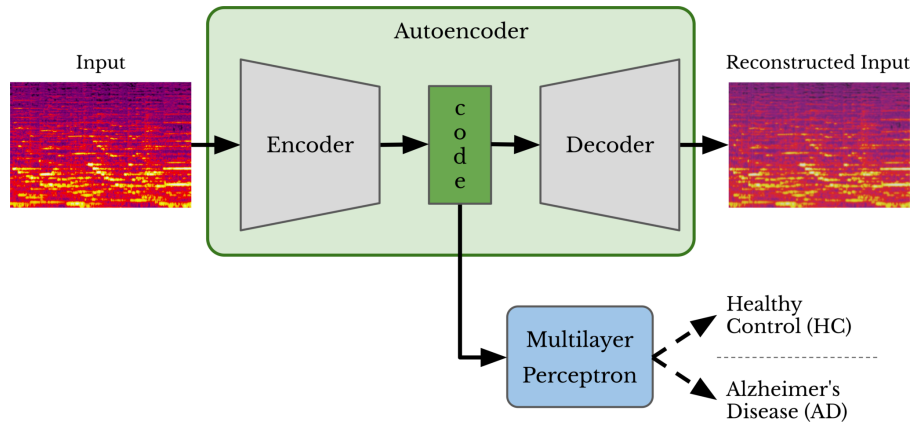


Figure 2: Architecture of the speech classifier for Alzheimer’s disease. The coding of the input data learnt during the encoding phase of the autoencoder allows the classification through a simple multilayer perceptron.

280 practice, we trained the autoencoder using the log mel spectrogram of the audio files, and then we used the encoded representation, that is a 128 dimensional vector, to feed the multilayer perceptron.

The architecture of the classifier shown in Figure 2 is characterized by four different phases. The **preprocessing** is the first one, where each raw audio file 285 is converted in the visual representation, that is the log mel spectrogram. In order to maximize the performance of the recurrent neural network, we set the extraction parameters (i.e., window size, overlapping and number of frequency bands) following the description and the suggestions provided by the *auDeep* authors. In particular, we used a 160 *ms* window size with an overlap of 80 *ms* 290 and 256 frequency bands. Whereas, to remove the background noise, we used a threshold between -45 *dB* to -60 *dB*. From each raw audio file, we extracted a set of 5 seconds long spectrograms: the longer slices were cut to the required length while the shorter slices were padded with silence. The **training** is the second phase, where the extracted spectrograms are used to train the autoencoder. 295 The dimensional reduction process allows to learn the features characterising the audio file to reconstruct the sample as close as possible to the original input. In particular, we used a unidirectional encoder and bidirectional decoder.

The unidirectional encoder can learn from the past state (i.e., the backwards learning propagation), while the bidirectional decoder can learn from the past and the future states (i.e., the forward learning propagation), simultaneously. Both encoder and decoder contain two layers with 256 gated recurrent unit cells. This made it possible to reach a good balance between network depth, classification performance and training time. The training was done setting a batch size of 64 for 128 epochs and a learning rate of 0.001. Whereas, the dropout rate was set to 20% for all hidden units. In the **features extraction** phase, the learnt representation of each spectrogram, that is a 128 dimensional vector, is extracted from the hidden layer of the *auDeep* network to feed the multilayer perceptron. Finally, in the **evaluation** phase, we used a multilayer perceptron with softmax output to binary classify the subjects. In particular, the multilayer perceptron contains 4 hidden layers, each of which has 128 hidden rectifier linear units (ReLU), and the training was performed for 400 epochs setting a learning rate of 0.001 and a dropout rate of 20% for all hidden units. The training of the speech classifier was performed using the Adam optimizer and a joint loss function. In particular, the root mean square error and the cross-entropy were used for the training of the autoencoder and the multilayer perceptron, respectively.

5. Results and Discussion

In this section, we present the results of our method based on automatic speech analysis to classify Alzheimer’s disease subjects. The proposed method was run on the *Google Colab* platform using a *12GB NVIDIA Tesla K80 GPU*, and the binary classification ability was demonstrated using the following well-known performance measures: precision, recall, accuracy and F1 score. We applied the 20-fold Cross-Validation technique to validate the stability and the performance of the classifier. Thus, the performance measures reported are then the average of the values computed in the Cross-Validation loop. In particular, for each fold, the dataset was apportioned into training and test sets without

significant differences in terms of characteristics, with an 80-20 split ratio, and making sure that data from the same subject was not contained simultaneously in the train and test set.

330 The comparison with the state-of-the-art approaches was also performed retrieving the same performance measures provided by the authors in their papers. Moreover, we reimplemented the method based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) proposed in [45], in order to evaluate the classification capability using a promising full automatic approach
335 for speech analysis that models the temporal properties of vocal production. In particular, the model proposed by the authors in [45], called DepAudioNet, consists of a serial combination of CNN and LSTM. The CNN includes a convolution layer of 40 cell units, a batch normalisation layer of 128 cell units and a max-pooling layer of 128 cell units. While an LSTM layer of 128 cell units
340 is stacked at the end of the architecture. The binary cross-entropy and the Stochastic Gradient Descent algorithm are used for training. A comprehensive description of the used parameters can be found in [45]. The reasons behind the selection of this methodology is that the automatic speech-based method proposed by the authors achieved better results in comparison with reference
345 approaches by using the log mel spectrogram to capture vocal characteristics. Moreover, a similar approach for speech emotion recognition, that involves both CNN for extracting high-level features from spectrograms and LSTM for aggregating long-term dependencies, was presented in [46].

Table 2 outlines the classification results of the proposed method based on
350 autoencoder in comparison to the DepAudioNet, that is the method based on CNN and LSTM presented in [45]. We evaluated the two methods, by using the original dataset and considering the data augmentation approach, in the two-class classification task, that is control subjects and Alzheimer’s disease subjects. It is worth noting that the capability to model the temporal properties of vocal production allows to DepAudioNet to obtain better results on the
355 original dataset, but the same DepAudioNet method partially benefits from the introduction of the data augmentation approach. Considering the data augmen-

Table 2: Automatic classifiers results (macro-averaged precision and recall): the DepAudioNet method compared with the proposed method.

Method	Precision	Recall	Accuracy	F1 score
DepAudioNet	0.751	0.751	0.751	0.751
DepAudioNet + Augment.	0.845	0.845	0.845	0.845
Our method	0.613	0.641	0.739	0.621
Our method + Augment.	0.907	0.865	0.933	0.885

tation approach, the proposed method outperforms the DepAudioNet method with average results for precision, recall, accuracy and F1 score 6.21% higher. The classification of the original dataset and considering the data augmentation approach resulted also in a significantly different training time. The proposed method took 13 minutes and 46 seconds for the original dataset and 6 minutes and 28 seconds for the augmented one. This is due to the fact that the shorter input data points, although in greater numbers, allow the method to converge faster. Whereas this does not occur with the DepAudioNet that took 9 minutes and 38 seconds and 16 minutes and 21 seconds, respectively. In particular, our method with data augmentation converges faster than the DepAudioNet, which requires the same number of training epochs even with the data augmentation approach. In particular, it can be noted that despite the DepAudioNet method improves using the data augmentation approach, the proposed method with data augmentation achieved high values of precision, recall, accuracy and F1 score, confirming the effectiveness of the model.

Table 3 summarises the selected state-of-the-art approaches reporting the associated used method, that is Logistic Regression (LogR), Naive Bayes (NB), K-Nearest Neighbor (kNN), Support Vector Machine (SVM), Random Forest (RF), a combination of different Machine Learning methods (ML), Neural networks (NN), and Bidirectional Encoder Representations from Transformers (BERT). It is worth noticing that not all the reported methods used the same dataset

Table 3: Comparison between the state-of-the-art methods for Alzheimer’s disease detection and the proposed method based on autoencoder and data augmentation.

Method	Precision	Recall	Accuracy	F1 score
LogR [23]	0.744	0.766	0.750	0.755
NB [35]	0.722	0.542	0.619	0.619
NB [36]	-	-	0.860	-
kNN [23]	0.727	0.708	0.721	0.717
SVM [29]	-	0.800	0.830	-
SVM* [30]	-	0.770	0.720	-
SVM [31]	0.857	0.720	0.800	0.783
SVM [35]	0.750	0.750	0.714	0.750
SVM [47]	-	-	-	0.745
RF [34]	-	-	-	0.680
RF [35]	0.731	0.792	0.714	0.760
RF [47]	-	-	-	0.703
ML* [48]	0.798	0.768	0.787	0.783
NN [23]	0.767	0.754	0.760	0.760
NN [32]	1.000	0.490	0.750	0.658
NN* [49]	0.859	0.904	0.869	0.876
BERT* [33]	0.860	0.880	0.833	0.840
BERT* [50]	0.906	0.843	0.881	0.872
Our method + Augment.	0.907	0.865	0.933	0.885

and the direct comparison could be unfeasible. For instance, the dataset used
380 is not always made publicly available by the authors or the authors do not provide details about the criteria for selecting the subset of data used. However, we highlighted the most promising methods tested on the same dataset that is considered to be a benchmark for this task. In particular, in Table 3 the * sym-

bol after the keyword of the method identifies studies that used the *Pitt Corpus*
385 (or a subset) for the evaluation. It is worth to notice that, BERT is actually a
deep neural network with millions of parameters recently published by Google,
but we decided to present it as a separate item because it achieved state-of-
the-art performance on many natural language understanding tasks [51]. Most
390 and automatic features extractions techniques, and in some cases, the authors
achieved good results training the proposed method only on the most signifi-
cant features. In particular, the extracted features range from simple linguistic
features [31], acoustic features [32, 35, 47] and syntactic features [34] used sep-
arately, to combinations of lexical and syntactic features [30, 48] and linguistic
395 and acoustic features [49]. Furthermore, some advanced approaches introduced
syntactic features [23], semantic features [33] and physical features (e.g., eye-
tracking) [29, 36]. In this work, we let the autoencoder identify the most relevant
features independently, avoiding the introduction of any audio artefact in the
learning phase. The rationale behind the proposed approach is that manual
400 audio transcription, annotation and features extraction may be prone to error
and lead to misinterpretation, and inappropriate data augmentation approaches
may introduce artefact penalizing for the classification. However, conventional
feature selection schemes used to extract the audio characteristics could help to
bridge the gap between good classification results and the lack of explainabil-
405 ity and interpretability of the deep learning-based models. We believe that is
of the utmost importance to proceed towards standardisation of the automatic
methods for cognitive impairment evaluation.

The methods based on traditional machine learning techniques usually obtain
good results, especially when the size of the dataset is a limiting factor for the
410 application of methods based on the deep learning approach. For instance in
[31], the authors proposed a method based on SVM able to achieve a precision
of 0.857 and a F1 score of 0.783. Recently, methods based on neural networks
have started to show their potential. In particular, the method presented in [49]
exploits linguistic and acoustic features to train a hierarchical attention network

415 architecture and exhibited a precision of 0.859 and a F1 score of 0.876. It is
worth noticing that the BERT-based solution proposed in [50] achieved high
classification values compared to the previous approach, however, the authors
confirmed that even using data augmentation approaches, the largest currently
available dataset for Alzheimer’s disease prediction is still insufficient in size for
420 unsupervised fine-tuning, and the large number of parameters requires a very
high pre-training/training time. By combining a neural network approach and
a data augmentation technique, in this work, we have overcome the problem of
the size of the dataset, and we have proposed a method able to outperform the
state-of-the-art approaches exhibiting high classification results. In particular,
425 the method proposed in this study achieved an accuracy of 93.30% in detecting
Alzheimer’s disease subjects and it does not require any pre-trained phase that
at the moment seems absolutely necessary to BERT-based models to achieve
good results. Moreover, in comparison to the BERT-based solutions proposed
in [33] and [50], our method requires much less training data and has a short
430 learning time, due to the absence of the pre-training phase, and requires the
fine-tuning of fewer parameters, only 24 million compared to 345 million of the
BERT-based models.

6. Conclusions

The ageing population is posing a challenge to all developed countries from
435 social, financial, and economic perspectives. The automatic and standardised
identification of those alterations related to Alzheimer’s disease represents a
crucial research problem and has proven to be a winning factor to provide timely
treatment. According to previous studies, speech has been known to provide an
indication of a person’s cognitive state. In this paper, we proposed an automatic
440 method to analyse the spontaneous speech productions in English of the subject,
and it proved to be a promising approach to distinguish between Alzheimer’s
disease subjects and healthy control subjects. We used a deep recurrent neural
network combined with a simple multilayer perceptron to extract the audio

features from the spectrogram and classify the spontaneous spoken production of
445 the subjects. We trained the model exploiting a specialized data augmentation
approach. The proposed method can discriminate healthy controls subject from
Alzheimer’s disease patients, exhibiting an accuracy of 93.30% and an F1 score
of 88.50%, which improves the state-of-the-art by 5.93%. The strengths of this
study include the use of a data augmentation approach that can be successfully
450 exploited for speech analysis in the detecting dementia context and the use of
a fairly simple model that has a short training time and does not require to be
e pre-trained on a large dataset.

This study confirms the relevance of acoustic features of spontaneous speech
for Alzheimer’s disease detection in the context of dementia diagnosis. More-
455 over, the proposed method paves the way for a future automatic and stan-
dardised approach for massive screening of dementia. The use and automatic
evaluation of routinely collected audio data minimize the required resources and
greatly reduce the potential risk of referral and diagnostic biases. The obtained
results are very encouraging and suggest that a fully automatic approach is
460 feasible and can achieve better results in detection and prediction tasks than
manual and semi-automatic approaches based on transcription and manual fu-
tures extraction.

Finally, it is important to notice that the description of the pathological
verbal productions through the acoustic and lexical features represents a com-
465plementary approach to the proposed method, and it might facilitate the devel-
opment of more explainable and interpretable models. As future research direc-
tion, we also plan to compare different data augmentation methods comparing
approaches that work on the raw audio files with those that work on images.
Moreover, it might strongly support the extension of the proposed method to
470 other languages with appropriate training and transfer learning approaches.

References

References

- [1] W. H. Organization, et al., Global action plan on the public health response to dementia 2017–2025.
- 475 [2] A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, A. M. Prina, B. Winblad, L. Jönsson, Z. Liu, M. Prince, The worldwide costs of dementia 2015 and comparisons with 2010, *Alzheimer's & Dementia* 13 (1) (2017) 1–7.
- [3] R. C. Petersen, Clinical practice. mild cognitive impairment., *The New England journal of medicine* 364 (23) (2011) 2227.
- 480 [4] A. E. Budson, P. R. Solomon, *Memory Loss E-Book: A Practical Guide for Clinicians*, Elsevier Health Sciences, 2011.
- [5] S. T. Farias, D. Mungas, B. R. Reed, D. Harvey, C. DeCarli, Progression of mild cognitive impairment to dementia in clinic-vs community-based cohorts, *Archives of neurology* 66 (9) (2009) 1151–1157.
- 485 [6] R. N. Kalaria, G. E. Maestre, R. Arizaga, R. P. Friedland, D. Galasko, K. Hall, J. A. Luchsinger, A. Ogunniyi, E. K. Perry, F. Potocnik, et al., Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors, *The Lancet Neurology* 7 (9) (2008) 812–826.
- 490 [7] F. Bertini, G. Bergami, D. Montesi, G. Veronese, G. Marchesini, P. Pandolfi, Predicting frailty condition in elderly using multidimensional socio-clinical databases, *Proceedings of the IEEE* 106 (4) (2018) 723–737.
- [8] L. Calzà, D. Beltrami, G. Gagliardi, E. Ghidoni, N. Marcello, R. Rossini-Favretti, F. Tamburini, Should we screen for cognitive decline and dementia?, *Maturitas* 82 (1) (2015) 28–35.
- 495

- [9] D. Beltrami, G. Gagliardi, R. Rossini Favretti, E. Ghidoni, F. Tamburini, L. Calzà, Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?, *Frontiers in aging neuroscience* 10 (2018) 369.
- 500 [10] R. L. Handels, C. A. Wolfs, P. Aalten, M. A. Joore, F. R. Verhey, J. L. Severens, Diagnosing alzheimer’s disease: a systematic review of economic evaluations, *Alzheimer’s & Dementia* 10 (2) (2014) 225–237.
- [11] A. Abbott, Dementia: a problem for our age, *Nature* 475 (7355) (2011) S2–S4.
- 505 [12] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, S. F. Cappa, Connected speech in neurodegenerative language disorders: a review, *Frontiers in psychology* 8 (2017) 269.
- [13] S. Abel, W. Huber, G. S. Dell, Connectionist diagnosis of lexical disorders in aphasia, *Aphasiology* 23 (11) (2009) 1353–1378.
- 510 [14] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid, et al., Automatic speech analysis to early detect functional cognitive decline in elderly population, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 212–216.
- 515 [15] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, J. Ogar, Aided diagnosis of dementia type through computer-based analysis of spontaneous speech, in: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.
- 520 [16] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, B. Schuller, audeep: Unsupervised learning of representations from audio with deep recurrent neural networks, *The Journal of Machine Learning Research* 18 (1) (2017) 6340–6344.

- [17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: A simple data augmentation method for automatic speech recognition, arXiv preprint arXiv:1904.08779.
- [18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, K. L. McGonigle, The natural history of alzheimer’s disease: description of study cohort and accuracy of diagnosis, *Archives of Neurology* 51 (6) (1994) 585–594.
- [19] D. M. Low, K. H. Bentley, S. S. Ghosh, Automated assessment of psychiatric disorders using speech: A systematic review, *Laryngoscope Investigative Otolaryngology* 5 (1) (2020) 96–116.
- [20] D. G. Clark, P. M. McLaughlin, E. Woo, K. Hwang, S. Hurtz, L. Ramirez, J. Eastman, R.-M. Dukes, P. Kapur, T. P. DeRamus, et al., Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment, *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* 2 (2016) 113–122.
- [21] A. König, A. Satt, A. Sorin, R. Hoory, A. Derreumaux, R. David, P. H. Robert, Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people, *Current Alzheimer Research* 15 (2) (2018) 120–129.
- [22] V. Vincze, G. Gosztolya, L. Tóth, I. Hoffmann, G. Szatlóczki, Detecting mild cognitive impairment by exploiting linguistic information from transcripts, *Association for Computational Linguistics*, 2016.
- [23] D. Beltrami, L. Calzà, G. Gagliardi, E. Ghidoni, N. Marcello, R. R. Favretti, F. Tamburini, Automatic identification of mild cognitive impairment through the analysis of italian spontaneous speech productions, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 2016, pp. 2086–2093.
- [24] K. Fraser, K. Lundholm Fors, M. Eckerström, C. Themistocleous, D. Kokki-

- nakis, Improving the sensitivity and specificity of mci screening with linguistic information, in: LREC workshop: RaPID-2. Miyazaki, Japan, 2018.
- [25] Q. Wei, A. Franklin, T. Cohen, H. Xu, Clinical text annotation—what factors are associated with the cost of time?, in: AMIA Annual Symposium Proceedings, Vol. 2018, American Medical Informatics Association, 2018, p. 1552.
- [26] J. J. Meilán, F. Martínez-Sánchez, J. Carro, J. A. Sánchez, E. Pérez, Acoustic markers associated with impairment in language processing in alzheimer’s disease, *The Spanish journal of psychology* 15 (2) (2012) 487–494.
- [27] B. Yu, T. F. Quatieri, J. R. Williamson, J. C. Mundt, Cognitive impairment prediction in the elderly based on vocal biomarkers, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.
- [28] C. Themistocleous, D. Kokkinakis, M. Eckerström, K. Fraser, K. L. Fors, Effects of mild cognitive impairment on vowel duration.
- [29] K. C. Fraser, K. Lundholm Fors, M. Eckerström, F. Öhman, D. Kokkinakis, Predicting mci status from multimodal language data using cascaded classifiers, *Frontiers in aging neuroscience* 11 (2019) 205.
- [30] K. C. Fraser, K. L. Fors, D. Kokkinakis, Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment, *Computer Speech & Language* 53 (2019) 121–139.
- [31] G. Gosztolya, V. Vincze, L. Tóth, M. Pákási, J. Kálmán, I. Hoffmann, Identifying mild cognitive impairment and mild alzheimer’s disease based on spontaneous speech using asr and linguistic features, *Computer Speech & Language* 53 (2019) 181–197.
- [32] C. Themistocleous, M. Eckerström, D. Kokkinakis, Identification of mild cognitive impairment from speech in swedish using deep sequential neural networks, *Frontiers in neurology* 9 (2018) 975.

- [33] A. Balagopalan, B. Eyre, F. Rudzicz, J. Novikova, To bert or not to bert:
580 Comparing speech and language-based approaches for alzheimer’s disease
detection, arXiv preprint arXiv:2008.01551.
- [34] K. L. Fors, K. C. Fraser, D. Kokkinakis, Automated syntactic analysis of
language abilities in persons with mild and subjective cognitive impair-
ment., in: MIE, 2018, pp. 705–709.
- 585 [35] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti,
M. Pákási, J. Kálmán, A speech recognition-based solution for the au-
tomatic detection of mild cognitive impairment from spontaneous speech,
Current Alzheimer Research 15 (2) (2018) 130–138.
- [36] K. C. Fraser, K. L. Fors, D. Kokkinakis, A. Nordlund, An analysis of eye-
590 movements during reading for the detection of mild cognitive impairment,
in: Proceedings of the 2017 Conference on Empirical Methods in Natural
Language Processing, 2017, pp. 1016–1026.
- [37] N. Kanda, R. Takeda, Y. Obuchi, Elastic spectral distortion for low resource
speech recognition with deep neural networks, in: 2013 IEEE Workshop on
595 Automatic Speech Recognition and Understanding, IEEE, 2013, pp. 309–
314.
- [38] N. Jaitly, G. E. Hinton, Vocal tract length perturbation (vtlp) improves
speech recognition, in: Proc. ICML Workshop on Deep Learning for Audio,
Speech and Language, Vol. 117, 2013.
- 600 [39] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen,
R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., Deep speech: Scal-
ing up end-to-end speech recognition, arXiv preprint arXiv:1412.5567.
- [40] A. Raju, S. Panchapagesan, X. Liu, A. Mandal, N. Strom, Data aug-
605 mentation for robust keyword spotting under playback interference, arXiv
preprint arXiv:1808.00563.

- [41] T. Ko, V. Peddinti, D. Povey, S. Khudanpur, Audio augmentation for speech recognition, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.
- [42] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, M. Bacchiani, Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. 610
- [43] H. Goodglass, E. Kaplan, S. Weintraub, BDAE: The Boston Diagnostic Aphasia Examination, Lippincott Williams & Wilkins Philadelphia, PA, 2001.
- [44] J. T. Becker, F. Boller, J. Saxton, K. L. McGonigle-Gibson, Normal rates of forgetting of verbal and non-verbal material in alzheimer’s disease., Cortex: A Journal Devoted to the Study of the Nervous System and Behavior. 615
- [45] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: An efficient deep model for audio based depression classification, in: Proceedings of the 6th international workshop on audio/visual emotion challenge, 2016, pp. 620 35–42.
- [46] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, B. Schmauch, Cnn+ lstm architecture for speech emotion recognition with data augmentation, arXiv preprint arXiv:1802.05630.
- [47] L. Calzà, G. Gagliardi, R. R. Favretti, F. Tamburini, Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia, Computer Speech & Language 65 (2020) 101113. 625
- [48] F. Haider, S. De La Fuente, S. Luz, An assessment of paralinguistic acoustic features for detection of alzheimer’s dementia in spontaneous speech, IEEE Journal of Selected Topics in Signal Processing 14 (2) (2019) 272–281. 630
- [49] W. Kong, H. Jang, G. Carenini, T. S. Field, Exploring neural models for predicting dementia from language, Computer Speech & Language 68 (2021) 101181.

- [50] A. Roshanzamir, H. Aghajan, M. S. Baghshah, Transformer-based deep
635 neural network language models for alzheimer’s disease risk assessment
from targeted speech, *BMC Medical Informatics and Decision Making*
21 (1) (2021) 1–14.
- [51] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we
640 know about how bert works, *Transactions of the Association for Compu-
tational Linguistics* 8 (2021) 842–866.