Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia

Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia

FLAVIO BERTINI, DAVIDE ALLEVI, GIANLUCA LUTERO, and DANILO MONTESI, Department of Computer Science and Engineering, University of Bologna, Italy Q1
LAURA CALZÀ, Interdepartmental Centre for Industrial Research in Health Sciences and Technologies, University of Bologna, Italy and and Department of Pharmacy and Biotechnology, University of Bologna, Italy

The World Health Organization estimates that 50 million people are currently living with dementia worldwide and this figure will almost triple by 2050. Current pharmacological treatments are only symptomatic, and drugs or other therapies are ineffective in slowing down or curing the neurodegenerative process at the basis of dementia. Therefore, early detection of cognitive decline is of the utmost importance to respond significantly and deliver preventive interventions. Recently, the researchers showed that speech alterations might be one of the earliest signs of cognitive defect, observable well in advance before other cognitive deficits become manifest. In this article, we propose a full automated method able to classify the audio file of the subjects according to the progress level of the pathology. In particular, we trained a specific type of artificial neural network, called autoencoder, using the visual representation of the audio signal of the subjects, that is, the spectrogram. Moreover, we used a data augmentation approach to overcome the problem of the large amount of annotated data usually required during the training phase, which represents one of the most major obstacles in deep learning. We evaluated the proposed method using a dataset of 288 audio files from 96 subjects: 48 healthy controls and 48 cognitively impaired participants. The proposed method obtained good classification results compared to the state-of-the-art neuropsychological screening tests and, with an accuracy of 90.57%, outperformed the methods based on manual transcription and annotation of speech.

CCS Concepts: • **Applied computing** → **Health informatics**; • **Computing methodologies** → *Supervised learning by classification;*

Additional Key Words and Phrases: Dementia, mild cognitive impairment, classification, speech data augmentation, neural networks

ACM Reference format:
Flavio Bertini, Davide Allevi, Gianluca Lutero, Danilo Montesi, and Laura Calzà. 2021. Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia. *ACM Trans. Comput. Healthcare* 3, 1, Article 8 (August 2021), 11 pages.
https://doi.org/10.1145/3469089

All the authors contributed equally to this research.
This work was supported by the OPLON project (Opportunities for active and healthy LONgevity, Smart Cities, Ministero Università e Ricerca, SCN_00176). The study was approved by the Ethical Committee of Azienda Ospedaliera Reggio Emilia (No. 148 2013/0013438).
Authors' addresses: F. Bertini, D. Allevi, G. Lutero, and D. Montesi, Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni, 7, Bologna, Italy; emails: flavio.bertini2@unibo.it, {davide.allevi, gianluca.lutero}@studio.unibo.it, danilo.montesi@unibo.it; L. Calzà, Interdepartmental Centre for Industrial Research in Health Sciences and Technologies, University of Bologna, Via Tolara di Sopra, 41/E, Bologna, Italy, Department of Pharmacy and Biotechnology, University of Bologna, Via Belmeloro, 6, Bologna, Italy; email: laura.calza@unibo.it.

2637-8051/2021/08-ART8 $15.00
https://doi.org/10.1145/3469089

1 INTRODUCTION

In the last decade, life expectancy increased globally, leading to various age-related issues. Dementia is one of the most increasing pathologies among the elderly population, and the World Health Organization recognized it as a public health priority [34], with an estimated 50 million people affected by the disease and nearly 10 million new cases every year worldwide. Dementia is a syndrome that includes different diseases, inclusive of Alzheimer's disease, of a chronic or progressive nature that affects memory, thinking, behavior and ability to perform everyday activities. The deterioration in cognitive function over time is commonly preceded by deterioration in emotional control, social behavior, or motivation. The economic implications of dementia in terms of direct medical and social care costs are one of the big challenges for healthcare systems. In 2015, the worldwide costs of dementia were estimated at $818 billion, 86% of which occur in high-income countries, and it was estimated that the $1 trillion thresholds would have been crossed by 2018 [43].

The symptoms linked to dementia can be manifest at different severity levels, and most people undergo a gradual cognitive decline. **Mild Cognitive Impairment** (**MCI**) represents an early stage, that can be the prodromal phase of cognitive decline, characterised by cognitive changes that are serious enough to be assessed with neuropsychological assessment, but not so severe to interfere with everyday activities, while **early Dementia** (**eD**) manifests cognitive deficits that influence everyday life [36]. In [7], the researchers estimated that 70% of diagnosed MCI subjects progressed to dementia with an annual conversion rate from 10% to 15% clinic sample [13]. Although epidemiological studies have shown that people adopting a better lifestyle, such as avoiding reckless use of alcohol, avoiding smoking, eating a healthy diet, and getting regular exercise have a reduced risk of dementia symptoms, current treatments are only symptomatic for memory (in a short time window) and psychiatric symptoms, and no disease-modifying therapies are available for dementia. Thus, similarly to other pathologies for which there is no cure, such as frailty condition [5], 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 of utmost importance during extreme events, such as pandemic diseases, during which the older population is most vulnerable and fragile.

The diagnosis of cognitive decline is a challenging topic. Despite the extensive literature about the diagnosis of all the different types of dementia, the presymptomatic diagnosis or even "detection" raises both theoretical issues and ethical concerns [8]. However, the implementation of preventive measures requires one to have psychometric tests, with high accuracy, low cost, and suitable for large-scale use. Subjects affected by dementia manifest cognitive alterations in various domains: memory, attention, executive functioning, visuospatial skills, perceptual speed, and language also [4], and it has been proven that the commonest screening tools, for instance, the Mini-Mental State Examination [14], are largely inadequate for detecting early changes in cognition [33]. In particular, they are much less effective to track down the MCI.

Episodic memory impairment has always been one of the most common signs of dementia. However, recently language has been subjected to growing interest, and literature suggests that language impairment is a promising sign to reveal early signs of cognitive decline [6]. In particular, analysis of spoken production is an ecological and inexpensive approach to identify MCI and other alterations related to cognitive functionalities. In literature, several studies obtained good results in cognitive impairment detection using different language features, such as acoustic features [2], lexical features [24], speech errors [1], and a combination of rhythmic, acoustic, lexical, morphosyntactic, and syntactic features [4]. However, most of the proposed methods require a preprocessing stage that includes several manual activities such as transcription, annotation, and correction. That results in a non-standardised and time-consuming approach with the potential loss of useful information leading to a not scalable screening tool.

In this article, we propose a method based on a specific type of neural networks, that is, autoencoder, trained using the visual representation of the audio signal of the subjects. The method has proven effective results in classifying potential patients into three classes: **healthy control** subjects (**HC**), MCI subjects, and eD 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* [20], whose aim is the unsupervised feature extraction from audio data.

Typically, neural network and deep learning models require a large amount of annotated data; however, in some health contexts and certainly in our study, it is not possible to collect large amounts of data. To allow the use of a classifier based on neural networks, we adopted a data augmentation approach to enlarge the size of the input dataset [35]. In particular, through three different operations, that is, time warping, frequency masking, and time masking, each log mel spectrogram[1] has contributed 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.

To the best of our knowledge, this is the first study that successfully uses a neural network model combined with data augmentation to automatically classify a small dataset of audio file related to MCI and eD subjects. In particular, the strengths of our method include

- the early detection of the prodromal phase of cognitive decline classifying potential patients in three classes in a single shot, avoiding binary classification as proposed in [28];
- the capability to fully automatically process the audio files and extract the required features, avoiding any manual features selection and manipulation activities;
- the capability to use a neural network model despite the small size of the dataset;
- the use of a data augmentation approach that does not bias or distort the audio file, which is extremely important in this specific context, and that can be successfully used in future researches; and
- the possibility to standardize the screening of cognitive impairment using the audio files capturing the spontaneous speech of the subjects. In particular, our method paves the way toward a standard way of collecting and analysing the audio file that does not alter the data itself and avoids any manual activities that may lead to unfair and non-uniform evaluations.

The promising results confirm the strength of the linguistic approach and the proposed method allows easy scalability.

The rest of the article is organized as follows. In Section 2, we review the literature on methods to detect language disorder in early-stage dementia subjects. Section 3 summarises the main characteristics of the dataset. We present our automatic method for speech classification for dementia 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 neuropsychological screening tests and other manual and semi-automatic methods based on transcription and annotation of speech. Some concluding remarks are made in Section 6.

2 RELATED WORKS

There is extensive literature that confirms the worth of linguistic features in detecting health-related issues [30]. 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.

[1] In 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 color-coded.

2.1 Speech Analysis for Dementia Detection

While traditional clinical methods to diagnose the early stage of dementia are not ideal [33], language assessments provide an effective, simpler and more economic approach [11]. In particular, automatic speech analysis based on natural language processing, speech recognition, and machine learning techniques can provide objective and fast diagnostic results [28]. 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 [41], 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 the number of first person singular verbs in distinguishing MCI patients from healthy controls. Whereas, in [3] and [16], 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, and noun-verb ratio are statistically significant features. Syntactic features were explored in [4] and [42] 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 ratio, and silent pause, have been explored more recently. In [32], the authors found that voiceless segments produced by patients affected by dementia were highly correlated with fluency. Whereas, a classifier based on the total duration of the "s" phoneme, the pseudo-syllable rate, the average pause duration, the total count of "m" phonemes was proposed in [44]. In another study, the authors showed that MCI patients produce longer vowels during text reading tasks [39]. Linguistic features are usually used with learning models to facilitate and automate the diagnosis of dementia and researchers explored different approaches based on support vector machines [17, 19, 21], neural networks [38], random forest [15, 40], and Naive Bayes [18] classifiers. Typically all these approaches require a combination of different features and a significant amount of manual processing to extract and clean the features to be used.

2.2 Audio Data Augmentation

Neural network models usually require a large amount of data for training, improving the accuracy, and avoiding overfitting. Data augmentation is a 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 [25], the authors investigated three distortion methods, that is, vocal tract length distortion, speech rate distortion, and frequency-axis random distortion, to artificially augment training samples. Whereas in [23], the authors proposed a method to transform the spectrograms using a random linear warping along the frequency dimension. A different approach involves the superimposing of a generated noise signal to the original audio [22] or the mixing of the original audio signal with music and TV/movie audio [37]. Whereas, a method that changes the speed of the audio signal was proposed in [27]. In [26], the authors used an acoustic room simulator to generate simulated audio data for speech recognition. The *SpecAugment* method proposed in [35] is simple and computationally cheap and operates on the *log mel* spectrogram of the input audio. In particular, the proposed augmentation operations, inspired by computer 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 DATASET DESCRIPTION

The study was approved by the Ethical Committee of Azienda Ospedaliera Reggio Emilia (no. 2013/0013438). The cohort enrolled 96 participants,[2] and it is balanced in terms of gender (48 males and 48 females), age (from 50 to

[2]Given the particular kind of data employed for this study and the restrictions on them from the Italian legislation, unfortunately we cannot make the datasets publicly available.

Table 1. Characteristics of the Cohort

	HC subjects	**MCI subjects**	**eD subjects**
Neurological assessment and inclusion criteria	• MMSE ≥ 24 • MoCA ≥ 18 • No neurological pathologies • No sensory impairment • No intellectual disability • No familiarity with dementia	• MMSE ≥ 18 • No problem in daily living activities	• MMSE ≥ 18 • Need of support for daily living activities
Age	61.60 ± 6.93	64.34 ± 7.33	66.38 ± 6.70
Education *(years)*	13.00 ± 3.92	11.28 ± 4.35	9.38 ± 4.01

75 years old), and education (all the subjects have at least a junior high school certificate). In particular, the cohort included 48 healthy control subjects (HC) and 48 cognitively impaired subjects. Of the latter, 32 subjects with MCI and 16 with eD. All the subjects were requested to complete a cognitive assessment path and the traditional cognitive battery including verbal tasks, **the Mini-Mental State Examination** (**MMSE**) and the **Montreal Cognitive Assessment** (**MoCA**). Table 1 provides the main characteristics of the cohort. A comprehensive description of the cohort building process can be found in Beltrami et al. [4]. It is worth noting that the statistical analysis showed no differences for age between the three subgroups, while the level of education of the eD group is significantly lower than the HC group (p-value = 0.0171). However, adequate verbal comprehension and production were mandatory inclusion criteria to be enrolled in the pathological group.

In addition to the standard cognitive evaluation, all subjects were requested to record their spontaneous speech induced by these three questions: *Could you please describe this picture?* (the picture showed a living room with some characters during certain domestic activities) [10], *Could you please describe a typical working day?*, and *Could you please describe the last dream you remember?*. The audio files were recorded in a quiet room using common off-the-shelf equipment, that is, an Olympus-Linear PCM Recorder LS-5 (in WAV files; 44.1 KHz, 16 bit). The length of the resulting 288 audio files (i.e., three for each subject) varies between approximately 10 seconds and 9 minutes. In particular, there are no differences between the three classes in terms of minimum duration, while the audio files of the eD subjects have a maximum duration equal to one-third of the other two classes. However, all the audio files have a duration of 85 seconds on average and there are no significant differences between the three classes.

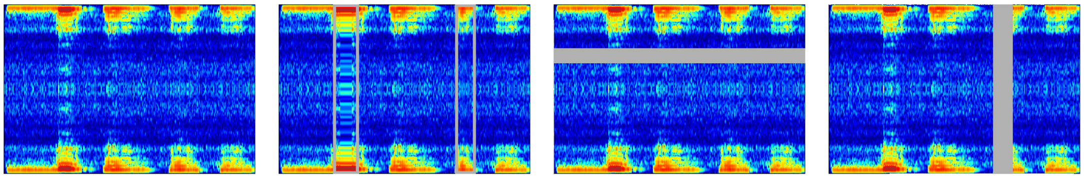
4 MILD COGNITIVE IMPAIRMENT AND EARLY DEMENTIA SPEECH CLASSIFIER

In this section, we present an automated method for speech analysis for classifying MCI and eD subjects. Firstly, we provide details about the data augmentation technique adopted. This is important as it will help in understanding the reasons behind the selection of a particular methodology. Secondly, we describe the architecture of the classifier based on a recurrent neural network and a multilayer perceptron.

4.1 Data Augmentation Technique

Typically, a large training dataset is a crucial aspect for the performance of the deep learning models; however, it is not always possible to collect new and labeled data. Data augmentation is an effective suite of techniques to enhance the size of the training dataset by applying random and realistic transformations, such as rotation, mirroring, translation, noise overlap, and hue, and saturation adjustment for images.

In the literature, there are several data augmentation approaches for audio files; however, due to the specific goal of this study, we avoided adopting techniques that heavily distort the original samples. Indeed, we were interested in the audio features characterizing dementia, and we could not afford to select data augmentation



(a) (b) (c) (d)

Fig. 1. The original log mel spectrogram without augmentation (a) and with the time warp (b), the frequency masking (c), and the time masking (d) applied. The grey box and band identify the wrapped region in (b) and the masked region in (c) and (d), respectively.

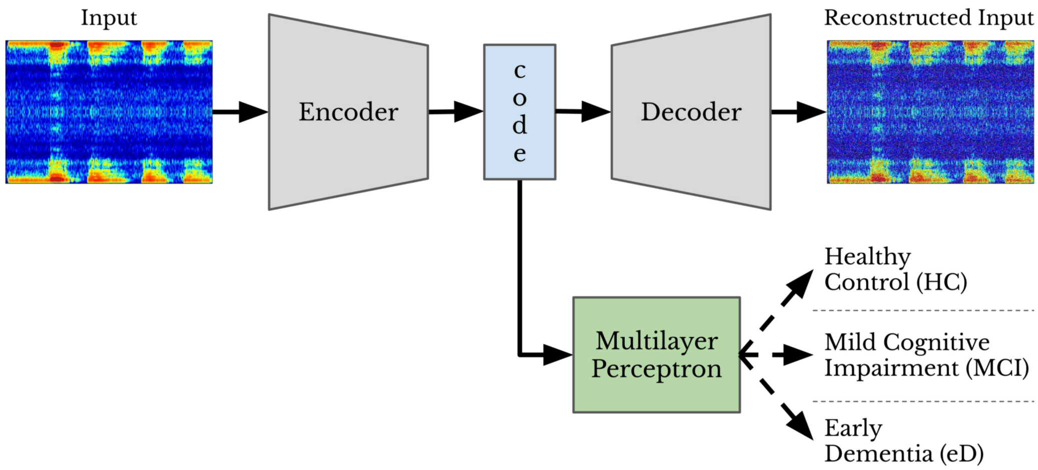


Fig. 2. Architecture of the speech classifier for dementia. The coding of the input data learnt during the encoding phase of the autoencoder allows the classification through a simple multilayer perceptron.

techniques introducing bias in the learning process. We used the *SpecAugment* suite presented in [35] that transforms the log mel spectrogram to increase the number of inputs. In particular, the suite consists of the three operations shown in Figure 1, that is, the time warping, which allows one to shift the spectrogram in time in a random direction (Figure 1(b)); the frequency masking, which masks a random slice of frequency steps (Figure 1(c)); and the time masking, which masks a random slice of time steps (Figure 1(d)). 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.

4.2 Autoencoder and Multilayer Perceptron Architecture

The proposed classifier for MCI and eD combines a recurrent neural network and a multilayer perceptron. Firstly, we used *auDeep* [20], which is a specific type of recurrent neural network called autoencoder, to learn efficient audio data coding in an unsupervised way. In particular, an autoencoder aims to reconstruct a given input through two complementary phases, that is, encoding and decoding. The dimensionality reduction that characterizes the first phase produces a code preserving only the most relevant features of the input. Then, we used that code to train a multilayer perceptron able to classify potential dementia subjects. Figure 2 shows the proposed architecture. In practice, we trained the autoencoder using the log mel spectrogram of the audio files, and we used the encoded representation to feed the multilayer perceptron.

The architecture of the classifier depicted in Figure 2 is characterized by the following four stages.

(1) **Preprocessing**. In this phase, the log mel spectrogram is extracted from each raw audio file. To maximize the performance of the autoencoder, we set the extraction parameters according to the description provided by the *auDeep* authors. In particular, since *auDeep* works better using quite tight windows with overlap and a relatively large number of frequency bands, we used a 160-ms window size with an overlap of 80 ms and 256 frequency bands. Moreover, we used a threshold between −45 dB and −60 dB to remove the background noise. This preprocessing made it possible to extract from each raw audio file a set of 5-seconds-long spectrograms: the shorter slices were padded with silence while the longer slices were cut to the required length. As a result, considering the different numerosity of the three groups of subjects, the differences in terms of duration between the audio files, and the split radio adopted for the training (80–20), the input data was composed of 1,958 samples for the HC subjects, 1,306 samples for the MCI subjects, and 653 samples for the eD subjects.

(2) **Training the autoencoder**. The extracted spectrograms were used to train the autoencoder that through a dimensional reduction process learned the features characterising the audio file to reconstruct the given input. In particular, we used a unidirectional encoder and a bidirectional decoder. The first can learn from the past state (i.e., the backwards learning propagation), while the latter can learn from the past and the future states (i.e., the forward learning propagation), simultaneously. Both the encoder and the 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. In practice, 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. Also, we applied the 20-fold Cross-Validation technique to validate the stability and the performance of the classifier.

(3) **Features extraction**. In this stage, the learnt representations of each spectrogram were extracted from the hidden layer of the autoencoder, to feed the multilayer perceptron.

(4) **Training the classifier**. In this final stage, we used a multilayer perceptron with softmax output to classify the subjects. In particular, the multilayer perceptron contains 4 hidden layers with 128 hidden rectifier linear units, 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.

5 RESULTS AND DISCUSSION

In this section, we present the results of our method based on automatic speech analysis to classify potential dementia subjects. The proposed method was run on the *Google Colab* platform using a *12 GB NVIDIA Tesla K80 GPU*, and the classification ability was demonstrated using the following performance measures: precision, recall, accuracy, and F1 score. In particular, we evaluated the method using the original dataset and considering the data augmentation approach, and we assessed the classification capability considering both the three-class classification task and the two-class classification task by merging in the **pathological subjects** (**PS**) group both MCI and eD subjects. The learning process was carried out by apportioning the data into training and test sets without significant differences in terms of characteristics, with an 80 − 20 split ratio. 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 [31], to evaluate the classification capability using a promising full automatic approach for speech analysis. The reason behind the selection of this methodology is that the automatic speech-based method proposed by the authors achieved better results in comparison with reference approaches by using the log mel spectrogram to capture vocal characteristics. Moreover, a similar approach for speech emotion recognition, which involves both CNN for extracting high-level features from spectrograms and LSTM for aggregating long-term dependencies, was presented in [12].

Table 2. Automatic Classifiers Results (Macro-Averaged Precision and Recall): the DepAudioNet Method Compared with the Proposed Method Based on Autoencoder

Dementia classes	Method	Precision	Recall	Accuracy	F1 score
HC / MCI / eD	DepAudioNet	51.02%	49.40%	52.90%	50.20%
	DepAudioNet + Augmentation	59.71%	60.90%	62.60%	60.30%
	Our method	58.59%	56.88%	65.23%	57.72%
	Our method + Augmentation	**86.19%**	**83.28%**	**86.98%**	**84.71%**
HC / PS	DepAudioNet	55.90%	59.20%	59.20%	57.50%
	DepAudioNet + Augmentation	68.74%	71.30%	71.30%	70.00%
	Our method	77.16%	77.35%	77.26%	77.25%
	Our method + Augmentation	**90.84%**	**90.56%**	**90.57%**	**90.70%**

Table 2 outlines the classification results of the proposed method based on autoencoder in comparison to DepAudioNet, which is the method based on CNN and LSTM presented in [31].

Firstly, we evaluated the performance of the two methods in the three-class classification task (i.e., HC, MCI, and eD) using the original dataset without data augmentation and with data augmentation. It is worthy to notice that our method based on autoencoder and without data augmentation equals in performance to the DepAudioNet method with data augmentation. However, the results improved on average by 40% when we introduced the data augmentation. In particular, the proposed method based on autoencoder achieved a precision, recall, accuracy, and F1 score of 86.19%, 83.28%, 86.98%, and 84.71%, respectively.
Then, we evaluated the two methods in the two-class classification task. In this case, the proposed method without data augmentation outperforms the DepAudioNet method with data augmentation with average results for precision, recall, accuracy, and F1 score 10% higher. The selected data augmentation approach turns out to be a good choice in the detecting dementia context allowing one to improve the results of the DepAudioNet-based classifier on average by 21.40%. However, the largest increase is obtained by coupling data augmentation with the autoencoder-based classifier. Indeed, the proposed method with data augmentation achieved a precision, recall, accuracy, and F1 score of 90.84%, 90.56%, 90.57%, and 90.70%, respectively.

It is worth noticing that the proposed method presents very short training times. In particular, the training time lasted 4 minutes and 25 seconds for the original dataset and 5 minutes and 47 seconds for the augmented one, both for the three and the two-class classification task. Moreover, the proposed method performs well in the three-class classification task avoiding the binary classification proposed in [29].

Table 3 summarizes the selected state-of-the-art approaches reporting the language of the used dataset and the associated used method, that is, **k-Nearest Neighbor** (**kNN**), **Logistic Regression** (**LogR**), **Neural networks** (**NNs**), **Random Forest** (**RF**), **Support Vector Machine** (**SVM**), and **Naive Bayes** (**NB**). In particular, we inserted only the best result when building a classifier for distinguishing control from pathological subjects, and the * symbol after the keyword of the method identifies studies that used the same dataset proposed in this work for the evaluation. Most of the cited works in Table 3 performed a lot of experiments applying manual and automatic features extraction techniques, and in some cases, the authors achieved good results training the proposed method only on the most significant features. In this work, we let the autoencoder identify the most relevant features independently, avoiding the introduction of any bias in the learning phase. We believe that the fully automatic analysis of the spontaneous speech of the subjects, avoiding any manual activities that may alter the screening, is of the utmost importance to proceed toward standardization 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 application of methods based on the deep learning approach. In [21], the authors proposed a method based on SVM able to achieve good results with a precision of 85.70% and

Table 3. Comparison between the State-of-the-Art Methods for MCI Detection and the Proposed Method Based on Autoencoder and Data Augmentation

Method	**Language**	**# classes**	**Precision**	**Recall**	**Accuracy**	**F1 score**
kNN* [3]	Italian	2	72.70%	70.80%	72.10%	71.74%
LogR* [3]	Italian	2	74.40%	76.60%	75.00%	75.48%
NN* [3]	Italian	2	76.70%	75.40%	76.00%	76.04%
RF* [9]	Italian	2	-	-	-	70.30%
SVM* [9]	Italian	2	-	-	-	74.45%
SVM [19]	Swedish	2	-	80.00%	83.00%	-
SVM [17]	Swedish	2	-	77.00%	72.00%	-
SVM [21]	Hungarian	2	85.70%	72.00%	80.00%	78.25%
NN [38]	Swedish	2	**100.00**%	49.00%	75.00%	65.77%
RF [15]	Swedish	2	-	-	-	68.00%
RF [40]	Hungarian	2	73.10%	79.20%	71.40%	76.03%
SVM [40]	Hungarian	2	75.00%	75.00%	71.40%	75.00%
NB [40]	Hungarian	2	72.20%	54.20%	61.90%	61.92%
NB [18]	Swedish	2	-	-	86.00%	-
Our method	Italian	2	90.84%	**90.56%**	**90.57%**	**90.70%**
Our method	Italian	3	86.19%	83.28%	86.98%	84.71%

an accuracy of 85.70%. Recently, methods based on NNs have started to show their potential. In particular, the method presented in [38] exploits acoustic features and metadata to train a deep NN architecture and exhibited a precision of 100.00% and an accuracy of 75.00%; however, it achieved a very low recall of 49.00%. By combining a NN 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, the method proposed in this study achieved on average 90.67% classification results in the two-class classification task and 85.29% classification results in the three-class classification task, which shows the effectiveness of our approach.

6 CONCLUSIONS

Aging is becoming a meaningful challenge for many countries from social, financial, and economic perspectives. Prompt detection of the early stages of dementia or even cognitive decline related to non-neurological conditions (systemic diseases such as renal dysfunctions, chronic pulmonary diseases, inappropriate pharmacological therapies, hypothyroidism, etc.) represents a crucial research problem. In this article, we proposed a method to detect MCI and eD conditions by analyzing subjects speech productions. Using a deep recurrent autoencoder combined with a specialized data augmentation approach, we can automatically extract and learn the features from audio data of the spontaneous speech of the subjects, avoiding any manual features selection and manipulation activities, and fully automatic discriminate healthy controls subject from MCI and eD subjects, exhibiting an accuracy of 86.98% and an F1 score of 84.63%.

The strengths of our study include the possibility to standardize and automate the massive screening of cognitive impairments. 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 feasible and can achieve better results in detection and prediction tasks than manual and semi-automatic approaches based on transcription and manual futures extraction.

Finally, it is worth noting that the language-specific profiling of pathological verbal productions proposed represents a complementary approach to the method proposed in this study and it can be very useful to drive the implementation of a valid and reliable dementia screening tool. Moreover, it might strongly support the extension of the proposed method to other languages with appropriate training and transfer learning approach.

ACKNOWLEDGMENTS

Daniela Beltrami and Enrico Ghidoni are gratefully acknowledged for subjects selection and interview recordings.

REFERENCES

[1] Stefanie Abel, Walter Huber, and Gary S. Dell. 2009. Connectionist diagnosis of lexical disorders in aphasia. *Aphasiology* 23, 11 (2009), 1353–1378.

[2] Emilia Ambrosini, Matteo Caielli, Marios Milis, Christos Loizou, Domenico Azzolino, Sarah Damanti, Laura Bertagnoli, Matteo Cesari, Sara Moccia, Manuel Cid, et al. 2019. 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'19)*. IEEE, 212–216.

[3] Daniela Beltrami, Laura Calzà, Gloria Gagliardi, Enrico Ghidoni, Norina Marcello, Rema Rossini Favretti, and Fabio Tamburini. 2016. Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. 2086–2093.

[4] Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà. 2018. Speech analysis by natural language processing techniques: A possible tool for very early detection of cognitive decline? *Frontiers in Aging Neuroscience* 10 (2018), 369.

[5] Flavio Bertini, Giacomo Bergami, Danilo Montesi, Giacomo Veronese, Giulio Marchesini, and Paolo Pandolfi. 2018. Predicting frailty condition in elderly using multidimensional socioclinical databases. *Proceedings of the IEEE* 106, 4 (2018), 723–737.

[6] Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F. Cappa. 2017. Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology* 8 (2017), 269.

[7] Andrew E. Budson and Paul R. Solomon. 2011. *Memory Loss E-Book: A Practical Guide for Clinicians*. Elsevier Health Sciences.

[8] Laura Calzà, Daniela Beltrami, Gloria Gagliardi, Enrico Ghidoni, Norina Marcello, Rema Rossini-Favretti, and Fabio Tamburini. 2015. Should we screen for cognitive decline and dementia? *Maturitas* 82, 1 (2015), 28–35.

[9] Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2020. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language* 65 (2020), 101113.

[10] Paola Ciurli, Paola Marangolo, and Anna Basso. 1996. *Esame Del Linguaggio-II*. OS.

[11] David Glenn Clark, Paula M. McLaughlin, Ellen Woo, Kristy Hwang, Sona Hurtz, Leslie Ramirez, Jennifer Eastman, Reshil-Marie Dukes, Puneet Kapur, Thomas P. DeRamus, et al. 2016. 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.

[12] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. 2018. CNN+LSTM architecture for speech emotion recognition with data augmentation. arXiv:1802.05630.

[13] Sarah Tomaszewski Farias, Dan Mungas, Bruce R. Reed, Danielle Harvey, and Charles DeCarli. 2009. Progression of mild cognitive impairment to dementia in clinic- vs community-based cohorts. *Archives of Neurology* 66, 9 (2009), 1151–1157.

[14] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. 1975. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12, 3 (1975), 189–198.

[15] Kristina Lundholm Fors, Kathleen C. Fraser, and Dimitrios Kokkinakis. 2018. Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment.. In *MIE*. 705–709.

[16] K. Fraser, K. Lundholm Fors, Marie Eckerström, Charalambos Themistocleous, and Dimitrios Kokkinakis. 2018. Improving the sensitivity and specificity of MCI screening with linguistic information. In *LREC Workshop: RaPID-2*.

[17] Kathleen C. Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2019. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language* 53 (2019), 121–139.

[18] Kathleen C. Fraser, Kristina Lundholm Fors, Dimitrios Kokkinakis, and Arto Nordlund. 2017. An analysis of eye-movements during reading for the detection of mild cognitive impairment. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1016–1026.

[19] Kathleen C. Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. 2019. Predicting MCI status from multimodal language data using cascaded classifiers. *Frontiers in Aging Neuroscience* 11 (2019), 205.

[20] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. 2017. auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research* 18, 1 (2017), 6340–6344.

[21] Gábor Gosztolya, Veronika Vincze, László Tóth, Magdolna Pákáski, János Kálmán, and Ildikó Hoffmann. 2019. 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.

[22] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv:1412.5567.

[23] Navdeep Jaitly and Geoffrey E. Hinton. 2013. Vocal tract length perturbation (VTLP) improves speech recognition. In *Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language*, Vol. 117.

[24] William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. 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*. 27–37.

[25] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. 2013. Elastic spectral distortion for low resource speech recognition with deep neural networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 309–314.

[26] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani. 2017. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. Interspeech 2017 (2017), 379–383.

[27] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *16th Annual Conference of the International Speech Communication Association*.

[28] Alexandra Konig, Aharon Satt, Alex Sorin, Ran Hoory, Alexandre Derreumaux, Renaud David, and Phillippe H. Robert. 2018. 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.

[29] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, et al. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1, 1 (2015), 112–124.

[30] Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology* 5, 1 (2020), 96–116.

[31] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. DepAudioNet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 35–42.

[32] Juan J. G. Meilán, Francisco Martínez-Sánchez, Juan Carro, José A. Sánchez, and Enrique Pérez. 2012. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *The Spanish Journal of Psychology* 15, 2 (2012), 487–494.

[33] Alex J. Mitchell. 2009. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *Journal of Psychiatric Research* 43, 4 (2009), 411–431.

[34] World Health Organization et al. 2017. Global action plan on the public health response to dementia 2017–2025.

[35] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. arXiv:1904.08779.

[36] Ronald C. Petersen. 2011. Clinical practice. mild cognitive impairment. *The New England Journal of Medicine* 364, 23 (2011), 2227.

[37] Anirudh Raju, Sankaran Panchapagesan, Xing Liu, Arindam Mandal, and Nikko Strom. 2018. Data augmentation for robust keyword spotting under playback interference. arXiv:1808.00563.

[38] Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. 2018. Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. *Frontiers in Neurology* 9 (2018), 975.

[39] Charalambos Themistocleous, Dimitrios Kokkinakis, Marie Eckerström, Kathleen Fraser, and Kristina Lundholm Fors. [n.d.]. Effects of mild cognitive impairment on vowel duration.

[40] László Tóth, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánréti, Magdolna Pákáski, and János Kálmán. 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research* 15, 2 (2018), 130–138.

[41] Veronika Vincze, Gábor Gosztolya, László Tóth, Ildikó Hoffmann, and Gréta Szatlóczki. 2016. Detecting mild cognitive impairment by exploiting linguistic information from transcripts. Association for Computational Linguistics.

[42] Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation-What factors are associated with the cost of time?. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 1552.

[43] Anders Wimo, Maëlenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, A. Matthew Prina, Bengt Winblad, Linus Jönsson, Zhaorui Liu, and Martin Prince. 2017. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & Dementia* 13, 1 (2017), 1–7.

[44] Bea Yu, Thomas F. Quatieri, James R. Williamson, and James C. Mundt. 2015. Cognitive impairment prediction in the elderly based on vocal biomarkers. In *16th Annual Conference of the International Speech Communication Association*.

Received November 2020; revised April 2021; accepted May 2021

AUTHOR QUERIES

Q1: AU: Please use affiliation where the author was affiliated when the work was done in the title line. New affiliations can be noted in the page 1 notes and the author address section on page 1.
Q2: AU: Please provide a url for all arXiv references.
Q3: AU: Please provide a url and retrieved date for all online documents and resources.

Print Window   Close Window

ACM Transactions on Computing for Healthcare

Please click on the "Yes, Please Send Corrections" button if you would like to mail your corrections to the Production Editor.

Article ID: HEALTH0301-08   **Filename:** HEALTH0301-08_LR.pdf   **Author email:** flavio.bertini@smartdata.cs.unibo.it

Corrections Submitted

Line No	Correction Text
1	Author's current address: F. Bertini, Department of Mathematical, Physical and Computer Sciences, University of Parma, Parco Area delle Scienze, 7/A, 43124, Parma, Italy; emails: flavio.bertini@unipr.it
156	"the Mini-Mental State Examination" without bold for "the"
396	Accessed August 28, 2021. https://gup.ub.gu.se/publication/270215?lang=en
387	Accessed August 28, 2021. https://www.who.int/publications/i/item/global-action-plan-on-the-public-health-response-to-dementia-2017---2025
332	Accessed August 28, 2021. https://www.giuntipsy.it/catalogo/test/esame-del-linguaggio-ii
337	https://arxiv.org/abs/1802.05630
360	https://arxiv.org/abs/1412.5567
389	https://arxiv.org/abs/1904.08779
392	https://arxiv.org/abs/1808.00563
401	In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 181–187, Berlin, Germany, August 2016. Association for Computational Linguistics.

Additionals

Yes, Please Send Corrections   No, I would like to modify Corrections