Smart Health 26 (2022) 100347

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Interpreting acoustic features for the assessment of Alzheimer's disease using ForestNet

Paula Andrea Pérez-Toro [a,b,*], Dalia Rodríguez-Salas [a], Tomás Arias-Vergara [a,b,c], Philipp Klumpp [a], Maria Schuster [c], Elmar Nöth [a], Juan Rafael Orozco-Arroyave [a,b], Andreas K. Maier [a]

[a] Pattern Recognition Lab. Friedrich-Alexander Universität Erlangen-Nürnberg, Martensstr. 3, Erlangen 91058, Germany
[b] GITA Lab, Facultad de Ingeniería, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín 050010, Colombia
[c] Department of Otorhinolaryngology, Head and Neck Surgery, Ludwig-Maximilians University, Pettenkoferstr. 4a, Munich 80336, Germany

ARTICLE INFO

Keywords:
Alzheimer's disease
Acoustic analysis
Interpretability
ForestNets

ABSTRACT

Nowadays, interpretable machine learning models are one of the most critical topics in the medical domain. The lack of interpretation leads to blind and unreliable models for clinicians, despite the fact that the aim is to support diagnosis through these tools. This problem has been increasing since the creation of large models such as those based on deep learning, which, despite providing good performance in prediction and classification tasks, are not transparent to human understanding. One of the increasingly prevalent clinical problems related to acoustic and linguistic disorders is Alzheimer's disease (AD), where one important challenge is to provide speech markers that help in supporting, understanding, and facilitating the diagnosis and monitoring of the disease. It motivates this study which proposes a methodology focused on analyzing acoustic features in AD and at the same time providing interpretation from the results. The proposed approach consists of using decision tree-based methods together with neural networks (ForestNet) for analyzing the classification results. Only features that can give interpretation were considered. Unweighted average recalls of up to 79% were achieved for discriminating AD patients. Then, we looked at the relevant features that provided most of the information for assessing AD, which were those related to rhythm, voiced rates, duration, and phone rates. This confirms that this kind of approach can be suitable for the discrimination of AD while maintaining a good performance.

1. Introduction

Dementia is a general term for conditions or diseases characterized by the progressive degeneration and death of the brain cells. Alzheimer's Disease (AD) is the most common form of dementia and affects two-thirds of the total cases of dementia. It is characterized by the progressive loss of neurons in the cerebral cortex and hippocampus (Prince, 2015), which causes symptoms related to memory loss, psychological, and behavioral alterations, and the deterioration of cognitive functions linked to deficits in communication (Association et al., 2013). The fluency of the patients' speech is affected by the difficulty in accessing the semantic information intentionally produced by abnormalities in language production (König et al., 2015). The cognitive function of the

* Corresponding author at: Pattern Recognition Lab. Friedrich-Alexander Universität Erlangen-Nürnberg, Martensstr. 3, Erlangen 91058, Germany.
E-mail addresses: paula.andrea.perez@fau.de (P.A. Pérez-Toro), dalia.rodriguez@fau.de (D. Rodríguez-Salas), tomas.arias@fau.de (T. Arias-Vergara), philipp.klumpp@fau.de (P. Klumpp), maria_elke.schuster@med.uni-muenchen.de (M. Schuster), elmar.noeth@fau.de (E. Nöth), rafael.orozco@udea.edu.co (J.R. Orozco-Arroyave), andreas.maier@fau.de (A.K. Maier).

https://doi.org/10.1016/j.smhl.2022.100347
Received 29 September 2022; Accepted 30 September 2022
Available online 7 October 2022

Table 1
Demographic information of the participants.

	AD patients F/M	HC subjects F/M
NumberL of subjects	97/42	51/ 39
Age [years]	72.7 (8.2)/69.8 (9.2)	62.8 (7.4)/62.4 (7.6)
MMSE	18.3 (4.1)/17.3 (4.3)	29.3 (1.0)/28.9 (1.1)

AD: Alzheimer's Disease. **HC**: Healthy Control. Values are expressed as mean (standard deviation). F: Female. M: Male.

patients is commonly evaluated according to the Mini-Mental State Examination (MMSE) (Folstein et al., 1983). This 30-point scale contains items related to language production, immediate memory, naming, and spatial attention. The scores of over 24 indicate normal cognition.

Even though most of the studies on dementia consider linguistic analysis, several studies have also considered acoustic analysis to support the diagnosis of AD. Commonly, this analysis is based on prosodic aspects and to a lesser extent on articulatory (Khodabakhsh et al., 2015; Martinc & Pollak, 2020; Pérez-Toro et al., 2021a; Syed et al., 2020). In addition, deep learning embeddings have become popular in the field (Campbell et al., 2020; Luz et al., 2020; Pappagari et al., 2020; Pérez-Toro et al., 2021a; Zhu et al., 2021), since they are compressed representations obtained from speech that can contain relevant acoustic information.

Prosody measures are mostly focused on temporal aspects, variation in Fundamental Frequency (F_0), intensity, voice periods, and interruptions. Articulatory features consider frequency-based analyses such as Mel-frequency cepstral coefficients and energy distributed in the Bark scale, formant frequencies, among others (Barragán Pulido et al., 2020) In previous studies, these features have reported accuracies of up to 80% for discrimination AD (Khodabakhsh et al., 2015; Luz et al., 2020; Martinc & Pollak, 2020; Pérez-Toro et al., 2021a; Syed et al., 2020).

Moreover, speaker and deep learning embeddings are also considered to assess AD such as i-vectors, x-vectors, and Wav2Vec with accuracies of up to 84% (Campbell et al., 2020; Pappagari et al., 2020; Pérez-Toro et al., 2021a; Zhu et al., 2021). Despite of showing good results, most of these kinds of approaches do not provide direct interpretation from the clinician's perspective, due to most of them coming from neural networks.

When using machine learning for clinical applications, interpretation is especially crucial in order to present the results in understandable terms to a human. Artificial Neural Networks (ANNs) aim to approximate a function by learning the value of the parameters that result in the best function approximation, which leads to models that cannot be directly interpreted with a mathematical formula. This problem was called "lack of transparency" by Roscher et al. (2020). In general, the main idea of interpretability in machine learning is trying to find a way to build understandable models to bridge this gap between machines and humans. Some of the most used methods to interpret these results are those based on decision trees since it is possible to estimate the importance of each decision variable. Recently, a model that combines the good performance of the ANNs and the interpretation provided by the decision tree called ForestNet was proposed in Rodríguez-Salas et al. (2020). It maps an ensemble of decision trees to a highly sparse Multi-Layer Perceptron (MLP) classifier, in which one can get the relative importance of each variable/feature.

We proposed the use of interpretable acoustic, phonemic, and emotional features for the assessment of AD. Furthermore, the reasons mentioned above motivated the use of decision tree methods such as eXtreme Gradient Boosting (XGBoost) and ForestNets, in order to look at the most important descriptors.

2. Data

The data consist of 229 semi-spontaneous speech recordings from a sub-set of the Pitt Corpus (Becker et al., 1994), where 139 are from AD patients and 90 are from Healthy Control (HC) subjects. The participants were asked to describe the Cookie Theft picture (Goodglass et al., 1983). Some of the corpus recordings were discarded due to high noise levels. The speech of the interviewer was removed according to the timestamps provided by the dataset. After segmentation, the average duration of the recordings is 61 ± 28 s for the AD patients and 60 ± 35 s for the HC subjects. The data was labeled by expert physicians according to the MMSE. Demographic information about the participants is included in Table 1.

3. Methods

The general methodology is shown in Fig. 1. Several acoustic features are considered to process information from speech recordings. Those are based on extracting prosodic, spectral, emotional, and phonemic information. Subsequently, a ForesNet is used in order to discriminate between AD patients and HC subjects and obtain the importance of each variable in the classification process.

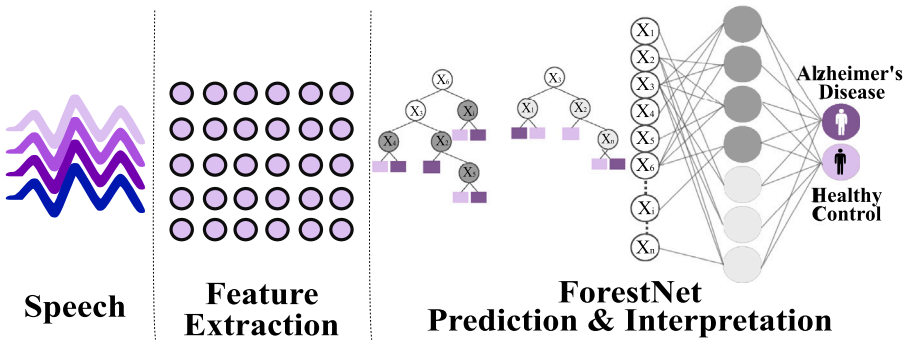

Fig. 1. Scheme of the general methodology addressed in this study using ForesNet.

3.1. Features

3.1.1. Pleasure arousal dominance (PAD)

This pre-trained model aims to extract emotional information according to 3 different dimensions such as valence (pleasant–unpleasant), arousal (calm-agitated), or dominance (dominant-submissive) (Mehrabian, 1996). Three different deep learning models are trained in order to address three binary-classification problems using the IEMOCAP database (Busso et al., 2008): (1) active vs. passive arousal, (2) positive vs. negative valence, and (3) strong vs. weak dominance. The model is composed of: (1) 3 channel Mel-Spectrogram for the input, (2) convolution layers together with attention maps for the feature extraction of the network, (3) a Bidirectional Gated recurrent Unit (Bi-GRU) to capture prosodic information. The input is a multi-channel log-Mel spectrogram formed by different resolutions (16 ms, 25 ms, and 45 ms) and considering sequences of 500 ms. These kinds of features have been previously used in related applications such as in the prediction of AD using the generated pre-trained embeddings from the last hidden layer (Pérez-Toro et al., 2021a) and to discriminate depression in Parkinson's Disease using the information extracted from the posterior probabilities (Pérez-Toro et al., 2021b) by using a Sigmoid activation function.

In this study, we use the information extracted from the regular and Log-Likelihood Ratios (LLR) of the posterior probabilities to facilitate the interpretation. Four functionals were computed across the posteriors obtained for each sequence (mean, standard deviation, minimum and maximum) to form a static vector.

3.1.2. VAD duration features

Those features were also used in previous related AD studies, showing good performance (Pérez-Toro et al., 2021a). This set consists of several duration-based descriptors that were extracted using an energy-based Voice Activity Detection (VAD) algorithm, where the speech and pause segments were identified. The descriptors include: the number of pauses per second, number of speech segments per second, the ratio between the number of speech segments and the number of pauses, and six functionals (mean, standard deviation, kurtosis, skewness, minimum, and maximum) for the duration of the pauses and speech segments.

3.1.3. Phonemic features

The set of phonemic features includes the phoneme posterior probability, average duration, rate, and standard deviation of duration. For this, a multilabel recurrent network with Long-Short Term Memory cells (LSTM) is used for the automatic recognition of phonemes (Arias-Vergara et al., 2021). The model was trained to predict three main dimensions manner of articulation (stop, nasal, lateral, trill, fricative, approximant, and vowels), place of articulation (labial, alveolar, velar, palatal, postalveolar, central, front, and back), and voicing (silence, voiced, and voiceless). In summary, the architecture of the network is as follows: Two convolution layers process the input tensors (Mel-spectrograms) with ReLU activation functions, two max-pooling layers, and dropout. The resulting feature maps are concatenated to form the sequence of feature vectors processed by two stacked bidirectional LSTMs. Then, a sigmoid activation function is used to compute the sequence of phoneme posterior probabilities. The network was trained with the TIMIT corpus, a dataset with time-aligned phonetic transcriptions of speech recordings from 630 American English native speakers (Garofolo et al., 1993).

3.1.4. Acoustic

A set of acoustic features that have shown good performance in the detection of AD is also considered (Barragán Pulido et al., 2020). These features were not mixed with the VAD Duration, since these are computed and based on energy contours and F_0 estimation. Chunks of 25 ms, with a hop size of 10 ms were taken for computing the voiced rate, the energy, and the F_0 contours. For the F_0 only the voiced segments were considered. From these descriptors, four statistical functionals were computed (mean, standard deviation, kurtosis, and skewness) per utterance. Additionally, jitter, shimmer, and the Harmonicity to Noise Ratio (HNR) were also considered. For the HNR, the same four functionals were computed.

Table 2
Classification performance using the XGBoost for each group of features and their combination.

Features	Number of features	UAR (%)	Sen (%)	Spe (%)	F1Score	AUC
PAD	48	54	56	51	0.53	0.54
Duration	15	56	52	59	0.54	0.55
Phonemic	**97**	**65**	**66**	**64**	**0.65**	**0.69**
Acoustic	9	65	62	69	0.64	0.70
All	169	60	57	63	0.59	0.65

3.2. Classification

3.2.1. eXtreme Gradient Boosting

It is a classifier based on decision trees and gradient boosting algorithms, which is more stable when compared to regular gradient boosting since it is optimized to offer a good balance between variance and bias (Chen et al., 2015). The training is based on the concept of weak learning, which aims to combine the information from those weak models to create a powerful one. XGBoost builds decision trees iteratively, where each tree is a "weak learner". It continues to sequentially build more "weak learners", which correct the previous tree until a stopping condition, such as the number of trees, is reached.

Since this is a decision tree-based technique, we can retrieve importance scores for each feature. The importance is calculated for a single decision tree by the amount each attribute improves the performance and then weighted by the number of observations in the node. The purity measure, in this case, the Gini index (Loh, 2011), is used to select split points. Finally, all of the decision trees average the obtained scores.

3.2.2. ForestNet

It is a special case of an MLP classifier and has the characteristic of being built using information given by an ensemble of trees (see Fig. 1). Given the number of trees as well as their maximum depth, the tree ensemble can be built and the architecture of a ForestNet can be fully defined with information extracted from the trees. ForestNet consists of a single hidden layer with the number of neurons equal to the total number of nodes that are the father of leaves in the tree ensemble; therefore, each hidden neuron has associated with a node in the tree. Each hidden neuron is only fed using the features which are involved in the path from the tree root to the node associated with the neuron. Such a node reaches at least one leaf, and each leaf has associated a class; the hidden neurons only feed the output neurons related to the class which their associated nodes reach. This generates a highly sparse MLP architecture.

Since each feature in the input layer feeds only a fraction of the hidden neurons, the importance of each feature can be measured as the relative frequency in which this feature feeds a different neuron in the hidden layer. We scaled the importance [0,1] for direct comparisons with XGBoost.

4. Experiments and results

Two different classifiers were considered for comparison purposes: an XGBoost and a ForestNet classifier. The classifiers were optimized following a nested 5-fold cross-validation strategy. The optimal parameters of the XGBoost were found through a grid search, where for the number of decision trees $\in \{5, 10, 15, \ldots, 100\}$ and the maximum depth $\in \{1, 3, 5, \ldots, 50\}$. For each combination of feature sets, a ForestNet was built and trained. In all cases, 60 decision trees with a maximum depth of 4 were used for building the ensemble using the Extra Trees Classifier algorithm (Geurts et al., 2006). Each ensemble was mapped to a ForestNet. Once built, the networks were trained during 250 epochs, with a batch size of 40, using the Adam optimizer with a learning rate of 1×10^{-3}, a weight decay of 4×10^{-4} and the Cross-Entropy Loss as criteria.

For the combination of the different feature sets, an early fusion strategy was applied by merging sets of features before performing the classification and making the final decision. The classification performance was evaluated according to the Unweighted Average Recall (UAR), Sensitivity (Sen), Specificity (Spe), F1Score, and Area Under the receiving operator Curve (AUC) as shown in Tables 2, 3, and 4. The sensitivity corresponds to the recall for the AD patients and specificity to the one for the HC subjects.

For the XGBoost classifier, the classification results are summarized in Table 2. The phonemic and acoustic features together produce the highest results for AD and HC identification. Despite of achieving a higher AUC than the phonemic features, the acoustic showed less of a balance between sensitivity and specificity than the phonemic. Notice that, for the XGBoost, the combination of all features did not improve the performance.

Conversely, for the ForestNet classifier, the most accurate results of each feature separately are obtained with the duration-based features (UAR= 69), the combination of all descriptors improves the performance by 11% (see Table 3). It might indicate that ForestNet is more robust for mapping higher dimensions than XGBoost.

The top 3 combinations according to the performance are summarized in Table 4. Notice that the combination of features provides higher performance results for the classification. In the case of XGBoost, the fusion of phonemic and acoustic improved the result by 3%. However, ForestNet obtained the most accurate results by combining the duration, phonemic and acoustic features (UAR= 79), and outperforming XGBoost by 11%.

Table 3
Classification performance using ForestNet for each group of features and their combination.

Features	Number of features	UAR (%)	Sen (%)	Spe (%)	F1Score	AUC
PAD	48	62	79	46	0.76	0.70
Duration	**15**	**69**	**81**	**57**	**0.80**	**0.75**
Phonemic	97	66	78	53	0.77	0.74
Acoustic	9	66	69	62	0.73	0.75
All	**169**	**78**	**80**	**77**	**0.83**	**0.86**

Table 4
Classification performance of the top 3 combinations for XGBoost and ForestNet.

Features	Number of features	UAR (%)	Sen (%)	Spe (%)	F1Score	AUC
XGBoost						
Phon-Ac	**106**	**68**	**69**	**67**	**0.67**	**0.70**
Dur-Phon	112	66	63	69	0.65	0.69
PAD-Phon-Ac	160	65	63	68	0.64	0.68
ForestNet						
Dur-Phon-Ac	**121**	**79**	**81**	**77**	**0.84**	**0.86**
All	169	78	80	77	0.83	0.86
PAD-Dur-Ac	72	77	78	76	0.82	0.85

Ac: Acoustic. Phon: Phonemic. Dur: Duration.

5. Discussion

After training the models, the importance of each feature was measured as described in Section 3.2. Fig. 2 illustrates the top 3 features sorted by importance for each feature set and classifier.

Notice that in most cases, there was no agreement in the ranking of features between the two models. However, the features related to pauses provided relevant information for both classifiers, which was expected considering that previous works highlighted that pausing is a possible marker of cognitive changes in dementia (Sluis et al., 2020; Zhu et al., 2022). In addition, the correlation of the top 10 features to compare both classifiers was computed. PAD, Duration, and Acoustic got a moderate Spearman's correlation coefficient ($0.4 < \rho < 0.7$), while for the phonemic a weak correlation ($\rho < 0.3$) was obtained. Thus, even if the two models have different feature rankings, the information that these features provide is similar.

Regarding the duration features, the maximum duration of the pauses highly contributed to both XGBoost and ForestNet. Similarly, for the classification using phonemic features and XGBoost, the posterior probability of the pauses was the first descriptor in the ranking. For features in the top 10, both classifiers mostly agree on the functionals drawn from the duration of the voiced segments, which also provide information related to the rhythm.

In the PAD model, the weak dominance and the negative valence were more important. The results concur with the posterior values, which were higher for the patients than for the HC subjects, i.e., the patients tend to be more "negative" and more "submissive". The standard deviation of the negative valence may suggest that the patients' speech tends to be more monotonous, producing fewer variations while speaking since those posteriors are lower for the patients. Additionally, these results may also be consistent with energy contour and periodicity measures which provide information on speech variations.

Previous studies reported that Alzheimer's patients have problems related to the production of phonemic cues (Cerbone et al., 2020). Despite obtaining significant results related to phonemic features, they are difficult to interpret since the participants are not saying the exact words or performing a verbal fluency task. One hypothesis could be that velar and postalveolar phonemes are present in common words that describe the Cookie Theft picture (Goodglass et al., 1983), such as the velar /k/ in *cookie*, and the postalveolar /j/ in *jar*.

The average importance of each feature set is displayed in Fig. 3. One reason why the acoustic features provide the highest average importance is the number of features (9); however, according to the results (Section 4), the duration features highly contributes to the classification performance of the ForestNet (UAR= 69%). The PAD features contributed similarly in both classifiers.

6. Conclusion

This study leverages the interpretation provided by decision tree-based methods and ANNs for analyzing different sets of acoustic features for the assessment of AD. Different sets of state-of-the-art features have been considered, in addition to novel emotional and phoneme-based features. XGBoost and ForestNet were used in order to classify and obtain the importance of each feature. The highest performance was achieved when the combination of the different sets of features was considered (UAR=79%). While both classifiers allow us to obtain feature importance, ForestNet provided the most accurate results over XGBoost with an improvement of 11%. Overall, the features that are providing more information in the discrimination are those related to rhythm, duration as well as voiced and phone rates.

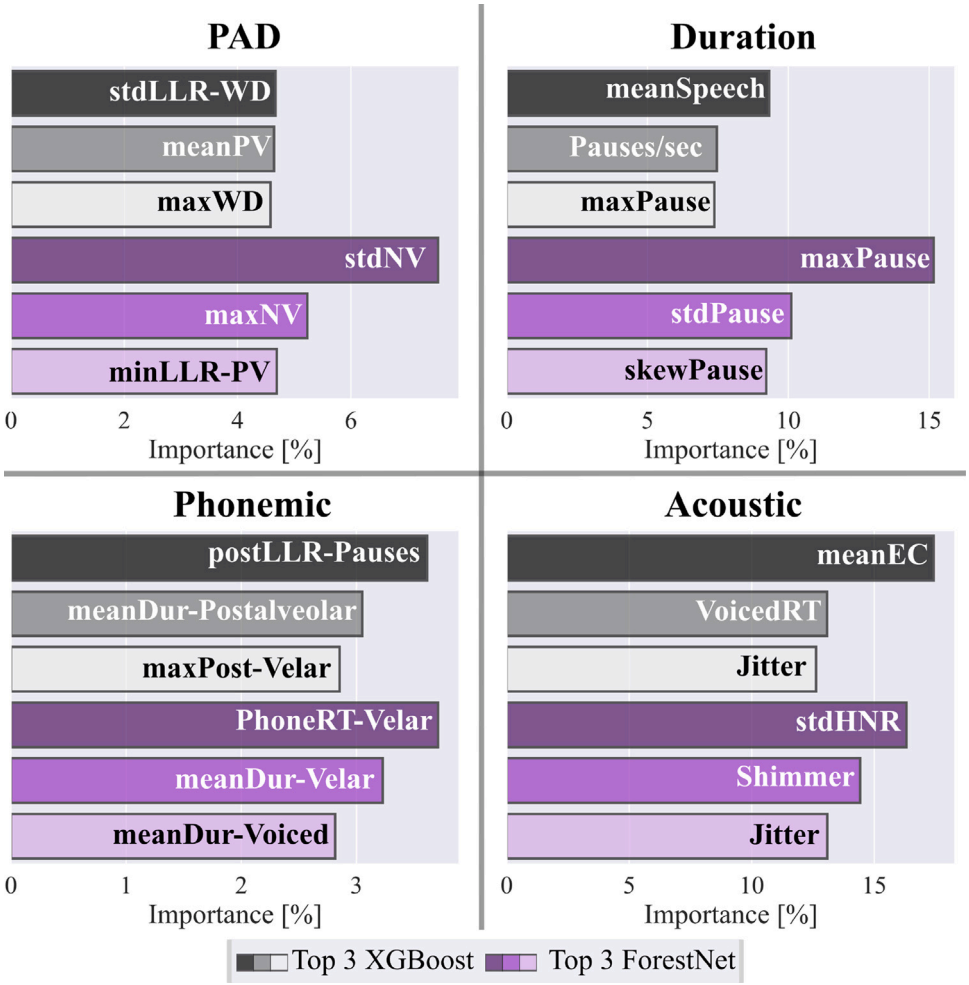

Fig. 2. Top 3 sorted features by importance for each classifier. stdLLR-WD: standard deviation of the weak dominance LLR posterior probability. PV: Positive Valence. NV: Negative Valence. Pause/Sec: Number of pauses per second. post: posterior probability. Dur: Duration. RT: Rate. HNR: Harmonicity to Noise Ratio. EC: Energy Countour.

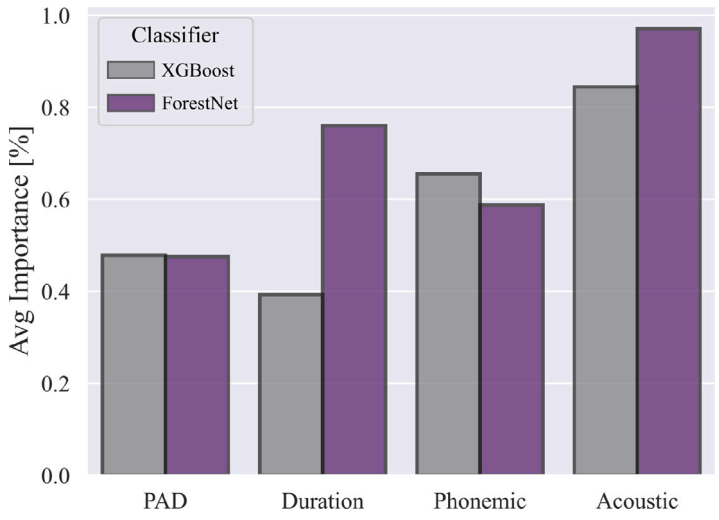

Fig. 3. Average (Avg) importance per feature set for the model that considers the combination of all features.

Note that the considered methods and results here may be interpretative from the clinical point of view, which makes our model suitable for supporting the clinical diagnosis of AD. However, we suggest further research for a better interpretation in relation to phonemic features, including the analysis of verbal fluency tasks and text analysis. Moreover, future work will consider different acoustic and linguistic features that may be useful for this analysis.

CRediT authorship contribution statement

Paula Andrea Pérez-Toro: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **Dalia Rodríguez-Salas:** Software, Writing – review & editing, Investigation. **Tomás Arias-Vergara:** Software, Writing – review & editing, Investigation. **Philipp Klumpp:** Writing – review & editing. **Maria Schuster:** Supervision, Writing – review & editing, Investigation. **Elmar Nöth:** Supervision, Conceptualization, Writing – review & editing. **Juan Rafael Orozco-Arroyave:** Supervision, Conceptualization, Writing – review & editing. **Andreas K. Maier:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 766287, and partially funded by CODI at UdeA grant # PRG2020-34068. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS.

References

Arias-Vergara, T., Klumpp, P., Vasquez-Correa, J. C., Nöth, E., Orozco-Arroyave, J. R., & Schuster, M. (2021). Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications*, *24*(2), 423–431.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Barragán Pulido, M. L., Alonso Hernández, J. B., Ferrer Ballester, M. Á., Travieso, C. M., Mekyska, J., & Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Systems with Applications*, *150*, Article 113213.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, *51*(6), 585–594.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335.

Campbell, E. L., Docío-Fernández, L., Jiménez Raboso, J., & García-Mateo, C. (2020). Alzheimer's dementia detection from audio and text modalities. arXiv preprint arXiv:2008.04617.

Cerbone, B., Massman, P. J., Woods, S. P., & York, M. K. (2020). Benefit of phonemic cueing on confrontation naming in Alzheimer's disease. *The Clinical Neuropsychologist*, *34*(2), 368–383. http://dx.doi.org/10.1080/13854046.2019.1607904, arXiv:10.1080/13854046.2019.1607904 PMID: 31030619.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1–4.

Folstein, M. F. (1983). The mini-mental state examination. *Archives of General Psychiatry*, *40*(7), 812.

Garofolo, J. S., Lamel, L., & Fisher, W. M. (1993). Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. http://dx.doi.org/10.1007/s10994-006-6226-1.

Goodglass, H., Kaplan, E., & Barressi, B. (1983). Cookie theft picture. In *Boston diagnostic aphasia examination*. Philadelphia, PA: Lea & Febiger.

Khodabakhsh, A., Yesil, F., Guner, E., & Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1), 1–15.

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 112–124.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In *Proc. interspeech 2020* (pp. 2172–2176). http://dx.doi.org/10.21437/Interspeech.2020-2571.

Martinc, M., & Pollak, S. (2020). Tackling the ADReSS challenge: a multimodal approach to the automated recognition of alzheimer's dementia. In *Proc. interspeech 2020* (pp. 2157–2161).

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*(4), 261–292.

Pappagari, R., Cho, J., Moro-Velazquez, L., & Dehak, N. (2020). Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. In *INTERSPEECH* (pp. 2177–2181).

Pérez-Toro, P. A., Bayerl, S. P., Arias-Vergara, T., Vásquez-Correa, J. C., Klumpp, P., Schuster, M., Nöth, E., Orozco-Arroyave, J. R., & Riedhammer, K. (2021). Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo challenge. In *Interspeech* (pp. 3785–3789).

Pérez-Toro, P. A., Vasquez-Correa, J. C., Arias-Vergara, T., Klumpp, P., Schuster, M., Nöth, E., & Orozco-Arroyave, J. R. (2021). Emotional state modeling for the assessment of depression in Parkinson's disease. In *International conference on text, speech, and dialogue* (pp. 457–468). Springer.

Prince, M. J. (2015). *World alzheimer report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International.

Rodríguez-Salas, D., Mürschberger, N., Ravikumar, N., Seuret, M., & Maier, A. (2020). Mapping ensembles of trees to sparse, interpretable multilayer perceptron networks. *SN Computer Science*, *1*, http://dx.doi.org/10.1007/s42979-020-00268-y.

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216.

Sluis, R. A., Angus, D., Wiles, J., Back, A., Gibson, T., Liddle, J., Worthy, P., Copland, D., & Angwin, A. J. (2020). An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer's Disease & Other Dementias®*, *35*, Article 1533317520939773.

Syed, M. S. S., Lech, M., & Pirogova, E. (2020). Automated screening for Alzheimer's dementia through spontaneous speech. In *Proc. interspeech 2020* (pp. 1–5).

Zhu, Y., Obyat, A., Liang, X., Batsis, J. A., & Roth, R. M. (2021). Wavbert: Exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection. In *Proc. interspeech 2021* (pp. 3790–3794).

Zhu, Y., Tran, B., Liang, X., Batsis, J. A., & Roth, R. M. (2022). Towards interpretability of speech pause in dementia detection using adversarial learning. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing* (pp. 6462–6466). IEEE.