Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

An overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition through Spontaneous Speech

Saturnino Luz[1], Member, IEEE, Fasih Haider[2], Member, IEEE
Davida Fromm[3], Ioulietta Lazarou[4], Ioannis Kompatsiaris[4], and Brian MacWhinney[3]

[1]Usher Institute, Edinburgh Medical School, The University of Edinburgh, UK
[2]School of Engineering, The University of Edinburgh, UK
[3]Department of Psychology, Carnegie Mellon University, USA
[4]Information Technologies Institute, CERTH, Thermi-Thessaloniki, Greece

Corresponding author: S Luz (email: s.luz@ ed.ac.uk).

[Funding acknowledgements???]

ABSTRACT The ADReSS-M Signal Processing Grand Challenge was held at the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2023. The challenge targeted a difficult automatic prediction problem of great societal and medical relevance, namely, the detection of Alzheimer's Dementia (AD). Participants were invited to employ signal processing and machine learning methods to create predictive models based on spontaneous speech data. The ADReSS-M challenge was designed to assess the extent to which predictive models built based on speech in one language (English) generalise to another language (Greek). To the best of our knowledge no work had investigated acoustic features of the speech signal in multilingual AD detection. This paper describes the context of the ADReSS-M challenge, its data sets, its predictive tasks, the evaluation methodology we employed, our baseline models and results, and the top five submissions. The paper concludes with a summary discussion of the ADReSS-M results, and our critical assessment of the future outlook in this field.

INDEX TERMS Biomedical signal processing, Medical conditions, Alzheimer's disease, Human disease biomarkers, Speech processing, Natural language processing, multilingual Alzheimer's dementia detection.

I. INTRODUCTION

THERE has been a great increase in interest in signal processing and machine learning methods for the detection of Alzheimer's and other forms of dementia through analysis of speech [1]. Several approaches for disease detection and prognostic assessment have been proposed, often lacking [2] standardisation and common benchmarks against which the different approaches and models could be compared. This situation has improved somewhat in recent years with the increasing availability of speech and language data sets for dementia research [3]–[5], and the advent of machine learning shared tasks (grand challenges) in Alzheimer's detection through spontaneous speech [6], [7]. While many of the approaches proposed in the context of those challenges produced high accuracy results based on the analysis of spontaneous speech [8], [9], the data were limited to American English data, and even where classification and regression methods were based on acoustic, as opposed to language-dependent features, it was unclear whether such acoustic analysis approaches generalise across languages [10]. In order to investigate this question, we organised the ADReSS-M Challenge at ICASSP 2023. which targeted dementia detection across two languages.

Alzheimer's Dementia (AD) is a category of neurodegenerative syndromes that entails a long-term and usually gradual decrease of cognitive functioning. To diagnose and assess disease progression as well as cognitive decline, biomarkers are often employed. A biomarker (or biological marker) is, in the U.S. Food and Drug Administration (FDA) definition, "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention" [11]. Unfortunately, most existing biomarkers for AD are either costly (neuroimaging methods such as positron emission tomography, PET, or magnetic resonance imaging, MRI) or invasive (such as analytes extracted from cerebrospinal fluid, which involve a lumbar puncture procedure). Alternative assessment methods, such as standardised cognitive tests, often suffer from ceiling effects, and are subject to daily fluctuations in cognition and executive function.

As cost-effective and accurate biomarkers of neurodegeneration have been sought in the field of dementia research, speech-based "digital biomarkers" have emerged as a promising possibility. Speech seems particularly well suited for this task, as speech and language convey much information about one's cognitive function, and can be collected in natural settings and over time thus overcoming the daily fluctuations caused by fatigue, low mood, short-term illnesses and text anxiety, which tend to affect cognitive test performance. However, as noted, the general applicability of speech-based digital biomarkers depends on whether they can be deployed in different linguistic contexts. This question has been under-researched in this emerging field. The "ADReSS-M: Multilingual Alzheimer's Dementia Recognition through Spontaneous Speech" challenge enabled the investigation of this issue by defining prediction tasks whereby participants trained their models on English speech data and assessed those models' performance on spoken Greek data. The models submitted to the challenge focused on acoustic and linguistic features of the speech signal whose predictive power were partially preserved across these languages.

ADReSS-M provided a platform for contributions to the application of signal processing and machine learning methods for two tasks: multilingual Alzheimer's dementia detection and cognitive score test predictions. The challenge also stimulated the discussion of machine learning architectures, novel signal processing features, feature selection and extraction methods, and other topics of interest to the growing community of researchers interested in investigating the connections between speech and dementia. A total of 24 research teams from 14 different countries (Belgium, Canada, China, Denmark, India, Finland, Germany, Greece, Poland, Spain, South Korea, Sweden, UK and USA) took part in the challenge, with the majority (17) creating models for both tasks. The approaches adopted by the various research groups that entered the challenge were quite diverse. Feature extraction approaches included acoustic feature extraction using standard feature sets such as eGeMAPS [12], to transcript generation through automatic speech recognition followed by linguistic feature extraction through pretrained multilingual word embedding models, to task-specific feature engineering (to represent speech intelligibility and different pause features, for instance), and combinations of these approaches, sometimes followed by further dimensionality reduction methods. Machine learning approaches included transfer learning using deep learning architectures, conventional machine learning algorithms such as support vector machines, logistic regression, random forests, gradient boosting, and late fusion methods involving combinations of these approaches. Feature fusion combining acoustic, paralinguistic and linguistic features was also often employed.

In what follows we describe the ADReSS-M challenge's modelling tasks, along with their evaluation metrics and ranking procedure, present the data sets in detail, describe our baseline models for the task, present the challenge's results, including a ranking table with the five top-scoring submissions, along with brief descriptions of the methods and approaches used by each of these submissions and a summary of the discussion of the results, and discuss future prospects for this area.

II. The ADReSS-M tasks

The ADReSS-M challenge consisted of two prediction tasks to be attempted by the participants, namely:

1) a classification task (AD detection), where the model will aim to distinguish speech of participants with normal cognition (NC, or control condition) from speech of participants with AD or mild cognitive impairment (MCI), and
2) a cognitive test score prediction (regression) task, where participants were asked to create models for inferring the speaker's Mini-Mental State Examination (MMSE) score based on speech data [REF: Folstein, MF, Folstein, SE, McHugh, PR (1975) Mini mental state: A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12, 188-199.]. The MMSE is a short, psychometrically sound screening tool for measuring cognitive functioning (e.g., orientation, attention, memory, language, visuospatial abilities) with a maximum score of 30 points.

Both tasks involved processing the raw spontaneous speech signal, extraction of features, using whatever preprocessing methods the participant wished to use, and creating the predictive models. No speech segmentation or transcription were provided.

Participants could choose to do one or both tasks. They were provided with a training set and, two weeks prior to the paper submission deadline, with test sets on which they could test their models. Up to five sets of results were allowed for scoring for each task per participant. All attempts had to be submitted together.

As the broader scientific goal of ADReSS-M was to gain insight into the nature of the relationship between speech and cognitive function across different languages, we encouraged participants to upload papers describing their approaches and results to a pre-print repository such as arXiv or medRxiv regardless of their ranking in the Challenge, and to share their code through a publicly accessible repository, if possible using a literate programming environment.

III. The data sets

The ADReSS-M data sets can be downloaded from DementiaBank at https://dementia.talkbank.org/ADReSS-M/, upon request. The training data set consists of spontaneous speech samples corresponding to audio recordings of picture descriptions produced by cognitively normal subjects and patients with an AD diagnosis, who were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination test [13]. The participants were all native speakers of English, and were asked to describe the picture shown in Figure 1.

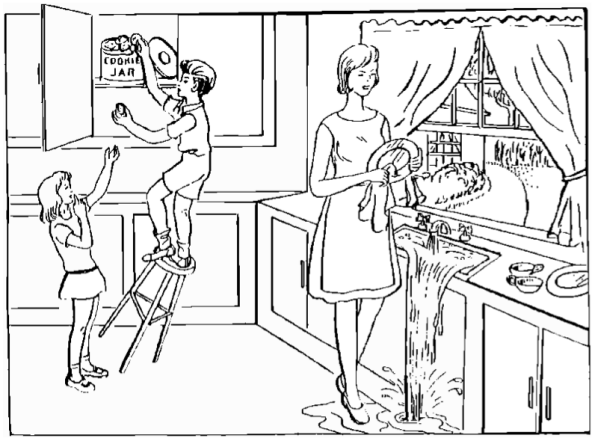

FIGURE 1. Cookie Theft picture from the Boston Diagnostic Aphasia Examination test, used to elicit connected speech for the English language data set.

The test set consists of spontaneous (connected) speech descriptions of a different picture, in Greek. The recordings were made in one of these languages. Participants were initially allowed access only to the training data (in English) and some sample Greek data (8 recordings) for development purposes.

The Greek recordings assess participants' verbal fluency and mood using a picture that the participant describes while looking at it. The assessor first shows the participant a picture representing a lion lying with a cub in the desert while eating, as shown in Figure **??**. The assessor then asks the participants to give a verbal description of the picture in a few sentences. The purpose of this task was to evaluate the participant's ability to generate coherent and descriptive language while also gaining insights into their mood as well as cognitive and emotional responses. By analyzing the language used to describe the picture, researchers can assess the participant's verbal fluency, vocabulary, syntax, and overall linguistic capabilities. Additionally, the context in which the data were collected is crucial to understanding the significance of the task and its findings. This particular task was conducted as part of psychological/ linguistic research study, in order to examine language processing, cognitive abilities, emotional responses and mood-related factors and explore potential connections between language and cognitive states through this assessment.

FIGURE 2. "Male Lion and Cub Chitwa South Africa Luca Galuzzi 2004 edit1" by Luca Galuzzi (Lucag) Edited (noise reduction) by: Arad is licensed under CC BY-SA 2.5.

The training data set was balanced with respect to age and gender in order to eliminate potential confounding and bias. As we employed a propensity score approach to matching we did not need to adjust for education, as it correlates with age and gender, which suffice as an admissible for adjustment (see [14, pp 348-352]). The data set was checked for matching according to scores defined in terms of the probability of an instance being treated as AD given covariates age and gender estimated through logistic regression, and matching instances were selected. All standardized mean differences for the covariates were below 0.1 and all standardized mean differences for squares and two-way interactions between covariates were below 0.15, indicating adequate balance for those covariates. The empirical quantile-quantile (eQQ) plots for the original and balanced data sets [15] are shown in Figure 3. The matched data eQQ shows instances near the diagonal and clear separation of the nominal variables, which indicate good balance.

The mean age, MMSE, and ratios of NC to AD participants are shown in Table 2.

TABLE 1. Descriptive statistics for the ADReSS-M training set (English) by diagnostic category (Dx) and sex. Abbreviations: n = number of participants, sd = standard deviation, MMSE = Mini-Mental State Examination.

Dx	Sex	n	Age (sd)	MMSE (sd)
NC	Female	75	65.6 (6.22)	29.0 (1.29)
NC	Male	40	67.7 (7.12)	28.9 (0.91)
AD	Female	70	69.9 (6.40)	17.4 (5.10)
AD	Male	40	68.4 (7.68)	18.7 (6.08)

IV. Evaluation metrics

The classi cation task is evaluated in terms of accuracy (Ac), speci city (Sp), sensitivity ( ) and F_1 scores. These metrics were computed according to equations (1)-(5).

$$A = \frac{T_n + T_p}{N} \quad (1)$$

$$Sp = \frac{T_n}{T_n + F_p} \quad (2)$$

$$F_1 = 2 \frac{}{+} \quad (3)$$

where N is the number of patients, T_p is the number of true positives, T_n is the number of true negatives, F_p is the number of false positives, F_n is the number of false negatives. The F_1 scores is the harmonic mean of sensitivity and positive predictive value, or precision (noted), computed as shown in equations (4) and (5).

$$= \frac{T_p}{T_p + F_n} \quad (4)$$

$$= \frac{T_p}{T_p + F_p} \quad (5)$$

For the regression task (MMSE prediction), the metrics used are the coef cient of determination and root mean squared error, as set out in equations (6) and (7), respectively, where where $\hat{y}$ is the predicted MMSE score, y is the patient's actual MMSE score, and $\bar{y}$ is the mean score.

$$R^2 = 1 \quad \frac{\sum_{i=1}^{N} (\hat{y}_i \quad y_i)^2}{\sum_{i=1}^{N} (\hat{y}_i \quad \bar{y})^2} \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i \quad y_i)^2}{N}} \quad (7)$$

The ranking of submissions was based on accuracy scores for the classi cation task (task 1), and on RMSE scores for the MMSE score regression task (task 2). The top 5 models comprised:

1) The two top performing (most accurate) teams for the classi cation task
2) The two top performing (least RMSE) teams for the MMSE regression task
3) The team that performed best on average for the two tasks, chosen according to the formula set out in equation (8), where T_i is the total score of team i and T is the total number of teams in the challenge. If a team chose not to submit results for a task, its score for that task was set to 0.

$$T_i = \frac{A_i}{\sum_j^T A_j} + 1 \quad \frac{RMSE_i}{\sum_j^T RMSE_j} \quad (8)$$

Ties were broken by averaging performance over all attempts. These criteria were applied so that the rank resulted in 5 different teams. So, if one team was selected as a top

FIGURE 3. eQQ plots for the original data set and corresponding balanced training data set.

The test set had similar statistical characteristics, but slightly higher average ages and MMSE scores for each category. The detailed composition of the test set is shown in

TABLE 2. Descriptive statistics for the ADReSS-M test set (Greek) by diagnostic category (Dx) and sex.

Dx	Sex	n	Age (sd)	MMSE (sd)
NC	Female	18	66.5 (6.66)	29.0 (1.03)
NC	Male	6	63.5 (9.38)	28.7 (1.63)
AD	Female	17	72.5 (6.97)	20.5 (4.61)
AD	Male	5	72.4 (8.08)	20.8 (4.66)

The training set audio recordings were distributed in MPEG audio layer 2/3 format, with a sample rate of 44,100 Hz and bitrate of 128 kb/s. The test set audio was encoded in 16-bit Signed Integer PCM format, with a sample rate of 22,050 Hz.

team under one of the criteria, it was selected as a top team in another. In such cases, the next top-performing team would be selected.

V. Baseline models

We created baseline models for each task to give the participants an idea of what the use of standard signal processing and machine learning methods could achieve for these tasks on the provided data sets.

In creating these models, we first normalised the volume of the audio files using FFMPEG's EBU R128 scanner filter. A sliding window of 1 s, with no overlap, was then applied to the audio, and eGeMAPS features were extracted over these frames. The eGeMAPS feature set [12] is a basic set of acoustic features designed to detect physiological changes in voice production. It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, totalling 88 features per frame.

Given the eGeMAPS features, we applied the active data representation method (ADR) [16] to generate a frame level acoustic representation for each audio recording. The ADR method has been used previously to generate large scale time-series data representation. It employs self-organising mapping to cluster the original acoustic features (C dimensions that are the number of neurons/ clusters of SOM). Then computes histogram representation of C (as shown in equation 9) for each audio file (i.e. A_i) and first-order derivate features (mean and std i.e. 2 features [16] where the rate of change is given by an approximation of derivative.

$$vADR_{Ai} = \frac{@cADR_{Ai}}{@t}$$

This method is entirely automatic in that no speech segmentation or diarisation information is provided to the algorithm.

$$nADR_{Ai\ norm} = \frac{nADR_{Ai}}{knADR_{Ai}\ k_1} \quad (9)$$

For the AD detection task (task 1), we employed a Naïve Bayes classifier with kernel smoothing estimation. The ADR for feature extraction was optimised using grid search ($C = 5; 10; 15; 20; 25$). We achieved accuracies of 75.00% and 73.91% on sample and test data respectively using 15+2 ADR, age and gender features per recording. On the test set, specificity was 79.2%, precision was 75%, sensitivity was 68.2%, and F_1 was 71.4%. The feature to training audio ratio was 19:237.

For the MMSE regression task (task 2), we employed a support vector machine (SVM) model with a RBF kernel with box constraint of 1, and sequential minimal optimization solver. The ADR for feature extraction was optimised using a grid search ($C = 5; 10; 15; 20; 25$). This model achieved a root mean squared error (RMSE) of 3.887 ($r = 0:348$) and 4.955 ($r = 0:273$) on sample and test data respectively using 25+2 ADR, age and gender features per recording. The feature to training audio recordings ratio was also 29:237.

The source code for data set generation and for the baseline system is available at https://gitlab.com/luzs/madress-2023, with access granted upon request.

FIGURE 4. The ADReSS-M baseline system architecture

VI. Rank of submissions

The submissions were ranked according to the procedure described in section IV. The scores for the top-5 teams (excluding the baseline system) are shown in

The top scoring team, from the Dept of Computer Engineering at Konkuk University and VOINOSIS Inc, South Korea, employed a novel complementary and simultaneous ensemble algorithm (CONSEN) on acoustic and disfluency features, exploring correlations between AD and MMSE predictions to improve performance [17]. The second place employed a mixed-batch transfer learning approach for both tasks, applied to eGeMAPS acoustic features [18]. The third highest scoring team explored a wider number of acoustic feature extraction methods, employing an XGBoost classifier for the classification task and SVM and XGBoost regressors for MMSE prediction [19]. The fourth ranked team employed an automatic speech recognition system to derive speech intelligibility features based on confidence scores assigned by the system which along with word-level duration and pause features formed the input for logistic regression and SVM regression models for tasks 1 and 2, respectively [20]. The fifth place team fused linguistic and acoustic features extracted through speech recognition and pre-trained word embedding and acoustic embedding models and employed neural networks consisting of two fully connected layers and SVMs for classification and regression [21].

The overall accuracy ranking for the participants is shown in Figure 5. It can be observed in this dot chart that there is

TABLE 3. Ranking of teams results by overall composite (T) scores (combined classification and regression results).

Rank	Team	Overall (T)	Detection (A)	MMSE (RMSE)
1	Dept of Computer Engineering at Konkuk University and VOINOSIS Inc, South Korea	1.011	0.870	3.727
2	Katholieke Universiteit Leuven, Belgium	1.002	0.826	4.345
3	University of Science and Technology of China	0.994	0.739	4.610
–	University of Edinburgh Baseline	0.990	0.739	4.955
4	University of Alberta, Canada; ILSP, Athena Research Center, Greece	0.989	0.696	4.769
5	Tsinghua University , China	0.989	0.696	4.788

a considerable gap between the two top-scoring teams and the remaining teams.

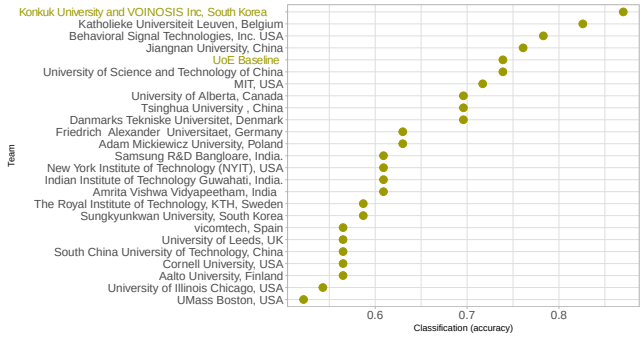

FIGURE 5. AD detection accuracy results.

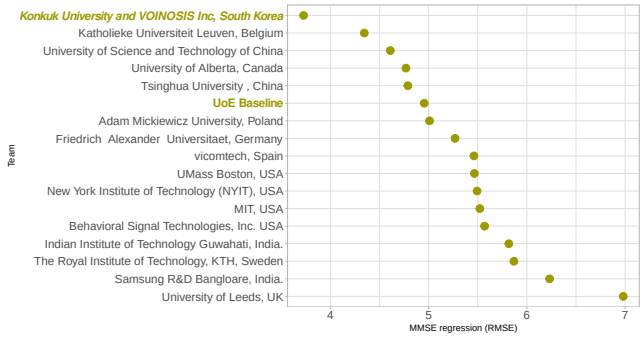

FIGURE 6. MMSE regression results.

A similar pattern can be discerned in the chart depicting the regression results (Figure 6) where the gap between the top scoring team and the remaining teams is even more pronounced. This underscores the effectiveness of the approach of using learning of MMSE scores to leverage classification learning, employed by the winning team.

VII. Brief descriptions of the top-5 submissions

Jin et al. [17] conducted a series of experiments using acoustic, disfluency and fusion of acoustic and disfluency features. They showed that the disfluency feature provides better results than acoustic features and generalizes well across languages. They proposed an ensemble algorithm (CONSEN) and achieved the best-performing results using the fusion of disfluency and acoustic features with an accuracy of 87.0% in AD detection and 3.727 RMSE in MMSE prediction. The unique feature of this top-scoring approach was its leveraging of MMSE prediction as a means to improve AD detection accuracy. While this approach would not be feasible where training data for cognitive testing is not available, it suggests an interesting way of combining speech-based cognitive assessment with better established tests of cognitive function currently in clinical use.

Tamm et al. [18] created models using a sequence of acoustic features and covariates (age, gender education). The models were first trained in English, then transferred to Greek using mixed-language batches and parameter averaging. Results yielded 82% accuracy for AD detection and a RMSE of 4.345 for MMSE score prediction on the test set. For the classification task, the best model had 91.7% specificity, 88.9% precision, 72.7% sensitivity and F1-score of 80.0%. The distinguishing characteristic of Tamm et al.'s approach is their use of the same deep learning architecture for both tasks. Their network architecture consisted of batch normalisation of input features, attention weights computed by two feed-forward layers with dropout and ReLU activation.

Mei et al. [19] provide insights into the methodologies, techniques, and algorithms employed by the USTC team to tackle the ADReSS-M Challenge. It discusses the system's architecture, data preprocessing, feature extraction methods, and machine learning or deep learning models used for emotion recognition in speech. The unique characteristics of the approach described are the use of a 10-dimensional feature set for distinguishing among pauses, following the method proposed in a previous AD detection challenge [22], the fusion of several low-level paralinguistic descriptors used for extraction and fine-tuning of a pre-trained wav2vec2 model [23]. The XGBoost classifier achieved 73.9% accuracy, and the pre-trained bilingual model achieved up to 87.5% in validation against the Greek language samples provided for training. The results indicate that using balanced, low-pass filtered, bilingual speech data in pre-training could be beneficial to multilingual AD detection.

Shah et al. [20] investigated language-agnostic speech representations, which are speech features or characteristics that can be effectively applied across different languages, without requiring language-specific adaptations. The researchers focused on using domain knowledge, likely related to the specific characteristics of AD, to develop and evaluate these speech representations for the purpose of detecting

the early cognitive changes across the AD spectrum. The study explored various machine learning techniques to learn meaningful representations from speech data, considering language-agnostic aspects to ensure the model's generalization across multiple languages. The findings of this research could contribute to the development of robust and language-independent diagnostic tools for AD, making it easier to identify potential patients regardless of their native language. The paper presents a concise overview of the researchers' methodology, experimental results, and implications for future research directions in the domain of speech-based AD detection.

Chen et al. [21] made use of three processing streams. For the extraction of paralinguistic features, they used three different feature sets from the openSmile toolkit. They applied SVM to each separately to perform classification and prediction. The best F1 score for these three analyses was 0.72 for the IS10-Paralilnguistics feature set. For an analysis based on pre-trained acoustic features, they used the XLSR-53 model in openSmile. Although that model has been trained on 53 languages, it does not include Greek and this could have led to a weaker performance for this method. Using the Whisper speech recognition model, they produced English texts from the Greek audio which they used to train a RoBERTa model. This method produced a lower F1 score of 0.55 due to inconsistencies between the pictures described in Greek and those for English. Features from both the XLSR-53 model and the RoBERTa model used a two level fully connected network to generate values for classification and regression.

VIII. Discussion

Computational analysis of spontaneous connected speech has the potential to enable novel applications for speech technology in longitudinal, unobtrusive monitoring of cognitive health. By focusing on AD recognition using spontaneous speech, the ADReSS-M signal processing grand challenge provided a platform for the investigation of alternative to neuropsychological and clinical evaluation approaches to AD detection and cognitive assessment. Furthermore, the multilingual setting provided by ADReSS-M allows the investigation of features that might generalise across languages, extending the applicability of the models. In keeping with the objectives of AD prediction evaluation, the ADReSS-M challenge provided a statistically matched data set so as to mitigate common biases often overlooked in evaluations of AD detection methods, including repeated occurrences of speech from the same participant, variations in audio quality, and imbalances of gender, age and educational level. We hope this might serve as a benchmark for future research on multilingual AD assessment.

ADReSS-M attracted the participation of a large number of participants from leading research labs from across the world, evidencing the relevance of the emerging field of research on speech-based digital biomarkers for AD in general, and on methods that generalise across languages in particular. The diversity of approaches presented by the participanting teams, including proposals for novel acoustic feature sets, the use of pre-trained models, the combination of automatic speech recognition and multilingual embedding models, the use of transfer learning, and a novel ensemble learning method that combines the diagnosis and the cognitive score prediction learning tasks, will hopefully open new avenues for further explorations in this area.

REFERENCES

[1] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic Alzheimer's disease detection from speech and language," *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1784–1797, 2020.

[2] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech and language processing approaches to monitoring Alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, 2020.

[3] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, "DementiaBank: Theoretical rationale, protocol, and illustrative analyses," *ASHA Wire*, Feb. 2023.

[4] A. W. Toga, M. Phatak, I. Pappas, S. Thompson, C. P. McHugh, M. H. Clement, S. Bauermeister, T. Maruyama, and J. Gallacher, "The pursuit of approaches to federate data to accelerate Alzheimer's disease and related dementia research: GAAIN, DPUK, and ADDI," *Frontiers in Neuroinformatics*, vol. 17, pp. 1175689, 2023.

[5] AD Workbench, "Alzheimer's disease data initiative," Web site, 2020, Retrieved from https://www.alzheimersdata.org/.

[6] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 202.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge," in *Proc. Interspeech 2021*, 2021, pp. 3780–3784.

[8] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for detection of Alzheimer's disease," *Frontiers in Computer Science*, vol. 2, pp. 57, 2021.

[9] Z. Shah, J. Sawalha, M. Tasnim, S.-a. Qi, E. Stroulia, and R. Greiner, "Learning language and acoustic models for identifying Alzheimer's dementia from speech," *Frontiers in Computer Science*, vol. 3, pp. 4, 2021.

[10] S. Luz, F. Haider, D. Fromm, and B. MacWhinney, Eds., *Alzheimer's Dementia Recognition Through Spontaneous Speech*, Frontiers Media SA, 2021.

[11] FDA-NIH Biomarker Working Group, *BEST (Biomarkers, EndpointS, and other Tools) Resource*, Food and Drug Administration (US), Bethesda, MD, USA, 2016.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., "The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing," *IEEE Trans Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[13] J. Becker, F. Boller, O. Lopez, J. Saxton, and K. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[14] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2nd edition, 2009.

[15] D. Ho, K. Imai, G. King, and E. A. Stuart, "MatchIt: Nonparametric preprocessing for parametric causal inference," *Journal of Statistical Software, Articles*, vol. 42, no. 8, pp. 1–28, 2011.

[16] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J Sel Top Signal Process*, vol. 14, no. 2, pp. 272–281, 2020.

[17] L. Jin, Y. Oh, H. Kim, H. Jung, H. J. Jon, J. E. Shin, and E. Y. Kim, "CONSEN: Complementary and simultaneous ensemble for Alzheimer's disease detection and MMSE score prediction," in *Procs*