



A Comparative Analysis of Federated Learning for Speech-Based Cognitive Decline Detection

Stefan Kalabakov^{1,*}, Monica Gonzalez-Machorro^{1,2,3,5,*}, Florian Eyben², Björn W. Schuller^{2,3,4}, Bert Arnrich¹

¹Hasso Plattner Institute, University of Potsdam, Germany ²audEERING GmbH, Germany ³MRI, Technical University of Munich, Germany ⁴Imperial College, UK ⁵Munich Center for Machine Learning, Germany

Abstract

Speech-based machine learning models that can distinguish between a healthy cognitive state and different stages of cognitive decline would enable a more appropriate and timely treatment of patients. However, their development is often hampered by data scarcity. Federated Learning (FL) is a potential solution that could enable entities with limited voice recordings to collectively build effective models. Motivated by this, we compare centralised, local, and federated learning for building speech-based models to discern Alzheimer’s Disease, Mild Cognitive Impairment, and a healthy state. For a more realistic evaluation, we use three independently collected datasets to simulate healthcare institutions employing these strategies. Our initial analysis shows that FL may not be the best solution in every scenario, as performance improvements are not guaranteed even with small amounts of available data, and further research is needed to determine the conditions under which it is beneficial.

Index Terms: cognitive decline, speech analysis, federated learning, personalisation, non-iid data

1. Introduction and Related Work

Cognitive decline (CD), i.e., a decrease in mental abilities, is commonly caused by neurodegenerative diseases such as Alzheimer’s Disease (AD) [1]. Despite the lack of a known cure for it, early diagnosis is crucial for interventions like cognitive training [2]. To aid early diagnosis, Mild Cognitive Impairment (MCI) was established as a cognitive state preceding AD [1] as approximately 30% of individuals with MCI develop dementia within five years [3], [4].

Unfortunately, screening for these diseases often involves lengthy, not entirely accurate tests [5]. As language impairment is a common symptom, recent interest has been shown in using speech technology to aid in recognising CD [6]. Several studies have used speech analysis to differentiate between healthy individuals and those with AD [7, 8, 9, 10]. For example, [7] achieved an accuracy of 82% using low-level descriptors of speech, while [8] reported a 70% F1 score when combining acoustic features and embeddings.

Despite these recent contributions, there has been less research on detecting intermediate cognitive states like MCI. This issue is partially caused by the scarcity of MCI data and the difficulty that even clinicians face in identifying patients with MCI [6]. For example, the most commonly used dataset in this field, the ADReSS dataset, includes only 156 participants who are either healthy or have AD. While some studies, such as [11],

have attempted to distinguish between MCI, AD, and Healthy Control (HC)—achieving a 50.7% accuracy in differentiation using a Support Vector Machine (SVM) with temporal speech features—these efforts are still limited.

A potential solution to the challenge of data scarcity is for institutions to share extracted knowledge rather than their data. This aligns with patient privacy laws such as GDPR and HIPPA. Recently, Federated Learning (FL) has emerged as a solution that fits this need [12]. In FL, clients exchange locally computed model weights to collaboratively build a Machine Learning (ML) model, bypassing the need to share private data. So far, FL has shown potential in distinguishing between HC and AD [13], [14]. For instance, [13] achieved an 84% accuracy rate using a simulated FL framework on the ADReSS dataset, employing linguistic and acoustic features.

The potential to recognise various stages of CD, such as MCI and AD, through acoustic information in the context of FL remains unexplored. By pooling knowledge about these often underrepresented classes from multiple clients/institutions, FL could potentially streamline and enhance the accuracy of cognitive disease testing procedures. This paper, therefore, investigates two main questions:

1. Can the inclusion of additional knowledge, from non-independent and identically distributed (non-iid) data, through FL improve model performance on a client’s data compared to models trained solely on that client’s data?
2. Does personalisation on a client’s data improve the performance of FL-produced models on that client’s data?

In essence, our main contribution is the realistic evaluation of FL for developing speech-based ML models capable of recognising intermediate CD states such as MCI, in addition to distinguishing between healthy individuals and those with AD. This evaluation is facilitated by: (i) using multiple datasets to simulate the non-iid data distribution across institutions in the real world, and (ii) comparing FL with paradigms such as Centralised and Local Learning (LL), each of which a possible routes institutions might take when considering the development of ML models.

The paper is organised as follows, Section 2 introduces the datasets, features, and the experimental and evaluation setup. Section 3 presents the results and provides a discussion of the most relevant findings. Section 4 draws conclusions from the results and outlines possible next steps.

2. Materials and Methods

Datasets. To simulate institutions with non-iid data, typically found in real-world deployments, we used four English datasets from DementiaBank [15]. The first, the ADReSS dataset [10], includes acoustic samples from AD and HC speakers. The sec-

Contact: stefan.kalabakov@hpi.de, monica.gonzalez@tum.de, fe@audEERING.com, shuller@tum.de, bert.arnrich@hpi.de

*These authors contributed equally and share first co-authorship

ond dataset contains only the MCI samples from the Pitt Corpus [16], collected under the same recording conditions as the ADReSS dataset. As the Pitt Corpus is a longitudinal study, we specifically chose the last speech samples from each speaker diagnosed with MCI. For our analyses, we combined the samples from the ADReSS dataset with those from the Pitt Corpus into a single dataset, henceforth referred to as the Pitt-ADReSS dataset. The third dataset, the Delaware Corpus [15], includes MCI and HC speakers. Lastly, the fourth dataset, the Lu dataset [15], features HC speakers and AD patients. All audio recordings correspond to the cookie theft picture description task [17].

Preprocessing. To ensure audio quality, we manually inspected all files and performed speaker diarisation as needed. All audio files were resampled to 16 kHz to standardize the sampling rate, and loudness normalization was applied using the normalise function from the pyaudb package (version 0.2.10). In the FL pipeline, each dataset represents a separate institution/client, so all preprocessing steps were performed locally.

We defined a speaker-disjunct held-out test set for each client, comprising 20% of their data. Using an algorithm introduced in [18], we generated 30 potential train-test splits. Each split's quality was assessed using the Jensen-Shannon divergence, measuring the distance between the diagnosis distribution in each split. We selected the split with the lowest distance between the train and test subsets. The selected split's testing subset had an information radius of 0.0 for Delaware, 0.007 for Pitt-ADReSS, and 0.05 for Lu. This results in a binary classification problem when considering the test set of Delaware and a multi-class classification problem when considering the test sets of the Pitt-ADReSS and Lu datasets. Table 1 presents the details of all the datasets used in this study, as well as several distributions across the chosen train-test splits.

Feature extraction. Three types of features were part of the file-level feature extraction: acoustic embeddings from a Wav2Vec2.0 model, the eGeMAPS feature set, and syllable nuclei-based rhythmic features.

We extracted the *acoustic embeddings* using a pre-trained Wav2Vec2.0 model available through HuggingFace ("wav2vec2-base-960h"). The model took raw audio segments and returned contextualised representations. These representations were averaged over the time dimension to produce 768-dimensional embeddings. Before the extraction process, we applied data augmentation techniques including the addition of Gaussian noise (50% probability, 0.001-0.015 noise amplification factor), pitch shifting (-4 to 4 semitones, 50% probability), time stretching (rate of 0.8-1.25, 50% probability), and audio shifting (fraction of -0.05-0.5, 50% probability). Because the dimensionality of these embeddings was large, we also constructed a version of this feature set using Principal Component Analysis (PCA), selecting the top 80 components and preserving 98.91% of the variance.

The *acoustic features from the eGeMAPS feature set* [19] were extracted using the Speech & Music Interpretation by Large-space Extraction (openSMILE) [20] feature extraction tool. This set comprises 88 extensively studied features related to prosody, voice quality, and articulation [21]. We aggregated the functionals of all features on the file level.

We also extracted *syllable nuclei-based rhythmic features*, as features related to pause information have previously demonstrated effectiveness in dementia detection [22], [23]. These features were extracted using Praat [24]. We identified syllable nuclei using peak detection with a dip in intensity between syllable peaks lower than 2 dB. Voiced peaks were considered syllable nuclei, with a silence threshold set to -25 dB and a min-

imum pause duration of 0.3 seconds [25], [26]. A total of 16 features, including the number of pauses, duration, phonation time, speech rate, and articulation rate, were calculated.

Lastly, we combined the interpretable features by merging the eGeMAPS features with syllable nuclei-based features from Praat, creating a 104-feature set we refer to as *the combined handcrafted feature set*.

2.1. Experimental Setup

In this paper, we explored the effectiveness of FL by comparing it with two other learning paradigms, specifically, Centralised Learning (CL) and LL. For each learning paradigm, we employed a ML pipeline that included either a Multilayer Perceptron (MLP) or an eXtreme Gradient Boosting (XGBoost) classification tree. The MLP comprised two fully connected hidden layers, initialised using the Kaiming He method [27]. We followed each of these hidden layers with a ReLU activation function [28] and a dropout layer [29].

In the CL paradigm, we trained and evaluated models using the pooled data of all clients. For LL, we trained and evaluated models on each client individually, using their respective training and testing subsets. In contrast, FL used the training subsets of all clients to train a shared model, ensuring that the data never left the respective client's infrastructure. We then evaluated this shared model on each client's testing subset.

More specifically, one round of shared model training in FL included the following steps: (i) a parameter server initialised a shared model, selected a random subset of clients to participate in the round of training, and sent the weights of the shared model to all participating clients, (ii) the involved clients computed the gradients/weight updates for that model based on their data and sent them back to the parameter server, (iii) depending on the type of FL strategy used, the parameter server performed an aggregation operation on the received gradients/weight updates and applied them to the shared model. After step (iii), the parameter server sent the weights of the updated shared model to the clients selected for participation in the next round. We used the Federated Averaging (FedAvg) algorithm [12] as the FL strategy for the MLP models, while XGBoost training was facilitated by a simple bagging aggregation algorithm available through the Flower framework version 1.7.0 [30].

In addition to standard FL, we also explored personalised FL models. We created these personalised FL models by fine-tuning the shared FL model for a few epochs on each client's training subset.

Hyperparameter tuning. To optimise the parameters of the models and the training process, we used Bayesian hyperparameter tuning on the training subsets of all clients. The tuning consisted of 100 iterations, each comprising a 5-fold Cross Validation (CV) repeated three times with different random seeds to mitigate stochastic initialisation. We averaged the metrics from these three runs to obtain the final results for each parameter combination. Importantly, we ran the optimisation for every combination of feature type, model type, and learning paradigm. The target of the search in each scenario was the weighted mean of the macro F1-scores across clients, with weights proportional to the client sizes.

The parameters we tuned and their ranges are detailed in Table 2. The chosen parameters for each combination of learning paradigm, feature type, and model type as well as the computing infrastructure used can be found in our repository¹.

¹<https://github.com/HPI-CH/FL4CDD-interspeech2024>

Table 1: The number of participants and their distributions with respect to gender, age, diagnosis, and audio length in the train and test splits of each of the three datasets.

Subset	Dataset	Total # participants	Gender distribution	Age distribution	Diagnosis distribution (# participants)	Audio length distribution per diagnosis (min)
Train	Delaware	44	24F, 20M	72.50 ± 8.56	MCI:28, HC:16	MCI:22.47, HC: 14.23
	Pitt-ADReSS	137	77F, 60M	66.03 ± 6.89	MCI:13, HC:62, AD:62	MCI:11.07, HC:36.83, AD:29.46
	Lu	40	24F, 16M	80.14 ± 10.59	HC:20, AD:20	HC: 21.71, AD: 19.54
Test	Delaware	11	9F, 2M	73.73 ± 6.0	MCI:7, HC:4	MCI: 5.78, HC: 2.23
	Pitt-ADReSS	35	18F, 17M	67.32 ± 7.03	MCI:3, HC:16,AD:16	MCI:3.01, HC:8.67, AD:8.27
	Lu	12	8F, 4M	77.17 ± 7.52	MCI:2,HC:5, AD:5	MCI:1.40, HC:4.72, AD:3.44

Table 2: The parameters and ranges considered during hyperparameter tuning.

	Parameter	Range	
General	class_weight_exponent	[0, 1.5]	
	C	[0.33, 0.66, 1]	
FL-specific	num_rounds	[1, 20]	
	num_per_epochs (MLP)	[1, 20]	
perFL-specific	per_n_estimators (XGBoost)	[1,20]	
	batch_size	[8, 16, 32, 64]	
MLP-specific	num_neurons_per_layer	[[16,8], [32,16], [32, 32], [64, 32], [64, 64], [128, 64], [512, 64]]	
	num_epochs	[1,15]	
	dropout_prob	[0.1, 0.4]	
	learning_rate	[1e-05, 0.1]	
	weight_decay	[1e-06, 1e-03]	
	XGB-specific	colsample_bytree	[0.01, 1]
		eta	[1e-05, 1]
n_estimators		[2, 200]	
max_depth		[2, 50]	
	reg_lambda	[1, 10]	
	subsample_ratio	[0.01, 1]	

Some parameters were not subject to tuning. Instead, their values were predetermined. These include: (i) the adoption of the AdamW optimiser, (ii) the application of an exponential learning rate scheduler with a gamma value of 1.0, and (iii) the utilisation of the “hist” tree method in the context of XGBoost.

Evaluation protocol. Upon determining the optimal hyperparameters, we evaluated the resultant models on a per-client basis using the held-out test sets. The evaluation was conducted across 10 runs, each with a unique random seed ranging from 0 to 10. For each run, we computed several per-client metrics, including accuracy (for comparison with related work), macro F1-score, Unweighted Average Recall (UAR), precision for each class, recall for each class, and the confusion matrix.

To compute per-client metrics across multiple runs, we calculated the mean and standard deviation of all per-client metrics over these runs. To obtain results that span across both clients and runs, we computed the mean and standard deviation of all per-client metrics across runs. A special case is the weighted average of the macro F1-score across clients and across runs, which is calculated by determining the weighted mean (proportionate to client size) of the macro F1-scores obtained by all clients across the 10 runs. This metric is used as the target in

the Bayesian hyperparameter tuning.

Finally, to generate a measure of statistical significance, we computed the 95% Confidence Interval (CI) for the per-client macro F1-scores across runs, as well as for the macro F1-score across clients and runs (across client and runs results are only available in our repository¹). The CIs were calculated using 1000 bootstrapping iterations [31].

3. Results and Discussion

Table 3 displays the outcomes for the best-performing parameters for each combination of feature type and learning strategy when an XGBoost classifier tree was used. Owing to space constraints, only the mean macro F1-Score per client and its 95% CI are included, the rest of the calculated metrics and the confusion matrices are available in our repository¹.

Firstly, our results indicate that an XGBoost classifier consistently outperforms an MLP in all scenarios, regardless of the learning strategy, feature set, or client. Secondly, the embeddings produced by the pretrained Wav2Vec2 model (both the full set and the PCA set) perform considerably worse on all clients compared to other feature types. This performance, coupled with their lack of explainability, renders the acoustic embeddings less suitable for this task than handcrafted features. Consequently, the results obtained using the MLP and these embeddings are not included, but are available in our repository¹.

Interestingly, different feature types seem better suited to different combinations of learning strategies and clients. For example, the eGeMAPS feature set seems more compatible with the CL and LL learning strategies, as it produced the best results for them in the case of the Pitt-ADReSS and Lu datasets/clients. Conversely, the combined handcrafted features demonstrated better performance when personalised FL was employed, producing the best personalised FL results on both the Delaware and Lu datasets/clients.

Additionally, each learning strategy — CL, LL, and FL — seems to produce the best overall results for one of the three datasets. Specifically, CL achieves the best outcome for the Pitt-ADReSS dataset/client, LL for the Lu dataset/client, and personalised FL for the Delaware dataset/client. This suggests that the optimal learning strategy depends on the dataset and chosen features, rather than a universal solution.

Further, we observe that in the case of the Pitt-ADReSS dataset/client, CL using the eGeMAPS feature set slightly improves results compared to LL, as it performs better for the HC and AD classes, while very slightly worsening the performance on the MCI class. However, this improvement may not justify the regulatory and logistical challenges associated with centralising healthcare data.

On the other hand, when comparing personalised FL with

Table 3: Per-client results on the held-out test set of the best-performing models for each combination of feature type and learning strategy when using an XGBoost classifier tree. The best overall centralised learning results, per dataset, are marked with an asterisk (*), the best local learning results with a dagger (†), and the best federated learning ones with a double dagger (‡). The best overall result is shown in **bold**.

Feature type	Learning strategy	Mean Macro F1-score Pitt-ADReSS (95%CI)	Mean Macro F1-score Delaware (95%CI)	Mean Macro F1-score Lu (95%CI)
eGeMAPS	CL	0.792 (0.743-0.835)*	0.515 (0.414-0.61)	0.454 (0.4-0.501)*
	LL	0.778 (0.744-0.811)†	0.607 (0.514-0.697)	0.495 (0.438-0.544)†
	FL	0.367 (0.319-0.417)	0.399 (0.337-0.455)	0.345 (0.268-0.427)
	perFL	0.754 (0.710-0.794)‡	0.728(0.634 - 0.811)	0.4 (0.335-0.463)
Syllable-based Rhythmic	CL	0.707 (0.663-0.747)	0.726 (0.642-0.805)*	0.222 (0.147-0.299)
	LL	0.716 (0.671-0.756)	0.643 (0.545-0.730)	0.301 (0.238-0.356)
	FL	0.449 (0.388-0.504)	0.380 (0.312-0.438)	0.327 (0.246-0.407)
	perFL	0.648 (0.584-0.700)	0.502 (0.411-0.589)	0.307 (0.247-0.369)
Combined Handcrafted	CL	0.774 (0.738-0.809)	0.657 (0.560-0.745)	0.430 (0.370-0.483)
	LL	0.767 (0.732-0.797)	0.718 (0.630-0.804)†	0.452 (0.387-0.508)
	FL	0.436 (0.382-0.492)	0.379 (0.315-0.437)	0.302 (0.228-0.374)
	perFL	0.722 (0.683-0.764)	0.798 (0.718-0.871)‡	0.413 (0.345-0.474)‡

LL using the combined handcrafted features on the Delaware dataset/client, we notice a larger performance boost favouring personalised FL. This improvement is primarily due to personalised FL substantially improving classifier performance on the HC class, while having a minimal effect on the MCI class.

When analysing the results for the Lu dataset/client, we find that leveraging knowledge from non-iid data through CL or FL does not improve model performance compared to LL with the eGeMAPS feature set. Despite the presence of the MCI class in the test set of the Lu dataset/client (absent in the training set), personalised FL fails to accurately predict any MCI instances and causes model performance deterioration on the HC and AD classes. However, the limited MCI data in the test subset could influence the significance of this observation. Comparing CL and LL with eGeMAPS features on this dataset shows a slight performance increase for the MCI class with CL, but a larger performance decrease for the HC class.

We also observe that personalising FL models consistently improves performance across various feature types and datasets, when compared to standard FL, except when personalising the FL model on the Lu dataset/client with rhythmic features.

Finally, to establish the legitimacy of our pipelines, our results for the most used dataset in the related work, i. e., the Pitt-ADReSS dataset, are competitive with those of related works that employ the ADReSS dataset, i. e., distinguish only between HC and AD [10]. To the best of our knowledge, limited research has been conducted on other datasets due to their recent introduction and small size.

Considering that, CD might be a spectrum with at least seven stages, and that MCI might not be a valid diagnosis [4], one limitation of this study is the use of three categories to represent participants’ cognitive states. Focusing on predicting the outputs of cognitive tests such as the Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA), rather than focusing solely on classification, might be a better option. However, this was not possible in this study, as the Lu dataset does not provide such outputs and the other datasets included result of differing cognitive tests. A further limitation was the small size of our datasets, reducing the generalisability of our observations. Despite our efforts to utilise all available English datasets from DementiaBank, including those with the picture description task, we obtained examples from merely 279

speakers. Finally, it is important to note that while this paper primarily examines speech-based analyses using acoustic information, incorporating linguistic information could enhance performance, as per previous studies [21]. Despite the potential of acoustic data [8], it may not fully capture the complexity of CD.

4. Conclusion and Future Work

This paper focused on the realistic evaluation of FL for developing speech-based ML solutions to detect CD. We used multiple datasets to simulate the non-iid data of institutions that may employ different learning strategies such as CL, LL, and (personalised) FL. Our results indicated that while (personalised) FL may not always be optimal, it produces superior results for certain combinations of datasets and feature types and the contexts in which it is useful need further exploration. Additionally, we also observed that personalising FL models consistently improved performance compared to standard FL, which means that the evaluation of more sophisticated personalisation methods should be an important task in future work.

Moreover, since this paper focuses solely on English speech data, in future work, we plan to explore the generalisation to other languages. In that context, selecting interpretable features, whether acoustic or linguistic, which are also language-agnostic, becomes particularly important. Finally, to mitigate the lack of data, future work should also consider including different speech tasks. Although the results from the picture description task show promise, they do not reflect natural speech.

5. Reproducibility

The data used in this paper can be requested from DementiaBank [15]. For reproducibility purposes, the train-test splits of all datasets, the best hyperparameters, the full set of results, and the code can be found in our repository¹.

6. Acknowledgements

This research was partially supported by the Munich Center for Machine Learning, with additional funding from the BMBF ”autoAAT” (”Automatische Auswertung von Spontansprachinterviews des Aachener Ahpasia Tests”) project (grant agree-

ment: 13GW0489A). The project was also partially funded by the BMBF project "Pre-Care ML" ("Vorhersage schwerwiegender kardiovaskulärer Ereignisse durch maschinelles Lernen", grant agreement: 01KU2207), under the frame of ERA PerMed.

7. References

- [1] J. Hugo and M. Ganguli, "Dementia and cognitive impairment: epidemiology, diagnosis, and treatment," *Clin. Geriatr. Med.*, vol. 30, no. 3, pp. 421–442, Aug. 2014.
- [2] M. Pais, L. Martinez, O. Ribeiro, J. Loureiro, R. Fernandez, L. Valiengo, P. Canineu, F. Stella, L. Talib, M. Radanovic, and O. V. Forlenza, "Early diagnosis and treatment of alzheimer's disease: new definitions and challenges," *Rev. Bras. Psiquiatr.*, vol. 42, no. 4, pp. 431–441, Aug. 2020.
- [3] A. Lanzi, S. E. Wallace, and M. Bourgeois, "Group external memory aid treatment for mild cognitive impairment," *Aphasiology*, vol. 33, no. 3, pp. 320–336, Mar. 2019.
- [4] J. Rasmussen and H. Langerman, "Alzheimer's disease – why we need early diagnosis," *Degener. Neurol. Neuromuscul. Dis.*, vol. 9, pp. 123–130, Dec. 2019.
- [5] A. T. Patocskaï, M. Pákáski, G. Vincze, M. Fullajtár, I. Szimjanovszki, G. Drótos, K. Boda, Z. Janka, and J. Kálmán, "Is there any difference between the findings of clock drawing tests if the clocks show different times?" *J. Alzheimers. Dis.*, vol. 39, no. 4, pp. 749–757, 2014.
- [6] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review," *J. Alzheimers. Dis.*, vol. 78, no. 4, pp. 1547–1574, 2020.
- [7] R. Chakraborty, M. Pandharipande, C. Bhat, and S. K. Koppurapu, "Identification of dementia using audio biomarkers," *ArXiv*, vol. abs/2002.12788, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211572771>
- [8] A. Balagopalan and J. Novikova, "Comparing Acoustic-Based Approaches for Alzheimer's Disease Detection," in *Proc. Interspeech 2021*, 2021, pp. 3800–3804.
- [9] K. Mei, X. Ding, Y. Liu, Z. Guo, F. Xu, X. Li, T. Naren, J. Yuan, and Z. Ling, "The usc system for adress-m challenge," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2.
- [10] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176.
- [11] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech Language*, vol. 53, pp. 181–197, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S088523081730342X>
- [12] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [13] S. I. Ali Meerza, Z. Li, L. Liu, J. Zhang, and J. Liu, "Fair and privacy-preserving alzheimer's disease diagnosis based on spontaneous speech analysis via federated learning," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, Jul. 2022.
- [14] X. Ouyang, "Design and deployment of multi-modal federated learning systems for alzheimer's disease monitoring," in *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, ser. MobiSys '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 612–614. [Online]. Available: <https://doi.org/10.1145/3581791.3597505>
- [15] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, "DementiaBank: Theoretical rationale, protocol, and illustrative analyses," *Am. J. Speech. Lang. Pathol.*, vol. 32, no. 2, pp. 426–438, Mar. 2023.
- [16] J. T. Becker, "The natural history of alzheimer's disease," *Arch. Neurol.*, vol. 51, no. 6, p. 585, Jun. 1994.
- [17] E. Kaplan, *Boston diagnostic aphasia examination booklet*. Lea & Febiger Philadelphia, PA, 1983.
- [18] M. Gonzalez-Machorro, P. Hecker, U. D. Reichel, H. N. Hammer, R. Hoepner, L. Pedrotti, A. Zmutt, H. Sagha, J. van Beek, F. Eyben, D. M. Schuller, B. W. Schuller, and B. Arnrich, "Towards Supporting an Early Diagnosis of Multiple Sclerosis using Vocal Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 1518–1522.
- [19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
- [20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: <https://doi.org/10.1145/1873951.1874246>
- [21] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only," in *Proc. Interspeech 2021*, 2021, pp. 3830–3834.
- [22] Y. Zhu, B. Tran, X. Liang, J. A. Batsis, and R. M. Roth, "Towards interpretability of speech pause in dementia detection using adversarial learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022.
- [23] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs," in *Proc. Interspeech 2021*, 2021, pp. 3820–3824.
- [24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2023. [Online]. Available: <http://www.praat.org>
- [25] N. H. de Jong, J. Pacilly, and W. Heeren, "PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically," *Assess. Educ.*, vol. 28, no. 4, pp. 456–476, Jul. 2021.
- [26] C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Towards Reference Speech Characterization for Health Applications," in *Proc. INTERSPEECH 2023*, 2023, pp. 2363–2367.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.
- [28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, p. 807–814.
- [29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: <http://jmlr.org/papers/v15/srivastava14a.html>
- [30] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, H. L. Kwing, T. Parcollet, P. P. d. Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.
- [31] L. Ferrer, "ConfidenceIntervals: Confidence interval computation for evaluation in machine learning using the bootstrapping approach."