2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871204

Fair and Privacy-Preserving Alzheimer's Disease Diagnosis Based on Spontaneous Speech Analysis via Federated Learning

Syed Irfan Ali Meerza[1], Zhuohang Li[1], Luyang Liu[2], Jiaxin Zhang[3] and Jian Liu[1]

Abstract— **As the most common neurodegenerative disease among older adults, Alzheimer's disease (AD) would lead to loss of memory, impaired language and judgment, gait disorders, and other cognitive deficits severe enough to interfere with daily activities and significantly diminish quality of life. Recent research has shown promising results in automatic AD diagnosis via speech, leveraging the advances of deep learning in the audio domain. However, most existing studies rely on a centralized learning framework which requires subjects' voice data to be gathered to a central server, raising severe privacy concerns. To resolve this, in this paper, we propose the first federated-learning-based approach for achieving automatic AD diagnosis via spontaneous speech analysis while ensuring the subjects' data privacy. Extensive experiments under various federated learning settings on the ADReSS challenge dataset show that the proposed model can achieve high accuracy for AD detection while achieving privacy preservation. To ensure fairness of the model performance across clients in federated settings, we further deploy fair aggregation mechanisms, particularly q-FEDAvg and q-FEDSgd, which greatly reduces the algorithmic biases due to the data heterogeneity among the clients.**

Clinical Relevance– **The experiments were conducted on publicly available clinical datasets. No humans or animals were involved.**

I. INTRODUCTION

Alzheimer's disease (AD), as the leading cause of dementia, significantly affects patients' memory, cognition, and behavior. As Alzheimer's progresses, the associated symptoms will grow severe enough to interfere with their daily tasks. Due to the irreversible nature and the long preclinical phase of AD, the initiation of intervention or preventive non-pharmacological strategies during the early stage of the disease is critical for reducing the risk of progression [1].

Due to the memory barriers caused by AD, people with AD have difficulty forming essential words or understanding discussions. Common symptoms include slurring, stammering, repeating, and the use of inappropriate words or phrases. Compared to traditional detection approaches (e.g., neuroimaging and cerebral spinal fluid (CSF) [2]), which are invasive and usually limited to clinical usage, speech-based AD screening is more flexible and has excellent potential for large-scale and long-term deployment as speech data can be gathered passively, organically, and continually throughout the day. Recently, there have been active research efforts in automatic AD screening via speech analysis and AI-based learning models. For instance, Weiner *et al.* [3] applied Linear Discriminant Analysis (LDA) and utilized acoustic features to detect people with Alzheimer's. Ambroini *et al.* [4] choose the pitch, voice breaks, shimmer, speech tempo, and syllable duration to detect AD using logistic regression, support vector machines, random forest, k-nearest neighbors, and Adaboost. With recent advancements of deep learning, Tifani *et al.* [5] used the Gated Convolutional Neural Network (GCNN) with speech audio and Mel-frequency cepstral coefficient (MFCC) features. Morteza *et al.* [6] combined the audio, lexical, and disfluency features to predict MMSE score using multi-modal Long short-term memory (LSTM) network.

Although the aforementioned studies can reasonably perform well in screening patients with AD, they all rely on centralized learning, which requires storing the private labeled data in a central server. This manner will inevitably raise the risks of severe data breaches and limited transparency on the system. To alleviate these concerns, federated learning (FL) [7], first formulated by Google in early 2017, is considered the most promising approach to privacy-preserving AI. Specifically, FL distributes the training process to end-user devices, enabling them to collaboratively learn/update a global model using the data kept locally on the device. In recent years, FL has been extensively used in healthcare to ensure the privacy of users' data. For example, Sadilek *et al.* augmented centralized models to federated settings for several diseases [8], and Weishan *et al.* used FL for COVID-19 detection [9]. Most of these studies primarily focus on physical health and medical image analysis. To the best of our knowledge, there has no prior research focusing on deploying speech-based AD diagnosis models in FL settings.

In this paper, to fill this gap, we propose the first federated learning framework for achieving fair and private Automatic Alzheimer's Diagnosis (AAD) through spontaneous speech analysis. By allowing the clients to keep their private speech data on their devices, the proposed framework can alleviate many privacy risks in the current centralized solutions. To better process the speech data and improve the diagnosis accuracy, our model is designed to contain two subnetworks: a Long Short-term Memory (LSTM) network to process the acoustic mel-frequency features and a feed-forward neural network to handle the linguistic feature (i.e., pause rate and duration). Despite its privacy benefits, the distributed nature of FL would introduce serious fairness issues due to the potential bias in heterogeneous training datasets against certain clients or demographic groups. As a result, although the diagnosis model can achieve reasonably high overall ac-

This work was supported in part by NSF CNS-2114161, ECCS-2132106, and CBET-2130643

[1]Syed Irfan Ali Meerza, Zhuohang Li, and Jian Liu are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA (e-mail: $smeerza@vols.utk.edu, zli96@vols.utk.edu, jliu@utk.edu$)

[2]Luyang Liu is with Google Research, Mountain View, CA, USA (e-mail: $luyangliu@google.com$)

[3]Jiaxin Zhang is with Oak Ridge National Laboratory, Oak Ridge, TN, USA (e-mail: $zhangj@ornl.gov$)

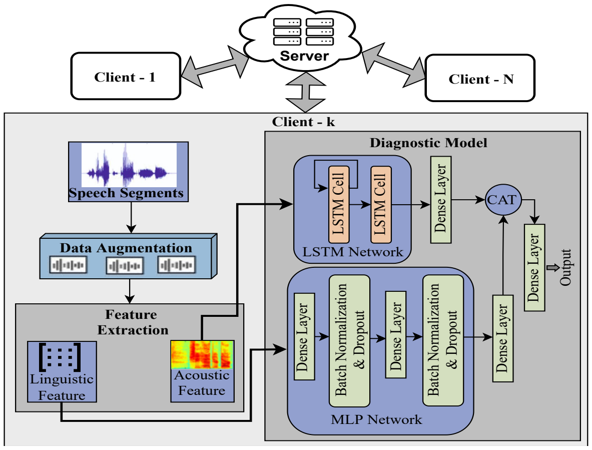

Fig. 1. System Overview.

curacy, there is no performance guarantee for each client due to the data heterogeneity (e.g., some clients with sufficient data may have a relatively high accuracy than others).

To address this issue, we further deploy two fair aggregation mechanisms, i.e., q-FedAvg and q-FedSGD [10]. By minimizing an aggregate reweighted loss specified by q, clients with more significant loss will be given a proportionately higher weight. This helps enable a client-level notion of fairness in the federated setting, which generalizes conventional accuracy parity. We conduct extensive evaluation using the ADDReSS challenge dataset [11] provided by DimentaiBank under different FL settings, and the results show that the proposed method can achieve comparable performance over existing centralized-learning-based approaches while ensuring privacy and fairness.

II. Automatic Diagnosis of Alzheimer's Disease via Federated Learning

Fig. 1 shows the overall architecture of our proposed automatic AD diagnosis system, consisting of *Data Augmentation*, *Feature Extraction*, and *Diagnostic Model*.

A. Data Augmentation

Augmenting data artificially is a systematic strategy for increasing the diversity of the dataset. To make our model robust against possible variations of the audio data, we create three new samples from each speech segment through the following audio transformations independently: (1) *Additive Noise*: White Gaussian Noise with a mean of 0 and standard deviation of 1 is added to create noisy samples to make our model robust against environmental noises; (2) *Pitch Shifting*: we also randomly modify the frequency of parts of the sound to change the octave of the audio samples; and (3) *Time Shifting*: we shift the audio by a factor of $\frac{f_s}{10}$ (f_s denotes the sampling rate) to make the audio signal shift right with a rollover.

B. Feature Extraction

Acoustic Feature. The Mel Frequency Cepstral Coefficients (MFCCs) of an audio signal are a set of features that concisely describe the overall shape of a spectral envelope. Due to its similarity with the human ear's auditory characteristics, it is a prime choice of features for speech-related tasks. Thus we derive 13-dimensional MFCCs using 26 filters in the Mel filter bank, with a window size of 25ms and a step size of 10ms from each speech segment. The range values of all the MFCC coefficients are normalized to avoid the non-uniformity arising from various range values of MFCCs.

Linguistic Feature. To further improve the model's performance, we also extract features associated with the fluency of the speech, such as the pause rate and length, which indicate speech impairments such as slurring and stuttering. To extract this kind of feature, we first split the audio file into 150ms chunks. We can use one of the following four cases to describe the speech/pause information of each short chunk: 1) only silence; 2) only speech; 3) silence then speech, and 4) speech then silence. Thus, we use one-hot encoding to denote all the possible features in a chunk. This creates a linguistic feature matrix with four columns and $A_d/150$ rows, where A_d denotes the audio duration in milliseconds. Finally, we add padding at the end of the feature matrix to make all the linguistic feature matrix to the same size.

C. Diagnostic Model

As shown in Fig. 1, the designed diagnostic classification model consists of two sub-network to process acoustic features and linguistic features, respectively. The final dense layers of this two sub-network will be concatenated as the final dense layer of the diagnostic model. Specifically, for handling time-frequency MFCC acoustic features, we use Long Short-term Memory (LSTM) network [12] to learn the temporal relations from the sequence of speech. LSTM has the ability to bridge long time lags which makes it suitable for our task. It learns to categorize AD by modeling the temporal relationships of speech and adding a feedback loop between the neural network's input and output, leveraging the ability to learn, retain and forget information in long dependencies. Our model structure contains 64 LSTM units in two layers, followed by two fully connected layers with 64 and 32 hidden units with ReLU activation function and a dropout layer with a rate of 0.2. For the linguistic-feature sub-network, we use a multilayer perceptron (MLP) with three fully connected layers with 16, 32, and 32 hidden units with ReLU activation function, batch normalization, and dropout layer. Both models are concatenated in the final layer to form a joint network. An output layer consisting of one neuron with Sigmoid activation is used to predict the label. The model is trained using the binary cross-entropy loss function and Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and momentum of 0.9.

D. Model Training in FL

Privacy-preserving Federated Learning. The FL framework is based on server-client communication. In the beginning, the server randomly initializes a global model and randomly selects a small subset of clients s_t with a fixed ratio to participate in that communication round t. The server shares the global model weights w^t with the participating clients, and the clients perform several training steps on the received model using their local data. All the participating clients send their local model updates to the server, where these updates will be averaged following a particular aggregation rule to update the shared global model to w^{t+1}.

Aggregation Rules. In this paper, we deploy and evaluate three aggregation rules, including the standard FedAvg, and two other fairness-oriented aggregations rules, namely, q-FedSGD, and q-FedAvg.

- *FedAvg* [7] works simply by applying E epochs of SGD locally to each chosen device and then averaging the resultant local models using $w_{t+1} = \sum_{k=1}^{N} \frac{n_k}{n} w_k^t$ where N is the number of participants in each round and w_k is the clients' local update, n is the total data sample of the participating clients and n_k is the number of data sample hold by client k. Unfortunately, this approach might introduce widely varying performance between different devices.
- *q-FedSGD* [10] is one of the fair aggregation methods we used. The server aggregates the local update using the upper bound of the local Lipschitz constants of the gradients and ∇F_k, F_k of the client k. F_k is the binary cross-entropy loss over local data. The server update is calculated according to
$$w^{t+1} = w^t - \frac{\sum_{k \in s_t} \Delta_k^t}{\sum_{k \in s_t} h_k^t}, \tag{1}$$
where local gradient $\Delta_k^t = F_k^q(w^t)\nabla F_k(w^t)$ and heuristic $h_k^t = qF_k^{q-1}(w^t)||\nabla F_k(w^t)||^2 + LF_k^q(w^t)$. w^t is the server model weights at round t, L is the Lipschitz constant, and q is a parameter that tunes the amount of fairness we wish to impose. ∇ denotes the gradient operator while Δ denotes the difference operator.
- *q-FedAvg* [10] is another mechanism to aggregate the updates considering the fairness. Similar to q-FedSGD, the step size is calculated dynamically using the upper bound of the local Lipschitz constants of the gradients. To extend the local updating technique of FedAvg, a heuristic where local updates are obtained by running SGD locally on device k is used. The server update is calculated using
$$w^{t+1} = w^t - \frac{\sum_{k \in s_t} \Delta_k^t}{\sum_{k \in s_t} h_k^t}, \tag{2}$$
where the change of weight $\Delta w_k^t = L(w^t - \bar{w}_k^{t+1})$, local gradient $\Delta_k^t = F_k^q(w^t)\Delta w_k^t$, and heuristic $h_k^t = qF_k^{q-1}(w^t)||\Delta w_k^t||^2 + LF_k^q(w^t)$.

III. Evaluation

A. Database

To evaluate our system, we use the ADReSS challenge dataset [11] provided by DimentiaBank, which consists of the recordings of 108 subjects performing spoken picture descriptions task known as Cookie Theft picture from the Boston Diagnostic Aphasia Exam [13]. Note that we only used 104 subjects' data as the rest audio files are only a few seconds. Among them, 53 subjects have been diagnosed with Alzheimer's disease. Each speech recording was segmented for voice activity with a maximum duration of 1 second per speech segment. The dataset we used contains 12,008 speech segments from 51 non-AD subjects and 17,672 speech segments from 53 AD subjects. We randomly choose 90 subjects' data for training and use the other 14 subjects' data for testing the model performance.

TABLE I

Comparison of different FL setups

Scenario	FL Setting	Accuracy	Precision	Recall	AUC
Centralized	-	0.891	0.935	0.923	0.932
FL IID	Cross-Silo-10	0.873	0.895	0.872	0.911
	Cross-Silo-15	0.821	0.869	0.882	0.912
	Cross-Silo-30	0.805	0.837	0.852	0.866
	Cross-Device-90	0.725	0.765	0.737	0.762
FL Non-IID	Cross-Silo-10	0.844	0.867	0.845	0.919
	Cross-Silo-15	0.818	0.849	0.864	0.904
	Cross-Silo-30	0.779	0.772	0.736	0.818
	Cross-Device-90	0.715	0.698	0.672	0.705

B. Experimental Setup

FL can be categorized in two major sets based on the participation behavior of the clients: (1) *Cross-Silo* setting, which infers learning across databases that contain data for multiple users, such as a healthcare analytics server federating across hospitals and clinics; (2) *Cross-Device* setting, is used for learning across user devices that include data produced by a single user. In addition, there are primarily two FL situations in terms of data distribution: (1) *IID* scenario, in which data is distributed independently and identically among clients; and (2) *Non-IID* scenario, in which each client stores data from various distribution patterns. In this paper, we evaluate the proposed AD diagnostic model under various FL settings. In the IID scenario, we have four settings: Cross-Silo-10, 15, 30, and Cross-Device-90 where we have 10, 15, 30, and 90 clients, respectively, and each client holds $1/N$ (where N is the number of clients) data that is randomly non-repetitively selected from the training set. In the Non-IID scenario, we have the same number of clients as in IID, but in Cross-Silo, each client contains 50% Non-AD subjects' and 50% AD subjects' data, while in Cross-Device, each client holds each subject's data.

C. Evaluation Metrics

To evaluate the model performance in different FL settings, we use the following metrics: (1) *Accuracy*: the ratio of correctly predicted speech segments to the total segments; (2) *Precision*: the ratio of correctly predicted positive segments to the total predicted positive segments; (3) *Recall*: the ratio of correctly predicted positive segments to all positive segments; (4) *AUC*: the total area underneath the ROC curve (receiver operating characteristic curve). To evaluate the model's fairness across clients, we use the following four metrics: (1) *Average Accuracy*: average accuracy of all the clients; (2) *Worst 10%*: average accuracy of the 10% worst-performing clients; (3) *Best 10%*: average accuracy of the 10% best-performing clients; and (4) *Standard Deviation (Std)*: measures how spread are the accuracies of the clients.

D. Experimental Results

Centralized Learning. We train our joint diagnostic model in a centralized manner as the baseline. As shown in Table I, the overall accuracy is 0.891 with high precision (0.935) and recall (0.923), which indicates the effectiveness of our designed speech-based AD diagnostic model.

Federated Learning. We first train on different FL setups using the FedAvg aggregation rule on the server. We find that, in the case of IID data distribution among the clients,

TABLE II
COMPARISON OF AGGREGATION TECHNIQUES

Algorithm	FL Setting	Avg. Accuracy	Worst 10%	Best 10%	Std
FedAvg	Cross-Silo-10	0.815	0.553	0.894	8.336
	Cross-Silo-15	0.799	0.588	0.880	9.545
	Cross-Silo-30	0.734	0.599	0.865	11.698
	Cross-Device-90	0.692	0.574	0.765	13.763
q-FedAvg	Cross-Silo-10	0.844	0.767	0.879	4.941
	Cross-Silo-15	0.819	0.752	0.841	5.123
	Cross-Silo-30	0.751	0.718	0.797	7.847
	Cross-Device-90	0.676	0.612	0.717	11.264
q-FedSGD	Cross-Silo-10	0.839	0.755	0.885	4.235
	Cross-Silo-15	0.832	0.741	0.882	6.436
	Cross-Silo-30	0.764	0.709	0.804	7.538
	Cross-Device-90	0.665	0.614	0.731	12.128

the results are comparable with centralized learning (CL) with the highest accuracy of 0.873 under the Cross-Silo-10 setting. The performance degradation compared to CL is due to the decentralization of the data. Moreover, in non-IID data distribution, the results are slightly lower compared to the IID setting, with the highest accuracy of 0.844 due to the data heterogeneity among clients, making the local model updates drift in different directions. In both cases, we find that the Cross-Device-90 setting achieves the lowest performance, i.e., the prediction accuracy of 0.725 and 0.715 for the IID and non-IID settings, respectively. This is because the non-AD client's audio sample is shorter than the AD clients, which makes the data imbalance between the classes, and a higher number of clients make the task more challenging as each client holds a smaller fraction of the data. This issue may downgrade the local models' performance and thereby further drop the global performance.

Fair Federated Learning. As the number of data samples for AD and non-AD subjects varies significantly, the clients in non-IID scenarios contain a diverse amount of data samples, which makes the local model of each client unstable, and intra-local model performance varies largely. It is clear from Table II that the accuracy of the client models while using FedAvg aggregation is more spread out (i.e., the measured Std is larger than 8 for all settings). This raises the issue of fairness in the classification task. To improve the fairness, we further deploy the q-FedAvg and q-FedSGD aggregations. It is evident from Table II that the average testing accuracy stays reasonably steady with the deployed fair aggregation rules: the results are more centered (i.e., fair) with reduced variance in accuracy compared to FedAvg. Most noticeably, in the case of Cross-Silo-10, the Std is reduced by 40.7% with q-FedAvg and 49.2% by q-FedSGD compared to FedAvg. While the average accuracy remains nearly the same, the fair aggregations rules can help improve the accuracies of worst 10% clients by over 0.2 at a slight cost of reducing the accuracies of the best 10% by less than 0.02 in the Cross-Silo-10 scenario. In the Cross-Device-90 scenario, the variance in the accuracy becomes more significant in all three aggregation rules due to the more significant drifts of the local models. Despite this, the fair aggregation rules can still help to reduce the Std and improve the accuracy of worst-performing clients by up to 0.4. We can observe from this evaluation that the proposed federated learning-based approach can achieve a fairer AD diagnostic model while maintaining both privacy and fairness.

IV. CONCLUSION

In this paper, we develop the first privacy-preserving framework for training a speech-based automatic AD diagnosis model leveraging federated learning. In comparison to traditional centralized learning, the suggested FL architecture can reduce numerous systemic privacy risks by enabling collaborative learning across different clients without asking them to divulge their private data. Moreover, different fair aggregation algorithms were deployed with the proposed models to make the classification model fair for every participant during the training despite the heterogeneity in their data. Our evaluation of the ADReSS challenge dataset shows that the proposed FL-based method can achieve comparably good performance over existing centralized approaches while ensuring data privacy and the model's fairness. In our future work, we seek to further refine the fairness notion in the proposed federated AD diagnosis framework improving its fairness among different demographic groups.

REFERENCES

[1] M. Prince, R. Bryce, C. Ferri *et al.*, "The benefits of early diagnosis and intervention," *World Alzheimer Report*, vol. 2011, 2011.

[2] T. L. Michaud, R. L. Kane, J. R. McCarten, J. E. Gaugler, J. A. Nyman, and K. M. Kuntz, "Using cerebrospinal fluid biomarker testing to target treatment to patients with mild cognitive impairment: a cost-effectiveness analysis," *PharmacoEconomics-open*, vol. 2, no. 3, pp. 309–323, 2018.

[3] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of alzheimer's disease in conversational german." in *INTERSPEECH*, 2016, pp. 1938–1942.

[4] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid *et al.*, "Automatic speech analysis to early detect functional cognitive decline in elderly population," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 212–216.

[5] T. Warnita, N. Inoue, and K. Shinoda, "Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data," in *Proc. Interspeech 2018*, 2018, pp. 1706–1710.

[6] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech," *arXiv preprint arXiv:2106.09668*, 2021.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[8] A. Sadilek, L. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie *et al.*, "Privacy-first health research with federated learning," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–8, 2021.

[9] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S. K. Lo, and F.-Y. Wang, "Dynamic fusion-based federated learning for covid-19 detection," *IEEE Internet of Things Journal*, 2021.

[10] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[11] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: the adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.