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/EMBC48229.2022.9871861

# Privacy-preserving Speech-based Depression Diagnosis via Federated Learning

Yue Cui[1], Zhuohang Li[1], Luyang Liu[2], Jiaxin Zhang[3], Jian Liu[1]

*Abstract*— **Mental health disorders, such as depression, affect a large and growing number of populations worldwide, and they may cause severe emotional, behavioral and physical health problems if left untreated. As depression affects a patient's speech characteristics, recent studies have proposed to leverage deep-learning-powered speech analysis models for depression diagnosis, which often require centralized learning on the collected voice data. However, this centralized training requiring data to be stored at a server raises the risks of severe voice data breaches, and people may not be willing to share their speech data with third parties due to privacy concerns. To address these issues, in this paper, we demonstrate for the first time that speech-based depression diagnosis models can be trained in a privacy-preserving way using federated learning, which enables collaborative model training while keeping the private speech data decentralized on clients' devices. To ensure the model's robustness under attacks, we also integrate different FL defenses into the system, such as norm bounding, differential privacy, and secure aggregation mechanisms. Extensive experiments under various FL settings on the DAIC-WOZ dataset show that our FL model can achieve high performance without sacrificing much utility compared with centralized-learning approaches while ensuring users' speech data privacy.**

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

## I. INTRODUCTION

As one of the most common mental illnesses, depression affects approximately 280 million people worldwide [1], and it has an enormous impact on the quality of life by not only affecting emotional well-being but also increasing the risk of physical health problems. An early diagnosis of depression, therefore, is essential for prompt treatment to help lessen depression symptoms and reduce any risk of suicide.

Existing studies [2] show that depressed individuals tend to have distinguishable speech characteristics, such as lower speech speed, frequent pauses, and smaller speaking rate, compared to non-depressed individuals. Thus, such an underlying difference in speech provides an alternative pathway to detect depression and assess its severity. Thanks to recent advances in deep learning techniques, speech-based Automatic Depression Detection (ADD), which relies on speech analysis learning models to help facilitate the early intervention of depression by patients themselves, has received considerable attention. Existing studies build certain classifiers or regressors on top of speech characteristics via centralized training for depression prediction [3]–[5]. While they provide reasonably good results in certain cases, their centralized training schemes require a large amount of speech data along with patients' diagnosis results to be gathered to a centralized server during the model development phase. Inevitably, this will not only raise serious privacy concerns about data breaches but also imposes limited transparency and provenance on the system, leading to a lack of trust from users and unwillingness to share their private speech data. Thus, a solution that can enable individual patients or medical facilities with medical records to contribute to the development of an accurate ADD model while protecting patients' privacy is highly desirable.

Towards this goal, we propose a federated framework for achieving privacy-preserving ADD via speech analysis. To the best of our knowledge, this is the first attempt to utilize federated learning (FL) [6] to enable collaborative learning of a speech-based ADD model across multiple clients while keeping the training data decentralized. Because private data is stored locally in FL, the proposed framework can mitigate many systemic privacy risks presented in the traditional centralized models. To further achieve robustness against attacks, we integrate defense mechanisms such as norm bounding, differential privacy, and secure aggregation rules into our framework. Evaluation is conducted under different federated settings on the DAIC-WOZ dataset and the results show that the proposed framework can reach a decent performance without sacrificing much utility compared with centralized models while achieving privacy preservation.

## II. RELATED WORK

**Depression Diagnosis Using Speech.** The characteristics of speech (e.g., loudness, pitch variation) have long been recognized as important cues for detecting depression [2]. To achieve speech-based ADD, various acoustic features have been explored, such as spectral features, glottis features, and voice quality features [7]. Early studies of speech-based ADD often pair up these features with machine learning classifiers and predict the depression state. Motivated by the recent prosperity in deep learning, another line of work propose to replace these traditional classifiers with deep neuron networks (e.g., autoencoder [3], long short-term memory network [4] and convolutional neural network (CNN) [5]) to improve detection accuracy. However, speech signals encode a rich body of privacy-sensitive information, including age,

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

[1]Yue Cui, Zhuohang Li, Jian Liu are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA (e-mail: *ycui22@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*)

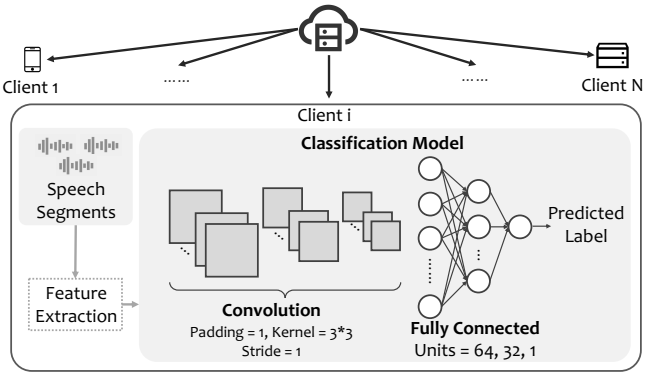


Fig. 1. Overview of the proposed approach.

gender, health conditions, and most importantly, biometric identity. Grounded on a centralized training framework, these methods require a large volume of speech data to be gathered, transmitted, and maintained on a central machine, which inevitably brings concerns on potential privacy breaches.

**Federated Learning for Healthcare and Speech Processing.** Federated learning (FL) is a training paradigm proposed to enable collaborative training of a machine learning model across distributed private datasets. Its strong emphasis on user privacy has recently attracted much attention in health research. A recent study by Nicola et al. [8] summarized several scenarios where FL could be used to provide better digital health. Sadilek et al. [9] demonstrated the successful use of FL in clinical research for the first time by comparing centralized and FL models across several diseases. However, these studies mainly focus on physical health and medical image analysis, leaving FL-based mental health analysis via speech signals still unexplored. Despite its attractive attributes, to date, there are only a few applications of FL for speech-related tasks. Gao et al. [10] presented the first study on Automatic Speech Recognition models under FL. Granqvist et al. [11] proposed to improve the centrally trained speaker verification system via FL. Hard et al. [12] demonstrated the feasibility of on-device FL training of the keyword-spotting model. Despite their initial success in applying FL to extract verbal information from speech, no study has investigated FL speech analysis models for privacy-preserving automatic depression diagnostics.

## III. PRIVACY-PRESERVING DEPRESSION DIAGNOSIS VIA FEDERATED LEARNING

### A. Depression Diagnosis Model

As depicted in Fig.1, the proposed model contains the following components:

**Feature Extraction.** The Mel Frequency Cepstral Coefficients (MFCCs) feature is the most commonly used audio feature in speech-related tasks because of its robustness in describing the variation of low frequencies signal and concentration on human perception. We thus 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. All MFCC coefficients are normalized to prevent its wide variation from impeding the training.

**Classification Model.** We design a speech-based ADD model utilizing convolutional neural network (CNN) to perform convolutions over the time dimension of mel-spectrograms for the participants. This special convolutional structure can help detect depression by extracting higher level information from the MFCC features. In specific, our model structure contains 3 convolution layers, consisting of 32, 64, and 128 filters of size $3 \times 3$. Each convolution layer is followed by a ReLU activation function. To reduce the dimensionality of the output feature maps, a max-pooling layer of size $2 \times 2$ is succeeded. The output feature is then fed to 2 fully connected layers with 64 and 32 hidden units, and both of them are followed by a dropout layer (the dropout rate is set to 0.1). Each fully connected layer is activated by the ReLu function. Finally, a neuron with Sigmoid activation is utilized to predict the binary labels: depressed or non-depressed. The model is trained on binary cross-entropy loss using SGD optimizer.

### B. Federated Depression Diagnostics

*1) FL Training Protocol:* At the beginning of the FL process, the central model is first randomly initialized with weights $w_0$. After initialization, the central server interacts with clients at each communication round repeatedly until the model converges. Specifically, a communication round at time $t \in [1, ..., T]$ contains the following steps: (1) The server selects a subset of clients to participate in the local training phrase from all $N$ clients with a fixed participate ratio $p$; (2) The central model $w_{t-1}$ is shared with the selected $p \times N = M$ clients; (3) The $M$ clients perform one or several training steps on the received central model using their local data; (4) All participated clients send back their model updates to the central server after finishing the local training; and (5) The central server computes an updated model with weight $w_t$ by an aggregation method based on the clients' individual updates $w_{t,i}, i \in M$. Unless mentioned otherwise, we adopt the most commonly used aggregator FedAvg [13], which aggregates the clients' updates according to: $w_t = \sum_{i=1}^{M} \frac{1}{M} w_{t,i}$.

*2) Improving Resilience by Integrating Defenses:* Recent studies have revealed that FL is vulnerable to various types of adversarial attacks [14]. We also evaluate our FL model with the presence of several following defense mechanisms.

**Norm Bounding.** To mitigate the negative impact of adversarial updates with large norms, we bound each participant's influence over the global model by clipping the L2 norm of the client's update gradient $\Delta w_{t,i} = w_{t,i} - w_{t-1,i}$ in communication round $t$ to a threshold $C$ before aggregation.

**Differential Privacy.** Differential Privacy (DP) can provide a guaranteed upper bound on the amount of information that can be leaked [11]. We implement the following types of DP: (1) *Local DP*: each local client adds Gaussian noise to its computed local update; (2) *Central DP*: the central server adds Gaussian noise to the aggregation result; and (3) *Layerwise DP*: Following an existing work from Rachel et. al [15] which shows that different layers in a deep neural network have various levels of sensitivity towards noises,

we inject different levels of Gaussian noise to clients' local updates according to the types of layers.

**Secure Aggregation.** We explore the following secure aggregation rules as a replacement for the conventional FedAvg aggregator: (1) *Median* [16]: at each round, the server sorts the $i$th parameter from $n$ local updates and takes the median value as updated $i$th parameter; (2) *Trimmed Mean* [16]: given a trimmed rate $\gamma$, the server sorts $i$th parameter from $n$ local updates, removes the smallest and largest $\gamma n$ values and computes the mean of the remaining $(1-2\gamma)n$ as the updated $i$th parameter; and (3) *Krum* [17]: suppose $f$ out of $n$ local clients are malicious. At each communication round, for each local update $w_k$, Krum computes the sum of Euclidean distances between $w_k$ and $n-f-2$ neighboring local updates that are closest to $w_k$ as its score. Then the $m$ local model updates with the smallest scores will be selected and their average will be computed as the global model update.

## IV. EVALUATION

### A. Experimental Setting

**Dataset.** Our experiments are conducted on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset [18]. The dataset contains audio recordings of clinical interviews from 189 participants in English, 56 of which are from depressed participants and the remaining are from non-depressed participants, each between 7-33 minutes (16 minutes on average) in length. Each participant is assessed using the Patient Health Questionnaire-8 (PHQ-8) metrics, which rates the severity of depression. The task is to classify the mental state of participants as depressed or not depressed. Each participant's audio recording is split into the training set and the test set, containing $80\%$ and $20\%$ of data respectively. Each sample is segmented to contain 15s of speech.

**Federated Learning Settings.** Regarding data partition, there are mainly two FL scenarios: (1) *IID* scenario, where data are independently and identically distributed across clients; and (2) *Non-IID* scenario, where each client holds data from different distribution patterns. In addition, according to the behavior of the participating clients, FL can be further categorized as (1) *Cross-Silo FL* that involves only a small number of relatively reliable clients, simulating FL across multiple organizations such as hospitals and research institutions; and (2) *Cross-Device FL* that involves a large number of mobile or edge devices with personal data.

In this work, we conduct experiments in all of the above scenarios. Specifically, the following 3 concrete settings are considered in the IID scenario: (1) *Cross-Silo*-8: 8 clients, each holding $\frac{1}{8}$ data that is randomly non-repetitively selected from the training set; (2) *Cross-Silo*-56: 56 clients, each holding $\frac{1}{56}$ data that is randomly non-repetitively selected from the training set; and (3) *Cross-Device*-189: 189 clients, each holding $\frac{1}{189}$ data that is randomly non-repetitively selected from the training set. In the Non-IID scenario, we study the following 3 settings: (1) *Cross-Silo*-8: 8 clients, each holding data from 7 depressed participants and 15 non-depressed participants; (2) *Cross-Silo*-56: 56 clients, each with data from 1 depressed participants and 2 non-depressed participants; and (3) *Cross-Device*-189: 189 clients, each containing data from 1 participant. The first two Non-IID settings are set to ensure data balance on each client, whereas the last Non-IID setting is naturally assigned by the number of participants in the dataset. The number of clients in IID setting is to make a fair comparison with non-IID settings. All of the above settings use a client participation rate of $p = 0.5$.

TABLE I
COMPARISON WITH DIFFERENT TRAINING SETTINGS.

| Scenario | Setting | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|---|
| Centralized | - | 0.968 | 0.937 | 0.923 | 0.930 | 0.918 |
| FL IID | Cross-Silo-8 | 0.906 | 0.860 | 0.823 | 0.841 | 0.865 |
| | Cross-Silo-56 | 0.887 | 0.809 | 0.820 | 0.814 | 0.849 |
| | Cross-Device-189 | 0.863 | 0.755 | 0.805 | 0.779 | 0.835 |
| FL Non-IID | Cross-Silo-8 | 0.890 | 0.855 | 0.763 | 0.806 | 0.764 |
| | Cross-Silo-56 | 0.879 | 0.830 | 0.752 | 0.789 | 0.722 |
| | Cross-Device-189 | 0.853 | 0.761 | 0.746 | 0.754 | 0.713 |

**Defense Parameter Selection.** (1) *Norm Bounding*: To maintain a good utility-robustness balance, the norm bound $C$ is empirically set to 1.5; (2) *Differential Privacy*: the level of noise can be measured by the mean ($\mu$) and variance ($\sigma$) of Gaussian noise. The Gaussian noise is set to $\mu = 0, \sigma = 10^{-3}$ for both Local DP and Central DP. In Layerwise DP, the noise is set to $\mu = 0, \sigma = 10^{-3}$ on the convolution layers and $\mu = 0, \sigma = 10^{-4}$ on the fully connected layers; and (3) *Secure Aggregation*: in Trimmed Mean, the trimmed rate $\gamma$ is set to 0.1, and we set $f = \frac{n}{10}$ in Krum.

**Evaluation Metrics.** We evaluate the performance of the model on the test set using the following five metrics: (1) *Accuracy*: the ratio of correctly predicted samples to the total samples; (2) *Precision*: the ratio of correctly predicted positive samples to the total predicted positive samples; (3) *Recall*: the ratio of correctly predicted positive samples to all positive samples; (4) *F1 score*: the harmonic mean of Precision and Recall; and (5) *AUC*: the total area underneath the ROC curve (receiver operating characteristic curve), measuring the performance of the model at distinguishing between the positive and negative classes.

### B. Depression Diagnosis via Federated Learning

**Centralized Learning (Baseline).** The first row of Table I shows the result of our model trained in a centralized manner. Our model achieves an accuracy, precision, and recall of $96.8\%$, $93.7\%$ and $92.3\%$, which are comparable to existing studies under the centralized training setting [3]–[5].

**Federated Learning with IID data.** As we can see from Table I, under the IID scenario, although the results are slightly lower than the centralized model due to the decentralization of the training data, our model can still maintain relatively high accuracy ($>86.3\%$) across all IID settings. In addition, we observe that the worst performance occurs in the Cross-Device-189 setting. It is possible that a larger number of clients will decrease the quantity of data held by clients, therefore degrading the accuracy of computed gradients.

**Federated Learning with Non-IID data.** As shown in Table I, the performance of models trained in the Non-IID scenario is decreased compared with centralized model and

TABLE II

COMPARISON OF DEFENSE MECHANISMS IN A NON-IID SETTING (CROSS-DEVICE-189).

| Defenses | | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|---|
| Norm Bounding | | 0.852 | 0.779 | 0.712 | 0.744 | 0.704 |
| Differential Privacy | Local DP | 0.818 | 0.723 | 0.640 | 0.679 | 0.667 |
| | Central DP | 0.840 | **0.765** | 0.677 | 0.718 | 0.722 |
| | Layerwise DP | **0.846** | 0.758 | **0.715** | **0.736** | **0.726** |
| Secure Aggregation | Median | **0.857** | **0.789** | **0.717** | **0.751** | **0.729** |
| | Trimmed Mean | 0.849 | 0.784 | 0.690 | 0.734 | 0.693 |
| | Krum | 0.845 | 0.774 | 0.686 | 0.727 | 0.686 |

models under the IID scenario. Such performance degradation is expected as the data heterogeneity across clients would cause the computed local model updates to drift towards different directions, resulting in sub-optimal server updates. Intuitively, a larger number of clients with more distinct client distribution may make the convergence of the global model more challenging. Therefore, in our work, momentum is applied at the server side to help maintain a more consistent update direction. Compared to the IID scenario, the model under the Cross-Device-189 setting only drops by $1.1\%$, $0.7\%$, and $7.3\%$ in accuracy, precision, and recall, respectively. This is due to the fact that the majority of clients hold data from non-depressed participants in this setting, which helps the model detect negative samples. We believe that the performance can be further improved by constraining the difference between client gradients and the global gradient or between client and global optimum values [19] when there are a massive number of clients in practical deployments.

### C. Federated Depression Diagnosis with Defenses

**Norm Bounding.** As shown in Table II, the model using norm bounding performs similarly to the model without defense. The recall and F1 score degraded slightly by $4.5\%$, and $1.3\%$, while the precision improved by $2.3\%$. These results show that applying a relatively small norm bound (i.e., $C$=1.5) has a limited impact on the model's performance.

**Differential Privacy.** As shown in Table II, among models with DP defenses, Central DP achieves better performance compared with Local DP. This result is expected since at each communication round, Central DP only injects noise to the aggregation result at the server side, while Local DP injects noise to every local update, resulting in a more negative impact on the model's utility. In addition, the Layerwise DP can form a more flexible defense mechanism while obtaining an acceptable utility because of the lower level of injected noise, whose performance only drops by $0.8\%$, $0.4\%$, $4.1\%$, $2.4\%$ and $1.7\%$ compared with the model without defenses in these metrics respectively.

**Secure Aggregation.** From the reported result in Table II, we can observe that Median has the best performance among the 3 aggregation rules in the given setting. One of the possible reasons behind this is that Median aggregation can provide better statistical robustness, making the model robust against a small fraction of clients with abnormal distribution.

## V. CONCLUSION

In this work, we proposed the first privacy-preserving framework via federated learning for training a speech-based depression diagnosis model. Compared with conventional centralized training schemes, the proposed framework can mitigate systematic privacy risks by enabling collaborative learning across multiple clients without sharing their private data. In addition, several defenses were investigated under the proposed framework to further improve the model's robustness. Experiments on the DAIC-WOZ dataset showed that the proposed FL-based method can ensure the clients' data privacy without scarifying much performance compared to the centralized approaches.

## REFERENCES


[1] World Health Organization, "Depression fact sheets," https://www.who.int/news-room/fact-sheets/detail/depression, 2021.

[2] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, 2012.

[3] M. Niu, K. Chen, Q. Chen, and L. Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *ICASSP*, 2021.

[4] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomed. Signal Process. Control*, 2022.

[5] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, "Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis," *MLWA*, 2020.

[6] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[7] H. Jiang, B. Hu, Z. Liu, G. Wang, L. Zhang, X. Li, and H. Kang, "Detecting depression using an ensemble logistic regression model based on multiple speech features," *Comput Math Methods Med*, 2018.

[8] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, 2020.

[9] 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*, 2021.

[10] Y. Gao, T. Parcollet, J. Fernandez-Marques, P. P. de Gusmao, D. J. Beutel, and N. D. Lane, "End-to-end speech recognition from federated acoustic models," *arXiv preprint arXiv:2104.14297*, 2021.

[11] F. Granqvist, M. Seigel, R. van Dalen, Á. Cahill, S. Shum, and M. Paulik, "Improving on-device speaker verification using federated learning with privacy," *arXiv preprint arXiv:2008.02651*, 2020.

[12] A. Hard, K. Partridge, C. Nguyen, N. Subrahmanya, A. Shah, P. Zhu, I. L. Moreno, and R. Mathews, "Training keyword spotting models on non-iid data with federated learning," *arXiv preprint arXiv:2005.10406*, 2020.

[13] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017.

[14] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," *NIPS*, vol. 34, 2021.

[15] R. Sterneck, A. Moitra, and P. Panda, "Noise sensitivity-based energy efficient and robust adversary detection in neural networks," *T-CAD*, 2021.

[16] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *ICML*, 2018.

[17] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *NIPS*, 2017.

[18] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014.

[19] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.