PRIVACY SENSITIVE SPEECH ANALYSIS USING FEDERATED LEARNING TO ASSESS DEPRESSION

Suhas BN and Saeed Abdullah

College of Information Sciences & Technology
Pennsylvania State University, University Park, USA
{suhas,saeed}@psu.edu

ABSTRACT

Recent studies have used speech signals to assess depression. However, speech features can lead to serious privacy concerns. To address these concerns, prior work has used privacy-preserving speech features. However, using a subset of features can lead to information loss and, consequently, non-optimal model performance. Furthermore, prior work relies on a centralized approach to support continuous model updates, posing privacy risks. This paper proposes to use Federated Learning (FL) to enable decentralized, privacy-preserving speech analysis to assess depression. Using an existing dataset (DAIC-WOZ), we show that FL models enable a robust assessment of depression with only 4–6% accuracy loss compared to a centralized approach. These models also outperform prior work using the same dataset. Furthermore, the FL models have short inference latency and small memory footprints while being energy-efficient. These models, thus, can be deployed on mobile devices for real-time, continuous, and privacy-preserving depression assessment at scale.

Index Terms— speech classification, depression, privacy, paralinguistics, mHealth

1. INTRODUCTION

Depression is a severe mental illness that affects millions of people worldwide [1]. Depression causes staggering individual and societal costs, including a higher risk of mortality [1]. There remains a significant treatment gap — individuals with depression often do not receive adequate treatment [2].

Addressing this treatment gap requires effective detection and monitoring of depression at scale. Toward this goal, recent studies have established that speech features including prosodic, articulatory, and acoustic signals can be indicative of depression onset and severity [3]. However, speech signals can be highly privacy-sensitive, and continuous monitoring can be challenging. Specifically, privacy-sensitive analysis is critical for assessing mental health issues. As such, recent studies have focused on using only privacy-preserving speech features (e.g., [4, 5]). However, using a subset of speech features leads to considerable information loss, negatively impacting model accuracy. Also, prior work requires a centralized approach to support continuous model updates from collected data. The need for data sharing across individuals, even for a subset of features, can lead to serious privacy risks. There has been an increasing focus on enabling decentralized training to reduce privacy risks in recent years. For example, Federated Learning (FL) aims to train models using local datasets and then merge those local models as necessary. Given that local datasets are not shared, FL significantly reduces privacy concerns. However, the resultant model accuracy and overhead can be a concern, particularly when deploying in mobile devices. To the best of our knowledge, there has been no work to establish the accuracy and performance overhead of using FL for privacy-preserving speech analysis to assess depression.

This paper aims to address this gap in two steps. First, we use an existing dataset (DAIC-WOZ) to establish that decentralized, privacy-preserving models can be used for robust assessment of depression using speech data. Second, we establish that the computational overhead of these models is low, and thus, they can be deployed to mobile devices for continuous and real-time monitoring. The paper makes the following novel technical contributions:

- We used transfer learning to assess depression and compared training overhead and accuracy using two FL frameworks: Federated Averaging (FedAvg) [6] and Federated Matching Averaging (FedMA) [7].
- The FL models achieve significantly better accuracy compared to the best-performing models in prior work using the DAIC-WOZ dataset (e.g., 87% accuracy for the combined dataset compared to 74.64% in [8]).
- We deploy the models in a smartphone to assess performance overhead for determining depression and show real-time and continuous assessment is possible.

2. RELATED WORK

Prior work has found that depressive states correspond to changes in prosodic and acoustic features (e.g., reduced pitch range, loudness, energy dynamics, and speaking rate) [3].

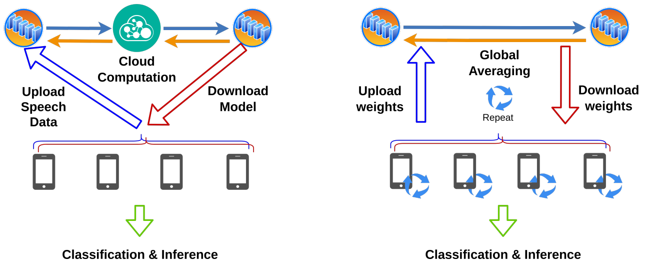


Fig. 1. The centralized (left) and federated learning methods (right) differ in model training and updating steps.

There has been an increasing focus on using speech as an objective biomarker of depression. Long et al. [9] used different speech features (e.g., short-time energy, intensity, formant frequencies, shimmer, jitter, and ZCR) on an in-house dataset and achieved 78.02% in detecting depression. Lalitha et al. [10] used deep neural networks to identify speech features that contain "emotional information". They achieved an average accuracy of 84.3% on the Berlin EmoDB database.

While DAIC-WOZ includes audio and video streams, we will focus on audio data, given the paper's scope. In recent work, Ma et al. [11] developed DepAudioNet, which leveraged Deep Convolutional Neural Networks and Long Short-Term Memory networks. The model achieved an F1 score of 0.52, precision of 0.35, recall of 1, and an approximated accuracy of 0.5. Hanai et al. [12] developed an LSTM model using audio features with an accuracy of 0.59. Srimadhur et al. [8] used an end-to-end network for classification from audio data with an accuracy of 74.64%. However, these previous studies do not focus on privacy-preserving audio analysis.

Prior work has explored the use of privacy-sensitive features from speech [4]. For example, Wyatt et al. [4] aimed to limit intelligibility and speech content reconstruction using a subset of features. However, reducing the feature space can lead to considerable information loss and negatively impact model accuracy. Furthermore, this approach still requires a central data repository to support model training and updating. Given that depression can be a lifelong condition, it is often necessary to continuously update models over depression to reflect different personal and lifestyle changes over time. The need for uploading and sharing data, even for a subset of features, can lead to serious privacy risks.

Several recent studies have used FL to reduce privacy risks in analyzing health data [13, 14]. Figure 1 compares the centralized and federated learning approaches. FL requires a trade-off regarding model performance and training overhead. Centralized models have access to more data, leading to better performance. The on-device training and merging model parameters can lead to considerable training time in FL. These trade-offs can be critical for systems focusing on real-time and continuous assessment of mental health issues, including depression. While FL has been used for different health applications, there is a lack of existing work that leveraged FL for privacy-preserving speech analysis to assess depression.

Gender	hasDepression	noDepression
Male	25	25
Female	25	25
Total	50	50

Table 1. The dataset used in this paper. We randomly selected a subset of the DAIC-WOZ dataset to ensure equal gender and depression class distribution.

Network	Depth	Parameters (M)	Input
GoogleNet	22	7	224 × 224
MobileNet v2	20	3.4	224 × 224
ResNet-18	18	11.7	224 × 224

Table 2. The characteristics of the network architectures used

This paper aims to address this gap. Specifically, we aim to establish a benchmark of accuracy and performance overhead of using FL to assess depression using speech signals from the DAIC-WOZ dataset.

3. DATA ATTRIBUTES AND PREPROCESSING

We have used the DAIC-WOZ dataset [15] to train and evaluate model performance. The dataset contains 189 interviews. Each interview has a corresponding PHQ-8 [16] score indicating depression severity. The original dataset is imbalanced regarding gender and the number of individuals with depression. To ensure balance, we randomly selected a subset of the dataset with equal gender and depression class distribution (25 individuals each) as shown in Table 1. The speech data has a sampling rate of 16 kHz. It contains speeches from both the participant and the virtual interviewer. The dataset provides start and stop timestamps indicating where the interviewer and the participant have spoken. We use this metadata to extract speech segments for the participants. Following prior work [17], we used an overlapping window length of 1s duration with a shift of 0.1s to extract log spectrogram features using the scipy.signal.spectrogram function, which is log-scaled as shown in Fig. 2. The images help model both temporal and harmonic structures of audio signals, leading to improved classification performance over existing methods.

4. EXPERIMENTAL SETUP

We used PyTorch on an NVIDIA Tesla V100-SXM2-32GB GPU to run the experiments. We use the Stochastic Gradient Descent optimizer (optim.SGD) with an initial learning rate (LR) = 0.001 and momentum of 0.9 for training the three networks. The 'LR' value is decayed every 7 epochs by a factor of gamma = 0.1 [18]. This decaying of the LR is useful in the case of FedAvg as it ensures convergence guarantees [19]. There has been work [20, 21] to alleviate high communication costs and performance drop, especially in the case of **non-I.I.D scenario** [22, 23]. Also, instead of training a single model on devices in FL environments, the authors suggest using a two-stream model, widely used in transfer learning [24]. We employ the Cross entropy function

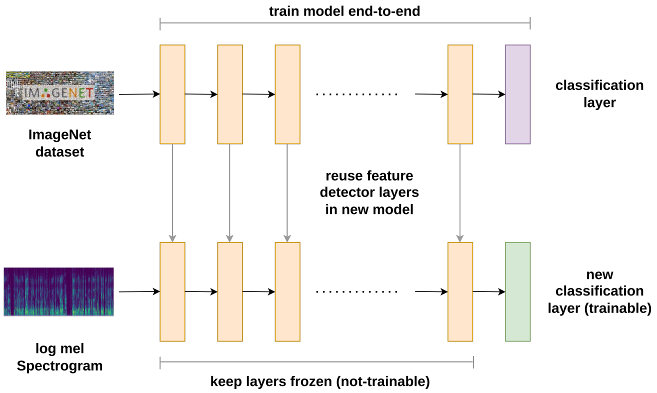

train model end-to-end
ImageNet dataset
classification layer
reuse feature detector layers in new model
log mel Spectrogram
new classification layer (trainable)
keep layers frozen (not-trainable)

Fig. 2. We used transfer learning strategies to efficiently train models. We started with pre-trained models and reused their feature detector layers. We then trained the output layer using log mel spectrograms from the dataset.

(nn.CrossEntropyLoss) as loss criterion. We minimized the risk of overfitting by using the five-fold cross-validation and early stopping criteria. The spectrogram features extracted are of dimensions 515x389 and are corrected to 224x224 using the transforms library from torchvision - where only the training images are augmented. Given our focus on using these models in smartphone devices, we chose three pre-trained models with a smaller number of parameters ($<$ 12 Million) as in Table 2.

5. FINDINGS

In this work, we focus on two binary classification tasks, namely - Depression classification (i.e., hasDepression vs. noDepression) and Depression severity classification (i.e., mild vs. severe). We consider a centralized and two FL approaches (FedAvg and FedMA). For FL approaches, we consider five devices on the same network.

5.1. Depression classification

For the depression classification task, we first assess model performance for gender-specific data. We then consider the combined dataset for model performance for comparison. We use five-fold cross-validation throughout the classification tasks. The resultant model performance is shown in Figure 3. A summary of the average training overhead is in Table 3. Figure 3 compares model performance for all the four sets of experiments. The centralized approach performs better than the federated methods by 6-10% across folds. The best avg. five-fold accuracy for the centralized approach is 0.934, while for the federated scheme, it is 0.91. From Table 3, we see that the centralized method is about 1.55-2.19x faster than federated schemes with ResNet-18 fastest for both centralized (155s) and federated schemes (327 & 340s respectively). Regarding the female subjects, the best average five-fold accuracy for centralized is 0.89, while for the FL scheme, it is 0.87. The centralized approach is about 1.78-2.91x faster than federated schemes. The centralized approach performs better by 4-6% across folds for the combined dataset. The best five-fold accuracy for the centralized approach is 0.885, while for FL, it is 0.87. The centralized method is about 1.48-2.26x faster than federated schemes.

Type	Male	Female	Combined	Severity
RN-18	155	167	253	222
GN	235	255	411	216
MN v2	204	223	359	191
RN-18 + FA	340	372	574	499
GN + FA	380	457	633	475
MN v2 + FA	345	415	595	411
RN-18 + FMA	327	487	554	470
GN + FMA	366	480	612	466
MN v2 + FMA	344	399	570	419

Table 3. Average time taken (s) to train different networks and algorithms. Note: RN-18 (ResNet-18), GN (GoogleNet), MNv2 (MobileNet v2), FA (FedAvg), FMA (FedMA)

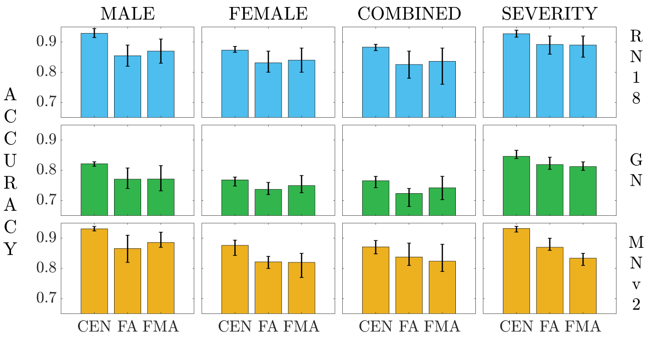


Fig. 3. Average five-fold cross-validation accuracy (has vs. no depression) for different algorithms (CEN: centralized, FA: FedAvg, FMA: FedMA). We also calculated performance across networks (rows) and subjects (columns).

5.2. Depression severity classification

Classifying mild and severe states can lead to informed clinical decision-making and pre-emptive care. We used the PHQ-8 score $>$10 as the threshold. This resulted in a binary classification task with low (PHQ-8 score $<$10) and high (PHQ-8 $\geq$10). The low class includes noDepression as well (i.e., PHQ-8 $\leq$4). We used the combined dataset for the classification task, resulting in a reasonably balanced class distribution (low: 28 subjects and high: 22 subjects). The results are shown in Figure 3. The best average five-fold accuracy in centralized and FL schemes are 0.93 and 0.88 (both in ResNet-18), respectively. The centralized method is about 2.11-2.24x faster than federated schemes, with MobileNet v2 training the fastest for both centralized and FL schemes (see Table 3).

5.3. Overall performance comparison

We used a ranking method to summarize overall performance across different approaches and networks. For each task category (e.g., Male), we assign a rank (1–3) to rate performance with rank 1 for the best performing model. By combining ranks across tasks (12 in total), we compare the robustness and consistency of different approaches. The results are summarized in Table 4 and 5 respectively. The centralized algorithm takes the least time to train and has the best accuracy. FedMA (2.33) performs slightly better than FedAvg (2.67) in time, but their accuracy ranking is the same, indicating similar performance across tasks. We compare network architecture performance in Table 5. MobileNet v2 performs well in both time and accuracy and offers the best of both worlds.

To summarize, the privacy-preserving FL models perform robustly with only 4-6% accuracy lost compared to a traditional centralized approach (*hasDepression* vs. *noDepression*). The FL models achieve comparable accuracy to the centralized approach for assessing depression severity (*high* vs. *low*). These findings establish the feasibility of using privacy-preserving FL models for depression assessment.

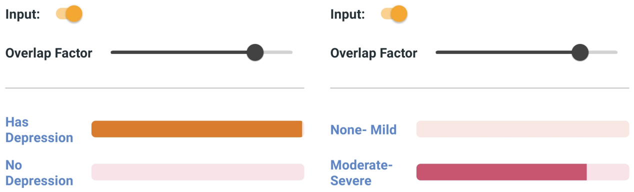
Time			Accuracy		
Method	Sum	Avg.	Method	Sum	Avg.
Central	12	1	Central	12	1
FedAvg	32	2.67	FedAvg	30	2.5
FedMA	28	2.33	FedMA	30	2.5

Table 4. Performance ranking (lower is better) for different approaches across twelve different task categories.

Time			Accuracy		
Network	Sum	Avg.	Network	Sum	Avg.
RN-18	**6**	**1.5**	RN-18	7	1.75
GN	11	2.75	GN	12	3
MN v2	7	1.75	**MN v2**	**5**	**1.25**

Table 5. Performance ranking for different network architectures (lower is better). The ranking is across twelve different task categories. MobileNet v2 provides the best trade-off.

6. DEPLOYMENT FEASIBILITY



(a) Speech with depression (b) Severity: Moderate-Severe

Fig. 4. The developed Android app. (a) It first aims to classify whether the speech shows signs of depression, and if it does, (b) it proceeds to rate its severity.

While the performance of the privacy-preserving, decentralized models is highly encouraging, it is critical to make sure that they can be deployed to devices with low computational resources. To assess the feasibility, we developed a smartphone app using TensorFlow Lite. The app classifies whether the speech shows signs of depression and, if it does, rates its severity. We used a Xiaomi Redmi Note 7 with 4GB RAM/64 GB ROM running Android 11 for deployment. For classification, we used an optimal two-fold model approach using the combined model and a model from the subject's gender (in real-time). Since the data capture phase is common and training takes little time - thus ensuring that the app works in real-time. Once we confirm depression exists through majority voting of the individual spectrogram image frames for both models, we use a depression severity model from the same model group. Thus, classification is more robust compared to a one-size-fits-all model. We used Android Profiler to collect real-time CPU, memory, network, and battery consumption data. The findings are shown in Table 6. For all models, the inference latency $<$ 100ms, which is essential for in-situ and real-time assessment. Each model requires $<$ 50MB of memory. The models are energy efficient — energy consumed/frame ranges from 0.26-0.37 joules, which translates to 72-102 μW/hr.

Type	IL (ms)	Memory (MB)	E (J)
RN-18	**48.2**	**23**	**0.26**
GN	64.1	26	0.37
MN v2	51.2	38	0.33

Table 6. The inference latency (IL), memory allocated (Memory), and the energy consumed per frame (E) by each model on a smartphone.

7. CONCLUSION

This paper explores the feasibility of using privacy-preserving, decentralized models for assessing depression and its severity using speech. Toward this goal, we use federated learning to enable on-device training. We used an existing dataset (DAIC-WOZ) to establish a performance benchmark for decentralized and privacy-preserving approaches. We show that the privacy-preserving FL models perform robustly with only 4-6% accuracy lost compared to a centralized approach. More importantly, these models achieve better accuracy than the best-performing models in prior work using DAIC-WOZ (e.g., 87% (FL) compared to 74.64% [8]). These findings establish the feasibility of using privacy-preserving FL models for depression assessment. We also explored the feasibility of deploying on devices with low computational resources. The FL models show small inference latency and a low memory footprint while being energy-efficient. As such, these models can be deployed to mobile devices to support continuous and real-time assessment of depression at scale.

8. REFERENCES

[1] World Health Organization, “Depression,” `who.int/health-topics/depression`, 2020.

[2] Graham Thornicroft, Somnath Chatterji, Sara Evans-Lacko, Michael Gruber, et al., “Undertreatment of people with major depressive disorder in 21 countries,” *The British Journal of Psychiatry*, vol. 210, no. 2, 2017.

[3] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.

[4] Danny Wyatt, Tanzeem Choudhury, Jeff Bilmes, and James A Kitts, "Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, 2011.

[5] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury, "Automatic detection of social rhythms in bipolar disorder," *Journal of the American Medical Informatics Association*, vol. 23, no. 3, pp. 538–543, 2016.

[6] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[7] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni, "Federated learning with matched averaging," *arXiv preprint arXiv:2002.06440*, 2020.

[8] NS Srimadhur and S Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," *Procedia Computer Science*, 2020.

[9] Hailiang Long, Zhenghao Guo, Xia Wu, Bin Hu, Zhenyu Liu, and Hanshu Cai, "Detecting depression in speech: Comparison and combination between different speech types," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017.

[10] S Lalitha, Shikha Tripathi, and Deepa Gupta, "Enhanced speech emotion detection using deep neural networks," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 497–510, 2019.

[11] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.

[12] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass, "Detecting depression with audio/text sequence modeling of interviews.," in *Interspeech*, 2018.

[13] Curtis P Langlotz, Bibb Allen, Bradley J Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S Cook, et al., "A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop," *Radiology*, vol. 291, no. 3, pp. 781–791, 2019.

[14] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[15] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al., "The distress analysis interview corpus of human and computer interviews.," in *LREC*, 2014, pp. 3123–3128.

[16] Robert L. Spitzer, Janet B.W. Williams, and Kurt Kroenke, "Patient health questionnaire (phq-8)," `https://tinyurl.com/h6h7uavx`, May 1991.

[17] B N Suhas, Jhansi Mallela, Aravind Illa, et al., "Speech task based automatic classification of ALS and Parkinson's Disease and their severity using log mel spectrograms," in *2020 International Conference on Signal Processing and Communications*. IEEE, 2020.

[18] Kaichao You, Mingsheng Long, Jianmin Wang, and Michael I Jordan, "How does learning rate decay help modern neural networks?," *arXiv preprint arXiv:1908.01878*, 2019.

[19] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020.

[20] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, et al., "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data," *arXiv preprint arXiv:1811.11479*, 2018.

[21] Xin Yao, Chaofeng Huang, and Lifeng Sun, "Two-stream federated learning: Reduce the communication costs," in *2018 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2018, pp. 1–4.

[22] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[23] Suhas BN, "Privacy-preserving assessment of depression using speech signal processing," M.S. thesis, 2021.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.