Privacy-Preserving Federated Learning for Depression Assessment

Murat Sahin, Melih Cosgun

***Abstract*—Depression is a serious illness affecting millions worldwide. However, the detection of depression has been a challenging area for machine learning (ML) researchers due to the lack of available medical data, which is often strictly private. To address the data sparsity problem and detect depression without compromising medical privacy, we aim to use federated learning. E-DAIC dataset is used in our study and audio modality is utilized. For the preprocessing, weighted random sampling method is used for balancing the data and MFCC coefficients are extracted afterwards. A centralized baseline model was initially developed and then transferred to the federated scheme. In the 4-client scenario, the performance loss from the centralized model was tolerable. However, in the 8-client scenario, the results were not satisfactory. It is important to note that in real-world scenarios, as the number of clients increases, that would result in more data and potentially different outcomes.**

***Index Terms*—Depression assessment, Deep learning, Speech processing, Federated learning**

I. Introduction

DEPRESSION is a disabling mental health condition that greatly affects human life. However, its effects can be reduced by detecting those conditions early on. In order to achieve this automatically, many machine-learning techniques have been proposed in recent years. However, as you know, those techniques are quite data-driven. They require a high amount of preferably high-quality data. Unfortunately, in a field like this, even finding a small public dataset is challenging. In order to solve this issue, Federated Learning [16] presents a promising approach, enabling clinics to maintain their private data while facilitating training of an ML model. The subject of our project is applying federated learning on this domain, without any data leak. Then, a comparison between the two approaches is made. We are using the E-DAIC [18] dataset, which is a dataset of nearly 300 recorded sessions, each is an average of 15 minutes.

The data was preprocessed by cropping 8-minute audio segments and extracting Mel-frequency cepstral coefficients (MFCCs) from them. We experimented with various learning approaches on these coefficients, including convolutional and temporal models, but ultimately chose a simple architecture with a single layer GRU. After evaluating the centralized case, we established our federated environment using the Flower framework [5] and conducted experiments in different scenarios. In our observation, the performance of a 4-client federated scheme is comparable to that of a centralized scheme. However, this is not the case for the 8-client scenario.

	Non-Depressive	Depressive
Training Samples	126	37
Validation Samples	44	12
Test Samples	39	17
Total Samples	209	66

TABLE I
E-DAIC Data Distribution

II. Dataset

Studies in this area motivated the creation of the Depression Sub-Challenge (DSC) of the Audio/Visual Emotion Challenge and Workshop (AVEC). This challenge has been organized regularly in the last decade, and the AVEC 2016 [19] and AVEC 2019 [18] challenges provided two important datasets to the literature, DAIC-WOZ and E-DAIC, respectively. They are the most commonly used datasets in the literature, as they contain nearly 300 samples (each is an average of 15 minutes) and multimodal modality. Note that E-DAIC is the extension of DAIC-WOZ to add new samples that replaces the human-controlled computer agent with a fully artificial intelligent agent. In our study, we are conducting our experiments over the E-DAIC dataset. See Table I for the detailed information about the dataset.

III. Background

A. Depression Assessment

Depression is a serious mental condition that affects millions of people in the world. It negatively affects the overall life; relation-wise, financially, academically and so on. Even though it is a terrible condition, it is curable after the correct diagnosis. This is why depression assessment is a serious topic. It has been done by the psychiatrists with the help of questionnaires and these questionnaires have been found to be successful.

Automatic depression recognition systems have been heavily researched in recent years due to the need for patient cooperation and trained professionals during these questionnaires. Initially, traditional machine learning models were used on top of hand-crafted features obtained from audiovisual cues. However, these features are subjective and require experienced professionals for designing. This is where deep learning comes into play with end-to-end architectures.

B. Federated Learning

Federated Learning (FL) [16] is a decentralized method of machine learning that enables training a machine learning model without accessing the data. This methodology tackles

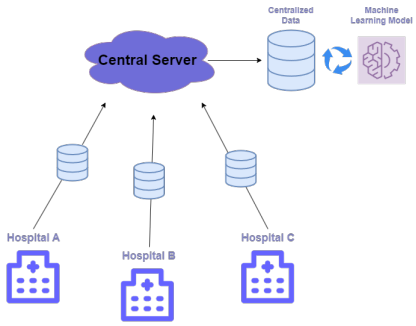


Fig. 1. Centralized Learning approach

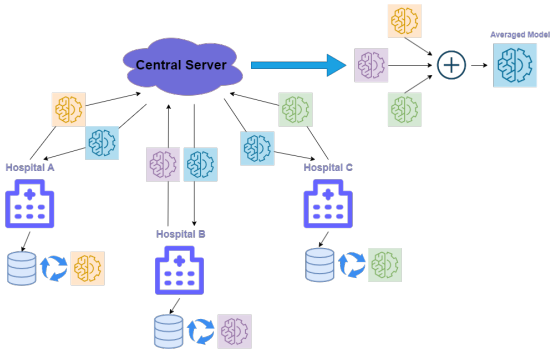


Fig. 2. Federated Learning approach

one of the major issues in training a generalized machine learning model, which is collecting the data, as shown in Figure 1. In Federated Learning, a centralized computation server sends a model architecture to the clients. Then, clients train the model on their local data with predetermined hyper-parameters. Afterward, clients send the calculated gradients for the model to the centralized server, which then computes the averaged gradients using a specific algorithm. Finally, the server sends the calculated gradients back to the clients, and this process repeats until the desired model is achieved. The main difference in this approach from the centralized one is that the clients' data never leaves their devices, as illustrated in Figure 2.

IV. Related Work

A. Depression Assessment via Deep Learning

There have been numerous works focused on utilizing deep learning for depression assessment. This has been done by utilizing various data types, shortly, video recordings, audio recordings, and transcripts.

[15] is the first instance in literature where an end-to-end network was used to assess depression. The architecture operates through a blend of CNN (for capturing short-term information) and LSTM (for capturing long-term information) architectures. The authors of this study performed random sampling on non-depressed samples to even out the size with the depressed samples within a mini batch, this leads to a network that is able to generalize well.

[12] developed an architecture that models both text-based and audio-based learning in LSTM networks, their model is composed of two LSTM branches whose output is merged into a final fully connected network. The major finding of the research is that combining text and audio features to create multimodal fusion may produce better results than using text or audio individually.

[21] attacked the data sparsity problem by the idea of transferring the attention information from another task to depression domain. By utilizing the attention mechanisms acquired during the speech recognition task, the authors aim to improve performance on the target task of depression detection. This strategy permits the model to capitalize on the knowledge gained in the speech recognition domain and apply it to depression recognition, thereby ultimately improving the model's capacity for depression assessment. Their approach lead them to obtain the state-of-the-art performance at that time (2020).

For the first time in the literature, [20] introduced the concept of utilizing Generative Adversarial Networks (GAN) to assess depression through speech tasks. Acoustic features are derived from speech segments using traditional techniques, and they attempt to augment these features for further training. They have surpassed the majority of results in the DAIC-WOZ dataset at that time by addressing the data sparsity problem in an unique way.

Table II compares the performance of these works. It demonstrates that accessing more data obviously improves the performance of deep learning models.

Methods	F1	Prec.	Rec.	MAE	RMSE
DepAudioNet [15]	.52	.35	1.0	-	-
Audio/Text Multimodal [12]	.77	.71	.83	5.10	6.37
Attention Transfer Network [21]	-	-	-	4.20	5.51
DCGAN Feature Generation [20]	-	-	-	4.634	5.520

TABLE II
COMPARING RESULTS OF RELATED CENTRALIZED WORKS (DAIC-WOZ)

B. Speech Based Federated Learning

Most speech-based federated learning (FL) research focuses on Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER). [14] conducted a study on the impact of Federated Learning (FL) on the SER task using the IEMOCAP dataset. The researchers analyzed varying local epoch numbers and batch sizes within the FL environment. The results revealed that increased batch sizes combined with more local epochs had a negative effect on performance.

[8] were pioneers in implementing Federated Learning (FL) with an ASR task. One of their most notable contributions was that they enhanced FedAvg algorithm by adding 'Hierarchical Optimization'. This means that the server retains a portion of the training data for local training, and during gradient aggregation, the server avoids the model from diverging too far from the intended task [8]. This technique has inspired future research [9].

One major flaw of this research was on their FL training platform, which named 'Federated Transpose Learning (FTL)'. They developed this platform solely as a simulation of Federated Learning without considering communication cost and

	Implementation	Federated Optimisation
Latif et al.	-	FedAvg
Gao et al.	Flower	FedAvg, Loss-based, WER-based
Dimitriadis et al.	FTL	Loss-based
Guliani et al.	Tensorflow	FedAvg

TABLE III
COMPARISON IN TERMS OF IMPLEMENTATION OF FL AND FEDERATED OPTIMIZATION ALGORITHM

security aspects, which are important issues in FL environment [11], [13]. [11] explicitly analyze the mentioned first issue through the introduction of a general cost function for Federated Learning, measuring computational cost alongside model performance. The researchers conduct experiments on the LibriSpeech dataset, however, [9] demonstrated that it is not optimally designed for non-iid ASR tasks.

The researchers point out that the dataset used in [11] and [8], namely LibriSpeech, is unrealistic for real-world FL scenarios since it only includes recordings from selected speakers and was conducted in a controlled environment with zero background noise [9]. To conduct more realistic 'ASR with FL' experiments, they use the CommonVoice dataset [3], which is a large and heterogeneous dataset. FL environment was implemented with Flower framework [5]. Table III displays the various implementations and Federated Optimization algorithms utilized in the research.

[10] highlight the privacy concerns associated with using speech data in a federated learning (FL) environment, which has not been addressed in mentioned research. The team conducted experiments on a vocal classification model, exploring four different cases of differential privacy (DP), the best-performing case was not using DP.

C. Speech Based Federated Learning for Depression Assessment

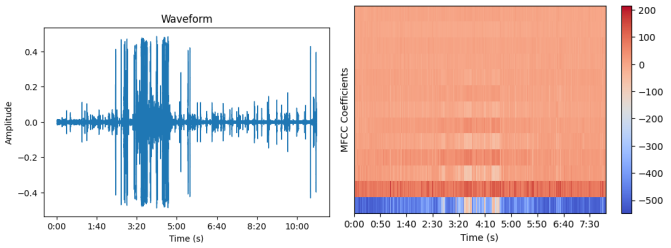
There are only two studies concerning Depression Assessment using Federated Learning [6], [7]. Both studies used speech data exclusively from the DAIC-WOZ dataset. [6] utilized transfer learning techniques in a FL environment to compare outcomes with centralized approaches. The researchers achieved 4-6% accuracy loss as compared to the centralized approach.

[7] attempts to increase privacy by utilizing three distinct privacy-preserving methods, involving Norm Bounding, Differential Privacy, and Secure Aggregation. The authors compared the performance of the vanilla approach with that of the privacy-preserving techniques implemented. In the vanilla FL approach, they experienced 7-8% accuracy loss compared to the centralized approach. However, after incorporating security parameters, they encountered 10-15% accuracy loss in the FL environment compared to the centralized environment.

V. METHODOLOGY

A. Preprocessing

The audio data can be fed directly into a neural network, as shown in Figure 3a. Alternatively, it can be pre-processed



(a) Waveform (b) MFCCs

Fig. 3. Example waveform and extracted MFCC features

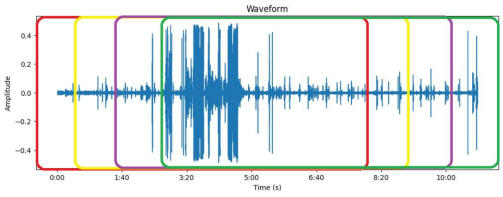


Fig. 4. Example windowed waveform

to extract higher-level audio features. We follow the latter approach and extract 13-dimensional Mel Frequency Cepstral Coefficients (MFCCs), as shown in Figure 3b. We consider 2048 points in each Fast Fourier Transform (FFT) and use a hop length of 512.

However, we do not extract these coefficients from the entire speech of an individual. Instead, we randomly select 8-minute windows, as shown in Figure 4, to sample and extract coefficients from within that window. This is done to increase the overall data and achieve uniformity among network inputs.

The number of windows placed for each individual audio sample is another consideration. To account for the data distribution, a weighted random sampling approach is used. The E-DAIC dataset is imbalanced, with approximately 75% of individuals being non-depressive. To balance the dataset, 10 windows were randomly selected from audio recordings of non-depressive individuals, and 30 windows were randomly selected from audio recordings of depressive individuals. Please refer to Table IV for the data distribution after this operation. Two training samples (357, 360) were excluded because they were shorter than 8 minutes.

B. Centralized Depression Assessment

After experimenting with various convolutional and recurrent architectures, we ultimately chose a simple architecture that consists of a single GRU layer followed by a fully-connected layer (refer to Figure 5). It is important to note that complex architectures are not desirable, as the training and

	Non-Depressive	Depressive
Training Samples	1240	1110
Validation Samples	440	360
Test Samples	390	170

TABLE IV
E-DAIC DISTRIBUTION AFTER SAMPLING

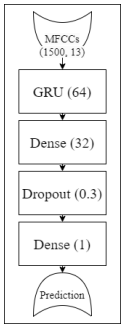


Fig. 5. GRU Architecture

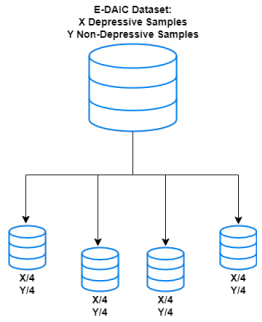


Fig. 6. Example of data partitioning for 4 clients.

Class	F1	Prec.	Rec.	Acc.
Non-Depressive	0.78	0.85	0.72	-
Depressive	0.60	0.52	0.71	-
Average	0.69	0.685	0.71	0.71

TABLE V
RESULTS ON CENTRALIZED

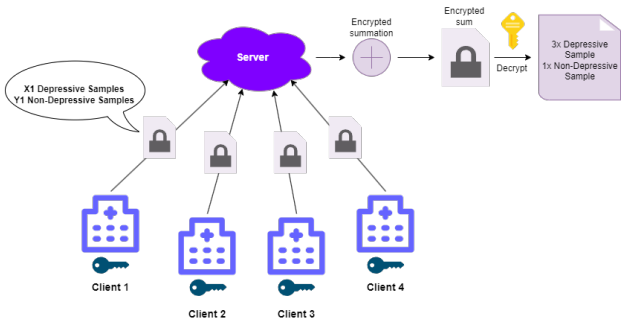


Fig. 7. Simplified illustration of Secure Random Sampling.

network complexities increase accordingly when transferred to a federated environment. Finishing tuning our optimizer on the validation loss, we decided to continue with following parameters: *Adam* optimizer, learning rate: *0.0003*, batch size: *32*, label smoothing: *0.1* and we used *early stopping* with validation data.

After tuning the hyperparameters, we trained our network 10 times and selected the model with the lowest validation loss. Then, the performance of this model on the test set is considered. See Table V for the results obtained from the final model of centralized learning. In the Evaluation section, we will also compare these results with those obtained from federated scenarios.

C. Federated Depression Assessment

FL methodology consists of three main steps, which are:

- Data Partitioning
- Secure Random Sampling
- FL Environment Creation

1) Data Partitioning: To partition the training data for N number of clients, first we applied data preprocessing steps in a centralized manner to avoid slow runtime. Then grouped the segmented data from the same samples. To simulate non-iid scenario, we ensured that no two clients have segments from the same sample. Afterwards, distributed both depressive and non-depressive segment groups distinctly. For a total of X depressive samples, each client would receive X/N depressive samples. Similarly, for a total of Y non-depressive samples, each client would receive Y/N non-depressive samples, as illustrated in the Figure 6.

2) Secure Random Sampling: Random sampling in a federated environment presents a challenge due to the inability of clients to view each other's data distribution. To apply random sampling in this binary label case, each data holder must possess two pieces of information: the number of segments to be extracted from one depressive sample and the number of segments to be extracted from one non-depressive sample. To address this issue, we propose the *Secure Random Sampling* method that uses Multiparty Homomorphic Encryption (MHE) [17], which is an encryption technique that provides computations on the encrypted data for multiple parties. This methodology has not been previously proposed in the literature, to the best of our knowledge. The method works as follows:

First, clients share the information about their data distribution by encrypting it with a common public key. The first slot of this ciphertext contains the number of depressive samples, while the second slot contains the number of non-depressive samples. These ciphertexts are then summed in an encrypted manner, with each slot being summed with its corresponding slot. The resulting ciphertext contains information on the total number of depressive and non-depressive samples separately. Clients then cooperatively decrypt this ciphertext using their own secret keys, allowing them to obtain the total number of depressive and non-depressive samples without seeing individual sample distributions. Finally, this information can be used to determine the ratio for segmentation when applying random sampling to individual local data.

Simulation of this methodology for 4 clients is applied by using Lattigo library [1]. Simplified illustration of the Secure Random Sampling is given in Figure 7.

3) FL Environment Creation: During federated learning, the training data is distributed to clients while the validation data remains on the server for centralized evaluation. In each round, the server calculates the centralized loss and accuracy over the validation data and stores the best performing weights of the model based on the validation loss. The server stops

after X number of epochs, and the model is determined by the best performing weights on the validation loss. The Flower framework [5] is utilized for simulating federated learning. The federated aggregation process employs the vanilla FedAvg algorithm [16], which operates by computing the weighted sum of gradients from each client.

The process of Federated Learning consists of several rounds. In each round, clients retrieve model weights from the server and perform one epoch of training on their local data. Afterward, clients send the updated weights back to the server. The server then calculates the aggregated weights using the FedAvg algorithm. With the calculated weights, the server performs central evaluation on the validation data. If the calculated weights perform better than any other round on the validation data, the server saves them. Finally, the server sends the calculated weights back to the clients, and this process repeats.

VI. Evaluation

The experiments were conducted in four different scenarios: a centralized environment, FL with four clients, FL with eight clients, and FL with four clients using differential privacy. The majority of hyperparameters were optimized for the centralized case, while only the learning rate and batch size hyperparameters were optimized for the federated cases due to their sensitivity to the amount of local data.

For the 4-client case, we utilized a learning rate of 0.0005 and a batch size of 32. Similarly, for the 8-client case, we found that a learning rate of 0.0005 and a batch size of 8 yielded the best performance. To evaluate the impact of additional defense mechanisms on the 4-client case, we applied differential privacy using the VectorizedDPKerasSGDOptimizer from the Tensorflow-privacy library [4], utilizing the default privacy parameters from the Flower framework GitHub repository [2]. The resulting confusion matrices for each case are shown in the Figure 8.

To evaluate our motivation for federated learning, we use a centralized case as a baseline. Detailed comparison can be observed from the Table VI. The FL case with four clients shows promise, with only a 3% loss in accuracy, a 7% loss in F1 score for depressive samples, and a 2% loss in F1 score for non-depressive samples. However, the FL case with eight clients performs significantly worse due to the smaller amount of data per client. For the case with DP, the introduced noise corrupted the correlation between data samples and got stuck in local minima by predicting only non-depressed. To overcome this, the noise parameters should be tuned.

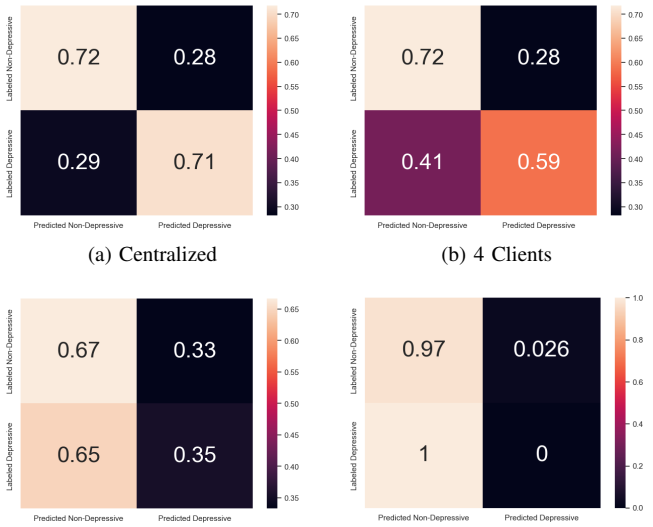
It is important to note that for each experiment case, the same dataset with the same amount of data is used. However, our motivation is to increase the amount of data used for training as the number of clients increases. This is because in a real-world scenario, each client will contribute their own data. Therefore, we expect to see an increase in performance as client sizes increase.

VII. Future Work

For future work on our project, we suggest implementing several improvements. In this work, we demonstrated the applicability of federated learning for depression assessment tasks. However, it is important to apply additional defense mechanisms for the federated environment, as vanilla federated learning is vulnerable to various attacks [13]. Additionally, by enhancing the centralized model, more robust comparisons can be made with federated learning. These improvements can be achieved through hyperparameter tuning or by utilizing different data modalities such as video and text. Additionally, for federated learning experiments, it is important to evaluate and present various metrics in a non-simulation environment, such as communication costs and privacy costs.

Case	F1	Prec.	Rec.	Acc.
Centralized	.60 (.78)	.52 (.85)	.71 (.72)	.71
4-Client	.53 (.76)	.48 (.80)	.59 (.72)	.68
8-Client	.33 (.68)	.32 (.70)	.35 (.67)	.57
4-Client DP	.0 (.81)	.0 (.69)	.0 (.97)	.68

TABLE VI
Results on 4 different cases.
Values in brackets represents the non-depressive class



(a) Centralized (b) 4 Clients

(c) 8 Clients (d) Differential Privacy (4C)

Fig. 8. Normalized Confusion Matrices
39 Non Depressive Samples, 17 Depressive Samples

VIII. Conclusion

Depression assessment is a crucial topic that, if done correctly, can improve the lives of millions. While deep learning techniques have shown success, they require large amounts of data. However, due to privacy concerns, collecting and sharing data in this field is challenging, resulting in a lack of data. Federated learning is a promising new field that addresses these concerns by its nature. By enabling clinics to collaboratively train a model without sharing their data, we can train models that use more data and, therefore, are able to generalize well. In this project, we demonstrated that even with a small amount of data, the 4-client federated scenario produced promising results close to centralized learning. In real-world scenarios, as the number of clients increases, unlike our experiments with a fixed dataset size, this would result in more data and potentially much better outcomes.

REFERENCES

[1] Lattigo v5. Online: https://github.com/tuneinsight/lattigo, November 2023. EPFL-LDS, Tune Insight SA.

[2] adap/flower. Differential Privacy. https://github.com/adap/flower/issues/605, 2023-12-16. Issue 605 on the adap/flower repository.

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2020.

[4] TensorFlow Authors. Tensorflow privacy, 2023.

[5] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. Flower: A friendly federated learning research framework, 2022.

[6] Suhas Bn and Saeed Abdullah. Privacy sensitive speech analysis using federated learning to assess depression. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6272–6276, 2022.

[7] Yue Cui, Zhuohang Li, Luyang Liu, Jiaxin Zhang, and Jian Liu. Privacy-preserving speech-based depression diagnosis via federated learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1371–1374, 2022.

[8] Dimitrios Dimitriadis, Kenichi Kumatani, Robert Gmyr, Yashesh Gaur, and Sefik Emre Eskimez. A Federated Approach in Training Acoustic Models. In *Proc. Interspeech 2020*, pages 981–985, 2020.

[9] Yan Gao, Titouan Parcollet, Javier Fernandez-Marques, Pedro Porto Buarque de Gusmão, Daniel Beutel, and Nicholas Lane. End-to-end speech recognition from federated acoustic models, 04 2021.

[10] Filip Granqvist, Matt Seigel, Rogier Dalen, Áine Cahill, Stephen Shum, and Matthias Paulik. Improving on-device speaker verification using federated learning with privacy. pages 4328–4332, 10 2020.

[11] Dhruv Guliani, Françoise Beaufays, and Giovanni Motta. Training speech recognition models with federated learning: A quality/cost framework. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3080–3084, 2021.

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

[13] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawit, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021.

[14] Siddique Latif, Sara Khalifa, Rajib Rana, and Raja Jurdak. Poster abstract: Federated learning for speech emotion recognition applications. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 341–342, 2020.

[15] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016.

[16] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[17] Christian Mouchet, Juan Ramón Troncoso-Pastoriza, and Jean-Pierre Hubaux. Multiparty homomorphic encryption: From theory to practice. *IACR Cryptol. ePrint Arch.*, 2020:304, 2020.

[18] Fabien Ringeval, Björn Schuller, Michel F. Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, M. Soleymani, and Maja Pantic. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[19] Michel F. Valstar, J. Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016.

[20] Le Yang, Dongmei Jiang, and Hichem Sahli. Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE Access*, 8:24033–24045, 2020.

[21] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Jun Deng, Nicholas Cummins, Haishuai Wang, Jianhua Tao, and Björn Schuller. Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. *IEEE Journal of Selected Topics in Signal Processing*, 14:423–434, 2020.