Federated learning for secure development of AI models for Parkinson's disease detection using speech from different languages

Soroosh Tayebi Arasteh[*,1,2,3], *Cristian David Rios-Urrego*[*,4], *Elmar Noeth*[1], *Andreas Maier*[1], *Seung Hee Yang*[2], *Jan Rusz*[5], *Juan Rafael Orozco-Arroyave*[1,4]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]Speech & Language Processing Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[3]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany
[4]GITA Lab, Faculty of Engineering, University of Antioquia, Medellín, Colombia
[5]Department of Circuit Theory, Czech Technical University in Prague, Prague, Czech Republic
soroosh.arasteh@fau.de, cdavid.rios@udea.edu.co

Abstract

Parkinson's disease (PD) is a neurological disorder impacting a person's speech. Among automatic PD assessment methods, deep learning models have gained particular interest. Recently, the community has explored cross-pathology and cross-language models which can improve diagnostic accuracy even further. However, strict patient data privacy regulations largely prevent institutions from sharing patient speech data with each other. In this paper, we employ federated learning (FL) for PD detection using speech signals from 3 real-world language corpora of German, Spanish, and Czech, each from a separate institution. Our results indicate that the FL model outperforms all the local models in terms of diagnostic accuracy, while not performing very differently from the model based on centrally combined training sets, with the advantage of not requiring any data sharing among collaborators. This will simplify inter-institutional collaborations, resulting in enhancement of patient outcomes.

Index Terms: federated learning, speech pathology, Parkinson's disease, deep learning, trustworthy speech processing

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that affects the nervous system, leading to the progressive deterioration of motor and non-motor functions, which contribute significantly to decreasing the quality of life of the patient's [1]. PD is characterized by resting tremor, rigidity, bradykinesia, postural instability, and other symptoms [2]. Most PD patients develop speech deficits which are grouped and called hypokinetic dysarthria where the speech is characterized by reduced loudness, monotonous pitch, and changes in voice quality [3,4]. Speech signals can be analyzed objectively to quantify the severity of the disease and track its progression over time, which can be useful in clinical research and treatment monitoring. Among the best motivations to consider the speech signals is that they can be easily collected and analyzed remotely, which can provide greater convenience to patients and reduce the need for frequent clinical visits [5]. This can be especially beneficial for patients who live in remote areas or have limited mobility. In addition, speech signals can provide a complementary source of information to clinical assessment and other diagnostic tests, which can improve the accuracy and reliability of PD diagnosis and treatment [6].

Recently, deep learning (DL)-based methods have particularly gained a lot of attention for analyzing PD speech signals [7, 8]. However, a major impediment to developing such robust DL models is the need for accessing lots of training data, which is challenging for many institutions. Thus, benefiting from data from different external institutions could solve this issue. However, strict patient data privacy regulations in the medical context make this infeasible in most cases in real-world practice [9–12]. Therefore, privacy-preserving collaborative training methods, in which participating institutions do not share data with each other are favorable. Federated learning (FL) [13–15], as the golden key to this issue, has been increasingly investigated by researchers and practitioners and received a lot of attention in the medical image analysis domain [11, 16–18] as it does not require sharing any training data between participating institutions in the joint training process. To the best of our knowledge, collaborative training methods based on FL have not been addressed in the literature on pathological speech signals yet, despite the availability of similar privacy regulations and restrictions as in the imaging domain [9], especially considering recent literature revealing the vulnerability of pathological speech signals in terms of patient data [19–21].

In this paper, for the first time, we investigate the applicability of FL in the privacy-preserving development of DL methods for PD detection using speech signals from three real-world language corpora, each from a separate and independent institution. We hypothesize that utilizing FL will substantially increase the diagnostic performances of networks for each local database while preserving patient privacy by avoiding data sharing between the institutions. Moreover, we assume that the FL model will perform relatively similarly, with only slight degradation compared to the hypothetical and non-privacy-preserving scenario where all the institutions could combine their training sets at a central location.

2. Material and Methods

2.1. Methodology

The methodology addressed in this study consists of the following main stages: data were acquired in different languages (Ger-

---

* STA and CDRU contributed equally to this work

Table 1: *Demographic and clinical information of the partici-pants. [F/M]: Female/Male. Values reported as mean $\pm$ std.*

	PD patients	HC subjects
Spanish		
Gender [F/M]	25/25	25/25
Age [F/M]	60.7±7/61.3±11	61.4±7/60.5±12
Range of age [F/M]	49-75/33-81	49-76/31-86
MDS-UPDRS-III [F/M]	37.6±14/37.8±22	
Speech item (MDS-UPDRS-III) [F/M]	1.3±0.8/1.4±0.9	
German		
Gender [F/M]	41/47	44/44
Age [F/M]	66.2 ±9.7/66.7±8.7	62.6±15.2/63.8±12.7
Range of age [F/M]	42-84/44-82	28-85/26-83
UPDRS-III [F/M]	23.3±12/22.1±10	
Speech item (MDS-UPDRS-III) [F/M]	1.2±0.5/1.4±0.6	
Czech		
Gender [F/M]	20/30	20/30
Age [F/M]	60.1±9/65.3±10	63.5±11/60.3±12
Range of age [F/M]	41-72/43-82	40-79/41-77
UPDRS-III [F/M]	18.1±10/21.4±12	
Speech item (MDS-UPDRS-III) [F/M]	0.7±0.6/0.9±0.5	

man, Spanish, and Czech), after, embeddings were extracted from speech signals for each participant using a pre-trained Wav2Vec 2.0 model [7, 8], then the extracted embeddings were utilized for the secure FL training of a classification architecture, and finally, a copy of the global model is sent back to each participating site for the classification of PD patients from healthy control (HC) subjects. This methodology is summarized in Fig. 1. Details of each stage are presented below.

2.1.1. Data

We considered speech corpora in three different languages: Spanish, German, and Czech; each database contains PD patients and HC subjects. The first corpus is PC-GITA which includes recordings of 50 PD patients and 50 HC subjects [22]. All participants were Colombian native speakers. The second corpus contained a total of 176 German native speakers (88 PD patients and 88 HC subjects) [23]. The last database contained recordings of 100 Czech native speakers divided into 50 PD patients and 50 HC subjects [24]. Specialized neurologists evaluated each patient according to the Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS-III) [25]. In addition, all recordings were captured in noise-controlled conditions, and the speech signals were downsampled to 16 kHz to feed a deep-learning model. The rapid repetition of the syllables /pa-ta-ka/ was considered in this study. This task allows the evaluation of specific movements required to produce stop consonants (/p/, /t/, /k/). Table 1 shows the demographic information of each database.

2.1.2. Feature Extraction

To create a representation for each recording, we used Wav2vec 2.0 architecture, a state-of-the-art topology based on transformers proposed in [8]. Wav2Vec 2.0 was trained using a self-supervised pre-training approach that allows the model to learn representations directly from the raw audio signal without additional annotations or labels. The training process involved two main steps. Firstly, the contrastive pre-training, where the model was trained to distinguish between two versions of the same audio signal including a positive sample (a randomly selected segment of the original audio signal) and a negative sample (a randomly selected segment of a different audio signal). The second stage was fine-tuned based on a specific automatic speech recognition (ASR) task. Particularly in this work, we used a Wav2Vec 2.0 model, pre-trained on 960 hours of un-

labeled audio from the LibriSpeech dataset [26], which was derived from English audiobooks and fine-tuned for ASR on the same audio with the corresponding transcripts. Due to the dynamic representation of 768 dimensions for each array with respect to time, we calculated a static vector for each participant from 6 different statistics (mean, standard deviation (std.), skewness, kurtosis, minimum, and maximum), building a speech representation of 4608 dimensions per recording.

2.1.3. Federated Learning

In order to speed up the collaborative training convergence, the FL process was performed merely for the classification network, i.e., after all the embeddings were locally extracted using the Wav2Vec 2.0 model. Of note, all the data pre-processing and feature extraction steps happened locally by every participating institution without sharing any data with other institutions.

Each institution performed a local training round of the classification network and transmitted the network parameters, i.e., the weights and biases, to a trusted server, which aggregated all the local parameters leading to a set of global parameters. In our implementation, we chose each round to be equal to one epoch of training with the full local dataset. Afterward, the server transmitted back a copy of the global network to each institution for another round of local training. The process continued until the convergence of the global network. It is worth mentioning that not only each institution did not have access to any training data from others, but also not even to the network parameters of others, rather only an aggregated network, without the knowledge about the contributions of other participating institutions to the global network. Once the training of the global classification network was converged, every institution could take a copy of the global network and locally utilize it for diagnosing its test data.

2.1.4. Classification and Evaluation

The classification network architecture contained 4 fully-connected layers with different sizes: 1024, 256, 64, and 2, respectively. Rectified linear unit (ReLU) activation and batch normalization [27] were considered in each layer, and a Softmax activation function was used at the output. The fully connected network was trained and evaluated following a stratified 10-fold cross-validation strategy. The process was repeated 5 times for a better generalization of the results. The He initialization scheme [28] was applied to all classification network weights and all the biases were initialized with zeros. Cross-entropy was chosen as the loss function and the models were optimized using the Adam optimizer [29] with a learning rate of 8×10^{-5} and weight decays of 5×10^{-6}. The classification networks were trained for 50 epochs in batches of size 16. Accuracy and area under the receiver-operator-characteristic curve (AUC) were chosen as the main evaluation metrics, while sensitivity and specificity were utilized as supporting metrics. Two-tailed paired t-test was employed for determining statistical significance. The significance threshold was set at p-value ≤ 0.05.

3. Experiments and Results

For each test database, we compared the diagnostic performances of the methods in three multicentric setups where the network was: i) locally trained using solely the training set of the corresponding database (Local), ii) trained utilizing the combination of all the training sets of different databases at a central location without privacy measures (Central), and iii)
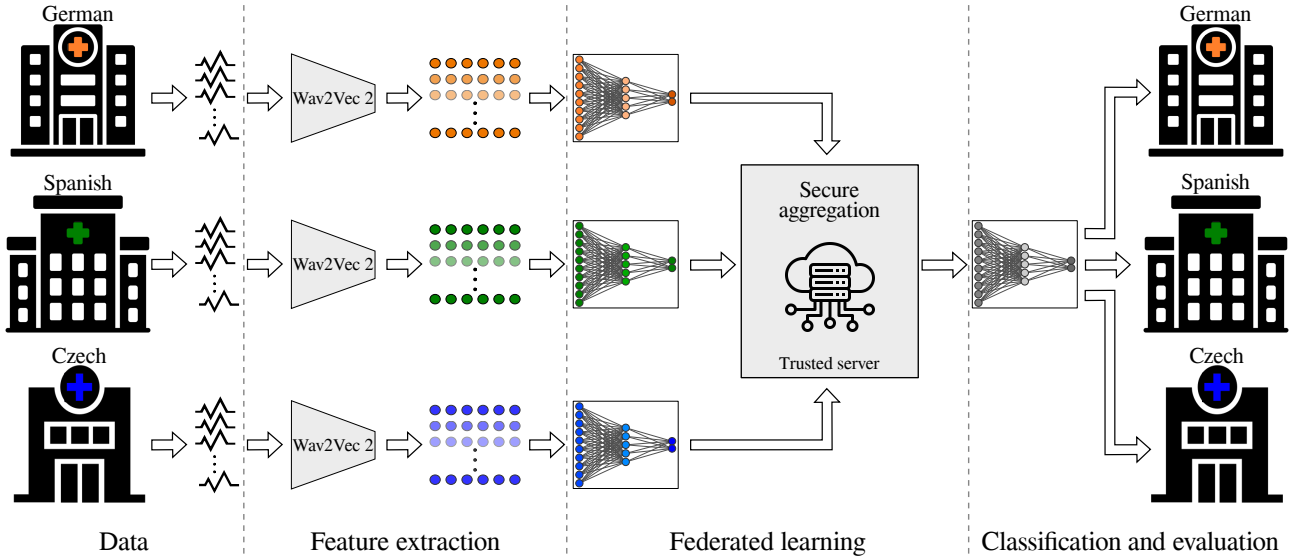
Figure 1: *General methodology: each institution pre-processes its local data, extracts the features using a Wav2Vec 2.0 model, and performs one epoch of the classifier network training locally, and transmits its local network parameters to a trusted server. The server aggregates all the parameters from all the institutions and transmits back the resulting global model to each institution for the next round of local training. In the end, each institution takes a copy of the final global model and performs its desired classification locally.*

trained with all the training sets of different databases based on FL, i.e., without sharing any data and preserving patient privacy information. Furthermore, due to the relatively small test sizes of each database, we repeated each experiment corresponding to each cross-validation step 5 times, including the training and evaluation of the classification network for all 3 setups. Considering the 5 repetitions and 10-fold cross-validation steps, a total of 50 values were obtained for statistical analysis of each experiment.

The average evaluation results are reported in Table 2 and details about diagnostic accuracy and classification performance are illustrated in Fig. 2-A. The accuracy of the FL method was significantly higher than local models for Spanish ($83.2 \pm 10.8\%$ vs. 77.0 ± 13.3; P-value = 0.001) and Czech ($76.0 \pm 12.2\%$ vs. 70.3 ± 14.6; P-value = 0.020) databases while it was only slightly higher for the German database ($75.8 \pm 8.3\%$ vs. 74.8 ± 9.1; P-value = 0.455) which contained the largest training set. These results suggest that combining the corpus of the same pathology but in different languages allows generalizing the architecture to classify pathological speech from healthy speech. Furthermore, comparing the non-private "Central" and the secure FL strategies, we can observe that the diagnostic accuracy of the FL method was not significantly different from the "Central" model for Spanish ($83.2 \pm 10.8\%$ vs. 82.0 ± 11.6; P-value = 0.436) and Czech ($76.0 \pm 12.2\%$ vs. 77.8 ± 9.2; P-value = 0.334) databases while it was for the German database ($75.8 \pm 8.3\%$ vs. 78.9 ± 8.3; P-value = 0.023). Moreover, Table 2 shows that the strategy proposed in this work obtained similar results to the state-of-the-art centralized training methods [30], with the advantage of patient privacy preservation by avoiding data exchange between local institutions using an FL strategy.

In addition, Fig. 2-B shows a visual comparison between the "Central" and the FL strategies from the receiver-operator-characteristic (ROC) curves and the corresponding AUC values obtained in each experiment. Again, when we compared each institution (language) separately, we can observe that the Cen-

Table 2: *Evaluation results for each database. "Local" represents solely using the training set of the target database, while "Central" means utilizing all training sets when combined with each other at a central location. Values are reported as mean $\pm$ std in percentages. The "P-value" is with respect to FL for each database for accuracy values.*

Training set	Accuracy	AUC	Sensitivity	Specificity	P-value
Spanish					
Local	77.0 ± 13.3	77.9 ± 13.1	71.2 ± 21.4	82.8 ± 19.4	0.001
Central	82.0 ± 11.6	80.6 ± 12.0	78.4 ± 17.5	85.6 ± 17.6	0.436
FL	83.2 ± 10.8	83.6 ± 11.8	77.2 ± 17.6	89.2 ± 14.1	-
German					
Local	74.8 ± 9.1	73.2 ± 8.6	76.3 ± 13.8	73.2 ± 14.2	0.455
Central	78.9 ± 8.3	77.5 ± 8.2	83.1 ± 13.0	74.8 ± 17.0	0.023
FL	75.8 ± 8.3	77.1 ± 7.4	90.8 ± 8.9	60.8 ± 18.0	-
Czech					
Local	70.3 ± 14.6	68.4 ± 14.4	64.8 ± 26.1	76.4 ± 23.1	0.020
Central	77.8 ± 9.2	77.6 ± 10.7	74.0 ± 18.6	82.0 ± 17.3	0.334
FL	76.0 ± 12.2	78.2 ± 10.7	62.0 ± 22.6	90.8 ± 13.5	-

tral and FL curves have the same trend and show no significant differences. It can also be observed that the Spanish language obtains the best result (AUC of 0.85), followed by German (AUC of 0.79) and Czech language (AUC of 0.78). Finally, Fig. 2-C shows the histogram and the probability density distributions obtained for the classification of German, Spanish, and Czech databases using the FL strategy. It can be observed that all three figures have the highest bins at their extremes, which corresponds to a high probability of the decision taken by the classifier. Moreover, it is possible to observe that in the case of Spanish and Czech, the highest bin is for the HC controls, which is related to the reported specificity (89.2% and 90.8%, respectively); while for Spanish, the highest bin corresponds to PD patients due to a higher sensitivity (90.8%).

4. Discussion

In this study, we showed the first successful application of cross-language federated learning for PD detection using three patho-
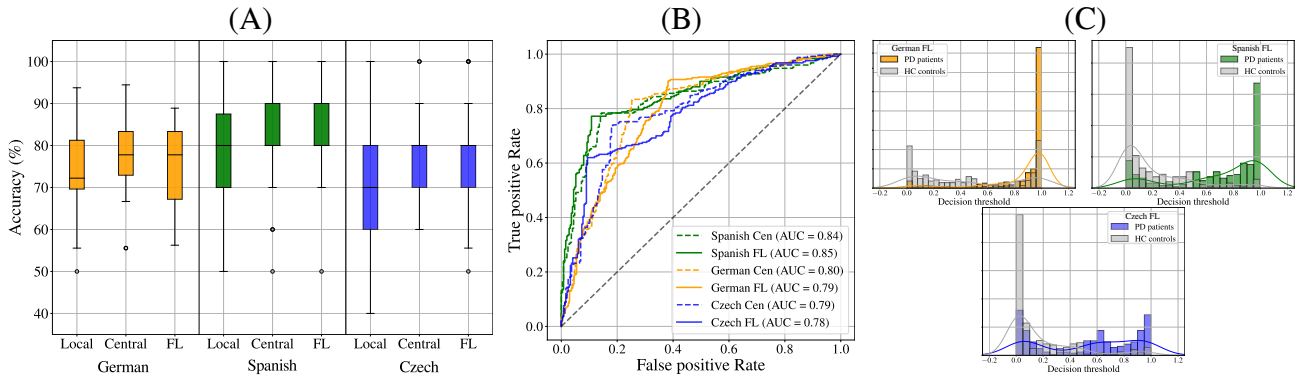
Figure 2: *Evaluation results. (A) Illustrates the final accuracy values for each test database using the 3 setups, where "Local" represents solely using the training set of the target database, while "Central" means utilizing all training sets when combined with each other at a central location. (B) Shows the receiver-operator-characteristic curves. (C) Shows the histogram and the probability density distributions obtained for the classification of German, Spanish, and Czech databases using the FL strategy.*

logical speech corpora, including a total of 188 PD and 188 HC subjects, covering Spanish, German, and Czech languages. We used a state-of-the-art topology namely the Wav2vec 2.0 [7, 8] for obtaining speech representations. We compared the performances in three multicentric setups where the architecture was: i) trained locally and separated by language, i.e., monolingual models, ii) trained utilizing the combination of all the training sets at a central location without privacy measures, i.e., cross-lingual model, and iii) trained with all the training sets of different databases based on FL strategy, i.e., without sharing any data and preserving patient privacy.

The results indicated that the FL model outperformed all the local models (mono-lingual models) for every test database in terms of diagnostic accuracy, while not requiring any data sharing between institutions. This result is very interesting and encourages the scientific community to further explore techniques for the generalization of models from databases of the same pathology, in different languages, without the need for sharing information between other institutions (cross-lingual model), which has been a major challenge. In addition, comparing the "Central" combination and FL strategies, we observed that in the majority of scenarios, the FL method was not significantly different from the Central method in terms of the model's diagnostic accuracy. This shows that the FL paradigm can considerably help the collaboration of institutions around the world in the creation of DL models with large amounts of data, cross-lingual, and preserving patient privacy by avoiding data exchange between local institutions, a major limitation in real-world practice that was not considered in current state-of-the-art cross-lingual approaches.

Our study has limitations. The collaborative FL training process was implemented in a proof-of-concept mode, i.e., using a single institutional network. Due to strict data protection regulations, the implementation of FL among different institutions would be challenging. However, we simulated a realistic setup where every database corresponded to a separate computing entity and we kept the data strictly independent from each other. As already mentioned, the parameter aggregation mechanism of the central server which was utilized in this study was direct averaging the individual network parameters of each participating database, i.e., the FedAvg algorithm [15], which is the simplest yet the most common aggregation mechanism. Furthermore, the databases utilized in this

study were non-independent-and-identically-distributed (non-IID). This was shown to be decreasing the performance of the global model in many different FL applications [31]. Consequently, future work could consider more advanced and task-specific aggregation methods such as [32–34] by accounting for the individual contribution of each participating site by analyzing their gradient updates in each FL training round before aggregation that could potentially increase the performance of the global model. In addition, we considered the most common task of PD detection, i.e., utilizing speech data containing the rapid repetition of the syllables /pa-ta-ka/ for the applicability of FL in pathological speech analysis in this study. In the future, we will extend this by considering further tasks and cross-pathology scenarios. As a side note, we could conclude that the characterization performed by the Wav2Vec 2.0 method is suitable to model different impairments for PD detection. This could be further investigated in the future with other controlled experiments such as at the level of phonemes, words, and phrases that could help interpret the features obtained by this model.

5. Conclusions

This paper shows that FL model yields similar or even better results compared to local approaches where mono-lingual models are created for every test database. FL offers the advantage of not requiring any data sharing between institutions, which we hope will encourage researchers and practitioners to improve scientific collaborations among different institutions around the world. The approach shows that FL allows for obtaining competitive results while preserving data privacy. We expect these results to promote simpler and more frequent collaborations between medical institutions, and subsequently, to further improve patient outcomes.

6. Acknowledgments

7. References

[1] J. Logemann *et al.*, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[2] A. McKinlay *et al.*, "A profile of neuropsychiatric problems and their relationship to quality of life for Parkinson's disease patients without dementia," *Parkinsonism & related disorders*, vol. 14, no. 1, pp. 37–42, 2008.

[3] S. Pinto *et al.*, "Treatments for dysarthria in Parkinson's disease," *The Lancet Neurology*, vol. 3, no. 9, pp. 547–556, 2004.

[4] K. A. Spencer and M. A. Rogers, "Speech motor programming in hypokinetic and ataxic dysarthria," *Brain and Language*, vol. 94, no. 3, pp. 347–366, 2005.

[5] J. Robin *et al.*, "Evaluation of speech-based digital biomarkers: Review and recommendations," *Digital Biomarkers*, vol. 4, no. 3, pp. 99–108, 2020.

[6] L. Moro-Velazquez *et al.*, "Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects," *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[7] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[8] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] A. Nautsch *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[10] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima Jr, J. Mancuso, F. Jungmann, M.-M. Steinborn *et al.*, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 473–484, 2021.

[11] G. A. Kaissis *et al.*, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[12] S. Tayebi Arasteh, A. Ziller *et al.*, "Private, fair and accurate: Training large-scale, privacy-preserving ai models in medical imaging," 2023. [Online]. Available: https://arxiv.org/abs/2302.01622

[13] J. Konecný *et al.*, "Federated optimization: Distributed machine learning for on-device intelligence," *ArXiv preprint arXiv:1610.02527*, 2016.

[14] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. [Online]. Available: https://arxiv.org/abs/1610.05492

[15] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[16] D. Truhn, S. Tayebi Arasteh *et al.*, "Encrypted federated learning for secure decentralized collaboration in cancer image analysis," *medRxiv*, 2022. [Online]. Available: https://www.medrxiv.org/content/early/2022/07/31/2022.07.28.22277288

[17] S. Tayebi Arasteh *et al.*, "Collaborative training of medical artificial intelligence models with non-uniform labels," *Scientific Reports*, vol. 13, p. 6046, 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-33303-y

[18] M. J. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[19] S. Tayebi Arasteh *et al.*, "The effect of speech pathology on automatic speaker verification – a large-scale study," *ArXiv preprint arXiv:2204.06450*, 2022.

[20] C. D. Rios-Urrego *et al.*, "Is there any additional information in a neural network trained for pathological speech classification?" in *Proceedings of TSD*. Springer Nature, 2021, pp. 435–447.

[21] N. Tomashenko *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[22] J. R. Orozco-Arroyave *et al.*, "New spanish speech corpus database for the analysis of people suffering from Parkinson's disease." in *Proceedings of LREC*, 2014, pp. 342–347.

[23] T. Bocklet *et al.*, "Automatic evaluation of Parkinson's speech-acoustic, prosodic and voice related cues," in *Proceedings of INTERSPEECH*, 2013, pp. 1149–1153.

[24] J. Rusz, "Detecting speech disorders in early Parkinson's disease by acoustic analysis," *Habilitation Thesis, Czech Technical University in Prague*, 2018.

[25] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[26] V. Panayotov *et al.*, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings of ICASSP*. IEEE, 2015, pp. 5206–5210.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML*. pmlr, 2015, pp. 448–456.

[28] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of ICCV*, 2015, pp. 1026–1034.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[30] J. C. Vásquez-Correa *et al.*, "Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages," *Pattern Recognition Letters*, vol. 150, pp. 272–279, 2021.

[31] A. Nilsson *et al.*, "A performance evaluation of federated learning algorithms," in *Proceedings of DIDL*, 2018, pp. 1–8.

[32] T. Li *et al.*, "Federated optimization in heterogeneous networks," *Proceedings of MLSys*, vol. 2, pp. 429–450, 2020.

[33] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," *Advances in neural information processing systems*, vol. 33, pp. 15 111–15 122, 2020.

[34] T. Feng and S. Narayanan, "Semi-FedSER: Semi-supervised Learning for Speech Emotion Recognition On Federated Learning using Multiview Pseudo-Labeling," in *Proc. Interspeech 2022*, 2022, pp. 5050–5054.