CROSS-DEVICE FEDERATED LEARNING FOR MOBILE HEALTH DIAGNOSTICS: A FIRST STUDY ON COVID-19 DETECTION

Tong Xia, Jing Han, Abhirup Ghosh, Cecilia Mascolo
Univeristy of Cambridge
tx229@cam.ac.uk

ABSTRACT

Federated learning (FL) aided health diagnostic models can incorporate data from a large number of personal edge devices (e.g., mobile phones) while keeping the data local to the originating devices, largely ensuring privacy. However, such a *cross-device* FL approach for health diagnostics still imposes many challenges due to both local data imbalance (as extreme as local data consists of a single disease class) and global data imbalance (the disease prevalence is generally low in a population). Since the federated server has no access to data distribution information, it is not trivial to solve the imbalance issue towards an unbiased model. In this paper, we propose *FedLoss*, a novel *cross-device* FL framework for health diagnostics. Here the federated server averages the models trained on edge devices according to the predictive loss on the local data, rather than using only the number of samples as weights. As the predictive loss better quantifies the data distribution at a device, *FedLoss* alleviates the impact of data imbalance. Through a real-world dataset on respiratory sound and symptom-based COVID-19 detection task, we validate the superiority of *FedLoss*. It achieves competitive COVID-19 detection performance compared to a centralised model with an AUC-ROC of 79%. It also outperforms the state-of-the-art FL baselines in sensitivity and convergence speed. Our work not only demonstrates the promise of federated COVID-19 detection but also paves the way to a plethora of mobile health model development in a privacy-preserving fashion.

Index Terms— Federated learning, Privacy-preserving, Mobile health, COVID-19 detection, Acoustic modelling

1. INTRODUCTION

Pervasive mobile devices along with on-device machine learning enable continuous sensing of individual health signals and cost-effective health screening at population scale [1]. However, traditional machine learning methods need the data from all the devices to be aggregated at a central server, raising privacy concerns as the health status and other personally identifiable information can potentially be leaked from the untrusted server or during data sharing [2]. Federated learning (FL) avoids aggregating the data and thus promise privacy by iteratively learning models at the participating devices using their local data and then aggregating the local models at a central server [3, 4]. This opens a new way for privacy-preserving diagnostic model development.

Most existing diagnostic FL frameworks consider cooperation among hospitals or health institutions with each participant containing clinical data from multiple individuals (also known as *cross-silo* FL setting) [5, 6, 7, 8]. While such settings have boosted accuracy over participating institutions learning in isolation and improved privacy over centralising the data from all institutes, they still fall short in scaling to more distributed settings where the data of each participant resides on their mobile devices. The *cross-silo* FL algorithms do not trivially transfer to *cross-device* FL settings mainly because the latter has many orders of magnitudes more client devices.

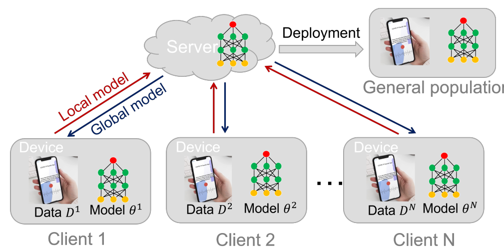


Fig. 1. *Cross-device* FL for mobile health, where models are trained on edge devices from private health sensing data, and the global model is aggregated from the clients' models.

In this paper, we push the envelope of decentralisation by considering *cross-device* FL, where the data resides in users' (clients') edge devices. The learning works in rounds and at every round, each client's edge device trains a model using locally collected health signals and disease labels, while the federated server aggregates the local models into a global one. Finally, the trained model is used for population health screening by any client device using its local sensing data (Fig. 1).

Cross-device FL imposes the following challenges: i) An individual's health status changes very slowly generally. Therefore, most personal devices will only present a single class, i.e., the current health status of the device owner. It is infeasible to balance the data distribution on the device, and thus learning from such data, the local model is likely to over-fit and be biased. ii) Due to the generally low disease prevalence, the data is also globally imbalanced, with a large proportion of healthy individuals. Without accessing the label distribution, the global aggregation could introduce an unwanted bias in the classification. Yet, failing to detect the disease may come at a heavy price in healthcare applications.

To address the local and global class imbalance, this paper proposes an efficient federated training algorithm, *FedLoss*. The novelty of *FedLoss* lies in its adaptive model aggregation: only a small number of clients are required to participate in each round, and their models are aggregated according to adaptive weights proportional to the predictive loss on their local data. Such an adaptive aggregation strategy alleviates the impact of data imbalance and speeds up global model convergence. The performance of *FedLoss* is validated in a COVID-19 detection task, where respiratory sounds (cough, breathing, and voice) and symptoms are leveraged to diagnose COVID-19. A dataset is crowd-sourced from around $3,000$ users through a mobile application [9, 10]. We learn a COVID-19 diagnostic classifier

This work was supported by ERC Project 833296 (EAR).

where the data stays on the devices, i.e., our experiments consider each user to be a single federated client.

There are two main contributions in this paper. First, we propose a novel federated training algorithm to enable *cross-device* FL for mobile health diagnostics and tackle the challenge resulting from data imbalance. Further, we conduct extensive experiments in a real-world COVID-19 detection task. Results demonstrate the superiority of our method over the start-of-the-art baselines.

2. RELATED WORK

Skewed label distribution across edge devices is natural in real-world applications, particularly in the healthcare domain [11]. It poses a challenge in FL: due to privacy constraints, class distribution cannot be handled by explicitly identifying the minority class [12] and thus it makes the solutions explored in classical centralised settings invalid. Some efforts have concentrated on client clustering [13, 14], adapting the global model based on auxiliary data [15], and adaptive client training by monitoring the loss from a global perspective [16, 12]. Yet, they either are inefficient when the number of clients is large or require additional centralised data. A close work to our study [13] (*FedCluster*) considered a *cross-device* setting in FL to diagnose arrhythmia from electrocardiograms. To improve the performance for the rare phenotype, *FedCluster* clusters the clients based on a global shared dataset. Then the local models are first merged within clusters and then cluster models are aggregated into the global model. On the contrary, we aim to solve the imbalance problem without any global data.

Cross-silo FL has been explored for health diagnostics including COVID-19. For example, Feki *et al.* proposed FL frameworks allowing multiple medical institutions to screen COVID-19 from Chest X-ray images without sharing patient data [5, 6, 7, 8]. Vaid *et al.* explored electronic medical records to improve mortality prediction across hospitals via FL [17, 18]. In these settings, the number of clients is small and the size of the local data is relatively large. To the best of our knowledge, we are the first to propose a *cross-device* federated learning framework for detecting COVID-19 from personal sounds and symptoms. This is more challenging than *cross-silo* FL due to the extreme data heterogeneity from the thousands of clients.

3. METHODOLOGY

3.1. Problem Formulation

Consider a system with N federated clients with each client, n owning a private local dataset $\mathcal{D}^n = \{(x_1^n, y_1^n), (x_2^n, y_2^n), ...\}$, where x_j^n is a health signal sample and y_j^n denotes the health status, i.e., if the associated disease is identified in the sample, $y_j^n = 1$, otherwise $y_j^n = 0$. y_j^n is locally extremely imbalanced with most clients presenting a single class, and it is also globally imbalanced with $y_j^n = 0$ (healthy) being the majority class. As shown in Fig. 1, we aim to train a federated model parameterised by θ that can predict y for any given x to achieve population health screening.

3.2. Basics of Federated Learning

Federated learning is an iterative process consisting of the following steps at every round: (1) At every round, t, each participating client, i receives a copy of the global model from the previous round, θ_{t-1} and updates it using its private local data to θ_t^i. (2) Each participating client sends updated model parameters, $g_t^i = \theta_t^i - \theta_{t-1}$ to the server. (3) The server updates the global model to θ_t by aggregating g_t^is. (4) Steps (1) to (3) are repeated until the global model converges.

The most popular aggregation strategy (step 3) is Federated Averaging (*FedAvg*) [19, 20, 21, 22], where the aggregation is an average of the model updates weighted by α_t^i, the fraction of the data samples at client i w.r.t. to the total samples available in the system,

$$\theta_t = \theta_{t-1} - \eta \sum_i \alpha_t^i g_t^i, \tag{1}$$

where η is the global updating rate.

Algorithm 1: FedLoss Algorithm

Data: Global model update rate η, global training rounds T, local update rate λ, local training epochs E, the number of clients each round M.
Result: Global model θ_T.

Server executes:
Initialise θ_0
for *each round t = 1,2,...,T* **do**
  $S_t \leftarrow$ A random set of M clients
  for *each client* $i \in S_t$ *in parallel* **do**
    $l_t^i, g_t^i \leftarrow i$-th client executes
  end
  $w_t = softmax(l_t^1, ..., l_t^M)$ # Different from FedAvg
  $\theta_t \leftarrow \theta_{t-1} - \eta \sum_{i=1}^M w_t^i g_t^i$
end

Client executes:
Received a global model θ_{t-1}
Initialise loss $l_t^i = 0$
for *sample* $j = 1, 2, ..., |\mathcal{D}^i|$ **do**
  $l_t^i \leftarrow l_t^i + CrossEntropy(\theta_{t-1}; \mathcal{D}_j^i)$ # Returning loss
end
Synchronise local model $\theta_{t,0}^i = \theta_{t-1}$
for *local epoch* $e = 1, 2, ..., E$ **do**
  $\theta_{t,e}^i \leftarrow \theta_{t,e-1}^i - \lambda \nabla_\theta CrossEntropy(\theta_{t,e-1}^i; \mathcal{D}^i)$
end
Calculate the overall update: $g_t^i = \theta_{t,E}^i - \theta_{t-1}$
Return l_t^i, g_t^i

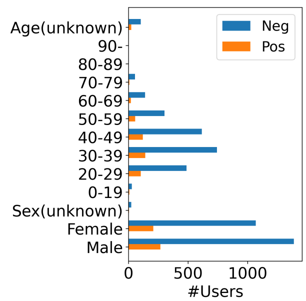
3.3. FedLoss

FedAvg is vulnerable to class imbalance as α_t^i ignores the label imbalance among the clients. To overcome this, we propose *FedLoss* (Algorithm 1) to achieve adaptive aggregation.

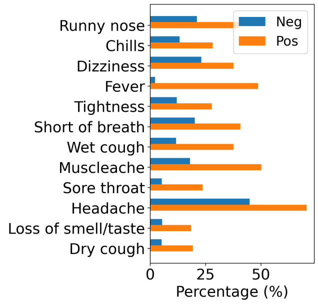
At each round of *FedLoss*, M clients are randomly selected to participate in training. Each selected client, i, optimises the received model for E epochs using the local data $\mathcal{D}^i$. The major difference between *FedLoss* and *FedAvg* is that at each round t in addition to sharing models, client i provides the predictive loss, l_t^i to support a weighted aggregation. l_t^i denotes the total cross-entropy loss incurred by the global model, θ_t on its local data, D^i. Note that l_t^i is computed prior to the local training step and thus it does not suffer from over-fitting at a client with small data.

Since unhealthy clients are under-represented (globally minority class), intuitively they are more likely to yield relatively higher predictive loss. Thus, *FedLoss* will assign a higher weight to their model updates, rendering the data on such clients to be more predictable by the global model. Finally, the server normalises the received losses using a *softmax* function to get the client-wise weights for aggregation. The adaptive aggregation in t-th round is denoted as,

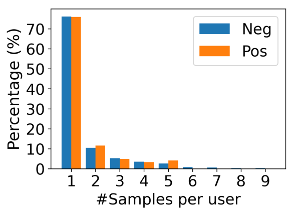
$$\begin{aligned} w_t &= softmax(l_t^1, ..., l_t^M), \\ \theta_t &= \theta_{t-1} - \eta \sum_{i=1}^M w_t^i g_t^i, \end{aligned} \tag{2}$$



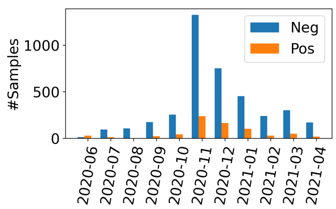
(a) Demographics.



(b) Symptoms distribution.



(c) User data amount.



(d) Monthly Data amount.

Fig. 2. Statistics of the data from 482 COVID-19 positive users and 2,478 negative users.

where w_t^i denotes the weight for the participating client i. The overall process is summarised in Algorithm 1.

4. EXPERIMENTAL SETUP

This section empirically evaluates *FedLoss* for COVID-19 detection.

4.1. Data Details

We use the data collected by a crow-sourced mobile application, *COVID-19 Sounds*[1]. At registration, the app assigns each user a unique anonymous ID. Users record their symptoms (cough, fever, etc.), three respiratory sound recordings (breathing, coughing, and speech), and the COVID-19 testing status on the corresponding day [9, 10]. After data cleaning (i.e., excluding non-English speakers, samples without COVID-19 test results and poor audio quality samples), there are 482 users with positive status and 2,478 users with negative status with a total of 4,612 samples. An overview of the statistics of the data is in Fig. 2: (a) The data represents a typical demographic distribution in a population. (b) There are more negative than positive users, with many asymptomatic positive users while a great proportion of the negative users report respiratory disease-related symptoms. (c) User data is sparse with over 70% of users only recording one sample. (d) The data accumulation procession spanned one year.

4.2. Backbone COVID-19 Detection Model

Following the previous works [10, 23], a VGGish framework is employed to extract acoustic features from the spectrogram of audio samples. Additionally, Han *et al.* reported that fusing the symptoms and acoustic features in an early stage of the deep model can achieve better COVID-19 detection performance than using a single modality. Inspired by this, we use a multi-modal deep learning model to predict COVID-19 status from audio and symptoms jointly, as illustrated in Fig. 3. Symptoms are represented by a multi-hot vector, which is concatenated with the dense feature from VGGish network outputs. The concatenated feature vector is then fed to a multi-layer fully connected network for classification. The final layer outputs a *Softmax* based binary class probabilities.

[1] www.covid-19-sounds.org

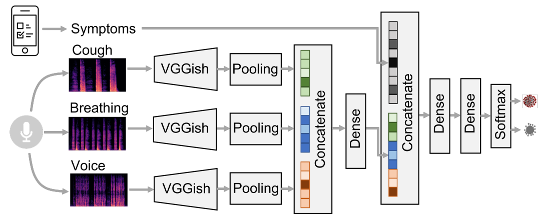


Fig. 3. A multi-modal model for COVID-19 detection.

4.3. Settings

This paper considers each app user as a federated client to examine *FedLoss*. Out of 2,960 users in the dataset we randomly held out 20% clients for testing and use the rest 80% of the clients for federated training. We experiment with two training settings:

- **Randomly**: The recorded data is assumed to be kept on the client device during the whole training period. We run $T = 2000$ federated rounds and $M = 30$ clients are randomly selected at each round.
- **Chronologically**: The recorded data is assumed to be cleared monthly by the user, which is practical. Regarding this, we design a multi-period training strategy: every month, only the clients with data recorded in this period can be selected and we run 100 rounds with each round sampling $M = 30$ clients for training (100 rounds can guarantee the convergence of the model on the incremental data).

All the experiments are implemented by Pytorch on a GPU with 64G memory. To avoid over-fitting on the client, a pre-trained VGGish is utilised [10], and the local training epoch is set to $E = 1$. A local learning rate of 0.008 for VGGish and 0.015 for the rest parameters are used for the SGD optimiser. The global update rate $\eta = 1$.

4.4. Baselines and Metrics

In addition to *FedAvg*, we also compare with *FedProx* [20]. *FedProx* handles non-identically distributed data across federated clients by regularising the local training loss at the clients so that the local models incur limited divergence from the global model.

For evaluation, we first use AUC-ROC (short for AUC) to show the overall rationality of the estimated diagnostic probability. Following the rule that for a sample if the predictive probability of being positive is larger than being negative, i.e., $p_{pos} > p_{neg}$, it will be diagnosed as positive, we also present sensitivity (SE) - the ratio between the correctly identified COVID-19 positive samples and overall positive samples, and specificity (SP) - the correct ratio for the healthy class. Additionally, we report sensitivity with a specificity of 80% (SE@80%SP) by tuning the decision threshold, i.e., a sample will only be diagnosed as positive when $p_{pos} > p_{neg} + \tau$, where τ is searched to guarantee a SP of 80%. A 95% Confidence Interval (CI) for all metrics is reported by using bootstrap [24].

5. RESULTS

5.1. Results and Discussion under Randomly Training Setting

COVID-19 Detection Performance. The overall performance comparison is summarised in Table 1. All federated learning based approaches achieve competitive AUC-ROC against centralised training. However, the federated baselines are unable to effectively detect COVID-19 positive users with sensitivity lower than 20%, although their specificity is very high. In contrast, our *FedLoss* yields a sensitivity of 50% while maintaining the specificity around 90%. In other words, *FedLoss* achieves the best trade-off for detecting positive and

Table 1. Performance comparison under *randomly* training setting. 95% CIs are reported in brackets.

	AUC	SE	SP	SE@80%SP
Centralised	0.79 (0.74 – 0.84)	0.46 (0.36 – 0.56)	0.93 (0.91 – 0.94)	0.62 (0.54 – 0.69)
FedAvg	0.80 (0.75 – 0.85)	0.11 (0.06 – 0.17)	1.00 (1.00 – 1.00)	0.59 (0.45 – 0.73)
FedProx	0.75 (0.69 – 0.80)	0.19 (0.12 – 0.27)	0.99 (0.99 – 1.00)	0.48 (0.31 – 0.63)
FedLoss (Proposed)	0.79 (0.73 – 0.83)	0.50 (0.40 – 0.59)	0.90 (0.88 – 0.92)	0.62 (0.50 – 0.70)

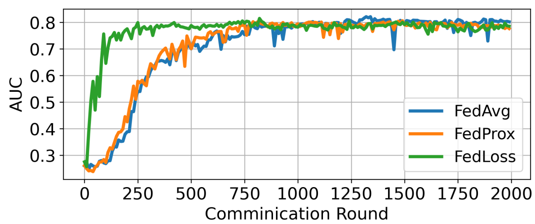


Fig. 4. Convergence analysis. AUC-ROC of testing set for every 10 round during training is displayed.

negative users, as proved by the highest average value of sensitivity and specificity (70%). When fixing the specificity of 80% uniformly, our *FedLoss* achieves sensitivity up to 62%, which is as good as the centralised model. All those validate the superiority of our weighted aggregation strategy in handling the data imbalance.

Convergence Comparison. System efficiency is another important metric for *cross-device* FL. To compare the convergence speed of *FedAvg*, *FedProx* and *FedLoss*, we show the testing AUC-ROC during the training process in Fig. 4. It can be observed that the AUC-ROC of our *FedLoss* gets converged significantly faster than the baselines: *FedLoss* needs about 250 rounds while *FedAvg* and *FedProx* requires about 1000 rounds. Therefore, *FedLoss* is 4× more efficient than baselines. Note that fewer communication rounds to convergence saves both computation and communication costs at the resource constraint edge clients.

Analysis of Weights. We conduct additional analysis on the adaptive weight during the training process. Since our *FedLoss* shows a superior sensitivity against the baselines, we particularly look at how the weights changed for COVID-19 positive and negative clients, for a comparison. Fig. 5 displays the average weight for positive and negative clients in each round. It is observed that in the beginning 100 rounds, the weight of positive clients is 4∼6 times of negative clients. This suggests the system can detect the potentially minority class as those clients are more difficult to predict. In the later rounds, the weights for positive and negative clients gradually become more balanced, since the global model has already learned the COVID-19 features, to a great extent.

5.2. Discussion under Chronologically Training Setting

The second setting aims to evaluate the performance of long-term FL with limited client participation in batches. As illustrated by the SE@80%SP in different periods in Fig. 6, all methods are inaccurate and unusable at the early stage with SE@80%SP lower than 50%. The poor performance is mainly attributed to the limited number of clients (i.e., the limited data), which leads to poor generalisation. Gradually, with more training rounds, from November 2020 our *FedLoss* starts to show a convergence trend with the SE@80%SP reaching 60%. Finally, our model achieves an AUC-ROC of 79%, a sensitivity of 45% and specificity of 90%, as summarised in Table 2.

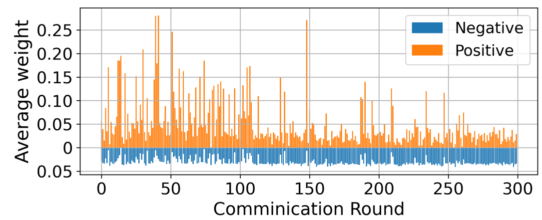


Fig. 5. Average weight for COVID-19 positive and negative clients per communication round. Note that the negative clients do not have negative weights, but the weights are just shown in negative direction for visualisation convenience.

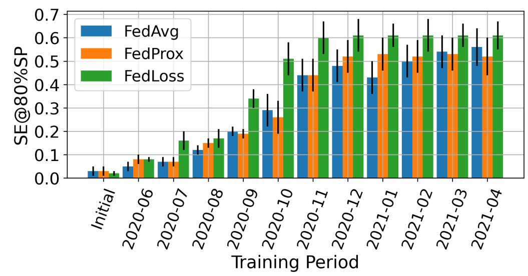


Fig. 6. Performance of the global model trained *chronologically*. Sensitivity with a specificity of 80% by the last round model in each month is displayed.

Table 2. Overall performance of the final model under *chronologically* training setting. 95% CIs are reported in brackets.

	AUC	SE	SP	SE@80%SP
FedAvg	0.79 (0.73 – 0.82)	0.20 (0.15 – 0.23)	0.99 (0.98 – 1.00)	0.56 (0.49 – 0.63)
FedProx	0.78 (0.75 – 0.81)	0.15 (0.10 – 0.20)	0.99 (0.98 – 1.00)	0.53 (0.44 – 0.60)
FedLoss (Proposed)	0.79 (0.74 – 0.84)	0.45 (0.39 – 0.53)	0.90 (0.89 – 1.0)	0.61 (0.55 – 0.64)

On the contrary, SE@80%SP of *FedAvg* and *FedProx* has slower convergence rate, converging two months later than *FedLoss*. We also note that in November 2020, all three approaches present a remarkable performance gain, which is mainly because the quantity of data reaches a peak in that month (refer to Fig. 2(d)). Overall, our final SE@80%SP (62%) significantly surpasses that of *FedAvg* (56%) and *FedProx* (53%), and our SE (45%) is quite competitive compared with centralised model (46%). The above comparison further verifies that our proposed *FedProx* can achieve a more generalised global model with fewer clients involved.

6. CONCLUSION

In this paper, we studied the feasibility of *cross-device* federated mobile health using a COVID-19 detection task as an example. To handle the natural challenge of data imbalance, a novel federated aggregation algorithm *FedLoss* has been proposed. Experimental results demonstrate the superiority of our approach in both effectiveness and efficiency. *FedLoss* aggregation scheme is general and can be extended to other mobile health applications, e.g., heart sound-based arrhythmia prediction, and smartwatch-enabled sleep quality monitoring. This paper also facilitates the change from traditional crowdsourcing of data to crowdsourcing of models on a large scale for privacy-preserving mobile health applications. While this study is a beginning of an exciting direction of *cross-device* federated mobile health, many challenges lie ahead, for example, the sparsity of labelled data at the devices, addressing which will be future work.

7. REFERENCES

[1] Steven R Steinhubl, Evan D Muse, and Eric J Topol, "The emerging field of mobile health," *Science Translational Medicine*, vol. 7, no. 1, pp. 283–295, 2015.

[2] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee, "Breathprint: Breathing acoustics-based user authentication," in *Proc. International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 278–291.

[3] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[4] Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys*, vol. 55, no. 3, 2022, 37 pages.

[5] Ines Feki, Sourour Ammar, Yousri Kessentini, and Khan Muhammad, "Federated learning for COVID-19 screening from chest X-Ray images," *Applied Soft Computing*, vol. 106, pp. 107330, 2021.

[6] Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala Al-Fuqaha, and Junaid Qadir, "Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 172–184, 2022.

[7] Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather HC Lee, Kevin Yu, et al., "Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study," *NPJ Digital Medicine*, vol. 4, no. 1, 2021, 11 pages.

[8] Qian Yang, Jianyi Zhang, Weituo Hao, Gregory P Spell, and Lawrence Carin, "Flop: Federated learning on medical datasets using partial networks," in *Proc. ACM Conference on Knowledge Discovery & Data Mining (SIGKDD)*, 2021, pp. 3845–3853.

[9] Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo, "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8328–8332.

[10] Tong Xia, Dimitris Spathis, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, et al., "COVID-19 sounds: A large-scale audio dataset for digital respiratory screening," in *Proc. Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021, 10 pages.

[11] M Mostafizur Rahman and Darryl N Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, pp. 224, 2013.

[12] Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro, "An agnostic approach to federated learning with class imbalance," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[13] Daoqin Lin, Yuchun Guo, Huan Sun, and Yishuai Chen, "Fedcluster: A federated learning framework for cross-device private ecg classification," in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2022, 6 pages.

[14] Felix Sattler, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, "On the byzantine robustness of clustered federated learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8861–8865.

[15] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu, "Addressing class imbalance in federated learning," in *Proc. AAAI*, 2021, vol. 35, pp. 10165–10173.

[16] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021.

[17] Akhil Vaid, Suraj K Jaladanki, Jie Xu, Shelly Teng, Arvind Kumar, Samuel Lee, Sulaiman Somani, Ishan Paranjpe, Jessica K De Freitas, Tingyi Wanyan, et al., "Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach," *JMIR Medical Informatics*, vol. 9, no. 1, pp. e24207, 2021.

[18] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.

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

[20] 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.

[21] Yan Gao, Titouan Parcollet, Salah Zaiem, Javier Fernandez-Marques, Pedro PB de Gusmao, Daniel J Beutel, and Nicholas D Lane, "End-to-end speech recognition from federated acoustic models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7227–7231.

[22] Meng Feng, Chieh-Chi Kao, Qingming Tang, Ming Sun, Viktor Rozgic, Spyros Matsoukas, and Chao Wang, "Federated self-supervised learning for acoustic event classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 481–485.

[23] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al., "Sounds of COVID-19: Exploring realistic performance of audio-based digital testing," *NPJ Digital Medicine*, vol. 5, no. 1, 2022, 9 pages.

[24] Thomas J DiCiccio and Bradley Efron, "Bootstrap confidence intervals," *Statistical Science*, vol. 11, no. 3, pp. 189–228, 1996.