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

Tong Xia, Jing Han, Abhirup Ghosh, Cecilia Mascolo
University 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 leaking from the untrusted server or during data sharing [2]. Federated learning (FL) avoids aggregating the data and thus promises 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.

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.
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, $i$ owning a private local dataset $D^i = \{x^i_1, y^i_1), (x^i_2, y^i_2), \ldots, \}$, where $x^i_j$ is a health signal sample and $y^i_j$ denotes the health status, i.e., if the associated disease is identified in the sample, $y^i_j = 1$, otherwise $y^i_j = 0$. $y^i_0$ is locally extremely imbalanced with most clients presenting a single class, and it is also globally imbalanced with $y^0_0 = 0$ (healthy) being the majority class. As shown in Fig. 1, we aim to train a federated model parameterised by $\theta$ 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, $\theta_{t-1}$ and updates it using its private local data to $\theta^i_t$. (2) Each participating client sends updated model parameters, $g^i_t = \theta^i_t - \theta_{t-1}$ to the server. (3) The server updates the global model to $\theta_t$ by aggregating $g^i_t$s. (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 $\alpha^i_t$, the fraction of the data samples at client $i$ w.r.t. to the total samples available in the system,

$$\theta_t = \theta_{t-1} - \eta \sum_i \alpha^i_t g^i_t,$$

where $\eta$ is the global updating rate.

3.3. FedLoss

FedAvg is vulnerable to class imbalance as $\alpha^i_t$ 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 $D^i$. The major difference between FedLoss and FedAvg is that at each round $t$ in addition to sharing models, client $i$ provides the predictive loss, $l^i_t$ to support a weighted aggregation. $l^i_t$ denotes the total cross-entropy loss incurred by the global model, $\theta_t$ on its local data, $D^i$. Note that $l^i_t$ 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 for aggregation. The adaptive aggregation in $t$-th round is denoted as,

$$w_t = softmax(l^1_t, \ldots, l^M_t)$$

$$\theta_t = \theta_{t-1} - \eta \sum_{i=1}^M w^i_t g^i_t,$$

where the data heterogeneity from the thousands of clients.

**Algorithm 1: FedLoss Algorithm**

**Data:** Global model update rate $\eta$, global training rounds $T$, local update rate $\lambda$, local training epochs $E$, the number of clients each round $M$.

**Result:** Global model $\theta_T$.

```
1 Server executes:
2 Initialise $\theta_0$
3 for each round $t = 1, 2, \ldots, T$ do
4     $S_t \leftarrow$ A random set of $M$ clients
5     for each client $i \in S_t$ in parallel do
6         $l^i_t, g^i_t \leftarrow i$-th client executes
7     end
8     $w_t = softmax(l^1_t, \ldots, l^M_t)$ # Different from FedAvg
9     $\theta_t \leftarrow \theta_{t-1} - \eta \sum_{i=1}^M w^i_t g^i_t$
10 end
```

```
symptoms distribution.

(b) Monthly Data amount.

[Image 58x504 to 166x714]

We use the data collected by a crowd-sourced mobile application, COVID-19 Sounds. Following the previous works [10, 23], a VGGish framework is employed to extract acoustic features from the spectrogram of audio samples, there are 482 users only recording one sample. (d) The data accumulation is chronologically presented in Fig. 2: (a) The data represents a typical demographic distribution in a population. (b) There are more symptoms 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 process spanned one year.

4. EXPERIMENTAL SETUP

This section empirically evaluates FedLoss for COVID-19 detection.

4.1. Data Details

We use the data collected by a crowd-sourced mobile application, COVID-19 Sounds. 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 shown 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 process 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.

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 to 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 specificity 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 $\tau$ 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
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.

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

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>SE</th>
<th>SP</th>
<th>SE@80%SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralised</td>
<td>0.79</td>
<td>0.54</td>
<td>0.62</td>
<td>(0.74 – 0.84)</td>
</tr>
<tr>
<td></td>
<td>(0.66 – 0.90)</td>
<td>(0.36 – 0.56)</td>
<td>(0.91 – 0.94)</td>
<td>(0.54 – 0.69)</td>
</tr>
<tr>
<td>FedLoss</td>
<td>0.59</td>
<td>0.11</td>
<td>0.09</td>
<td>(0.78 – 0.85)</td>
</tr>
<tr>
<td></td>
<td>(0.66 – 0.73)</td>
<td>(0.06 – 0.17)</td>
<td>(1.00 – 1.00)</td>
<td>(0.45 – 0.73)</td>
</tr>
<tr>
<td>FedProx</td>
<td>0.79</td>
<td>0.30</td>
<td>0.19</td>
<td>(0.73 – 0.80)</td>
</tr>
<tr>
<td></td>
<td>(0.40 – 0.59)</td>
<td>(0.08 – 0.92)</td>
<td>(0.50 – 0.70)</td>
<td>(0.48 – 0.63)</td>
</tr>
</tbody>
</table>

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


