Federated Remote Physiological Measurement with Imperfect Data

Xin Liu[1], Mingchuan Zhang[3], Ziheng Jiang[1], Shwetak Patel[1], Daniel McDuff[2]
Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA[1]
Microsoft Research, Redmond, USA[2]
School of Computer Science, Fudan University, Shanghai, China[3]
{xliu0, ziheng, shwetak}@cs.washington.edu,
damcduff@microsoft.com, mczhang18@fudan.edu.cn

Abstract

The growing need for technology that supports remote healthcare is being acutely highlighted by an aging population and the COVID-19 pandemic. In health-related machine learning applications the ability to learn predictive models without data leaving a private device is attractive, especially when these data might contain features (e.g., photographs or videos of the body) that make identifying a subject trivial and/or the training data volume is large (e.g., uncompressed video). Camera-based remote physiological sensing facilitates scalable and low-cost measurement, but is a prime example of a task that involves analysing high bit-rate videos containing identifiable images and sensitive health information. Federated learning enables privacy-preserving decentralized training which has several properties beneficial for camera-based sensing. We develop the first mobile federated learning camera-based sensing system and show that it can perform competitively with traditional state-of-the-art supervised approaches. However, in the presence of corrupted data (e.g., video or label noise) from a few devices the performance of weight averaging quickly degrades. To address this, we leverage knowledge about the expected noise profile within the video to intelligently adjust how the model weights are averaged on the server. Our results show that this significantly improves upon the robustness of models even when the signal-to-noise ratio is low.

1. Introduction

Federated learning (FL) enables distributed devices (e.g., cellphones) to collaboratively learn models without data leaving each device [15, 23]. While creating traditional machine learning systems involves uploading raw data and labels to a centralized location for training, FL can avoid this. A core premise is that a model trained from aggregated decentralized data can be more effective than training with the data that any one device has access to on its own. More specifically, federated learning leverages locally-computed updates (weights) from a large number of single devices to create a robust aggregated model that can then be shared. To summarize, federated learning has several useful properties, the ability to: 1) preserve privacy more easily by only sharing model weights instead of raw data and labels, 2) increase the diversity and generalizability of a model by aggregating a diverse population's data, 3) reduce the bandwidth and storage resources required when uploading raw data to a centralized server.

The benefits of FL are particularly attractive in applications in which models rely on sensitive data that are also personally identifiable. This is very true in contexts that involve biometric, physiological and health data. The growing need for technology that supports remote healthcare has been acutely highlighted by the COVID-19 pandemic [37, 41]. One such technology that can support remote care is low-cost, on-device, camera-based vital sign measurement [8, 18, 19, 26, 35, 36]. These systems use ubiquitously available webcams and smartphone cameras to measure important physiological vital signs such as the cardiac pulse [27], breathing rate [26] and blood oxygen saturation [32] of a patient without the data leaving the device. The methods rely on capturing subtle variations in light reflected from the body that capture volumetric changes in blood (the photoplethysmogram/PPG) and mechanical motions resulting from cardiac and respiratory function (e.g., the ballistocardiogram/BCG) [21]. Democratizing (or scaling) camera-based physiological sensing in this way has much potential. For example, to help in screening for atrial fibrillation and other forms of arrhythmia [5] which are predictors of stroke risk.

Video recordings that contain the necessary fidelity to capture physiological changes contain *both* private health data *and* personally identifiable information. The physiological signals themselves have personally identifiable features [12] and the video frames may also contain visually

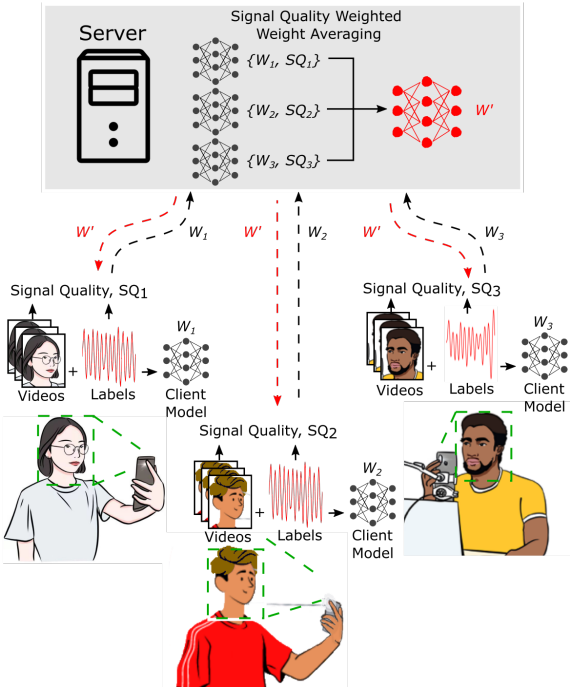

Figure 1. We present a privacy preserving federated system for on-device, camera-based physiological sensing. We propose a novel weight averaging approach that significantly improves on model robustness in the presence of noisy videos and labels. W_N represents the weights from each client, SQ_N represents the signal quality score either for the video, labels or both, and W' represents the server weights after weight averaging.

recognizable body parts (e.g., the face). Furthermore, to effectively measure the very subtle changes in the body associated with these physiological processes, the videos should not be compressed too heavily as motion-compression algorithms typically remove the signals of interest [22]. As such, the recordings contain sensitive data and are often large; therefore, they ideally would not be transferred or stored in great volumes in the cloud.

When building models for measuring physiological vital signs, it is critical that the learned representations are not corrupted because of "bad" data (either features or labels) from a few devices. However in the context of FL where the server does not have access to the data itself, how do we ensure that that this does not happen? Ideally, during weight aggregation it would be possible to adapt to, or *exploit*, client weights that were derived from cleaner rather than noisier data. At the same time, we do not want to completely ignore weights from a given client as every client will have access to data from a subject that was not "seen" by other clients and generally we would want a model to *explore* and maximize the diversity of our observations.

As shown in Fig. 1, in our scenario we have individuals collecting video on their own mobile devices alongside reference sensor measurements for training (as in [20]). In this case, there could be different levels of video noise resulting from camera sensor quality and automatic gain calibration. There could also be noise in the reference label, for example if a person was moving during the calibration period or did not attach the reference sensor correctly. Fortunately, both video and the physiological signals of interest (i.e., the PPG signal) have been studied extensively. We have strong statistical priors about the nature of these signals. In this work, to demonstrate our approach clearly we perform experiments assuming knowledge about the signal-to-noise ratios in the videos and labels. However, we could equally leverage domain knowledge to automatically calculate weight contributions from different devices. Our method does not discard the weights from clients with noisy data, but rather includes all weights while accounting for signal quality.

The contributions of this paper are: 1) to introduce the first federated camera-based remote physiological measurement system, 2) to show that this system can match the performance of a traditional supervised learning approach, 3) to introduce a critical averaging approach that accounts for the signal quality and diversity of samples. 4) to provide an on-device mobile training and inference implementation. Our code, models, and video figures are provided in the supplementary materials.

2. Related Work

Federated Learning in Healthcare. Federated learning enables training machine learning models from a set of distributed remote devices (e.g., mobile devices) while storing data only on the individual clients. Early work established optimization principals on how to perform non-convex optimization on distributed client's model weights [23]. Due to federated learning's unique characteristics in protecting privacy, it has been used and studied in healthcare applications. The volume of training data in healthcare applications is often smaller than in many traditional machine learning tasks. Therefore, aggregating as much data as possible from decentralized clients' could help boost the performance of machine learning applications in healthcare while not leaking sensitive information or violating HIPAA guidelines [29, 38]. Brisimi et al. [3] proposed to use federated learning to train a supervised classification model for cardiac events. More specifically, they develop a federated learning based framework to enable multiple data holders (i.e., hospitals) to collaborate and converge to a centralized model. More recently, [9] proposed a framework that leveraged federated learning to perform transfer learning for wearable sensors called FedHealth. In this framework, when the clients receive the updated model weights from the server all the layers in the neural network are frozen except for the last two fully connected dense layers. They claim that fine-tuning the last two layers on the client side can help build personalized models

for each user or organization. FedHealth was evaluated on a Parkinson's disease dataset. The application of federated learning in COVID-19 has also been investigated. Qayyum et al. [28] explored the use of federated learning in automatic diagnosis of COVID-19. They demonstrated improvements on results of X-ray and Ultrasound datasets after using federated learning. In the field of physiological measurement, Brophy et al. [4] investigated the use of federated learning and generative adversarial networks to estimate continuous blood pressure from the PPG signal. This work is quite distinct from ours as it uses contact sensor based PPG measurements while our work is focused on deriving the PPG signal and heart rate from facial videos.

Machine Learning in Remote Physiological Measurement. Remote physiological measurement or camera based physiological measurement is an emerging field. Early research established signal processing based methods for extracting physiological signals (in particular the cardiac pulse) from light reflection capture by the camera [10, 18, 26, 31, 34–36]. For example, Independent Component Analysis (ICA) was proposed to demix RGB channel information to recover a source containing the blood volume pulse (BVP) [26]. Wang et. al further extended this by calculating a projection plane orthogonal to the skin-tone based on physical principles [36]. Similar to many other vision tasks, deep learning has also helped boost the performance of remote physiological sensing, making models more robust to sources of noise seen in real-world applications including head motions and ambient lighting changes. A two-branch convolutional attention neural network was first proposed [8]. To model spatial and temporal information from the videos simultaneously, a 3D convolutional neural network was presented to further improve performance [39]. More recently, an on-device Temporal Shift Convolutional Attention Network (TS-CAN) was proposed to address the gap between efficiency and accuracy [19]. TS-CAN achieved state-of-art accuracy while dramatically reducing the computational cost and enabling real-time demonstrations on an embedded system at a high frame rate. Researchers have also investigated meta learning as a way to perform few-shot adaption for personalizing camera-based physiological sensing models [16, 20].

3. Method

Traditional supervised learning approaches to camera-based physiological sensing have been trained on large-scale centralized video datasets and physiological labels [8, 20, 39]. There are several drawbacks to this. First, the data are highly identifiable containing appearance (e.g., faces) and physiological information. Second, these data consume considerable data storage resources (data for each subject often excess 1GB). For these reasons it would be desirable to have a solution that only involves analyzing videos on the client (so that videos need not be shared) and ideally in distributed manner. In this paper, we explore the use of federated learning in camera-based video-based physiological measurement. We leverage domain knowledge about the expected noise profile within our data to intelligently dynamically adjust how the model weights are averaged on the server. Our results empirically show that approach creates a more accurate physiological estimation model.

Algorithm 1 FedWeight: Federated Remote Physiological Measurement with Signal Quality Weighting

Require: S: Subject-wise video data
1: **Server Update:** with an initialization W_0
2: **for** each round $t = 1, 2, 3...$ **do**
3: $S_t \leftarrow$ random select a set of clients
4: **for** each client k in S_t **do**
5: $\omega_t^k, b_t^k, \sigma_k = ClientUpdate(k, W_t)$
6: **end for**
7: $W_t = \frac{\sigma_k}{\sum \sigma_k} \cdot (\omega_t^k + b_t^k)$
8: **end for**
9: **Client Update: (k,** θ**)**
10: **for** each batch B in **do**
11: $\omega_t^k, b_t^k \leftarrow \theta - \beta \nabla_\theta \mathcal{L}(f(\theta))$
12: $\sigma_k \leftarrow$ assessing signal quality of client k based on noisy levels
13: **end for**

Federated Learning based Video-based Physiological Measurement. FL is a decentralized training schema where clients (i.e., smartphones) perform local training and upload trained model weights to a centralized server (e.g., the cloud). This training mechanism minimizes the risks associated with leaking identifiable or sensitive data. In the health and physiological sensing domain, federated learning has significant potential. Specifically in our scenario, FL means that facial video data and physiological gold-standard signals can remain on the mobile device and/or be processed in real-time and not transferred to any cloud storage. By only updating model parameters to the centralized server, we can learn a shared model through aggregating a large diverse population without collecting their own data.

As a baseline, we use FedAvg [23], the most commonly used federated learning algorithm. As Fig. **??** illustrates, each client uses video recordings and reference PPG signals captured by the owner of the device. These are used to train models local to each client. The model weights are then uploaded to a centralized server to execute model aggregation. FedAvg [23] uses an iterative model averaging approach to updating the model server's model's weights. This approach has been shown to be effective on image classification tasks so we start with this technique as a baseline for creating camera-based physiological measurement models in a federate manner.

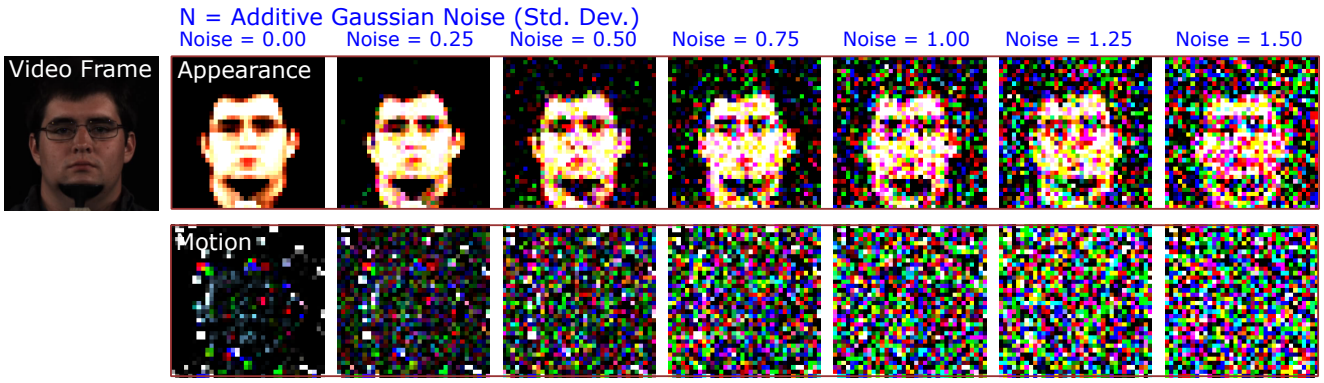

Figure 2. In our experiments we simulate camera sensor noise by adding Gaussian noise to the images. Here we illustrate the impact on the appearance and motion inputs to the two branch convolutional attention network.

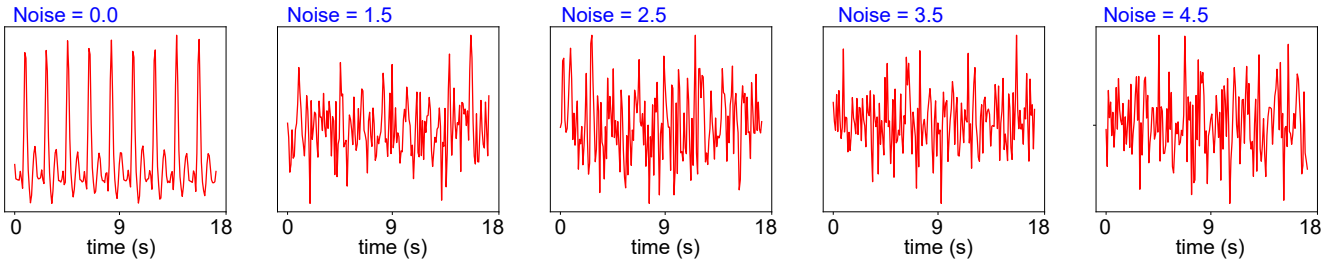

Figure 3. In our experiments we simulate contact reference PPG sensor noise by adding Gaussian noise to gold-standard contact sensor measurements. Here we illustrate the impact on training labels.

Noise Weighted Federated Learning. When training video-based physiological measurement algorithms, the goal is to recover physiological changes from very subtle (often sub-pixel) variations in image intensity. As we shall see training with FedAvg is effective if the training data from every client is "clean" (i.e., not corrupted). However, in reality it is much more likely to be the case that the quality of the training data on some individual devices will be better than others. This could be due to camera noise (e.g., quantization error) which can be most severe in poor lighting conditions when the gain is increased or user error in collecting and synchronizing the videos and reference physiological signals.

Treating the weights from every client equally is naive and does not appear to be the best way to solve optimization if the quality of the data from some devices is worse than that from others. We would prefer to have a method that promotes weights from clients with less noisy data (exploitation) while still considering weights from all clients to promote diversity (exploration). In this paper, we propose a simple but effective version of federated averaging, called FedWeight, by leveraging knowledge about the signal quality from each client. The centralized server model weight is calculated as in Equation 1 where k is the index of a layer, σ_i is the signal quality of client i, ω_i^k is the client i's model weights in the layer k , b_i^k is the bias in the client model weights in the layer k.

$$W_{server}^k = \frac{\sigma_i}{\sum \sigma_i} \cdot (\omega_i^k + b_i^k) \tag{1}$$

Our proposed signal-based aggregation is outlined in Algorithm 1. We first have an initialized centralized model weight W_0. Within each round of federated training, we randomly select a subset of clients for training. For each selected client, we then run a one-step optimization. After finishing local training for all the selected clients, we then perform signal-quality based aggregation as Equation 1 does. The output of each round in federated training is an aggregated model based on signal quality of selected clients' weights. Unlike FedAvg, which treats weights from all clients equally during model aggregation, our proposed leverages the fact that signal quality has a big impact on model performance to perform a more adaptive form of aggregation.

4. Experiment

4.1. Datasets

AFRL [11]: There is a total of 300 videos from 17 male participants and 8 female participants. The resolution of each video is 658 x 492 and the sampling rate is 120 fps. We down-sampled resolution to 36 x 36 [8] and resampled

the video to 30 fps. A fingertip reflectance medical-grade photoplethysmograms (PPG) device was provided to record ground-truth PPG signal for training the network and for evaluating the performance of our proposed system. During the data collection, every participant was asked to keep stationary for the first two tasks and perform head motion tasks in the subsequent four tasks. These motion tasks include rotating their head along the vertical axis, horizontal axis as well as orienting their head randomly to one of nine predefined locations. For the vertical and horizontal rotations, participants were asked to rotate in an angular velocity of 10 degrees/second, 20 degrees/second, 30 degrees/second, respectively. The six recording were repeated twice with two backgrounds. This data collection protocol was approved by the institutions IRB.

MMSE-HR [40]: 40 participants were recruited to join the data collection, and there is a total of 102 videos at resolution of 1040 x 1392 and sampling rate of 25 fps. The ground-truth PPG signal was recorded by a Biopac2 MP150 system[1] at 1000 fps. These size of this dataset is smaller than AFRL, but it include more spontaneous motions videos such as emotions. This data collection protocol was approved by the institutions IRB.

UBFC [2]: A total of 42 videos from 42 participants were recorded at resolution of 640 x 480 and sampling rate of 30 fps. UBFC has a similar volume as MMSE, which is also smaller than AFRL. All the videos are recorded at uncompressed 8-bit RGB format. The medical-grade pulse oximeter (CMS50E transmissive pulse oximeter) was used to record PPG signal for evaluation. All the participants were asked to keep stationary during the experiments. This data collection protocol was approved by the institutions IRB.

4.2. Implementation Details

We implemented our system in PyTorch [25], and all the experiments were conducted on an Nvidia 2080Ti GPU. We chose TS-CAN [19] as our backbone network to evaluate how FL works in remote physiological measurement since TS-CAN is the state-of-the-art neural network and can process frames in real-time on mobile platforms. To briefly summarize, TS-CAN is a two-branch neural network for on-device camera-based physiological measurement. The network contains an appearance branch that takes a sequence of normalized frames as inputs and generates attention masks to guide TS-CAN's motion branch. The motion branch takes a sequence of normalized difference frames (difference between every two consecutive frames). TS-CAN also leverages tensor shift modules to efficiently model temporal relationships which helps extract the subtle physiological signals in the videos. More details can be found in [19].

We first implemented TS-CAN with a window size of 20 frames instead of 10 frames because prior work has empirically shown a larger window size leads to better overall performance [20]. In this work, we focus on cross-dataset evaluation since the performance on cross-dataset evaluation is substantially worse than within-dataset evaluation using current state-of-the-art methods [8, 19]. We conducted all the federated training on the AFRL dataset [11] and evaluated the aggregated model on UBFC [2] and MMSE [40] datasets. For the federated training, we chose the Adam optimizer [14] with an learning rate of 0.001 on the client updates. We trained all the federated experiments for seven rounds until convergence. We followed the same training schema to replicate the traditional supervised performance of TS-CAN [19, 20].

To simulate different levels of noise in our training data (AFRL), we first sampled a subject noise level, σ_s, for each of the 25 subjects in the dataset from a Gaussian distribution with a mean equal to the experiment noise level (e.g. 0.25) and standard deviation of 0.1. During the training, to add noise to the videos we added Gaussian pixel noise from another distribution with mean of zero and standard deviation at the subject's noise level, σ_s. To add noise to the labels we added a vector of Gaussian noise from a distribution with mean of zero and standard deviation at the subject's noise level, σ_s. These noise samples were then were added to each video frames or ground-truth label vector, respectively, as the Fig. 2 and 3 illustrate. In the federated weighting process, the signal quality score was assigned to σ_s after normalizing across all subjects. As Fig. 2 and 3 show, we performed experiments adding six levels of noise to the videos [0.25, 0.50, 0.75, 1.00, 1.25, 1.50], and four levels of noise to the ground-truth labels [1.5, 2.5, 3.5, 4.5], respectively.

Since our network is trained on the derivative of the PPG signal [8]. We applied standard post-processing steps to extract the heart rate estimate: 1) calculating cumulative sum and using a detrending function [33] (λ=10) to convert the signal to the PPG waveform; 2) dividing the estimated and ground-truth values for each participant into 360-frame non-overlapping moving windows (approximately 12 seconds); 3) applying a 2nd-order Butterworth filter with a cutoff frequency of 0.75 and 2.5 Hz which represents a realistic range of heart rates for adults. Following those steps, we then computed three metrics for each window including the mean absolute error (MAE) in heart rate frequency between the predicted signal and the reference contact PPG, signal-to-noise ratio (SNR) [10] of the waveform and the Pearson correlation coefficient between the heart rate estimates and the those from the reference contact PPG. For heart rate estimation the frequency of the heart rate was determined by selecting the frequency with maximum power in the range [40Hz, 150Hz].

To explore the efficiency of end-to-end deployment in on-device training and inference, we also conducted experiments on a quad-core Cortex-A72 Raspberry Pi 4B to evalu-

[1] https://www.biopac.com/

ate the model's performance on an edge device. We trained the model and performed inference 10 times to get a reliable averaged on-device training and inference time.

5. Results & Discussion

How does FL compare to regular supervised training? The results of regular supervised training and FedAvg FL are summarized in Table 1. For the UBFC dataset, FL outperforms regular supervised training. On the other hand, regular supervised training outperforms FL on the MMSE dataset. Through this comparison, we observe that the differences are small and that there is not a consistent accuracy difference between the two. However, FL has several additional benefits compared to regular training as have been discussed. Therefore, our results point to a promising future for FL in privacy preserving camera-based cardiac measurement.

How does video and label noise impact FL? Next, we examine how the performance of FL is affected by noise in the videos and labels. Tables 2 and 3 and Fig. 4 show that the performance of the camera-based pulse measurement and heart rate estimation degrades significantly when using a naive weight averaging when some of the data is corrupted by noise. For example, in the noisy video experiments, we observed that the HR MAE increases by 19% and 20% when the noise level was increased from 0.25 to 0.5 and from 0.5 to 0.75 (UBFC dataset). However, a different pattern was found in the noisy label experiments described in Table 3. The MAE results remain similar across different noise levels, which indicates that noisy label does not significantly affect the performance of training and could be used as a regularization technique during training. Overall, the label noise had a much less severe impact on performance. In summary, simple federated averaging struggles with either noisy data or noisy labels in remote physiological measurement.

What is the impact of FedWeight? For the video noise level of 0.25, 0.5, 0.75, 1.0, 1.25 and 1.5, FedWeight improves 20%, 30%, 24%, 20%, 6% and 38% in MAE respectively, when compared to FedAvg. A similar pattern was also observed in the MMSE dataset where FedWeight leads to a reduction of errors by 5%, 15%, 17%, 18%, 13% and 11% respectively. Moreover, our proposed FedWeight achieved comparable results as FedAvg in the case of noisy labels on the UBFC dataset. FedWeight helped achieve slightly better results in the MMSE dataset, but we still argue that noisy labels don't significant affect the performance of federated training or traditional supervised training. To summarize, intelligently combining weights using a signal quality weighted averaging method leads to a considerably more robust model if the features (videos) are corrupted by noise. We believe that this result would likely by consistent for many other computer vision and machine learning tasks.

How to automate signal quality measurement? In this paper, we assume the noise level and signal quality

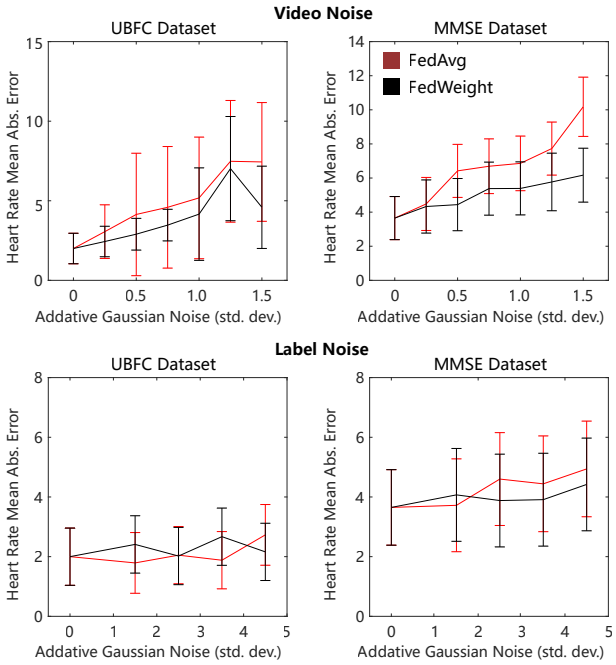

Figure 4. The heart rate mean absolute error for FedAvg and FedWeight at different video/label noise levels in the UBFC and MMSE datasets. Error bars reflect standard error where N is the number of videos.

are available to the centralized server. This could be the case if clients were able to provide a data quality report based on their knowledge of their individual sensor noise profiles. However, automating signal quality measurement would be preferred in many real-world scenarios. We are aware of this limitation and actively working on building an range of automatic signal quality metrics to test. Inspired by the metric in the task of super resolution, we argue that Peak Signal-to-Noise Ratio (PSNR) could be one way of measuring image noise level and quality. Moreover, we are also actively studying using the patterns of training loss and the quality of estimated PPG signal to assess the quality of videos.

Can we create an on-device FL prototype?

We deployed our FL system on-device as part of our experimentation. The average on-device inference time was 24.5ms per frame while the on-device training time was 105ms per frame. Based on these results, the training time is almost five time the inference time. Deploying models like our on edge devices is non-trivial. Most deep learning frameworks [1, 6, 24] focus on training on server machines, leaving inference to edge devices [13, 17]. To enable efficient federated learning on edge devices, several challenges need to be solved: the underlying framework needs to allow efficient local training on the heterogeneous device; the runtime has to be small enough to fit on to a resource-constrained device; flexible communication patterns should

Table 1. Comparison between traditional supervised training and FL with noise level of 0. Bold numbers reflect better performance.

	UBFC			MMSE		
Method	MAE↓	SNR↑	Pearson↑	MAE↓	SNR↑	Pearson↑
Supervised Training [19]	2.31	4.34	0.93	**2.99**	**2.42**	**0.79**
Federated Training	**2.00**	**4.38**	0.93	3.65	1.45	0.77

MAE = Mean Absolute Error in HR estimation, SNR = BVP Signal-to-Noise Ratio, ρ = Pearson Correlation in HR estimation.

Table 2. Comparison between FedAvg and FedWeight with different levels of video noise.

		MAE (beats/min)↓		SNR (dB)↑		Pearson↑	
Dataset	**Noise**	**FedAvg**	**FedWeight**	**FedAvg**	**FedWeight**	**FedAvg**	**FedWeight**
	0	2.00	2.00	4.38	4.38	0.93	0.93
	0.25	3.06	**2.44**	2.33	**3.53**	0.83	**0.92**
	0.50	4.14	**2.90**	1.69	**2.01**	0.76	**0.89**
UBFC	0.75	4.59	**3.47**	0.02	**2.07**	0.76	**0.87**
	1.00	5.18	**4.16**	-1.18	**0.4**	0.75	**0.81**
	1.25	7.48	**7.02**	**-2.77**	-3.22	0.66	**0.79**
	1.50	7.44	**4.59**	-2.33	**-0.03**	0.66	**0.79**
	0	3.93	3.93	2.29	2.29	0.80	0.80
	0.25	4.58	**4.33**	0.67	**0.84**	0.65	**0.67**
	0.50	5.22	**4.44**	0.07	**0.41**	0.57	**0.68**
MMSE	0.75	6.46	**5.38**	-0.51	**0.07**	0.46	**0.54**
	1.00	6.58	**5.39**	-1.12	**0.01**	0.44	**0.56**
	1.25	6.61	**5.77**	-0.90	**-0.64**	0.44	**0.53**
	1.50	6.92	**6.17**	-2.29	**-1.79**	0.43	**0.55**

MAE = Mean Absolute Error in HR estimation, SNR = BVP Signal-to-Noise Ratio, ρ = Pearson Correlation in HR estimation.

Table 3. Comparison between FedAvg and FedWeight with different levels of label noise.

		MAE (beats/min)↓		SNR (dB)↑		Pearson↑	
Dataset	**Noise**	**FedAvg**	**FedWeight**	**FedAvg**	**FedWeight**	**FedAvg**	**FedWeight**
	0	2.00	2.00	4.38	4.38	0.93	0.93
	1.5	**1.79**	2.41	**4.72**	4.70	**0.96**	0.93
UBFC	2.5	2.05	**2.02**	**4.83**	4.69	0.94	**0.96**
	3.5	**1.88**	2.67	**4.28**	3.44	**0.96**	0.93
	4.5	2.73	**2.16**	3.94	**4.97**	**0.96**	0.94
	0	3.93	3.93	2.29	2.29	0.80	0.80
	1.5	**3.72**	4.07	**0.97**	-0.27	**0.78**	0.73
MMSE	2.5	4.60	**3.88**	-0.28	**0.36**	0.73	**0.79**
	3.5	4.44	**3.91**	-0.34	**0.38**	0.74	**0.79**
	4.5	4.94	**4.42**	-0.81	**-0.78**	0.72	**0.74**

MAE = Mean Absolute Error in HR estimation, SNR = BVP Signal-to-Noise Ratio, ρ = Pearson Correlation in HR estimation.

be supported and simple to implement for different aggregation algorithms. We are actively exploring this direction based on deep learning compilation techniques [7, 30], including extending current deep learning compilers to training workload and optimize kernels for heterogeneous devices automatically.

6. Limitations

Although our proposed FedWeight improves on the performance of federated camera-based physiological measurement in the presence of noise, there are still a few limitations. First, we picked six representative video noise levels and four label noise levels. However, these noise levels do not represent the entire spectrum of real-world noise. We plan to run greedy search experiments to explore more noise levels in the future. Second, we assume the "ground-truth" noise levels are available to the centralized server during model aggregation. In the future, we plan to develop a system to automatically measure noise levels and signal quality using domain knowledge (e.g., skewness of PPG signal and PSNR in the image) in imaging and physiology as discussed in

section 5. Finally, we performance experiments on datasets that are not fully representative of all physical appearances. Before similar sensing algorithms are deployed they would require further validation and clinical evaluation.

7. Broader Impact

Ubiquitous computing offers a lot of potential for improving access to healthcare. For those that find it difficult to, or cannot, travel to a physician easily would benefit from technology that provides reliable measurement of physiological vital signs. If measurement can be performed from only a video, what happens if we detect a health condition in an individual when analyzing a video for other purposes. When and how should that information be disclosed? If the system fails in a context where a person is in a remote location, it may lead them to panic. For example, non-contact camera-based vital sensing can be used to measure a person's stress level without any notification. Especially during this pandemic, video conference meeting has become the major way to communicate between people. Non-contact physiological sensing could be easily plugged in softwares such as Zoom or Teams. Employer could easily sense their employees' health status during the meeting if we don't have the law enforcement for th is technology.

In the United States, a high standard was set by the Health Insurance Portability and Accountability Act (HIPAA) to protect sensitive patient data. We believe non-contat camera-based physiological measurement also should be under HIPPA compliance. Given the unique characteristic of camera-based physiological measurement, it even includes more sensitive information (e.g., long facial videos) than many other healthcare technology. We argue that a special protection of data transferring should be enforced to minimizing the risk of data leaking. A better way to do this is to store and run inference on local mobile devices. However, how to collect large-scale physiological and video data to train a "super" model still remains challenge due to the concerns of data leaking and management. In this paper, we have successfully demonstrated how federated learning interplays with non-contact physiological sensing. Even without uploading a single raw video or physiological data to centralized server, it is still possible to attain a "super" aggregated model for everyone to use.

8. Conclusion

In this paper, We present a federated learning system called FedWeight that accounts for training imperfect data such as noisy data or noisy labels. We apply this to the task of camera-based remote physiological measurement. Our results show that traditional federated weight averaging degrades quickly if the data on some of the clients is corrupted by noise, our proposed method is more robust to corruption particularly video noise. Federated learning has many attractive properties for camera-based health monitoring where it not only protect sensitive information but also provides a way to aggregate large scale clients to train a robust model. We envision federated learning and FedWeight will have a big potential in various applications in mobile health, especially in remote physiological measurement.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

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

[4] Eoin Brophy, Maarten De Vos, Geraldine Boylan, and Tomas Ward. Estimation of continuous blood pressure from ppg via a federated learning approach. *arXiv preprint arXiv:2102.12245*, 2021.

[5] Pak-Hei Chan, Chun-Ka Wong, Yukkee C Poh, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Ming-Zher Poh, Daniel Wai-Sing Chu, and Chung-Wah Siu. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association*, 5(7):e003428, 2016.

[6] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems, 2015.

[7] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, et al. {TVM}: An automated end-to-end optimizing compiler for deep learning. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pages 578–594, 2018.

[8] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

[9] Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

[10] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[11] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1462–1469. IEEE, 2014.

[12] Javier Hernandez, Daniel J McDuff, and Rosalind W Picard. Bioinsights: Extracting personal data from "still" wearable motion sensors. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6. IEEE, 2015.

[13] Ziheng Jiang, Tianqi Chen, and Mu Li. Efficient deep learning inference on edge devices. *ACM SysML*, 2018.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[16] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[17] Juhyun Lee, Nikolay Chirkov, Ekaterina Ignasheva, Yury Pisarchyk, Mogan Shieh, Fabio Riccardi, Raman Sarokin, Andrei Kulik, and Matthias Grundmann. On-device neural net inference with mobile gpus, 2019.

[18] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014.

[19] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020.

[20] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. Metaphys: Few-shot adaptation for non-contact physiological measurement. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 154–163, 2021.

[21] Daniel McDuff. Camera measurement of physiological vital signs. *arXiv preprint arXiv:2111.11547*, 2021.

[22] Daniel J McDuff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 63–70. IEEE, 2017.

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

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[26] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[27] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010.

[28] 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. *arXiv preprint arXiv:2101.07511*, 2021.

[29] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

[30] Nadav Rotem, Jordan Fix, Saleem Abdulrasool, Garret Catron, Summer Deng, Roman Dzhabarov, Nick Gibson, James Hegeman, Meghan Lele, Roman Levenstein, Jack Montgomery, Bert Maher, Satish Nadathur, Jakob Olesen, Jongsoo Park, Artem Rakhov, Misha Smelyanskiy, and Man Wang. Glow: Graph lowering compiler techniques for neural networks, 2019.

[31] Chihiro Takano and Yuji Ohta. Heart rate measurement based on a time-lapse image. *Medical engineering & physics*, 29(8):853–857, 2007.

[32] L Tarassenko, M Villarroel, A Guazzi, J Jorge, DA Clifton, and C Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014.

[33] Mika P Tarvainen, Perttu O Ranta-Aho, and Pasi A Karjalainen. An advanced detrending method with application to hrv analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002.

[34] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under

realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016.

[35] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[36] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[37] Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, et al. Pathological findings of covid-19 associated with acute respiratory distress syndrome. *The Lancet respiratory medicine*, 8(4):420–422, 2020.

[38] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[39] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *Proc. BMVC*, pages 1–12, 2019.

[40] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.

[41] Ying-Ying Zheng, Yi-Tong Ma, Jin-Ying Zhang, and Xiang Xie. Covid-19 and the cardiovascular system. *Nature Reviews Cardiology*, 17(5):259–260, 2020.