ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446004

IN-THE-WILD PHYSIOLOGICAL-BASED STRESS DETECTION USING FEDERATED STRATEGY

Po-Chen Lin, Jeng-Lin Li, Woan-Shiuan Chien, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Continuously identifying day-to-day mental stress can be realized by accessing wearable devices to measure physiological indicators. However, the nature of bodily signals raises issues of privacy and data heterogeneity. Recent federated learning scheme provides a promising direction to alleviate the privacy concern, but the large inter-client differences can lead to a sub-optimal model performance. In this work, we propose a client-aware aggregation strategy to customize the global model forked by each client to conduct mutual learning in federated setting. Our proposed mixture Federated Mutual Learning (*mixFML*) weighs the distances of local models to generate a unique mixture of global model per client. We evaluated our method on the public TILES-2018 and an in-house Firefighters dataset for stress detection using HRV. Our proposed *mixFML* achieved 8.0% and 1.8% MCC improvement on two datasets compared to federated mutual learning.

Index Terms— federated learning, stress detection, heart rate variability, mutual learning

1. INTRODUCTION

The pervasiveness of wearable devices and advancement of biometric sensors technologies have enabled the next generation smart health systems. These technologies are key in providing in-time and unobtrusive physiological state monitoring. Stress is a critical risk factor for diverse diseases and can lead to degraded working performances. Automatic stress detection using bodily signals is an emerging frontier (e.g., [1, 2]). However, transferring physio-data from each user's wearable sensor for centralized model training is being heavily scrutinized with the uprising awareness on user privacy. Additional complexity emerges as variability from devices and idiosyncratic nature of an individual adds to the heterogeneity nature of these bodily signals [3]. For example, while it is known that physio-indicators, such as electrodermal activity and heart rate variability (HRV), reflect stress state, there exists a large inter-subject variability due to geometrical and physiological factors [4]. Hence, given the proliferation of wearable devices, privacy concerns and signal heterogeneity are two major hurdles in realizing a real-world stress detection model for *in-the-wild* setting.

Recently, research effort in publicly releasing TILES-2018 [5] dataset uniquely provides a large scale, longitudinal, and highly close-to-life data to help advance physio-signal modeling and stress detection in real workplaces. Recent studies on TILES-2018 have largely focused on sophisticated physio-feature designs to improve detection accuracy [6, 7]. However, their centralized training approaches will likely hinder the real-world applicability due to privacy concern. The emergence of Federated learning (FL) provides a modified learning paradigm that operates by aggregating each client's local model to collectively learn a global model. This prevents transferring of data and eliminates data privacy issue when deploying in everyday life [8]. However, a known issue of FL is its inability to handle heterogeneous clients.

Several studies have extended the initial FL approach using personalized strategies to deal with heterogeneous clients. For example, given prior knowledge of client characteristics, performing client selection can help FL to converge to a better global model [9, 10]. Another line of research handles heterogeneity by devising strategies to train client model *locally*. Research in this direction is inspired by knowledge transfer techniques, e.g., knowledge distillation [11] aiming to transfer knowledge between teacher-student models in an ensemble manner resembles the server-client FL setup. Specifically, Federated Mutual Learning (*FML*) [12] generates well-trained local client models by mutually learning with a global server model. While *FML* is emerged as a current SOTA for local model training in FL setting, a single global model is assumed to facilitate each local client training. We argue that this is not optimal for handling the client heterogeneity especially for bodily signals collected in the real-world setting.

In this work, we propose a mixture FML (*mixFML*) to relax the single global model constraint using a distance weighting mechanism to generate *customized* mutual model for each client. We evaluate our approach for the task of stress detection using HRV features on two datasets, the public TILES-2018 and an in-house Firefighters dataset. We compare the performance with other federated methods and obtain a 8% and 1.8% MCC improvement on the two datasets respectively compared to the SOTA *FML* method. The rest is organized as follows: section 2 includes database description and our proposed method; section 3 details our experimental setup and results; finally, we conclude in section 4.

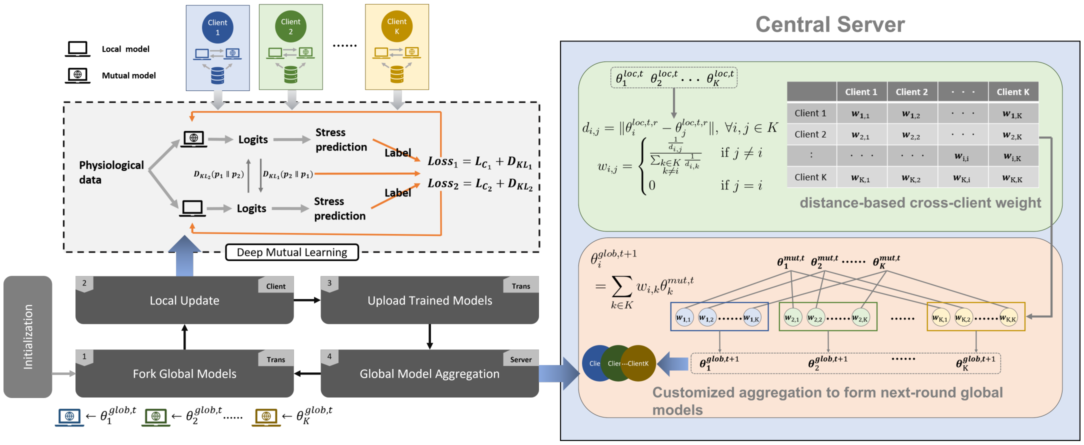
Local model
Mutual model
Client 1
Client 2
......
Client K
Central Server
Physiological data
Logits
Stress prediction
Label
$Loss_1 = L_{C_1} + D_{KL_1}$
$Loss_2 = L_{C_2} + D_{KL_2}$
$D_{KL_2}(p_1 \| p_2)$
$D_{KL_1}(p_2 \| p_1)$
Deep Mutual Learning
Initialization
1 Fork Global Models Trans
2 Local Update Client
3 Upload Trained Models Trans
4 Global Model Aggregation Server
$\leftarrow \theta_1^{glob,t}$ $\leftarrow \theta_2^{glob,t}$ ...... $\leftarrow \theta_K^{glob,t}$
$\theta_1^{loc,t}\ \theta_2^{loc,t} \cdots \theta_K^{loc,t}$
$d_{i,j} = \|\theta_i^{loc,t,r} - \theta_j^{loc,t,r}\|, \forall i,j \in K$
$w_{i,j} = \begin{cases} \frac{\frac{1}{d_{i,j}}}{\sum_{k \in K, k \neq i} \frac{1}{d_{i,k}}} & \text{if } j \neq i \\ 0 & \text{if } j = i \end{cases}$
Client 1 Client 2 · · · Client K
distance-based cross-client weight
$\theta_i^{glob,t+1} = \sum_{k \in K} w_{i,k} \theta_k^{mut,t}$
$\theta_1^{mut,t}\ \theta_2^{mut,t}$ ...... $\theta_K^{mut,t}$
$\theta_1^{glob,t+1}$ $\theta_2^{glob,t+1}$ ...... $\theta_K^{glob,t+1}$
Customized aggregation to form next-round global models

Fig. 1: *mixFML*: 1) Server and clients initialize global and local models. 2) Each client forks global model from server as mutual model and conduct DML using mutual and local model during local update. 3) Mutual and local model are uploaded to server to calculate distance-based weights. 4) Server aggregates customized next-round global models for each client.

Table 1: An overview of HRV features extracted from [6, 14].

Feature Group	Features
Time(8)	meanNN, sdNN, coefficient of variation, mean of 1^{st} diff., RMSDD, standard deviation of absolute of 1^{st} diff., pNN50, normalized mean of absolute 1^{st} diff.
Frequency(6)	High frequency power(HF), Low frequency power(LF), HF/LF, normalized HF, normalized LF, Very low frequency power(VLF)
Multi-scale(47)	**For *RR* and *dRR* calculate:** *MSPE*, *MSmPE*, PE_{dw}, *d(s1,:)*, *d(s2,:)*, 1^{st} diff. of *d(s1,:)* and *d(s2,:)*, sum of *d(s1,:)* and *d(s2,:)* **Additional**: total asymmetry index

2. RESEARCH METHODOLOGY

2.1. Dataset

TILES-2018 database consists of *in-the-wild* physiological measurements and survey responses from 212 real hospital workers, which were collected longitudinally over ten weeks while participants carried out their daily work/life as usual. These participants were equipped with wearable devices to track daily activities and physiological measurements; various dimensions of their well-being over time were assessed using self-reported surveys. In this work, we utilize the RR time series from OMSignal smart-shirt as input and self-rated stress survey (1-5) as labels. The dataset has 58.2% stress-labeled data after binarization with its global mean ($\mu = 1.8$), following the same setup in [13]. Since the amount of available data varies among participants, we select 25 balanced participants, each having around 40 samples for this study.

To expand on similar research in high-stress workplaces, we replicate the TILES-2018 by collecting data from real firefighters over ten weeks. Heart rate is measured using Fitbit's photoplethysmography, and stress responses are collected via the same survey. Label binarization uses the TILES-2018 threshold of 1.8, resulting in 67.3% stress-labeled data. This database comprises 23 participants, each contributing 30 to 40 samples.

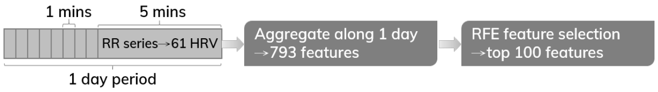
1 mins
5 mins
RR series→61 HRV
1 day period
Aggregate along 1 day →793 features
RFE feature selection →top 100 features

Fig. 2: The process of HRV feature extraction from RR series.

2.2. Heart Rate Variability (HRV) Feature Extraction

OMSignal and Fitbit provide heart rate records every 1 and 5 seconds respectively. We then calculate the RR intervals to form RR time series based on recorded heart rate. Heart Rate Variability (HRV) features are known to correlate with stress level in the literature [15]. An overview of the HRV feature extraction process is shown in Fig. 2.

Similar to prior works on TILES-2018 [6, 14], we extract 14 conventional time/frequency domain HRV features (using open toolbox[1]) and 47 advanced multi-scale HRV features (listed in Table 1) from RR series in every 5-minute window. These features are then aggregated at a day-level using 13 statistical functions, and Recursive Feature Elimination (RFE) is applied to select top 100 features. Specifics about the extraction process can be found in previous research [14].

2.3. Federated Learning (FL) Strategy

2.3.1. Problem Formulation

Suppose we have K clients with index set $\mathbb{K}$ and a central server in the federated system, each client $k \in \mathbb{K}$ keeps its own dataset $D_k := \{X_n^k, y_n^k\}_{n=1}^{N_k}$ of N_k samples. Our objective is to find an optimal local model $f_k(\theta_k^{loc})$ of parameters θ_k^{loc} for each client to solve local tasks (stress detection) using only local data (HRV features) and share information without any data transfer (federated learning setting).

[1] https://aura-healthcare.github.io/hrv-analysis/

2.3.2. *Federated Mutual Learning (FML)*

Our proposed FL strategy is primarily motivated by Federated Mutual Learning (*FML*) [12], where global model and each client's local models transfer knowledge by introducing KL divergence as an additional loss calculated on each other's logits output. That is, during round t in a *FML* training, each client k forks global model from central server as its mutual model: $\theta_k^{mut,t} \leftarrow \theta^{glob,t}$, and conduct deep mutual learning (DML) with its local model using client-side private data. The trained mutual models after r rounds of local epochs will be uploaded to the server. Then, the next-round of new global model will be formed by equal-weighted aggregation from the each client's mutual model: $\theta^{glob,t+1} = \frac{1}{K}\sum_{k \in K} \theta_k^{mut,t,r}$. The loss function of DML can be formulated as the following:

$$\begin{aligned} L_{loc} &= \alpha L_{C_{loc}} + (1-\alpha) D_{KL}(p_{mut} \parallel p_{loc}) \\ L_{mut} &= \beta L_{C_{mut}} + (1-\beta) D_{KL}(p_{loc} \parallel p_{mut}) \end{aligned} \tag{1}$$

where loc is the private local model, mut is the mutual model, L_C is the classification loss function (cross entropy loss in this work) and D_{KL} is the KL divergence between two prediction logits p_{loc} and p_{mut}. α and β serve as the proportion of knowledge transferred between the two models.

2.3.3. *Mixture of FML (mixFML)*

While the original *FML* has shown outstanding capability when handling heterogeneous clients via collectively learning of both the global and local models in FL setting, using a single global model, i.e., derived by equal contribution from all clients, as the mutual model to perform DML across all clients limits each client's local model's capacity. For example, mutual learning relies on proper transfer of knowledge from "relevant clients". As the clients number increases (as often seen in real-world applications), relevant information might be diluted due to the use of simple averaging, and it can negatively impact the target local model's performance, especially when large heterogeneity (e.g., *in-the-wild* learning) is expected. To deal with this issue, we propose a mixture Federated Mutual Learning (*mixFML*) that enables the system to customize the mutual model for each client. Each client additionally push its local model parameters to the server to represent its local characteristics, and the server generates a *uniquely customized* mutual model by means of weighted aggregation. The weights are computed based on the distances between the local models that are formulated as:

$$\begin{aligned} d_{i,j} &= \|\theta_i^{loc,t,r} - \theta_j^{loc,t,r}\|, \\ w_{i,j} &= \frac{1/d_{i,j}}{\sum_{\substack{k \in \mathbb{K} \\ k \neq i}} 1/d_{i,k}}, \text{ if } j \neq i \text{ else } 0, \ \forall i,j \in \mathbb{K} \end{aligned} \tag{2}$$

where $d_{i,j}$ is the parameter-based euclidean distance between client i, j, $\theta_i^{loc,t,r}$ is the trained local model parameter at round t after r local epochs, and $w_{i,j}$ is the weight for client j when generating the mixture weight for client i. The customized global model $\theta_i^{glob,t+1}$ for client i are derived by the following aggregation formula:

$$\theta_i^{glob,t+1} = \sum\nolimits_{k \in K} w_{i,k} \theta_k^{mut,t} \tag{3}$$

The mixture excludes client i since its own traits are captured in the local model already. The larger $w_{i,j}$ implies that client i and j are relatively similar (as measured in terms of distances of the parameter space).

2.4. Evaluation Metrics

To simulate a real-world use case, we split each participant's data as training, evaluation and testing sets in a 70/10/20 ratio. The performance is reported after 5-fold cross validation. As for evaluation metrics, we use balanced accuracy (BACC), F1, and Matthews correlation coefficient (MCC). MCC has recently been argued as a more reliable statistical measure that would produce a high score only if a binary classifier obtain good results in all 4 quadrants of a confusion matrix [16].

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

In our experiment, we use two real-world datasets to evaluate our proposed *mixFML* on stress detection using HRV when comparing with other federated methods.

- *Exp1*: Performance comparison among FL methods for stress detection. In TILES-2018 dataset, we randomly assign 1 participant to each client for federated model training, where the number of clients is set to the range of [5/15/25], and the same process repeats on Firefighters with the range of [5/15/23].
- *Exp2*: We further examine the influence of data quantity by setting the availability of training data quantity from 20% to 80% of total data in both datasets and compare the performance among different FL methods.

The compared FL alternatives in both experiments include conventional *FedAvg* [17] that is the most basic form of FL; *FedProx* [18] is proposed to handle heterogeneity (the proximal term μ set to 0.01). The above two alternatives involve FL scheme in deriving a global model used for every client. *Local* model indicates every client has its private model trained on its own data solely. *FML* is a SOTA learning scheme that adopts mutual learning to train every local private model to be used for each client. Our proposed model is termed as *mixFML*. Both *FML* and our proposed *mixFML* have α and β set to 0.5. For all FL models, a MLP network with two hidden layers of size 64 and 16 followed by ReLU activation layer and a softmax for the output layer is used for global and local models; dropout layer and L2 parameter are applied to avoid overfitting. The federated models are trained for 100 rounds with 5 local epochs and learning rate set to 0.001.

Method	**TILES-2018**									**Firefighters**								
	5 Clients			**15 Clients**			**25 Clients**			**5 Clients**			**15 Clients**			**23 Clients**		
	BACC	**F1**	**MCC**	**BACC**	**F1**	**MCC**	**BACC**	**F1**	**MCC**	**BACC**	**F1**	**MCC**	**BACC**	**F1**	**MCC**	**BACC**	**F1**	**MCC**
FedAvg	0.518	0.654	0.038	0.536	0.621	0.073	0.601	0.622	0.201	0.521	0.801	0.048	0.481	0.839	0.065	0.593	0.780	0.211
FedProx	0.509	0.555	0.018	0.531	0.635	0.065	0.626	0.639	0.250	0.561	0.794	0.124	0.543	0.801	0.089	0.597	0.776	0.214
Local	0.558	0.648	0.115	0.617	0.654	0.232	0.639	0.673	0.278	0.730	0.864	0.449	0.740	0.870	0.463	0.761	0.835	0.514
FML	0.564	**0.661**	0.127	0.620	0.669	0.239	0.632	0.667	0.264	**0.742**	0.858	0.458	0.740	0.878	0.474	0.754	0.830	0.537
mixFML	**0.579**	0.653	**0.156**	**0.649**	**0.712**	**0.301**	**0.672**	**0.697**	**0.344**	0.733	**0.878**	**0.473**	**0.748**	**0.883**	**0.492**	**0.783**	**0.847**	**0.555**

Table 2: Performance comparison among FL variants for personal stress detection on TILES-2018 and Firefighters.

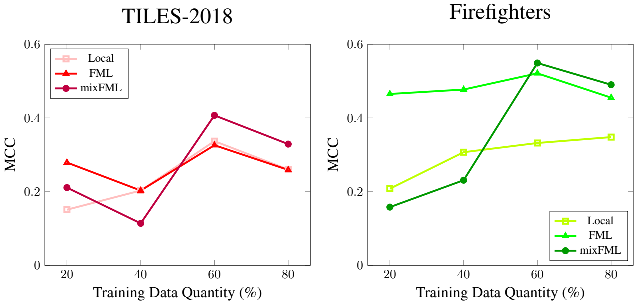

Fig. 3: MCC scores among personalized FL methods as training data quantity increases on TILES-2018 and Firefighters.

3.2. Results and Analysis

3.2.1. Performance Comparison

The left side of Table 2 summarizes the performance comparison among various federated learning variants for stress detection on the TILES-2018 dataset. Our proposed model *mixFML* consistently outperforms the other methods across multiple evaluation metrics considering different number of clients enrolled in the FL setting. Specifically, *mixFML* achieves the highest BACC (1.5%, 2.9%, 4.0% improvements against *FML*) and MCC (2.9%, 6.2%, 8.0% improvements against *FML*) values for all client configurations. Comparatively, those using a single global model (*FedAvg* and *FedProx*) exhibit low performances, with *FedProx* slightly surpassing *FedAvg* in terms of F1 scores and MCC, particularly as the number of clients increases. This is likely due to the high heterogeneity observed for the *in-the-wild* setting where a single global model can not well capture the nuances of the diverse participants. In fact, the use of *Local* model for each client demonstrates that even without mutual learning, the client-specific model outperforms the single global model, but still falls short of the performance achieved by *mixFML*.

We again verify the proposed *mixFML* on the in-house Firefighters dataset. The experimental results are shown in the right side of Table 2. *MixFML* continues to be the best performing model over all other FL methods for all client configurations, achieving 1.5%, 1.8%, 1.8% MCC improvements against *FML*. An obvious gap exists among those with/without personalized local models, indicating the critical need of client-specific personalization for stress detection using HRV *in-the-wild*. In both experiments, *mixFML* method emerges as a consistently strong performer for stress detection, the larger the number of clients (i.e., the larger the client heterogeneity), its capability is further demonstrated. The ability to maintain high BACC, F1, and MCC values suggests its adaptability and reliability in real-world scenarios.

3.2.2. Influence of Data Quantity

We further systematically evaluate the performances as a function of data quantity. We increase the training data quantity from 20% to 80% of all data, leaving the 10% for evaluation and 10% for testing. We experiment with the three personalized federated learning alternatives (*FML*, *mixFML*, and *Local* for comparison) using the highest number of clients (25, 23) in both datasets. Fig 3 summarizes the experimental results. In general, all methods intuitively experience gradual improvements as more training data become available, except 80% setting. In both datasets, *FML* showcases its stability of maintaining competitive performances and consistently outperforms *Local* in Firefighters. Our proposed *mixFML* shows the best performances particularly when the amount of data quantity reaches between 40% to 60%, that is 15 to 20 samples (about 2-3 weeks) given a client has about 35 to 40 samples (about 5-6 weeks of data). We observe under limited data scenario, our method can be sub-optimal. It is likely due to the fact there is not enough data yet to well-train the local models that can be used in customizing the unique client-specific global model for mutual learning.

4. CONCLUSION

In this work, we propose a novel distance-based mixture weights to tailor the global model used in mutual learning for enhanced private local model learning in a federated scheme. This method is evaluated on stress detection using HRV in two *in-the-wild* datasets. Our proposed approach outperforms other FL alternatives, demonstrating its capability in handling heterogeneous clients better. We further shows its capability of maintaining stably high performance when dealing with an increasing number of clients. We further examine the effective data quantity required to utilize this FL strategy. In our future work, we would continue to explore the relationship across clients and investigate the effect of composition in deriving the customized global model, such as reducing the number of clients in forming the global model, which could also lead to a more efficient learning scheme operable on much more clients with less demand on data quantity.

5. REFERENCES

[1] Muhammad Majid, Aamir Arsalan, and Syed Muhammad Anwar, "A multimodal perceived stress classification framework using wearable physiological sensors," *arXiv preprint arXiv:2206.10846*, 2022.

[2] Juan Antonio Castro-García, Alberto Jesús Molina-Cantero, Isabel María Gómez-González, Sergio Lafuente-Arroyo, and Manuel Merino-Monge, "Towards human stress and activity recognition: A review and a first approach based on low-cost wearables," *Electronics*, vol. 11, no. 1, pp. 155, 2022.

[3] Ahmed M Abdelmoniem, Chen-Yu Ho, Pantelis Papageorgiou, and Marco Canini, "Empirical analysis of federated learning in heterogeneous environments," in *Proceedings of the 2nd European Workshop on Machine Learning and Systems*, 2022, pp. 1–9.

[4] Rudi Hoekema, Gérard JH Uijen, and Adriaan Van Oosterom, "Geometrical aspects of the interindividual variability of multilead ecg recordings," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 5, pp. 551–559, 2001.

[5] Karel Mundnich, Brandon M Booth, Michelle l'Hommedieu, Tiantian Feng, Benjamin Girault, Justin L'hommedieu, Mackenzie Wildman, Sophia Skaaden, Amrutha Nadarajan, Jennifer L Villatte, et al., "Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, vol. 7, no. 1, pp. 1–26, 2020.

[6] Abhishek Tiwari, Shrikanth Narayanan, and Tiago H Falk, "Stress and anxiety measurement" in-the-wild" using quality-aware multi-scale hrv features," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 7056–7059.

[7] Arthur Pimentel, Abhishek Tiwari, Shrikanth Narayanan, and Tiago H Falk, "Human mental state monitoring in the wild: Are we better off with deeperneural networks or improved input features?," *CMBES Proceedings*, vol. 44, 2021.

[8] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[9] Minxue Tang, Xuefei Ning, Yitu Wang, Jingwei Sun, Yu Wang, Hai Li, and Yiran Chen, "Fedcor: Correlation-based active client selection strategy for heterogeneous federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10102–10111.

[10] Zheng Chai, Ahsan Ali, Syed Zawad, Stacey Truex, Ali Anwar, Nathalie Baracaldo, Yi Zhou, Heiko Ludwig, Feng Yan, and Yue Cheng, "Tifl: A tier-based federated learning system," in *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, 2020, pp. 125–136.

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[12] Tao Shen, Jie Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Kun Kuang, Fei Wu, and Chao Wu, "Federated mutual learning," *arXiv preprint arXiv:2006.16765*, 2020.

[13] Abhishek Tiwari and Tiago H Falk, "New measures of heart rate variability based on subband tachogram complexity and spectral characteristics for improved stress and anxiety monitoring in highly ecological settings," *Frontiers in Signal Processing*, vol. 1, pp. 737881, 2021.

[14] Abhishek Tiwari, Isabela Albuquerque, Mark Parent, Jean-François Gagnon, Daniel Lafond, Sébastien Tremblay, and Tiago H. Falk, "Multi-scale heart beat entropy measures for mental workload assessment of ambulant users," *Entropy*, vol. 21, no. 8, pp. 783, 2019.

[15] Rossana Castaldo, Paolo Melillo, Umberto Bracale, M Caserta, Maria Triassi, and Leandro Pecchia, "Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[16] Davide Chicco and Giuseppe Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[17] 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*. PMLR, 2017, pp. 1273–1282.

[18] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.