Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Federated model with contrastive learning and adaptive control variates for human activity recognition*#

Ignatius IWAN, Bernardo Nugroho YAHYA‡, Seok-Lyong LEE

Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, Yongin 17035, Republic of Korea

E-mail: ignatiusiwan@hufs.ac.kr; bernardo@hufs.ac.kr; sllee@hufs.ac.kr

Received Sept. 13, 2024; Revision accepted Jan. 8, 2025; Crosschecked May 22, 2025

Abstract: Recent attention to privacy issues demands a communication-safe method for training human activity recognition (HAR) models on client activity data. Federated learning (FL) has become a compelling technique to facilitate model training between the server and clients while preserving data privacy. However, classical FL methods often assume independent and identically distributed (IID) data among clients. This assumption does not hold true in practical scenarios. Human activity in real-world scenarios varies, resulting in skewness where identical activities are executed uniquely across clients. This leads to local model objectives drifting away from the global model objective, thereby impeding overall convergence. To address this challenge, we propose FedCoad, a novel federated model leveraging contrastive learning with adaptive control variates to handle the skewness among HAR clients. Model contrastive learning minimizes the gap in representation between global and local models to help global model convergence. During local model updates, the adaptive control variates penalize the local model updates with respect to the model weight and the rate of change from the control variates update. Our experiments show that FedCoad outperforms state-of-the-art FL algorithms on HAR benchmark datasets.

Key words: Federated learning (FL); Human activity recognition (HAR); Contrastive learning; Deep learning
https://doi.org/10.1631/FITEE.2400797 **CLC number:** TP391

1 Introduction

Sensor-based human activity recognition (HAR) using wearable devices is critical in user-centered applications like healthcare (Wu et al., 2022), smart environments (Bianchi et al., 2019), and fall detection (Mrozek et al., 2020). Federated learning (FL) (McMahan et al., 2017) offers a decentralized and privacy-preserving solution for training models across devices distributed among clients. The FL method works well when the client data are independent and identically distributed (IID) as most of the clients have similar quantities, labels, and features of data. However, humans perform various activities and even have unique patterns. Therefore, real-world HAR data are inherently non-IID since they exhibit some skewness that can degrade model performance.

There are three types of HAR skewness in real-world conditions (Presotto et al., 2022): feature distribution skew, where two clients perform the same activity with different patterns (e.g., walking patterns differ between younger and older individuals); quantity distribution skew, where clients have imbalanced labeled data volumes; label distribution skew, where two clients have different sets of labels. For example, an athlete has more activity labels related to sports

‡ Corresponding author
* Project supported by the Hankuk University of Foreign Studies Research Fund (2025-01) and the International Cooperative R&D Program of Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) (No. P0022316)
# Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2400797) contains supplementary materials, which are available to authorized users
ORCID: Ignatius IWAN, https://orcid.org/0009-0002-4447-0058; Bernardo Nugroho YAHYA, https://orcid.org/0000-0002-7121-2436; Seok-Lyong LEE, https://orcid.org/0000-0002-8630-5395

than an office worker does. These conditions lead to divergent local objectives during training and hinder global model optimization (Karimireddy et al., 2020; Li X et al., 2020).

Existing methods have sought to address these non-IID challenges. Some methods focus on the model aggregation stage to handle non-IID issues, such as implementing momentum for model aggregation normalization (Hsu et al., 2019) or clustering clients with similar data distributions for grouped updates (Presotto et al., 2022; Guo JL et al., 2024). However, normalization during model aggregation may fail, as the local updates are already skewed and clustering operations require significant computational resources, especially as the number of clients grows. Other methods focus on the local training stage, such as FedProx, which applies L2-norm regularization to constrain local updates (Li T et al., 2020), and SCAFFOLD, which uses control variates to reduce update variance (Karimireddy et al., 2020). While effective in some cases, these methods risk local model overfitting due to the limited diversity of client data in HAR.

Fig. 1 illustrates the problem where two heterogeneous clients are trained together in an FL manner compared to a centralized training update. Unlike centralized training, which uses a comprehensive dataset to achieve the global optimum, FL relies on locally skewed datasets. Local models trained on heterogeneous data diverge from the global objective, and the aggregated global model often fails to generalize. In FL, local clients learn only from their own dataset and overlook other clients' underlying data distributions, leading to updates that are misaligned with the true global optima.

An additional approach for mitigating the local model overfitting issue is to provide an auxiliary perspective, such as supplementary data for generalization or convergence direction correction, which helps prevent local model overfitting. Recently, contrastive learning has emerged as a promising solution to enhance generalization by encouraging models to align similar representations. It can be used in unsupervised or supervised settings. Unsupervised approaches (Guo CR et al., 2024) leverage unlabeled data, while supervised methods (Li CL et al., 2021) teach the model to create different representations for data with distinct labels. However, both types are affected by the amount of available data that can be stored in wearable devices, and each client's data reflect only individual patterns. Another approach, MOON (Li QB et al., 2021), introduces model contrastive learning that uses the global model as an external resource for the local training guide to prevent local model overfitting. Despite its advantages, MOON focuses solely on the training phase and is susceptible to misdirection by the biased global model.

To investigate the assumption above, we conducted experiments using the MotionSense dataset by training a one-dimensional convolutional neural network (1D CNN). T-distributed stochastic neighbor embedding (T-SNE) was used to visualize the hidden vectors. Fig. 2a shows that a centralized model achieves well-separated representations, capturing diverse activity classes. However, unlike the result of Li QB et al. (2021), the global model in Fig. 2b reveals overlapping representations in the global model trained with FL, caused by the local model overfitting to skewed datasets (Fig. 2c). Therefore, solely relying on global model aggregation to correct local client updates may fail to guide local model training.

To address these limitations, we propose FedCoad, a method that seamlessly unifies model contrastive learning with dynamic control variates to enhance local model training stability and achieve better global

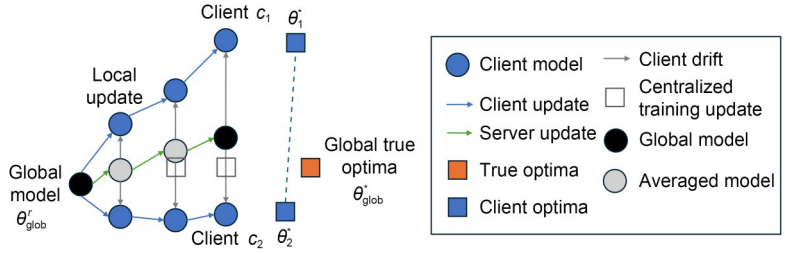

Fig. 1 Effect of client update variance on the global model to reach true global optima. References to color refer to the online version of this figure

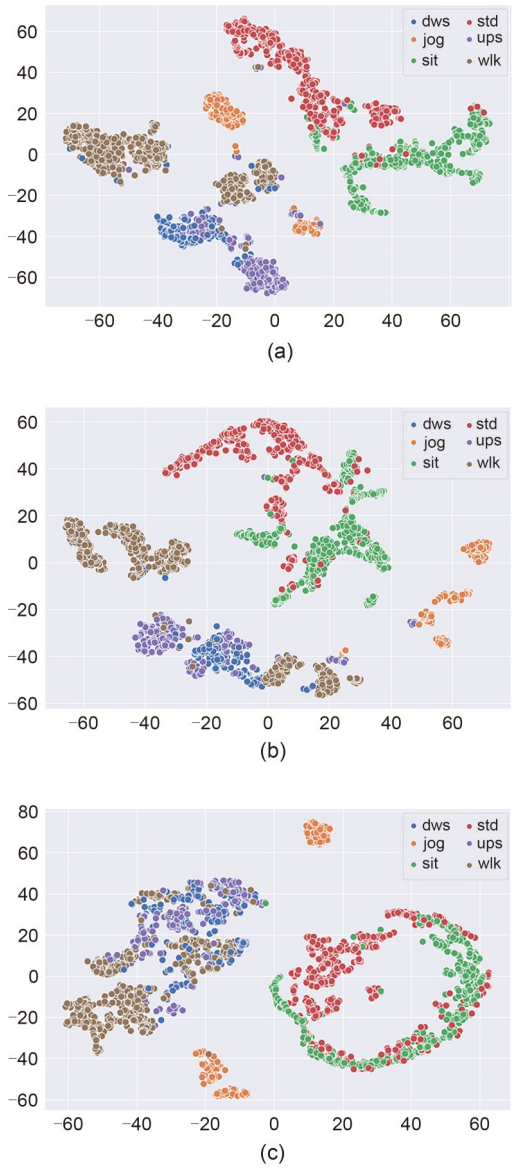

Fig. 2 T-SNE visualization in the MotionSense dataset for the centralized model (a), FedAvg (b), and local model (c). dws: walking downstairs; jog: jogging; sit: sitting; std: standing; ups: walking upstairs; wlk: walking. References to color refer to the online version of this figure

model convergence. Unlike existing methods, which address local model divergence in either the training phase (e.g., MOON and FedProx) or the parameter update phase (e.g., SCAFFOLD), FedCoad addresses both simultaneously. During the training phase, model contrastive learning aligns local and global model representations, mitigating discrepancies while maintaining local objectives. In the parameter update phase, adaptive control variates penalize deviations from the global update trajectory, preventing overfitting and ensuring convergence. By addressing the divergence in both the training and parameter update phases, FedCoad ensures that local models avoid overfitting and achieve superior performance in non-IID environments. To summarize, the contributions of this work are as follows:

1. We propose a new FL method that maximizes the synergy of model contrastive learning and control variates to prevent local model divergence in a non-IID environment.

2. The proposed FedCoad handles the local model divergence in both the training and parameter update phases.

3. Our proposed method can significantly reduce the communication exchange cost compared to the state-of-the-art methods for HAR.

4. We perform extensive experiments and ablation studies in popular HAR benchmark datasets and demonstrate the superiority of the proposed FedCoad.

2 Related studies

In this section, we give an overview of popular FL algorithms, as shown in Table 1. FedAvg (McMahan et al., 2017) was the first and is the most popular FL algorithm for training a global model with multiple clients. The popularity of FedAvg has led to the adoption of FL technology within the realm of HAR research, as shown by recent studies (Sozinov et al., 2018), which have started to explore the applicability of FL to HAR for distributed training. However, one of the problems in FL for HAR is that each client's data are different and follow non-IID patterns. A recent study (Li X et al., 2020) found that the non-IID condition affects FedAvg and decreases its performance because each local model is trained on its own skewed dataset. Subsequently, some studies attempted to improve FedAvg by handling non-IID conditions with improvement categorizations such as aggregation (Hsu et al., 2019; Wang et al., 2020) and local training (Karimireddy et al., 2020; Li T et al., 2020; Li QB et al., 2021).

The work that handles the global model aggregation step involves the server normalizing the aggregation updates during the global model update. FedMA (Wang et al., 2020) matches and averages weights in

Table 1 List of state-of-the-art FL algorithms and current limitations

Reference	Methodology	Limitation
McMahan et al., 2017	FedAvg	Performance decreases due to the clients' skewness in non-IID environments
Li T et al., 2020	FedProx	The penalty from regularization may affect the local model convergence
Hsu et al., 2019	FedAvgM	Encountered issues when local model is skewed before global model aggregation
Karimireddy et al., 2020	SCAFFOLD	The training objectives lack a previous local model as reference to prevent overfitting
Li QB et al., 2021	MOON	Global model may fail to guide the local model learning in some cases
Li CL et al., 2021	Meta-HAR	Limited labeled data may introduce bias to local model representation
Guo CR et al., 2024	ModCL	Client devices have limited capacity to store large amounts of data

a layer-wise manner using a Bayesian non-parametric method. FedAvgM (Hsu et al., 2019) introduces momentum when aggregating the local model weights for the global model updates. However, normalizing the weights during model aggregation may fail, as their local model updates already overfit their skewed local environment before aggregation. Another way is to separate the clients based on their characteristics and provide FL training for similar clients. For instance, the server can cluster the clients (Presotto et al., 2022; Guo JL et al., 2024) based on performance comparisons or similarity metrics. Instead of having a single global model, each cluster has a cluster model, which is exchanged and updated by the clients in the respective cluster. However, clustering-based methods have a computational issue since the clustering operation performs iterative comparisons between clients. Moreover, when the number of clients increases, the computation requirements scale proportionally.

The work that handles local training involves the clients controlling the local updates to prevent the local model from being too skewed to their local condition by using a regularization term. FedProx (Li T et al., 2020) presents a penalty term for the local training objective. The term uses the L2 norm to calculate the distance between the local model weights and the global model weights during local model training. Rather than regularizing the local model during training, SCAFFOLD (Karimireddy et al., 2020) controls the local model parameter updates by using control variates. Both a server and clients have control variates. After training, the client uses the difference between those control variates to adjust the parameters of the local model. Generally, all methods in the second category try to control the local model update, but their ability to prevent the local model from overfitting is limited, as they have a perspective only from their skewed dataset. Additionally, both methods regularize the local model only in either the training phase or parameter update phase.

Instead of directly learning from client-labeled data, which can introduce bias, some FL studies adopted contrastive learning (Li CL et al., 2021; Guo CR et al., 2024) that compares data points and obtains a general representation inside the underlying structure of the data. In general, contrastive learning methods can be deployed in unsupervised and supervised settings. The unsupervised setting uses the unlabeled data of the clients to guide the model training. For example, ModCL (Guo CR et al., 2024) is one of the HAR state-of-the-art methods that pre-trains the model on unlabeled data using inter- and intra-modality consistency in a contrastive manner before adapting to the clients. Although successful, this method requires the client to have a lot of unlabeled data because learning can be unstable with small amounts of data. Clients' wearable devices often have little storage and insufficient computations to train with massive data over a long period. The supervised setting uses the labeled data information and guides the model to learn representations that can differentiate them through contrastive learning. For example, Meta-HAR (Li CL et al., 2021) pre-trains a global encoder model with pairwise loss before fine-tuning the global encoder into the clients. However, there are usually only small amounts of client-labeled data, and they are often biased toward individual patterns.

In the image domain, the MOON technique involves model contrastive learning using the global model as another local training guide to prevent local model overfitting and promote generalized representations.

Even though MOON (Li QB et al., 2021) falls into the supervised contrastive learning category, this model tries to handle this issue by looking at the global model representation as another perspective to adjust the local model updates. As the global model learns from various clients, it contains general knowledge that can handle heterogeneity. MOON proposes to conduct model contrastive learning by comparing the global model and the previous round of local model output representation as drift control. This approach is successful in the image domain, but the global model may fail to guide the local model update in HAR, as the global model itself is the average result of skewed client local models.

Our proposed approach falls into the second category, which handles non-IID by improving the client's local training. To handle the limitation of model contrastive learning, this approach penalizes the local training loss by regulating it with control variates that take the previous local and global model updates into consideration.

3 Methodology

This section introduces the problem statement in this study and describes the FedCoad framework. Table 2 lists the notations that are used throughout this paper.

3.1 Preliminaries

In FL, there are N clients, denoted as $\mathcal{C}=\{C_1, C_2, \cdots, C_N\}$. Client C_i has a local dataset $\mathcal{D}_i=\{(x,y)\}$, where x represents the feature data and y represents the activity label. Let us consider a set of datasets $D=\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_N\}$ where each dataset corresponds to a client in $\mathcal{C}$. D is non-IID if there exists a pair of datasets $\mathcal{D}_i, \mathcal{D}_k \in D$ that follows one of the conditions below:

Feature distribution skew: $P_{\mathcal{D}_i}(x) \neq P_{\mathcal{D}_k}(x)$ is a condition where there exists an inequality between the probability distributions of $\mathcal{D}_i$ and $\mathcal{D}_k$ since the feature data in $\mathcal{D}_i$ have a significantly marginal distribution compared to the feature data in $\mathcal{D}_k$. This can happen in HAR, since everyone may perform activities in different ways. For example, an elderly person may have a slower walking pattern than an adolescent person.

Label distribution skew: $P_{\mathcal{D}_i}(y) \neq P_{\mathcal{D}_k}(y)$ is a condition where there exists an inequality between the probability distributions of $\mathcal{D}_i$ and $\mathcal{D}_k$ because the labels in $\mathcal{D}_i$ have a significantly marginal distribution compared to those in $\mathcal{D}_k$. In HAR, each person may have different daily activities. For example, office workers may spend more time sitting than athletes.

Quantity distribution skew: The total amount of data $|\mathcal{D}_i|$ is significantly different from $|\mathcal{D}_k|$. In HAR, since each person has a different schedule for the same activity, they have a different number of samples.

Table 2 Notations used in this paper

Symbol	Description	Symbol	Description
$\mathcal{C}$	Set of clients	x	Feature instances of a dataset
C_i	Client at index i, $C_i \in \mathcal{C}$	y	Label instances of a dataset
N	Total number of clients	ρ_{glob}	Control variate of the global model
D	Set of datasets	ρ_i	Control variate of client at index i
$\mathcal{D}_i$	Dataset of client at index i, $\mathcal{D}_i \in D$	Δ_i	Rate of change between the previous and updated control variates of client at index i
θ	Model weight		
θ_i^r	Local model weight of client i at round r	δ_{sup_i}	Supervised cross entropy loss of client at index i
θ_{glob}^r	Global model weight at round r	δ_{con_i}	Model contrastive loss of client at index i
$F_\theta(\cdot)$	Output of the whole model with weight θ	δ_{local_i}	Total local loss of client at index i
$J_\theta(\cdot)$	Output of the model with weight θ before the output layer	δ_{reg_i}	Regularized total local loss of client at index i
		α	Learning rate
$P_{\mathcal{D}_i}(x)$	Feature distribution of dataset $\mathcal{D}_i$	μ	Hyperparameter for controlling the influence of model contrastive loss
$P_{\mathcal{D}_i}(y)$	Label distribution of dataset $\mathcal{D}_i$		
r	Round of federated learning	τ	Temperature of model contrastive loss

With the conditions above, the objective is to build a model with weight θ over the dataset $D \triangleq \bigcup_{i\in\{1..N\}} \mathcal{D}_i$ without exchanging any of the client's personal training data. This work also assumes that each client learns only classes available during local training because updating model parameters for unseen classes may harm the performance of FL (Diao et al., 2021). Therefore, the objective is to minimize loss according to Eq. (1):

$$\min_{\theta} \mathcal{L}(\theta) = \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{|D|} L_i(\theta), \tag{1}$$

where $L_i(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\delta_i(\theta; (x,y))\right]$ is the empirical loss of C_i.

3.2 Method

This subsection explains FedCoad components such as the model architecture used for model contrastive learning, the inner workings of FedCoad during the local model training and parameter update phase, and the FL workflow in FedCoad.

3.2.1 Model architecture

The model has three components (Fig. 3): the base encoder which is responsible for extracting representations from data inputs; the projection head which is an additional layer that maps the encoder representation to another space with a determined dimension; an output layer that predicts the class label based on projection head output. To formalize the formula, we use the notations $F_\theta(\cdot)$ to denote the whole model and $J_\theta(\cdot)$ to denote the model before the output layer. For example, $J_\theta(x)$ refers to the mapped representation of data input x.

3.2.2 Local training phase

In local model training, the local model needs to optimize two losses (Fig. 3): classification loss and model contrastive loss. Classification loss refers to cross-entropy loss, and model contrastive loss refers to metric learning loss, which operates on the similarity between the representations of the current local model and the global model along with the previous local model. When a client C_i starts to conduct local training, it downloads the global model θ^r_{glob} from the server and trains its local model θ^r_i at round r. For every data input x, the model obtains the representation of x from the current local model being updated $g = J_{\theta^r_i}(x)$, the representation of x from the global model $g_{\text{glob}} = J_{\theta^r_{\text{glob}}}(x)$, and the representation of x from the local model at the previous round $g_{\text{prev}} = J_{\theta^{r-1}_i}(x)$. With the assumption that the global model can produce better representations (Li QB et al., 2021), the goal is to minimize the distance between g and g_{glob} while increasing the dissimilarity between g and g_{prev}. Therefore, to encourage g to have hard-to-learn distinctions with g_{prev}, the calculation of client C_i model contrastive loss (δ_{con_i}) follows Eq. (2):

$$\delta_{\text{con}_i} \leftarrow -\log \frac{\exp\left(\frac{\text{sim}(g, g_{\text{glob}})}{\tau}\right)}{\exp\left(\frac{\text{sim}(g, g_{\text{glob}})}{\tau}\right) + \exp\left(\frac{\text{sim}(g, g_{\text{prev}})}{\tau}\right)}, \tag{2}$$

where τ denotes the temperature parameter that controls the penalty strength on the distance between the local model and the previous local model. The total local loss of client C_i is the sum of its classification loss (δ_{sup_i}) and model contrastive loss (Eq. (3)):

$$\delta_{\text{local}_i} = \delta_{\text{sup}_i}(\theta^r_i; (x,y)) + \mu\delta_{\text{con}_i}(\theta^r_i; \theta^{r-1}_i; \theta^r_{\text{glob}}; x), \tag{3}$$

where μ is a hyperparameter that controls the influence of contrastive loss.

3.2.3 Local parameter update phase

As clients are trained only with their local data, their updates tend to drift from the true global optima and pursue their local minima (Karimireddy et al., 2020). Therefore, assigning a new parameter directly after local training hinders global model convergence. To handle this, clients need to regularize their local parameter updates after the local training. Karimireddy et al. (2020) used control variates to adjust the local model update to minimize client drift. Following their footsteps, in this study, we incorporate control variates such as local control variates denoted as $\rho_1, \rho_2, \cdots, \rho_N$ and a global control variate denoted as ρ_{glob}.

The purpose of the control variate is to adjust the local model update so that it does not pursue local minima. The local control variate represents the update speed or the rate of change between a client-trained local model and the global model. As the global control variate is the average of the local control variates, it

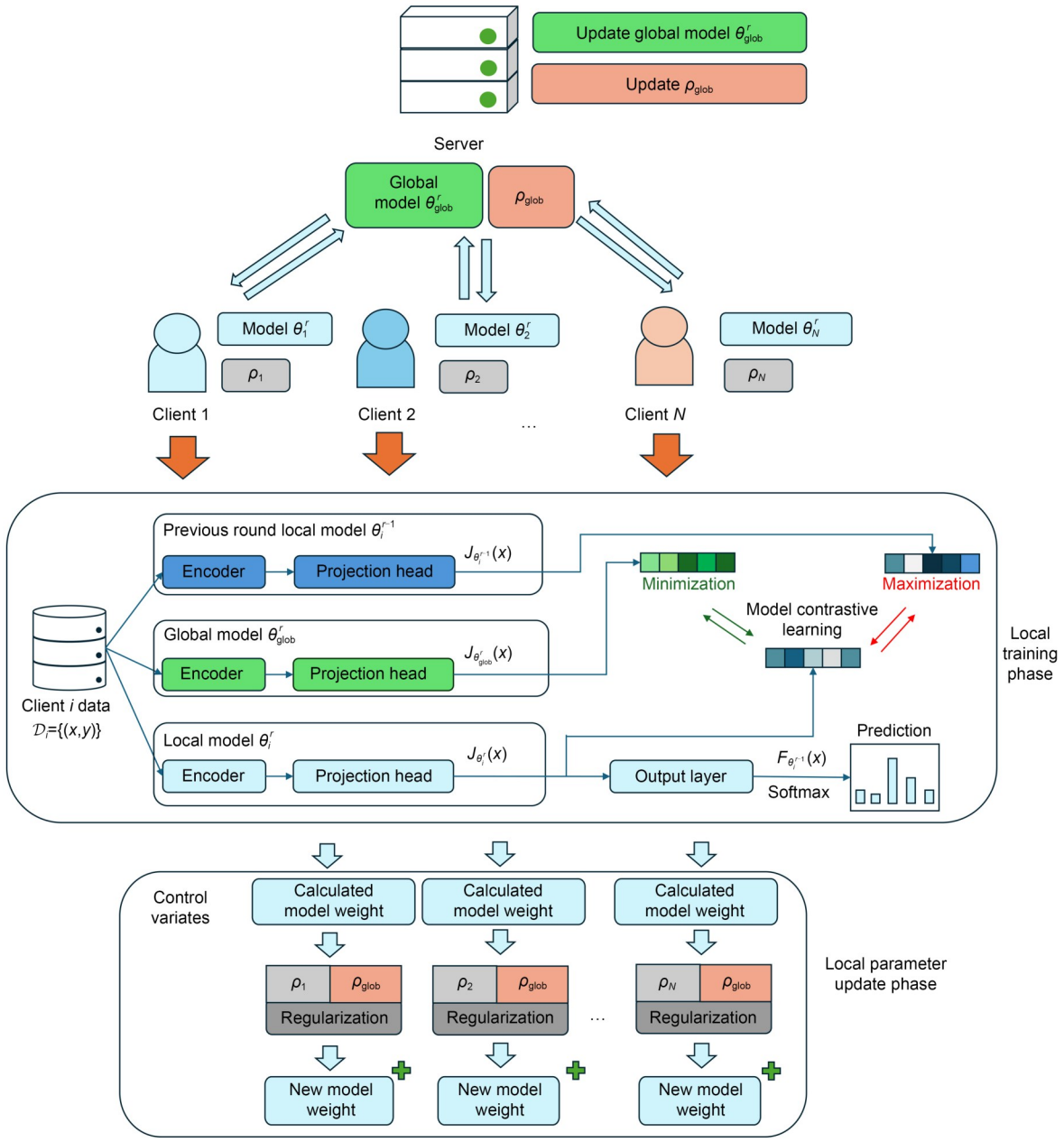

Fig. 3 Framework of FedCoad

embodies the overall client update speed. Therefore, FedCoad prevents local clients from converging too fast to their local minima by factoring the divergence between overall clients (global control variates) and their update speed (local control variates).

In the first round, the server initializes the global control variate using random weights and makes copies for the client's local control variate. Therefore, the regularized local loss for the local client updates follows Eq. (4):

$$\delta_{\mathrm{reg}_i}=\delta_{\mathrm{local}_i}+\rho_{\mathrm{glob}}-\rho_i. \tag{4}$$

The local objective follows the expression below:

$$\min_{\theta_i^r}\mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\,\delta_{\mathrm{sup}_i}(\theta_i^r;(x,y))+\mu\delta_{\mathrm{con}_i}(\theta_i^r;\theta_i^{r-1};\theta^r_{\mathrm{glob}};x)+\rho_{\mathrm{glob}}-\rho_i\,]. \tag{5}$$

For every C_i, the parameter weight updates follow Eq. (6):

$$\theta_i^r \leftarrow \theta_i^r - \alpha \nabla \delta_{\mathrm{reg}_i}, \tag{6}$$

where α is the learning rate and $\nabla\delta_{\mathrm{reg}_i}$ is the local model update. Next, C_i updates ρ_i according to Eq. (7):

$$\rho_i \leftarrow \rho_i - \rho_{\mathrm{glob}} + \frac{\theta_{\mathrm{glob}}^r - \theta_i^r}{E\alpha}, \tag{7}$$

where E is the number of local epochs. After all clients send their local model updates and their local control variate rate of change Δ_i, the server updates ρ_{glob} according to Eq. (8):

$$\rho_{\mathrm{glob}} = \sum_{k=1}^{N} \Delta_k. \tag{8}$$

3.2.4 FL algorithm

Using all the components mentioned above, FedCoad aims to build a global model by using a server to arrange FL training with clients. Details of the workflow are shown in Algorithms 1 and 2. In the first round ($r=1$), the server initializes the first global model weight θ_{glob}^0 and server control variates ρ_{glob} (Algorithm 1, lines 2 and 3). Then, the server lists N clients that will join the training and sends θ_{glob}^0 and other hyperparameters to $\mathcal{C}$ (Algorithm 1, line 6). After receiving the server transmission, each client C_i performs "LocalTraining" using their local model θ_i^r and ρ_{glob} (Algorithm 1, line 7). Additionally, if it is the first round ($r=1$), C_i will initialize their local control variates ρ_i (Algorithm 2, line 3). Inside "LocalTraining," client C_i calculates the prediction loss δ_{sup_i} using cross-entropy and model contrastive loss δ_{con_i}. To obtain δ_{sup_i}, C_i runs θ_i^r on their local data $\mathcal{D}_i$ to obtain local model prediction $F_{\theta_i^r}$ and compares it with the true label y (Algorithm 2, line 8). For the δ_{con_i}, C_i runs θ_i^r, θ_{glob}^r, and θ_i^{r-1} on $\mathcal{D}_i$ to obtain g, g_{glob}, and g_{prev}, respectively, before C_i calculates δ_{con_i} (Algorithm 2, lines 9–12). Next, C_i sums δ_{con_i} and δ_{sup_i} as $\delta_{\mathrm{local}_i}$ (Algorithm 2, line 13), which will then be penalized using ρ_{glob} and ρ_i (Algorithm 2, line 14). The penalized loss δ_{reg_i} is then used for updating θ_i^r (Algorithm 2, line 15). After "LocalTraining," C_i updates the ρ_i values and their rate of change Δ_i (Algorithm 2, lines 18–20). Subsequently, each C_i sends the updated θ_i^r and Δ_i to the server (Algorithm 2, line 21). Finally, the server receives a list of $\{\theta_1^r, \theta_2^r, \cdots, \theta_N^r\}$ and $\{\Delta_1, \Delta_2, \cdots, \Delta_N\}$ to update θ_{glob}^r and ρ_{glob} (Algorithm 1, lines 8 and 9).

Algorithm 1 FedCoad server workflow

Input: number of communication rounds R, clients $\mathcal{C}$, client datasets D, and learning rate α
Output: final global model θ_{glob}^r

1 Server:
2 initialize θ_{glob}^0
3 initialize ρ_{glob}
4 **for** $r = 1, 2, \cdots, R$ **do**
5 **for** $i = 1, 2, \cdots, N$ **do**
6 send θ_{glob}^r and ρ_{glob} to C_i
7 $\theta_i^r, \Delta_i \leftarrow$ LocalTraining $(i, \theta_i^r, \rho_{\mathrm{glob}})$
8 **end for**
9 $\theta_{\mathrm{glob}}^{r+1} = \sum_{k=1}^{N} \frac{|\mathcal{D}_k|}{|D|} \theta_k^r$
10 $\rho_{\mathrm{glob}} \leftarrow$ update global control variates using Eq. (8)
11 **end for**
12 return θ_{glob}^R

Algorithm 2 Local training workflow

Input: client index i, client model weight θ_i^r, number of local epochs E, batch size B, and global control variate ρ_{glob}
Output: updated model weight θ_i^r and local control variate rate of change Δ_i

1 LocalTraining $(i, \theta_i^r, \rho_{\mathrm{glob}})$:
2 **if** $r=1$ **then**
3 $\rho_i \leftarrow \rho_{\mathrm{glob}}$
4 **end if**
5 $\theta_i^r \leftarrow \theta_{\mathrm{glob}}^r$
6 **for** epoch $e=1, 2, \cdots, E$ **do**
7 **for** each batch $b=\{x, y\}$ of $\mathcal{D}_i$ **do**
8 $\delta_{\mathrm{sup}_i} \leftarrow \mathrm{CrossEntropy}(F_{\theta_i^r}, y)$
9 $g = J_{\theta_i^r}(x)$
10 $g_{\mathrm{glob}} = J_{\theta_{\mathrm{glob}}^r}(x)$
11 $g_{\mathrm{prev}} = J_{\theta_i^{r-1}}(x)$
12 $\delta_{\mathrm{con}_i} \leftarrow$ loss calculated by using Eq. (2)
13 $\delta_{\mathrm{local}_i} \leftarrow$ sum of local loss calculated by using Eq. (3)
14 $\delta_{\mathrm{reg}_i} \leftarrow$ penalized loss calculated by using Eq. (4)
15 update local model weight θ_i^r using Eq. (6)
16 **end for**
17 **end for**
18 $\rho_{\mathrm{temp}} = \rho_i$
19 $\rho_i \leftarrow$ local control variates updated by using Eq. (7)
20 $\Delta_i = \rho_i - \rho_{\mathrm{temp}}$
21 return θ_i^r and Δ_i to the server

4 Experiments

This section describes the experiment and analysis results. FedCoad was evaluated on popular HAR benchmark datasets and compared with methods from previous studies. The supplementary materials cover other details such as the performance measure, comparison with pretraining methods, embedding visualization, and communication exchange comparison (supplementary materials, Sections 1.1–1.4).

4.1 Datasets

Three benchmark datasets were used in this study:

MotionSense (Malekzadeh et al., 2019) includes results from a study aimed at advancing the development of privacy preservation in sensor data transmission systems. There were 24 subjects, including 10 women and 14 men, in the experiment. The subjects' ages ranged from 18 to 46 years, and their body weights ranged from 48 to 102 kg. An iPhone 6s was used as the mobile device and placed in each subject's front trouser pocket. The subjects performed activities, including walking upstairs, sitting, walking, standing, jogging, and walking downstairs. Fig. 4a shows the distribution of activities in the dataset. This study used only the accelerometer signal from the mobile device for compatibility.

Wireless sensor data mining (WISDM) (Kwapisz et al., 2011) was used to inspect problems in collecting sensor data embedded in mobile devices. There were 36 subjects who performed daily living activities such as standing, walking, sitting, walking upstairs, jogging, and walking downstairs. Fig. 4b shows the distribution of activities in the dataset. The subjects used a smartphone in the front pocket of their trousers while performing each activity at various times. The data were gathered from the accelerometer inside the smartphone with a sampling frequency of 20 Hz.

Heterogeneity HAR (HHAR) (Stisen et al., 2015) was used to investigate the effects of heterogeneous mobile devices in terms of sensors, device models, and workloads on the performance of activity recognition. HHAR contains data from nine subjects aged between 25 and 30 when performing daily living activities, including walking, biking, walking upstairs, walking downstairs, sitting, and standing. Fig. 4c shows the distribution of activities in the dataset. Smartwatches were strapped around the subjects' arms, and smartphones of variable frequency were placed on their waists. The following devices were used during the HHAR study: Samsung Galaxy S3 Mini (100 Hz), Samsung Galaxy S Plus (50 Hz), LG Nexus 4 (200 Hz), Samsung Galaxy S3 (150 Hz), Samsung Galaxy Wear (100 Hz), and LG G (200 Hz). During the data collection, all the subjects were asked to perform each activity for 5 min. This study used only the triaxial accelerometer signals from the smartphone for compatibility with other datasets.

4.2 Experimental setup

The experiment was performed on a computer with Intel® Core™ i9-9900K 3.60 GHz, GPU NVIDIA RTX 2080, and the neural network was built using Python and PyTorch libraries. This study compared FedCoad with the state-of-the-art approaches FedAvg (McMahan et al., 2017), FedAvgM (Hsu et al., 2019), SCAFFOLD (Karimireddy et al., 2020), FedProx (Li T et al., 2020), and MOON (Li QB et al., 2021). The

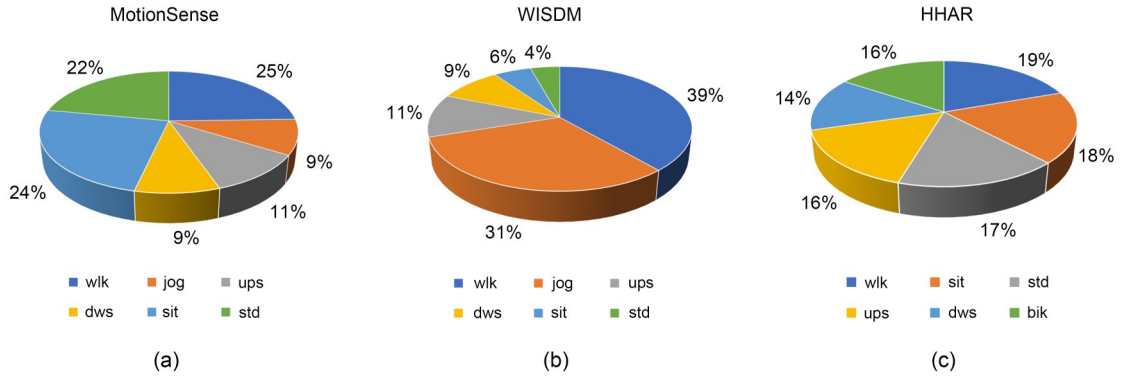

Fig. 4 Class distributions in the MotionSense (a), WISDM (b), and HHAR (c) datasets. bik: biking. References to color refer to the online version of this figure

model used 1D CNN as the base encoder, followed by two fully connected layers as the projection head and one fully connected layer as the output layer.

For data preprocessing, all the data were segmented into a time window with 50% overlap and a sequence length of 200. For example, MotionSense was captured at 50 Hz, so sensor readings with a duration of 4 s were taken as one sample window. During the evaluation, 80% of the subjects were used as training clients, while the rest were used for server private evaluation data. Each algorithm used Adam as the optimizer with a learning rate of 0.001 and a weight decay of 0.000 01. The number of batch size was set to 8, the number of communication rounds was set to 100, and the number of local epochs was set to 5. Additionally, the temperature parameter τ in methods that used contrastive learning was set to 0.5.

4.3 Performance comparison

In this subsection, we report the performance of all the listed algorithms for those clients with fully labeled data. All algorithms were run five times to account for variation. The experiment varied the hyperparameter μ={1.0, 5.0, 10.0} for both FedCoad and MOON according to the range used by Li QB et al. (2021). For FedAvgM, the momentum value β was set to 0.7 and 0.997 following the original paper (Hsu et al., 2019). Table 3 shows the accuracy and F1-score results of all the listed algorithms.

With respect to the accuracy metric, FedCoad outperformed MOON, which used model contrastive learning, and SCAFFOLD, which used control variates. For example, FedCoad (μ=1.0) outperformed MOON (μ=10.0) by 3.58% in MotionSense and outperformed SCAFFOLD by 1.16% in WISDM. FedCoad (μ=1.0) also slightly outperformed FedProx by 0.28% in MotionSense and 0.97% in WISDM. For the HHAR dataset, FedCoad (μ=5.0) achieved the highest accuracy and outperformed SCAFFOLD by 1.38% and FedProx by 4.46%.

Next, we used the F1-score, which accounts for class imbalance in HAR, to estimate the FedCoad improvement. From the F1-score results, methods that included only model contrastive learning failed to improve the performance on vanilla FedAvg. For example, MOON (μ=1.0) achieved 66.26% in terms of the F1-score while FedAvg was 67.69% in WISDM, which showed that model contrastive learning alone is insufficient for HAR. On the other hand, the methods that included regularization in the local training step, such as FedProx, SCAFFOLD, and FedCoad, outperformed FedAvg. Specifically, FedCoad, FedProx, and SCAFFOLD can perform competitively among other methods. For example, the F1-score of FedCoad (μ=1.0) outperformed that of FedAvg by 4.38% and 5.86% in MotionSense and WISDM, respectively.

For hyperparameter μ, FedCoad with μ=1.0 and μ=5.0 achieved the highest F1-score and accuracy on all benchmark datasets. For the MotionSense dataset, FedCoad with μ=1.0 achieved an accuracy of 95.52% and an F1-score of 92.70%, which surpassed those of other methods. On the other hand, FedCoad with μ=5.0

Table 3 Performance results of FedCoad and other FL algorithms on benchmark datasets

Method	Accuracy (%)			F1-score (%)		
	MotionSense	WISDM	HHAR	MotionSense	WISDM	HHAR
FedAvg	91.44	73.94	80.86	88.32	67.69	79.07
FedProx	95.24	75.61	79.85	92.16	70.33	77.63
SCAFFOLD	92.48	75.42	82.93	90.29	71.20	82.02
FedAvgM (β=0.7)	86.23	73.48	77.10	83.13	66.52	74.53
FedAvgM (β=0.997)	81.76	64.25	65.92	78.00	55.74	71.09
MOON (μ=1.0)	91.76	74.62	80.36	88.32	66.26	78.24
MOON (μ=5.0)	91.65	73.36	77.58	87.90	64.59	74.64
MOON (μ=10.0)	91.94	72.63	80.24	88.37	63.26	78.25
FedCoad (μ=1.0)	95.52	76.58	83.78	92.70	73.55	82.81
FedCoad (μ=5.0)	93.47	76.67	84.31	90.61	73.17	83.22
FedCoad (μ=10.0)	94.78	76.46	84.06	91.67	71.91	82.91

outperformed the other methods on the HHAR dataset, with an accuracy of 84.31% and an F1-score of 83.22%. For the WISDM dataset, μ=1.0 and μ=5.0 gave the same performance in terms of accuracy and F1-score. Therefore, μ=1.0 and μ=5.0 are suitable values for FedCoad.

Fig. 5 shows the correctness of FedCoad class predictions for all benchmark datasets. In MotionSense, FedCoad could correctly predict the true classes with a minimal amount of error. For the other datasets, FedCoad produced some classification errors for activities with similar locomotion. For example, in Figs. 5b and 5c, FedCoad showed some errors distinguishing between "upstairs" and "downstairs" in WISDM and between "sit" and "stand" in HHAR. However, FedCoad correctly predicted the true classes for most of the cases in both datasets.

According to the results of FedCoad, combining model contrastive learning with control variates improves the performance compared to the other methods. For example, FedCoad (μ=1.0), with F1-scores of 92.70% and 73.55%, outperformed SCAFFOLD by 2.41% and 2.35% in MotionSense and WISDM, respectively.

4.4 Convergence speed comparison

In this subsection, we compare all the methods against FedAvg in terms of their F1-score convergence rates. The number of rounds needed for all methods to achieve a similar F1-score as running FedAvg for 100 rounds is shown in Table 4. Fig. 6 compares the convergence speed of all methods. FL baselines with regularization can achieve significant speedup compared to FedAvg. For example, in Table 4, FedCoad and SCAFFOLD achieved a 12.5× speedup compared to FedAvg in WISDM. At the start of the round, the speed of F1-score improvement in FedCoad was similar to that in FedAvg and other FL methods. Nevertheless, FedCoad achieved a higher F1-score than FedAvg in later rounds. Furthermore, FedCoad achieved the highest speedup in two out of three datasets (MotionSense and WISDM), demonstrating its superiority.

On the other hand, all methods had unstable convergence in the HHAR dataset (Fig. 6c). Unlike the other datasets, the clients in the HHAR dataset had different devices, such as the Samsung Galaxy S3 or LG Nexus 4. Therefore, heterogeneity from the devices

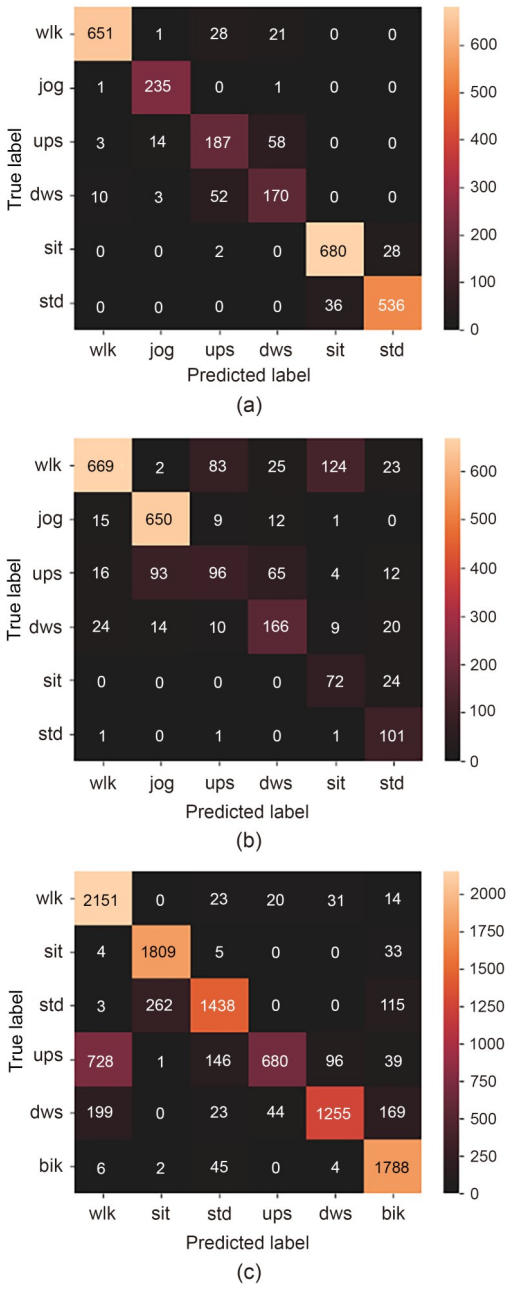

Fig. 5 FedCoad confusion matrix on MotionSense (a), WISDM (b), and HHAR (c) datasets. References to color refer to the online version of this figure

affects global model aggregation. Nevertheless, the proposed FedCoad had an upward and stable trend from round 80 to round 100 and achieved the highest F1-score performance in the end compared to the other methods.

For all benchmark datasets, FedCoad and SCAFFOLD performed competitively in terms of the convergence rate. For MotionSense, FedCoad was

Table 4 Number of rounds for FL methods to achieve a similar F1-score as running FedAvg for 100 rounds

Method	Number of rounds			Speedup		
	MotionSense	WISDM	HHAR	MotionSense	WISDM	HHAR
FedAvg	100	100	100	1×	1×	1×
FedAvgM	\	51	7	\	1.96×	14.28×
FedProx	11	19	11	9.1×	5.26×	9.1×
SCAFFOLD	9	8	9	11.11×	12.5×	11.11×
MOON (μ=1.0)	15	48	27	6.67×	2.08×	3.7×
MOON (μ=5.0)	73	94	32	1.37×	1.06×	3.12×
MOON (μ=10.0)	40	99	4	2.5×	1.0×	25×
FedCoad (μ=1.0)	7	8	10	14.3×	12.5×	10×

"\" means that the method fails to outperform FedAvg in terms of the convergence speed

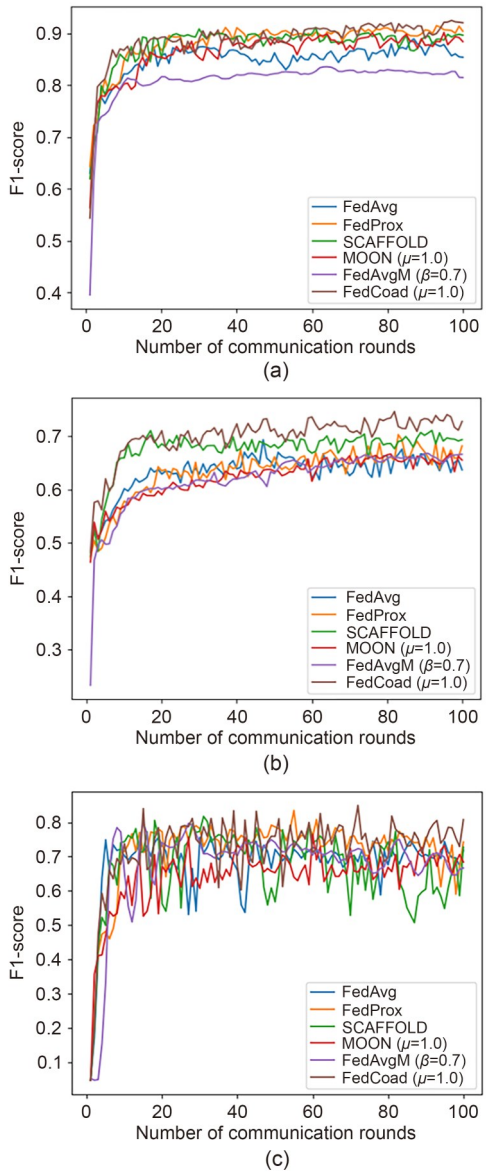

Fig. 6 F1-score for various numbers of communication rounds for MotionSense (a), WISDM (b), and HHAR (c) datasets. References to color refer to the online version of this figure

superior to all the other methods (14.3 times faster than FedAvg). However, in WISDM, SCAFFOLD and FedCoad were both 12.5 times faster than FedAVG.

4.5 Handling real-world HAR skewness

In real-world HAR implementation, three skewness scenarios (Presotto et al., 2022) are considered. In this study, we simulated skewness using the benchmark dataset that previously followed laboratory settings to make it follow non-IID in a real-world setting. Fig. 7 shows the effects of each skewness scenario on the clients' data distributions. The following steps were taken to simulate the skewness:

1. Feature distribution skew (FS): this experiment used the skewness already present in each dataset, as each client already had a different pattern.

2. Label distribution skew (LS): 0 to 2 labels were randomly removed to simulate the variation in the available activities of each client (Li CL et al., 2021).

3. Quantity distribution skew (QS): each client has only small amounts of labeled data in real life (Tang et al., 2021); thus, in the experiments, each client used only 10% of the available data as a training set.

4.5.1 Feature skew and label skew (FS+LS)

In the FS+LS condition, some of the clients' original training label sets were removed. Therefore, the clients had an uneven label set for training, and each client's local objectives can misguide local model training. Table 5 shows that the performance of all FL methods was reduced significantly in the FS+LS condition. For example, the performance of SCAFFOLD decreased by 5.24%, and that of MOON decreased by 2.82% on average across all datasets. A notable exception was FedProx, whose F1-score decreased

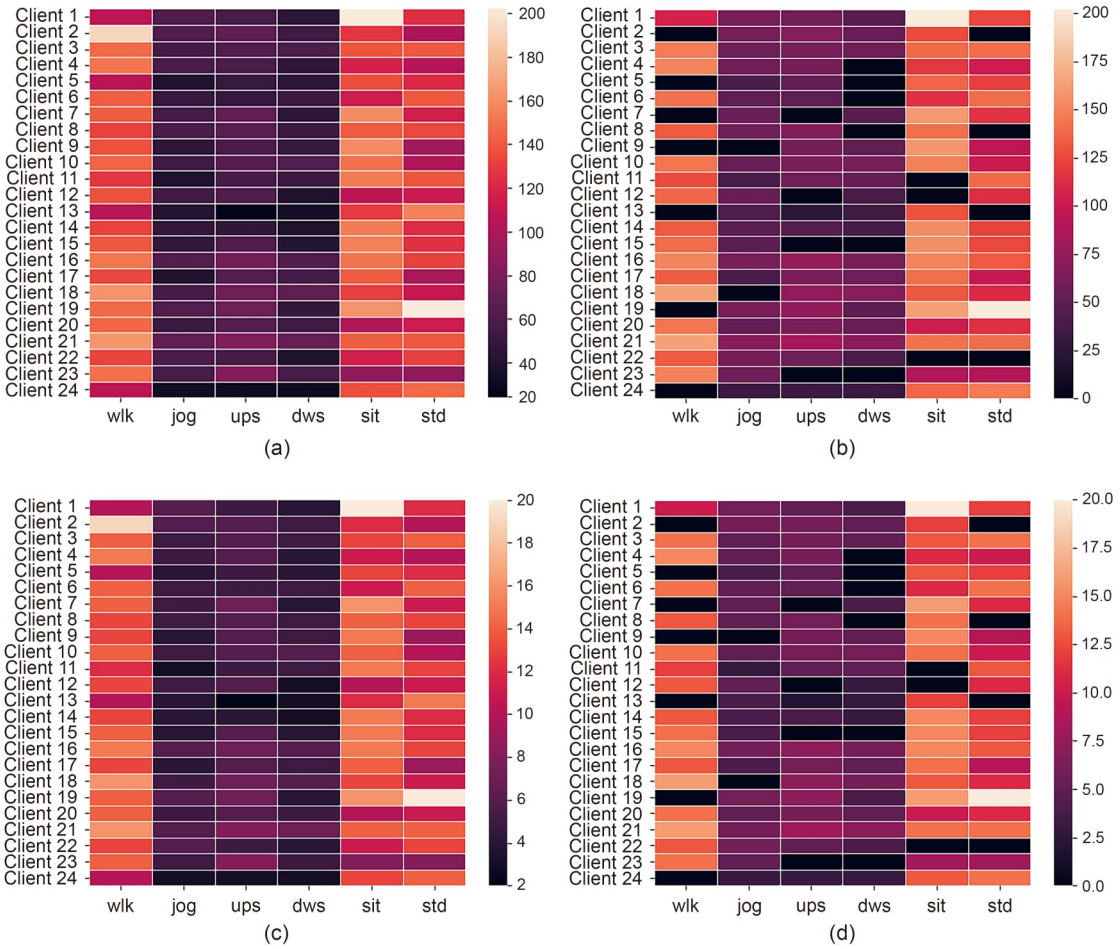

Fig. 7 Data distribution of each party in different skewness scenarios using the MotionSense dataset: (a) FS; (b) FS+LS; (c) FS+QS; (d) FS+LS+QS. References to color refer to the online version of this figure

Table 5 F1-score and reduction of FL methods in the FS+LS scenario

Method	F1-score (%)			Reduction (%)		
	MotionSense	WISDM	HHAR	MotionSense	WISDM	HHAR
FedAvg	86.80	66.58	71.32	1.52	1.11	7.75
FedProx	90.87	69.44	59.64	1.29	0.89	17.99
SCAFFOLD	88.10	66.42	73.28	2.19	4.78	8.74
FedAvgM (β=0.7)	82.10	68.01	68.21	1.03	−1.49	6.32
MOON (μ=1.0)	85.27	65.24	73.84	3.05	1.02	4.40
FedCoad (μ=1.0)	90.58	63.77	72.06	2.12	9.78	10.75

by only 1.29% in MotionSense and 0.89% in WISDM. However, FedProx had the worst performance in HHAR, showing a reduction of 17.99%. Even though FedCoad achieved competitive performance with the other methods in MotionSense and WISDM, it showed a severe performance reduction of 7.55% on average across all datasets.

4.5.2 Feature skew and quantity skew (FS+QS)

In the FS+QS condition, all clients retained their original training label set, but the quantity decreased to 10% of the original size. Therefore, the risk of overfitting in local model training increases since the amount of training data is reduced. Compared to the FL+LS condition, there were no significant performance reductions

among all FL methods. For example, FedCoad performance decreased by 2.04% on average across all datasets (Table 6). FedProx had the most significant performance reduction of 5.43% in MotionSense. However, there is an odd phenomenon in HHAR. Instead of decreasing, the performance of all FL methods increased over their performance in the FS condition. Nevertheless, FedCoad is robust enough to train a global model in the FS+QS condition.

4.5.3 All skewness scenarios combined (FS+QS+LS)

In the FS+QS+LS condition, some of the clients' original training label set was removed and the quantity of the remaining training data decreased by 10%. Most FL methods experienced a larger performance reduction than in the FS+QS condition (Table 7). For example, FedCoad performance was reduced by 4.24% on average across all datasets in FS+QS+LS compared to 2.04% in the FS+QS condition. It seems the FS+LS and FS+QS+LS conditions caused large performance reductions using all FL methods. Therefore, LS skewness was the main cause of performance reduction among other skewness scenarios. In terms of performance, FedCoad performed the best in the FS+QS+LS conditions. For example, it achieved the highest F1-score of 88.54% in MotionSense and 69.49% in WISDM.

5 Discussion

In this section, we delve into the experimental results and provide insights into the factors contributing to the observed outcomes. Without the simulated skewness, as described in Section 4.3, the proposed FedCoad achieved competitive performance compared to methods like MOON and SCAFFOLD, which address heterogeneity either during local training or through local model parameter updates. For instance, in the WISDM dataset, FedCoad (μ=1.0) outperformed MOON (μ=1.0) by 1.96% and SCAFFOLD by 1.16% in accuracy. This suggests that FedCoad and similar baseline methods are sufficient for training a global model in the FS condition. However, the F1-score results indicate that FedCoad (μ=1.0) predictions are more balanced across classes, with an F1-score improvement of 7.29% over MOON (μ=1.0).

When introducing simulated skewness (Section 4.5) that reflects real-world conditions, the performance of several FL baseline methods declined significantly. For example, in the MotionSense dataset under FS+LS+QS conditions, methods like MOON, which focuses on heterogeneity during local training, and SCAFFOLD, which focuses on parameter updates, experienced more severe degradation than FedCoad. MOON, which uses only model contrastive learning

Table 6 F1-score and reduction of FL methods in the FS+QS scenario

Method	F1-score (%)			Reduction (%)		
	MotionSense	WISDM	HHAR	MotionSense	WISDM	HHAR
FedAvg	85.07	64.08	80.11	3.25	3.61	−1.04
FedProx	86.73	66.64	87.04	5.43	3.69	−9.41
SCAFFOLD	90.12	67.97	85.87	0.17	3.23	−3.85
FedAvgM (β=0.7)	83.20	62.61	80.88	−0.07	3.91	−6.35
MOON (μ=1.0)	87.90	66.53	78.29	0.42	−0.27	−0.05
FedCoad (μ=1.0)	89.84	67.62	85.47	2.86	5.93	−2.66

Table 7 F1-score and reduction of FL methods in the FS+LS+QS scenario

Method	F1-score (%)			Reduction (%)		
	MotionSense	WISDM	HHAR	MotionSense	WISDM	HHAR
FedAvg	76.14	63.95	76.91	12.18	3.74	2.16
FedProx	85.03	65.35	86.96	7.13	4.98	−9.33
SCAFFOLD	84.69	67.33	75.51	5.60	3.87	6.51
FedAvgM (β=0.7)	77.62	61.48	78.92	5.51	5.04	−4.39
MOON (μ=1.0)	77.57	63.85	75.93	10.75	2.41	2.31
FedCoad (μ=1.0)	88.54	69.49	78.31	4.16	4.06	4.50

during local training, showed a 10.75% reduction in F1-score compared to its performance without simulated skewness. This highlights the insufficiency of handling heterogeneity solely during local training. Similarly, FedProx and SCAFFOLD, which address heterogeneity during parameter updates, showed F1-score decreases of 7.13% and 5.6%, respectively, indicating that regularizing local model weights after training mitigates some of the effects of HAR skewness, as it directly reduces the bias from skewed local training objectives. FedCoad, which addresses heterogeneity during both local training and local model parameter updates, showed superior performance under these conditions. It achieved the highest F1-score of 88.54% with the smallest performance decrease (4.16%) compared to other baseline methods. These results underscore FedCoad's resilience and robustness in handling non-IID environments. Additional experiments further validated its strengths, showing higher convergence rates and the ability to train model encoders that generate distinguishable feature embeddings (supplementary materials, Section, 1.3), which are essential for classification tasks.

Furthermore, the ablation studies described in the supplementary materials reveal the contribution of FedCoad's components (supplementary materials, Section 2.1) and their synergistic effect on performance. Experiments on client availability (supplementary materials, Section 2.3) also highlight FedCoad's scalability. For example, when client availability increased from 20% to 100%, FedAvg's F1-score was improved by only 1.41%, whereas FedCoad's improved by 4.44%.

Despite its advantages, FedCoad has limitations for real-world implementation. It currently does not use unlabeled data, which are abundant on client devices and could enhance local training and adaptation. Additionally, FedCoad incurs higher communication costs due to the exchange of global and local control variates during training. Nevertheless, its strong performance in simulated skewness environments suggests its potential for real-world applications. Future work should focus on addressing these limitations, such as incorporating semi-supervised learning to leverage unlabeled data and optimizing communication efficiency to reduce costs.

6 Conclusions

In this work, we study the problem of HAR in the FL setting and build a generalized global model that can be adapted for participating clients. The skewness among clients in the real world poses significant challenges to implementing FL for HAR. To address this problem, we propose FedCoad, a new approach to align global and local model representation using model contrastive learning to bridge and control variates to regularize the local model update. In our experiments, FedCoad outperformed other methods in tests of skewed dataset settings (non-IID) in the benchmark datasets. Even though FedCoad had disadvantages in the FS+LS conditions, it outperformed other methods in the FS+QS+LS conditions, which implies that FedCoad can be more robust in a real-world scenario. Furthermore, the performance of FedCoad was comparable to that of pretraining methods that require additional fine-tuning data (supplementary materials, Section 1.2). The results also showed that FedCoad can reduce the communication cost by up to 90.4% compared to the state-of-the-art ModCL (supplementary materials, Section 1.4).

Despite these improvements, FedCoad has a limitation in the unlabeled data partition and personalization scheme. It lacks the use of unlabeled data on the client side, which is beneficial for local training and adaptation steps to optimize the performance in client environments. Extreme heterogeneity in hardware capabilities or data imbalance among clients may also impact FedCoad's stability and performance. Moreover, FedCoad's additional memory requirements for control variates may pose challenges to low-powered edge devices with minimal storage.

In future work, this study can be extended to use unlabeled data with semi-supervised learning on the client's side for leveraging the unlabeled data to optimize performance and explore personalization strategies that could further tailor the model to individual client environments. The development of an adaptive weighting mechanism could also help in the server aggregation step to combine the local model updates from clients with extreme data imbalance. Additionally, optimizing or compressing control variates could reduce memory demands, making FedCoad more suitable for edge devices.

Contributors

Ignatius IWAN designed the research and drafted the paper. Bernardo Nugroho YAHYA and Seok-Lyong LEE provided conceptualization, supervision, and funding acquisition. All the authors revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The datasets in this paper can be acquired through the following links: https://archive.ics.uci.edu/dataset/507/wisdm+smartphone+and+smartwatch+activity+and+biometrics+dataset (WISDM), https://github.com/mmalekzadeh/motion-sense (MotionSense), and https://archive.ics.uci.edu/dataset/344/heterogeneity+activity+recognition (HHAR).

References

Bianchi V, Bassoli M, Lombardo G, et al., 2019. IoT wearable sensor and deep learning: an integrated approach for personalized human activity recognition in a smart home environment. *IEEE Int Things J*, 6(5):8553-8562. https://doi.org/10.1109/JIOT.2019.2920283

Diao EM, Ding J, Tarokh V, 2021. HeteroFL: computation and communication efficient federated learning for heterogeneous clients. Proc 9th Int Conf on Learning Representations.

Guo CR, Zhang YW, Chen YQ, et al., 2024. Modality consistency-guided contrastive learning for wearable-based human activity recognition. *IEEE Int Things J*, 11(12):21750-21762. https://doi.org/10.1109/JIOT.2024.3379019

Guo JL, Li ZT, Liu AF, et al., 2024. REC-Fed: a robust and efficient clustered federated system for dynamic edge networks. *IEEE Trans Mob Comput*, 23(12):15256-15273. https://doi.org/10.1109/TMC.2024.3452312

Hsu TMH, Qi H, Brown M, 2019. Measuring the effects of non-identical data distribution for federated visual classification. https://arxiv.org/abs/1909.06335

Karimireddy SP, Kale S, Mohri M, et al., 2020. SCAFFOLD: stochastic controlled averaging for federated learning. Proc 37th Int Conf on Machine Learning, p.5132-5143.

Kwapisz JR, Weiss GM, Moore SA, 2011. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explor Newsl*, 12(2):74-82. https://doi.org/10.1145/1964897.1964918

Li CL, Niu D, Jiang B, et al., 2021. Meta-HAR: federated representation learning for human activity recognition. Proc Web Conf, p.912-922. https://doi.org/10.1145/3442381.3450006

Li QB, He BS, Song D, 2021. Model-contrastive federated learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10713-10722. https://doi.org/10.1109/CVPR46437.2021.01057

Li T, Sahu AK, Zaheer M, et al., 2020. Federated optimization in heterogeneous networks. Proc 3rd Conf on Machine Learning and Systems.

Li X, Huang KX, Yang WH, et al., 2020. On the convergence of FedAvg on non-IID data. Proc 8th Int Conf on Learning Representations.

Malekzadeh M, Clegg RG, Cavallaro A, et al., 2019. Mobile sensor data anonymization. Proc Int Conf on Internet of Things Design and Implementation, p.49-58. https://doi.org/10.1145/3302505.3310068

McMahan HB, Moore E, Daniel R, et al., 2017. Communication-efficient learning of deep networks from decentralized data. Proc 20th Int Conf on Artificial Intelligence and Statistics, p.1273-1282.

Mrozek D, Koczur A, Małysiak-Mrozek B, 2020. Fall detection in older adults with mobile IoT devices and machine learning in the cloud and on the edge. *Inform Sci*, 537:132-147. https://doi.org/10.1016/j.ins.2020.05.070

Presotto R, Civitarese G, Bettini C, 2022. FedCLAR: federated clustering for personalized sensor-based human activity recognition. Proc IEEE Int Conf on Pervasive Computing and Communications, p.227-236. https://doi.org/10.1109/PerCom53586.2022.9762352

Sozinov K, Vlassov V, Girdzijauskas S, 2018. Human activity recognition using federated learning. Proc IEEE Int Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, p.1103-1111. https://doi.org/10.1109/BDCloud.2018.00164

Stisen A, Blunck H, Bhattacharya S, et al., 2015. Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. Proc 13th ACM Conf on Embedded Networked Sensor Systems, p.127-140. https://doi.org/10.1145/2809695.2809718

Tang CI, Perez-Pozuelo I, Spathis D, et al., 2021. SelfHAR: improving human activity recognition through self-training with unlabeled data. *Proc ACM Interact Mob Wear Ubiq Technol*, 5(1):36. https://doi.org/10.1145/3448112

Wang HY, Yurochkin M, Sun YK, et al., 2020. Federated learning with matched averaging. Proc 8th Int Conf on Learning Representations.

Wu JH, Sun J, Song J, et al., 2022. Health assessment method based on multi-sign information fusion of body area network. *Inform Sci*, 618:136-149. https://doi.org/10.1016/j.ins.2022.10.033

List of supplementary materials

1 Extension of the experiment section

2 Ablation studies

Table S1 Comparison of the performance of pretraining methods

Table S2 Communication exchange cost for one round in MotionSense

Table S3 Accuracy of FedCoad components in ablation studies

Fig. S1 T-SNE of the centralized setting and FedCoad in the MotionSense dataset

Fig. S2 T-SNE of the centralized setting and FedCoad in the WISDM dataset

Fig. S3 T-SNE of the centralized setting and FedCoad in the HHAR dataset

Fig. S4 Effect of temperature τ

Fig. S5 Effect of client availability on FedAvg and FedCoad performance