Breaking Data Silos in Parkinson's Disease Diagnosis: An Adaptive Federated Learning Approach for Privacy-Preserving Facial Expression Analysis

Meng Pang[1], Houwei Xu[1], Zheng Huang[1], Yintao Zhou[1], Wei Huang[1, 2*], Binghui Wang[3]

[1]Nanchang University
[2]Yichun University
[3]Illinois Institute of Technology
pangmeng1992@gmail.com, 419100230103@email.ncu.edu.cn, huangzheng@email.ncu.edu.cn,
yintaozhou@email.ncu.edu.cn, 060101@e.ntu.edu.sg, bwang70@iit.edu

Abstract

The early diagnosis of Parkinson's disease (PD) is crucial for potential patients to receive timely treatment and prevent disease progression. Recent studies have shown that PD is closely linked to impairments in facial muscle control, resulting in characteristic "masked face" symptoms. This discovery offers a novel perspective for PD diagnosis by leveraging facial expression recognition and analysis techniques to capture and quantify these features, thereby distinguishing between PD patients and non-PD individuals based on their facial expressions. However, concerns about data privacy and legal restrictions have led to significant "data silos", posing challenges to data sharing and limiting the accuracy and generalization of existing diagnostic models due to small, localized datasets. To address this issue, we propose an innovative adaptive federated learning approach that aims to jointly analyze facial expression data from multiple medical institutions while preserving data privacy. Our proposed approach comprehensively evaluates each client's contributions in terms of gradient, data, and learning efficiency, overcoming the non-IID issues caused by varying data sizes or heterogeneity across clients. To demonstrate the real-world impact of our approach, we collected a new facial expression dataset of PD patients in collaboration with a hospital. Extensive experiments validate the effectiveness of our proposed method for PD diagnosis and facial expression recognition, offering a promising avenue for rapid, non-invasive initial screening and advancing healthcare intelligence.

Introduction

Parkinson's disease (PD) is a prevalent neurodegenerative disorder characterized by subtle early symptoms that make diagnosis particularly challenging. Typically, when patients exhibit overt motor impairments, the disease has often progressed to an advanced stage, severely impacting their daily life and work capabilities (Jankovic and Tan 2020). It is noteworthy that PD is not exclusive to the elderly, with a wide range of onset ages from adolescents to the elderly. Indeed, the proportion of early-onset PD patients in the general population has reached 5% to 10% (Post et al. 2020), highlighting the practical significance of conducting intensive research on this disease.

*Corresponding Author

Figure 1: A comparison of six expressions between normal people and PD patients. To protect patient privacy, we obscure the patients' eyes.

To timely intervene in the progression of PD and delay its advancement, the primary task is to achieve accurate clinical diagnosis of early-stage PD. Currently, diagnostic methods for PD are mainly classified into two categories (Giuliano, Cerri, and Blandini 2021): in-vivo diagnosis and in-vitro diagnosis. In-vivo diagnosis (Khachnaoui, Mabrouk, and Khlifa 2020; Školoudík et al. 2020; Sioka, Fotopoulos, and Kyritsis 2010) relies on specialized medical imaging techniques like CT, MRI, TCS, and PET, enabling doctors to assess PD severity by comparing images of specific tissues or regions with those of non-patients. While accurate, in-vivo diagnosis also presents certain limitations: it requires professional equipment operated by trained professionals, and some patients, such as pregnant women or those with specific medical conditions, may be unsuitable for certain scans. Additionally, the high cost poses a burden on ordinary families. In contrast, in-vitro diagnosis (Tuncer, Dogan, and Acharya 2020; Lilhore et al. 2023; Balaji et al. 2021; Gomez et al. 2021), which analyzes external biomarkers like voice and gait signals, is more convenient and flexible. This type of method does not depend on expensive equipment, is easy to administer, and has fewer patient restrictions, making it increasingly popular among doctors and researchers. With technological advancements, in-vitro diagnosis is poised to become a crucial tool for early PD detection.

Facial expressions, as vital external biomarkers, contain rich health information within subtle changes, offering a new perspective for non-invasive disease diagnosis. Literature analysis (Bek, Poliakoff, and Lander 2020; Mattavelli et al. 2021) reveals that neurodegenerative diseases like PD

are closely linked to facial muscle control disorders, leading to stiffness, unnaturalness, or reduced expressions, manifesting as "masked faces" symptoms. Figure 1 compares six basic facial expressions of normal individuals and PD patients, highlighting PD patients'tendency to exhibit "masked faces" with nearly identical expressions across different emotions, thus enabling the identification of potential PD patients through remote facial expression recognition.

However, training PD diagnostic models based on facial expression recognition faces challenges due to data privacy protection and legal restrictions, leading to an obvious "data silos" phenomenon that hinders data sharing and the depth of "face-reading for diagnosis" research. Limited by small local datasets, existing models lack generalization and accuracy for clinical application. To address these, this paper proposes a novel adaptive federated learning (FL) method based on client contributions to collaboratively compute privacy using facial expression data from multiple medical institutions. It evaluates client contributions from gradient contribution, data quality, and learning efficiency perspectives, overcoming non-IID issues caused by possible data heterogeneity in multi-source environments or varying data sizes. Specifically, during federated model training, gradient contribution is measured by the angular deviation between local and global gradients, with higher weights assigned to clients with significant differences to ensure model robustness. Data quality is assessed by local data loss during training, and client learning efficiency is evaluated by monitoring the training loss descent rate. These multi-dimensional indicators guide the adaptive adjustment of each client's uploaded model weights, directing the global model to converge efficiently toward optimality. To demonstrate the practical value of the proposed method, we cooperated with the Second Affiliated Hospital of Nanchang University to create the Parkinson's Disease Facial Expression (PDFE) dataset, which is currently the largest known dataset of facial expressions in Parkinson's disease. It includes 95 patients, each with seven images depicting neutral emotions and six other basic emotions (surprise, fear, anger, sadness, happiness, disgust) for experimental verification.

The contributions of the paper are summarized as follows:

- Proposing a non-invasive in-vitro diagnostic technology for Parkinson's disease based on facial expression recognition, aiming to address accessibility challenges faced by patients, especially in resource-scarce or mobility-limited situations. This is the first attempt to apply FL to enhance diagnostic model training in this field.
- Design an innovative adaptive FL framework to tackle data silos in PD diagnosis. This framework collaboratively performs privacy computing on patient facial expression data from multiple institutions, comprehensively evaluating data contributions from three dimensions to overcome non-IID issues.
- Creating the largest facial expression dataset for in-vitro PD diagnosis research, featuring multiple expressions from 95 PD patients.
- Demonstrating the effectiveness of our method through extensive experiments, offering a promising approach for rapid, non-invasive initial screening and advancing healthcare intelligence.

Related Work

PD Diagnosis: The diagnosis of PD primarily falls into two categories: in-vivo and in-vitro methods. In-vivo diagnosis relies on direct measurements of physiological and biochemical indicators, offering high accuracy but entailing expensive costs and potential harm to patients. In contrast, in-vitro diagnosis is more convenient and flexible, auxiliarily diagnosing through analysis of external markers such as voice and gait signals. Research (Sakar et al. 2019) found that PD affects facial muscle activity and coordination, leading to early speech disorders, making voice signals a crucial diagnostic tool. Lilhore *et al.* combined Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models to identify voice characteristics of PD patients using Mel spectrograms (Lilhore et al. 2023), surpassing traditional methods in accuracy. However, the diagnostic effectiveness of voice signals varies among patients with different native languages (Hsu et al. 2017). Research (Bloem et al. 2004) found that PD patients also exhibit symptoms of bradykinesia and postural instability. Based on this, Balaji *et al.* used vertical ground reaction force sensors to collect gait signals and diagnosed PD through the Support Vector Machine algorithm (Balaji et al. 2021), albeit with potential additional burden on patients from wearing sensors.

Given the limitations of gait and voice signals in PD diagnosis, researchers are now exploring facial expressions—easily collected, universally applicable, and language-independent—as a novel diagnostic approach. Studies have shown that PD patients display subtle facial muscle movements, restricted expressions, and impaired eye movements, characterizing a "masked face" (Bandini et al. 2017). Researchers like Rajnoha *et al.* (Rajnoha et al. 2018) have used static facial images for decision tree classification, while Jin *et al.* (Jin et al. 2020) extracted features from facial smile videos. Huang *et al.* (Huang et al. 2023) proposed a method that leverages deep learning models, such as Swin Transformer, to address the binary classification of complex features in PD patients' six basic facial expressions.

Despite progress in PD's in-vitro diagnosis based on facial expression recognition, a pressing challenge remains: most existing diagnostic models are trained on local, small sample data, limiting model generalization and prone to overfit, thereby reducing actual diagnostic accuracy. This issue primarily stems from the privacy and legal protection of facial expression data of PD patients, making data collection extremely difficult. Simultaneously, concerns over data leakage prevent different medical institutions from sharing their facial expression datasets, exacerbating the "data silos" problem. To address this, this paper proposes, for the first time, a PD diagnosis method based on a FL framework for multi-source facial expression data. This method enables collaborative model training across multiple clients while protecting patient privacy, achieving a high PD diagnosis accuracy of 98.35%.

Federated Learning: FL is a machine learning approach that enhances model performance by leveraging distributed

data resources while ensuring privacy and security (Wang et al. 2022). It is crucial in the medical field due to the highly sensitive nature of medical data and strict privacy regulations like GDPR and HIPAA. FedAvg (McMahan et al. 2017) is a classic optimization algorithm in FL, which updates and maintains a global model by collecting and averaging local model parameters from participating nodes. FedProx (Li et al. 2020) introduces a regularization term based on FedAvg to limit the update range of model parameters during local training, enhancing global model stability. MOON (Li, He, and Song 2021) incorporates contrastive learning to align local model updates with global optimization, significantly improving global model performance and convergence speed. FedSol (Lee et al. 2024) proposes a new framework for federated stabilized orthogonal learning, aiming to identify stable parameter regions and ensure orthogonality between local and proximal gradients, allowing for effective model updates while minimizing global model perturbation. Distinct from these methods, we propose an adaptive FL approach based on contribution estimation, which evaluates client contributions and improves model aggregation performance. Our method safeguards patient privacy and achieves promising performance in PD diagnosis.

The Proposed Method

Problem Definition

Suppose there are K clients, where the private data distribution of client k is denoted by $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{n_k}$, with $n_k = |\mathcal{D}_k|$ indicating the number of samples. The local model parameters for client k are w_k, and its optimization objective is $F(w_k)$. In the joint training among hospitals for Parkinson's disease diagnosis, the collected data can significantly differ in feature distribution, labeling criteria, and sample size, resulting in different local objectives: $F(w_{k1}) \neq F(w_{k2})(k1 \neq k2)$. Therefore, coordinating the local objectives $F(w_k)$ and adaptively adjusting aggregation weights in the global aggregation poses a significant challenge. In global aggregation, defined as $F = argmin\mathcal{G}(F(w_{k=1}), ..., F(w_{k=K}))$, the proportion of each client's sample size is commonly used as the aggregation weight. The global optimization objective is given by: $\mathcal{G}(F(w_{k=1}), ..., F(w_{k=K})) = \sum_k \frac{|\mathcal{D}_k|}{\sum_k |\mathcal{D}_k|} F(w_k)$.

To address this challenge, we propose an adaptive FL method, illustrated in Figure 2. This method evaluates the contributions of each medical institution based on three factors: data, gradient, and learning efficiency. It uses these contributions to adjust each client's aggregation weight during model merging, promoting fairness and efficiency, and encouraging medical institutions to engage in joint training.

Overview of the Algorithm

Traditional FL approaches often mimic FedAvg (McMahan et al. 2017), using client sample sizes directly as aggregation weights, neglecting the distinction between sample quantity and data value. This bias favors clients with larger datasets, leading to unfairness in the global model. Focusing on data contribution and gradient contribution mitigates this issue by prioritizing clients with rare distributions or poor model

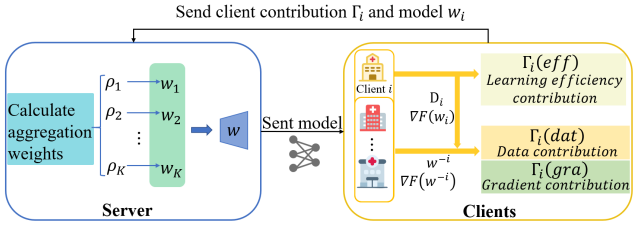

Figure 2: Flowchart of the proposed algorithm.

performance (Jiang et al. 2023). Furthermore, learning efficiency is vital in evaluating client value, with poor data quality impeding local model fitting and high-quality data expediting convergence(Fang and Ye 2022). In cases where participants contribute low-value or maliciously distorted data and gradients, learning efficiency can counteract these negatives, complementing data and gradient contributions. By integrating these three factors, the server can more accurately assess clients' true value, ensuring fair representation and enhancing the global model's robustness and effectiveness.

When evaluating client contributions, the primary consideration is assessing individual contributions to the whole. In the current federated contribution evaluation, the Shapley value is the most widely used measure, and the method for calculating the contribution of client i is as follows:

$$v(i; U, D, N) = \mathop{\mathbb{E}}_{M\sim[N], S\sim D^{M-1}}[U(S \cup i) - U(S)], \quad (1)$$

where D represents the data distribution, N denotes the number of participating clients, $D^N = \{D_i\}_i^N$ indicate the data distribution set of all subset length N, S represents the client set, if S set has M elements, the corresponding data distribution of all possible cases is denoted as D^M, namely $S \sim D^M, |S| = M, U$ represents the utility function, which is used to calculate the contribution value of set S.

Based on the Shapley value definition, calculating client i's contribution requires enumerating its marginal contributions across all subsets excluding i, which is computationally expensive. To simplify this, Wang *et al.* (Wang, Dang, and Zhou 2019) introduced the leave-one-out method, dividing the set into two parts. The contribution of a party is then measured as the value lost when it is excluded (often using test set accuracy as the utility function). Although this simplification may be imperfect and unfair for similar, interchangeable parties, it does not affect our evaluation scheme. Incorporating the leave-one-out concept, our contribution measurement strategy is as follows:

$$\hat{v}(i, \Gamma, N) = \Gamma_i(N \backslash \{i\}, \{i\}), \quad (2)$$

where $N\backslash\{i\}$ signifies the exclusion of client i from the total set of N clients. Γ_i represents our introduced utility function, which comprehensively evaluates the client's value across three dimensions: gradient, data, and learning efficiency. Given the varying significance of individual contributions in diverse real-world scenarios, we incorporate hyperparameters λ to adjust their respective weights, thereby enhancing the system's adaptability and efficiency.

Algorithm 1: Adaptive FL

Input: Communication rounds T, number of clients K, local datasets D_k,learning rate η, Contribution ratio λ, global batch size B.
Output: Final global model w_T.
Server:

Initialize server model w_0
for t=0,1,2,...,T-1 **do**
  $B_i = \frac{|\mathrm{D}_i|}{\sum_{i \in K} |\mathrm{D}_i|} \times B$ ▷ $Sec4.1$
  send: $w_{t,i}, b_i \leftarrow w_t, B_i$
  for $i \in K$ **do**
    $(\Gamma_{t,i}(gra), \Gamma_{t,i}(dat), \Gamma_{t,i}(eff)), w_{t,i}$ ←**Client Update**
    calculate $\rho_{t,i}$ ▷ $Eq.(3)\, and\, Eq.(4)$
    $w_{t+1} \leftarrow \sum_{i=1}^{K} \rho_{t,i} w_{t,i}$
  end for
end for

Client Update:

$\nabla F\left(w_t\right) = w_t - w_{t-1}$.
$w_t^{-i} = \frac{w_t - \rho_{t-1,i} w_{t-1,i}}{1-\rho_{t-1,i}}$.
(x_i, y_i): randomly sample b_i samples.
$\nabla F\left(w_{t,i}\right) = -\eta \nabla F\left(x_i, y_i, w_{t,i}\right)$
$w_{t,i} = w_{t,i} + \nabla F\left(w_{t,i}\right)$
$\nabla F(w_t^{-i}) = \frac{\nabla F(w_t) - \rho_{t-1,i} \nabla F(w_{t,i})}{1-\rho_{t-1,i}}$
$\Gamma_{t,i}(gra) = 1 - \cos\left(\nabla F\left(w_{t,i}\right), \nabla F\left(w_t^{-i}\right)\right)$ ▷ $Eq.(5)$
$\Gamma_{t,i}(dat) = \varepsilon\left(\mathrm{D}_i, w_t^{-i}\right)$ ▷ $Eq.(6)$
$\Gamma_{t,i}(eff) = \frac{\mathcal{L}_{i,t-1} - \mathcal{L}_{i,t}}{\mathcal{L}_{i,t}}$ ▷ $Eq.(7)$
Upload $(\Gamma_{t,i}(gra), \Gamma_{t,i}(dat), \Gamma_{t,i}(eff))$ and $w_{t,i}$ to server.

Ultimately, Γ_i is computed as follows:

$$\Gamma_i = \lambda_1 \Gamma_i(gra) + \lambda_2 \Gamma_i(dat) + \lambda_3 \Gamma_i(eff), \qquad (3)$$

where $\Gamma_i(gra)$, $\Gamma_i(dat)$, and $\Gamma_i(eff)$ are gradient contribution, data contribution, and learning efficiency contribution, respectively. λ_1, λ_2 and λ_3 are the hyperparameter used to balance the individual contributions and λ_1+λ_2+λ_3=1.

By weighting the three dimensions of the client's contribution (each contribution is normalized after the server receives the client's contribution, in the case of gradient contribution, $\Gamma_i(gra) = \frac{\Gamma_i(gra)}{\sum_{i=1}^{K} \Gamma_i(gra)}$, we obtain the true contribution of client i and use this contribution as a weight parameter for the central server aggregation model. We similarly normalize the Γ_i to obtain the final aggregated weight ρ_i, which is calculated as

$$\rho_i = \frac{\Gamma_i}{\sum_{i=1}^{K} \Gamma_i}. \qquad (4)$$

To achieve more precise predictions, we also leverage the client's past average contributions to dynamically adjust its current contribution. For clarity, the detailed algorithmic procedure is outlined in Algorithm 1.

Client Contribution Estimation

This section will describe in more detail the computation process of the three contributions we propose, namely gradient contribution, data contribution, and learning efficiency contribution.

Gradient contribution. The data distribution and feature differences among different clients can often be significant, and this heterogeneity affects the gradient direction of each client and its contribution to the global model. To leverage these differences, we dynamically adjust the weight assignment based on the gradient direction discrepancies between clients, aiming to optimize the performance of the global model. Specifically, we assess the similarities and differences in data characteristics and distribution of each client by calculating the gradient direction discrepancy between one client and all others. The process is formulated as

$$\Gamma_{t,i}(gra) \triangleq 1 - \cos\left(\nabla F\left(w_{t,i}\right), \nabla F\left(w_t^{-i}\right)\right), \qquad (5)$$

where $\cos(\cdot)$ represents cosine similarity, $\nabla F(w_{t,i})$ represents the parameter change after local training when client i is in round t, and $\nabla F(w_t^{-i})$ represents the gradient aggregation value excluding the gradient of client i when all client gradients are aggregated, that is, the gradient sum of other clients. When the gradient direction of one client significantly differs from others, it indicates that it may have explored data features or distributions not covered by other clients, thus deserving a higher level of information and contribution to the training of the global model. In this way, FL not only aggregates data from different locations but also leverages heterogeneity to enhance the robustness and generalization ability of the model, enabling it to better adapt to the constantly changing and complex data environment in the real world.

Data contribution. When dealing with multiple clients or data sources, some may exhibit unique features and data distributions. Thus, a model's poor performance on a specific client may not stem from the model itself but from the distinctiveness of the client's data. Consequently, we assess the client's data contribution as a metric by evaluating the error rate on their local training dataset when using a model aggregated from other clients. This error rate ε quantifies the client's data contribution, defined as follows:

$$\Gamma_{t,i}(dat) \triangleq \varepsilon\left(\mathcal{D}_i, w_t^{-i}\right), \qquad (6)$$

where w_t^{-i} denotes the model derived by excluding client i during the server's aggregation in round t. If w_t^{-i} performs poorly on the data from client i, it indicates that the server model inadequately captures the attributes of that client's data distribution. Hence, assigning a higher weight to client i during model training is necessary to ensure the model more accurately captures and reflects the characteristics of that data source. This strategy enhances not only the model's generalization ability in heterogeneous data environments but also its fairness across data sources.

Learning efficiency contribution. High-quality data enhances the model's learning efficiency, making learning efficiency a valuable metric for assessing client value. We measure learning efficiency by the rate of change in the model's

loss value within the client's local training set, calculated as

$$\Gamma_{t,i}(eff) \triangleq \frac{\mathcal{L}_{i,t-1} - \mathcal{L}_{i,t}}{\mathcal{L}_{i,t}}, \tag{7}$$

where $\mathcal{L}_{i,t}$ represents the loss value of client i on the local training set at round t. Poor data quality slows down the model fitting process, directly impacting model performance. Conversely, clients with high-quality data exhibit faster training and model convergence. This disparity influences not only individual client performance but also affects the optimization of the entire system. Thus, emphasizing learning efficiency and the underlying data quality is crucial for enhancing overall model training effectiveness.

Experiments

Experimental Setup

Datasets and Models In the task of diagnosing Parkinson's Disease (PD), we utilized five datasets: PDFE, Oulu-CASIA (Zhao et al. 2011), RaFD (Langner et al. 2010), CK+ (Lucey et al. 2010), and Tsinghua-FED (Yang et al. 2020). The PDFE dataset, newly created by our research team, consists of images of 95 PD patients (55 males and 40 females, with an average age of 62.7) captured using a digital camera, displaying neutral and six basic emotions. The data collection and usage were approved by all participating patients. The remaining four datasets are publicly available facial expression datasets featuring healthy individuals. To showcase the versatility and superiority of our proposed method compared to existing Federated Learning (FL) techniques, we benchmarked it against the CIFAR-10 (Krizhevsky 2009) and FMNIST (Xiao, Rasul, and Vollgraf 2017) datasets. For PD diagnosis, we employed the ResNet-18 model, whereas for CIFAR-10 and FMNIST, we utilized a standard 4-layer Simple-CNN (Luo et al. 2021). All experiments were conducted on a machine equipped with an Intel(R) Core i5-13400F processor, 16GB of memory, an NVIDIA 4060 GPU, Ubuntu 22.04.5 LTS, and PyTorch version 2.1.

Evaluation Protocol and Configuration For the task of PD diagnosis, we divided the PDFE dataset into five folds, using one as the test set (augmented with elderly data from Tsinghua-FED as normals) and the rest four for PD patient training. The training data was distributed to three fully participating clients with a batch size of 30. The final results were averaged over five experiments. For image classification task on CIFAR-10 and FMNIST datasets, we split the data into 25% test and 75% training sets. The training data was distributed to 20 fully participating clients with a batch size of 64, and each partitioned dataset was stored for consistency. With local epochs set to 1, a learning rate of 0.005 was suitable for most methods, while our proposed method used 0.05. We reported the highest stable accuracy after sufficient training rounds for all methods.

Notably, our proposed algorithm features two specific configurations distinct from conventional FL approaches:

- Firstly, we observed that increasing the local update rounds (the number of local model updates per client per communication round, typically set as $n_i/batch \times epoch$) does not necessarily improve results. While larger local update rounds can accelerate training, they can also lead to the overfitting of local data, reducing overall performance. Therefore, we cautiously set the local update rounds in our algorithm at 1.
- Secondly, unlike many FL methods that set a uniform local batch size for different clients, we tailor the batch size for each client based on their data proportion to the global dataset, set as $n_i/total \times B$, where n_i is the number of samples for client i, $total$ is the sum of samples across all clients, and B is the global batch size. This design aims to integrate the influence of data volume into the local client training process prior to model aggregation, which contrasts with many other FL methods relying on simple weighting in the final model aggregation stage. Our approach strives to better simulate centralized training, potentially leading to faster convergence towards the global optimum.

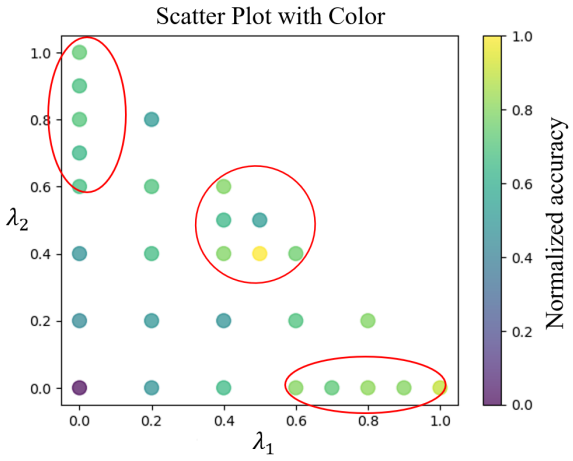

Figure 3: Coarse-grained grid search results of λ_1 and λ_2.

For the rationale behind these specific configurations, please refer to Parts 1 and 2 of the ablation experiments.

Hyperparameter Setting To balance the three contributions, we conducted a two-step grid search on the CIFAR-10 dataset to determine the hyperparameters λ_1, λ_2, and λ_3, ensuring λ_1+λ_2+λ_3=1. After setting λ_1 and λ_2, we calculated λ_3 as 1-λ_1-λ_2. Initially, a coarse-grained search with a 0.2 interval revealed high performance in three regions (refer to Figure 3), suggesting improved performance with greater weights on data and gradient contributions, alongside significant learning efficiency. Subsequently, a fine-grained search within these regions, using a 0.1 interval, identified optimal values as λ_1=0.5, λ_2=0.4, and λ_3=0.1. These values were then fixed for subsequent experiments.

Evaluation on Ordinary Image Classification

In this section, we evaluate the classification performance of our method on the CIFAR-10 and FMNIST datasets and compare it with current mainstream FL methods, including

Method	CIFAR-10			FMNIST		
	a=0.1	a=0.5	IID	a=0.1	a=0.5	IID
FedAvg	65.69	66.42	71.91	89.76	89.88	91.89
FedProx	65.56	66.25	71.89	89.79	89.87	91.85
MOON	65.41	66.44	71.91	89.75	89.91	91.82
FedSol	66.09	67.56	71.60	89.83	88.35	92.14
Ours	**72.87**	**71.45**	**72.95**	**91.49**	**91.46**	**92.31**

Table 1: Classification accuracy of different FL methods on the CIFAR-10 and FMNIST datasets.

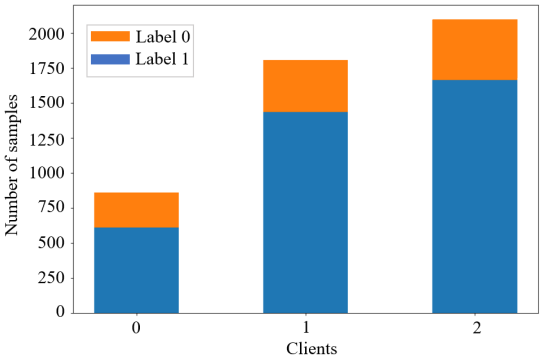

Figure 4: Label distribution in non-IID for the amount of data, Label 0 represents PD patients.

PD Diagnosis Method	Accuracy(%)↑
GLCM+SVM	46.60
DeiT-small	71.93
EfficientNetV2	71.25
InceptionResnetV1	84.07
ResNet18-DataAugment	95.43
Ours	**98.35**

Table 2: PD diagnosis accuracies (%) of different PD diagnosis methods on the PDFE dataset.

FedAvg (McMahan et al. 2017), FedProx (Li et al. 2020), Moon (Li, He, and Song 2021), and FedSol (Lee et al. 2024) in both IID and non-IID scenarios. To control the degree of non-IID, we adjust the parameter a of the Dirichlet distribution, with smaller values indicating more severe non-IID conditions (Lin et al. 2020). It is observed from Table 1 that our method achieves the highest classification accuracy on both datasets, especially in non-IID scenarios. For instance, when a=0.1, our method's classification accuracy is 6.78% higher than the second-best method, FedSol. The results demonstrate the rationality and effectiveness of our method in evaluating each client by measuring gradient contribution, data contribution, and learning efficiency contribution, which enhances the global aggregation performance of the model.

Evaluation on PD Diagnosis

In this subsection, we first evaluate the performance of our method in diagnosing PD under IID training scenario and compare it with the advanced PD diagnosis methods, including GLCM+SVM (Hou, Qin, and Su 2022), ResNet18-DataAugment (Huang et al. 2022), and three latest deep learning models: DeiT-small (Touvron et al. 2021), EfficientNetV2 (Tan and Le 2021), and InceptionResnetV1 (Szegedy et al. 2017). Table 2 presents the PD prediction accuracy of all comparative algorithms on the PDFE dataset. As shown in Table 2, our method attains the highest accuracy of 98.35%, surpassing the current state-of-the-art method, ResNet18-DataAugment, by 2.92%. These results demonstrate that our PD diagnosis model, trained using multi-source facial expression data based on FL, exhibits superior performance while also protecting patient privacy compared to existing methods.

Considering that the multi-source data distributions provided by different medical institutions in reality are likely to exhibit non-IID characteristics, which may arise from variations in data scale or image heterogeneity caused by different collection conditions. Therefore, we will evaluate the performance of our method in PD diagnosis under these two non-IID scenarios, respectively.

(1) Non-IID caused by data scale. To simulate the first scenario, we adjusted the parameter a of the Dirichlet distribution to induce a non-IID characteristic in the data distribution across clients. Figure 4 illustrates the data distribution for three clients when a=1, in which we observe that the amount of data varies across these three clients, and there are also significant differences in the label distribution on each client. Table 3 reports the PD diagnosis results of our method under the non-IID training scenarios when a takes values of 100, 10, and 1, respectively [1]. As shown in Table 3, our method maintains stable performance in PD diagnosis, as the degree of non-IID increases from a=100 to 1, with only a 0.36% decrease in PD diagnosis accuracy. The results verify the robustness of our method in the non-IID scenario caused by varying volumes of multi-source data.

(2) Non-IID caused by image heterogeneity. To simulate the second scenario, we applied transformations to the lightness, contrast, and saturation of the training images across different clients. Through observation, we found that altering the lightness resulted in the most significant differences in the distribution of image features, while changes in contrast and saturation had a much smaller impact on the vari-

[1]Due to the small size of PDFE, a should not be set below 1 to avoid clients without data assignment.

a	Accuracy	Variations	Accuracy
1	97.78	Lightness	97.20
10	98.15	Contrast	97.31
100	98.14	Saturation	98.03

Table 3: PD diagnosis accuracy (%) of the proposed method under non-IID training scenarios.

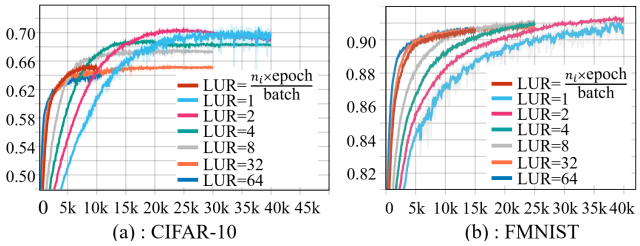

Figure 5: Impact of LUR on model performance.

ation in feature distribution. Table 3 reports the PD diagnosis results of our method in the non-IID scenario caused by image heterogeneity, demonstrating excellent performance across three cases. Notably, even in the lighting variation case where image heterogeneity is most pronounced, our method still achieves a PD diagnosis accuracy of 97.20%.

In summary, the experiments have demonstrated the effectiveness of our proposed adaptive FL method in PD diagnosis, with the potential to address the "data silos" issue in current facial expression-based models. Furthermore, they have proven the robustness of our method in non-IID training scenarios due to data scale and image heterogeneity, maintaining accurate diagnosis. The superiority is primarily owing to its ability to guide adaptive model weight adjustment based on multi-dimensional indicators (data, gradient, learning efficiency), enabling efficient convergence towards optimality.

Ablation Study

(1) Impact of local update rounds (LUR). We conducted ablation studies on CIFAR-10 (a = 0.1) and FMNIST (a = 0.1) datasets to examine the impact of various LUR, including 1, 2, 4, 8, 32, 64, and the default $n_i/batch \times epoch$ commonly used in FL methods, on model performance. As shown in Figure 5, our approach demonstrated good classification results on both datasets, notably with fewer LUR (e.g., LUR=1 or 2). While larger LUR can accelerate training, they may also lead to the overfitting of local data, reducing overall performance.

(2) Impact of local batch size (LBS). We conducted an ablation experiment on CIFAR-10 (a = 0.1) by varying the LBS. Initially, referencing traditional FL methods like FedAvg and TCT (Yu et al. 2022), we fixed the LBS for each client at 64, and compared it with our approach where LBS is set proportionally to the number of clients, specifically LBS=$n_i/total \times 64$. As illustrated in Figure 6, when the batch size is controlled according to our method, there is a slight improvement in model performance after convergence. This confirms the effectiveness of our approach to control LBS based on the proportion of samples among clients, mimicking centralized training.

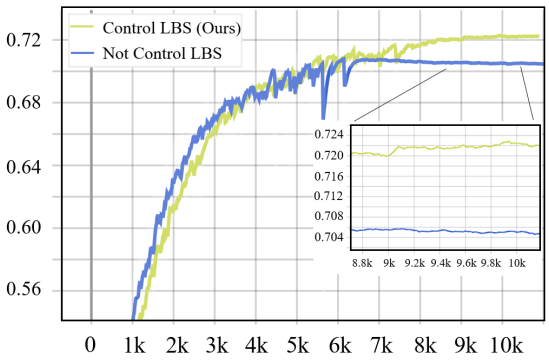

Figure 6: Impact of LBS on model performance.

Gradient	Data	Learning efficiency	Accuracy(%)
✓			96.89
	✓		96.62
		✓	96.14
✓	✓	✓	**98.35**

Table 4: Ablation results for each contribution factor.

(3) Ablation study on contribution factors. To evaluate the impact of various contributing factors on the diagnostic performance of PD, ablation experiments were conducted on the PDFE dataset, as detailed in Table 4. It is observed that solely relying on gradient, data, or learning efficiency contribution leads to a decrease in PD diagnosis accuracy by 1.46%, 1.73%, and 2.21%, respectively. The results indicate that utilizing a single contribution factor is insufficient, emphasizing the effectiveness of combining multiple factors to achieve superior PD diagnosis with our model.

Conclusion

This paper underscores the potential of facial expression recognition as a diagnostic tool for PD and proposes an adaptive FL model based on client contribution evaluation. The aim is to facilitate collaborative learning by integrating facial expression data from multiple sources of PD patients, addressing the "data silo" issue in PD diagnostic model training while effectively protecting patient data privacy. This represents the first attempt to utilize FL in this diagnostic context. Our method evaluates client contributions based on gradients, data, and learning efficiency, addressing challenges posed by non-IID data due to variations in data scale or heterogeneity. Experiments on CIFAR-10 and FMNIST showcase our method's superiority over other advanced FL approaches, and tests on our newly created PDFE dataset confirm its effectiveness in PD diagnosis, advancing intelligent and remote PD diagnosis.

Acknowledgments

This work is supported in part by Natural Science Foundation of China (62466036, 62271239), by Natural Science Foundation of Jiangxi Province (20232BAB212025), by High-level and Urgently Needed Overseas Talent Programs of Jiangxi Province (20232BCJ25024), by Jiangxi Double Thousand Plan (JXSQ2023201022), by Jiangxi Provincial Key Laboratory of Data Security Technology (20242BCC32026), and by Key Research and Development Program of Jiangxi Province (20243BBG71035).

References

Balaji, E.; Brindha, D.; Elumalai, V. K.; and Umesh, K. 2021. Data-driven gait analysis for diagnosis and severity rating of Parkinson's disease. *Med. Eng. Phys.*, 91: 54–64.

Bandini, A.; Orlandi, S.; Escalante, H. J.; Giovannelli, F.; Cincotta, M.; Reyes-Garcia, C. A.; Vanni, P.; Zaccara, G.; and Manfredi, C. 2017. Analysis of facial expressions in parkinson's disease through video-based automatic methods. *J. Neurosci. Methods*, 281: 7–20.

Bek, J.; Poliakoff, E.; and Lander, K. 2020. Measuring emotion recognition by people with Parkinson's disease using eye-tracking with dynamic facial expressions. *J. Neurosci. Methods*, 331: 108524.

Bloem, B. R.; Hausdorff, J. M.; Visser, J. E.; and Giladi, N. 2004. Falls and freezing of gait in Parkinson's disease: a review of two interconnected, episodic phenomena. *Mov. Disord.*, 19(8): 871–884.

Fang, X.; and Ye, M. 2022. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10072–10081.

Giuliano, C.; Cerri, S.; and Blandini, F. 2021. Potential therapeutic effects of polyphenols in Parkinson's disease: in vivo: and: in vitro: pre-clinical studies. *Neural Regener. Res.*, 16(2): 234–241.

Gomez, L. F.; Morales, A.; Orozco-Arroyave, J. R.; Daza, R.; and Fierrez, J. 2021. Improving parkinson detection using dynamic features from evoked expressions in video. In *CVPR*, 1562–1570.

Hou, X.; Qin, S.; and Su, J. 2022. Visual detection of Parkinson's disease via facial features recognition. In *CIAC*, 249–257.

Hsu, S.-C.; Jiao, Y.; McAuliffe, M. J.; Berisha, V.; Wu, R.-M.; and Levy, E. S. 2017. Acoustic and perceptual speech characteristics of native Mandarin speakers with Parkinson's disease. *J. Acoust. Soc. Am.*, 141(3): EL293–EL299.

Huang, W.; Xu, W.; Wan, R.; Zhang, P.; Zha, Y.; and Pang, M. 2023. Auto diagnosis of Parkinson's disease via a deep learning model based on mixed emotional facial expressions. *IEEE J. Biomed. Health.*, 28: 2547–2557.

Huang, W.; Zhou, Y.; Cheung, Y.-m.; Zhang, P.; Zha, Y.; and Pang, M. 2022. Facial Expression Guided Diagnosis of Parkinson's Disease via High-Quality Data Augmentation. *IEEE T. Multimedia*, 25: 7037–7050.

Jankovic, J.; and Tan, E. K. 2020. Parkinson's disease: etiopathogenesis and treatment. *J. Neurol., Neurosurg. Psychiatry*, 91(8): 795–808.

Jiang, M.; Roth, H. R.; Li, W.; Yang, D.; Zhao, C.; Nath, V.; Xu, D.; Dou, Q.; and Xu, Z. 2023. Fair federated medical image segmentation via client contribution estimation. In *CVPR*, 16302–16311.

Jin, B.; Qu, Y.; Zhang, L.; and Gao, Z. 2020. Diagnosing Parkinson disease through facial expression recognition: video analysis. *J. Med. Internet Res.*, 22(7): e18697.

Khachnaoui, H.; Mabrouk, R.; and Khlifa, N. 2020. Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: a review. *IET Image Process.*, 14(16): 4013–4026.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. In *Technical Report*.

Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D. H.; Hawk, S. T.; and Van Knippenberg, A. 2010. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.*, 24(8): 1377–1388.

Lee, G.; Jeong, M.; Kim, S.; Oh, J.; and Yun, S.-Y. 2024. FedSOL: Stabilized Orthogonal Learning with Proximal Restrictions in Federated Learning. In *CVPR*, 12512–12522.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *CVPR*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. In *MLSys*, 429–450.

Lilhore, U. K.; Dalal, S.; Faujdar, N.; Margala, M.; Chakrabarti, P.; Chakrabarti, T.; Simaiya, S.; Kumar, P.; Thangaraju, P.; and Velmurugan, H. 2023. Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease. *Sci. Rep.*, 13(1): 14605.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33: 2351–2363.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 94–101.

Luo, M.; Chen, F.; Hu, D.; Zhang, Y.; Liang, J.; and Feng, J. 2021. No fear of heterogeneity: classifer calibration for federated learning with non-IID data. In *NeurIPS*, 5972–5984.

Mattavelli, G.; Barvas, E.; Longo, C.; Zappini, F.; Ottaviani, D.; Malaguti, M. C.; Pellegrini, M.; and Papagno, C. 2021. Facial expressions recognition and discrimination in Parkinson's disease. *J. Neuropsychol.*, 15(1): 46–68.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 1273–1282.

Post, B.; Van Den Heuvel, L.; Van Prooije, T.; Van Ruissen, X.; Van De Warrenburg, B.; and Nonnekes, J. 2020. Young onset Parkinson's disease: a modern and tailored approach. *J. Parkinson's Dis.*, 10(s1): S29–S36.

Rajnoha, M.; Mekyska, J.; Burget, R.; Eliasova, I.; Kostalova, M.; and Rektorova, I. 2018. Towards identification of hypomimia in Parkinson's disease based on face recognition methods. In *ICUMT*, 1–4.

Sakar, C. O.; Serbes, G.; Gunduz, A.; Tunc, H. C.; Nizam, H.; Sakar, B. E.; Tutuncu, M.; Aydin, T.; Isenkul, M. E.; and Apaydin, H. 2019. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Appl. Soft Comput.*, 74: 255–263.

Sioka, C.; Fotopoulos, A.; and Kyritsis, A. P. 2010. Recent advances in PET imaging for evaluation of Parkinson's disease. *Eur. J. Nucl. Med. Mol. Imaging*, 37: 1594–1603.

Školoudík, D.; Mašková, J.; Dušek, P.; Blahuta, J.; Soukup, T.; Burgetová, A.; and Bártová, P. 2020. Digitized Image Analysis of Insula Echogenicity Detected by TCS-MR Fusion Imaging in Wilson's and Early-Onset Parkinson's Diseases. *Ultrasound Med. Biol.*, 46(3): 842–848.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 4278–4284.

Tan, M.; and Le, Q. 2021. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106. PMLR.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357.

Tuncer, T.; Dogan, S.; and Acharya, U. R. 2020. Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. *Biocybern. Biomed. Eng.*, 40(1): 211–220.

Wang, B.; Li, A.; Pang, M.; Li, H.; and Chen, Y. 2022. Graphfl: A federated learning framework for semi-supervised node classification on graphs. In *ICDM*, 498–507.

Wang, G.; Dang, C. X.; and Zhou, Z. 2019. Measure contribution of participants in federated learning. In *IEEE Big Data*, 2597–2604.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yang, T.; Yang, Z.; Xu, G.; Gao, D.; Zhang, Z.; Wang, H.; Liu, S.; Han, L.; Zhu, Z.; Tian, Y.; et al. 2020. Tsinghua facial expression database–A database of facial expressions in Chinese young and older women and men: Development and validation. *PloS one*, 15(4): e0231304.

Yu, Y.; Wei, A.; Karimireddy, S.; Ma, Y.; and Jordan, M. 2022. Tct: Convexifying federated learning using bootstrapped neural tangent kernels, 2022. *URL https://arxiv. org/abs/2207.06343*.

Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; and PietikäInen, M. 2011. Facial expression recognition from near-infrared videos. *Image Vis. Comput.*, 29(9): 607–619.