Federated Speech Encoder Fine-Tuning for Multilingual Cross Site Parkinson's Speech Classification: Performance Fairness Analysis

Shivani Kolekar[*], Jisoo Shin[*], Haewoon Nam[†], Kyungbaek Kim[*]
[*]*Department of Artificial Intelligence Convergence , Chonnam National University, Gwangju*
[†]*Department of Electronics and Communication Engineering, Hanyang University, Seoul*
South Korea
shivanikolekar@jnu.ac.kr, 0811jisoo@jnu.ac.kr, hnam@hanyang.ac.kr, kyungbaekkim@jnu.ac.kr

***Abstract*—Speech is a promising noninvasive biomarker for early Parkinson's disease (PD), but most PD voice classifiers are trained centrally on single corpora and reported only with global accuracy, obscuring variation across clinical sites and patient subgroups. We study multilingual cross site federated learning for PD versus healthy control (HC) speech classification, where each client is a clinical site that typically also corresponds to a language or multi task dataset, yielding strongly non IID features and heterogeneous label balances. We introduce a subpopulation aware aggregation rule that uses site by diagnosis performance statistics to shape client level contributions during automatic mixed precision fine tuning of a pretrained Wav2Vec2 speech encoder. The server tracks metrics for each site by diagnosis subgroup and upweights clients associated with weaker subpopulations while leaving the client side loop and encoder architecture unchanged. On multilingual PD speech from multi task recording sites, this aggregation strategy keeps global performance metrics closer to a strong FedAvg baseline, while explicitly steering training toward a more balanced performance distribution across sites and diagnosis subgroups.**

Index Terms-Medical AI, Federated learning, Fine-tuning, Speech Classification

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects motor control and speech, with prevalence estimates reaching up to 3% among individuals over 65 years of age. Voice and speech changes are among the earliest and most accessible biomarkers, and a large body of work shows that machine learning models can detect PD from sustained phonation and read or spontaneous speech with competitive accuracy [28]. Clinical speech datasets, however, are typically collected at individual hospitals or research sites, often in different languages and under different recording protocols. Regulatory and ethical constraints make it difficult to pool raw audio centrally, which motivates federated learning (FL) for speech-based PD detection across institutions and languages [3]–[5].

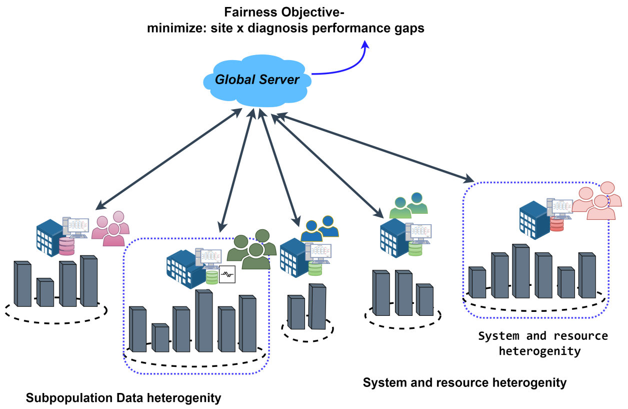
Fairness Objective-
minimize: site x diagnosis performance gaps
Global Server
System and resource heterogenity
Subpopulation Data heterogenity
System and resource heterogenity

Figure 1: Fed SAFE overview: a subpopulation aware federated speech encoder that aggregates site updates using fairness weights from site × diagnosis performance, under non IID PD vs HC speech distributions and heterogeneous network resources across sites.

Federated learning allows multiple sites to train a shared model without sharing raw data, but in practice naïve FL protocols can amplify existing heterogeneity. Differences in language, microphone setup, disease stage, and patient demographics lead to highly non IID partitions. Standard FedAvg, which optimizes an average loss over clients, can exhibit strongly unequal performance across sites and subgroups, even when the global accuracy looks acceptable [12], [16]. This has motivated a growing literature on fair FL, including client-level objectives such as q Fair Federated Learning (q-FFL) that trade off mean accuracy and uniformity across clients [16], fair averaging rules and client reweighting, and healthcare-specific frameworks that unify client-, group-, and sample-level fairness or target particular demographic gaps [12]. Recent surveys emphasize that fairness in FL is multi-dimensional and that group-level criteria are particularly important in high-stakes domains such as digital healthcare.

For PD speech, existing FL studies focus primarily on feasibility and global performance under cross-site heterogeneity. Sarlas et al. report that FL can match or surpass site-specific models while respecting data locality [3]; Tayebi Arasteh et al. train FL models across three European language corpora and show that federated PD detection can approach centrally trained baselines without sharing raw recordings [5]; and Quan et al. propose FedOcw, an optimized FL framework for cross-lingual PD speech that improves convergence and cross-language transfer [4]. These works, however, primarily evaluate performance at the level of sites or languages and do not explicitly model or optimize fairness for clinically meaningful subpopulations such as the combination of site and diagnosis.

In this work we study subpopulation aware fairness for federated PD speech models. Rather than treating each hospital as a single unit, we use *site by diagnosis* cells as basic units and track their error and balanced accuracy, viewing fairness as how quickly and how uniformly these cells reach acceptable accuracy under a fixed communication budget. Building on q style reweighting [16] and healthcare FL frameworks that reweight clients or groups by loss or error [12], we define a subpopulation aware q fair, site by diagnosis based aggregation family that applies bounded emphasis to high error cells. All configurations share the same pretrained speech encoder, local optimizer, and protocol, and differ only in the strength of this emphasis. Our main configuration, FedSafe, uses a moderate q value with a bounded fairness factor derived from PD and HC errors; we compare it to ablations that weaken or remove this factor. Following fixed budget FL comparisons [12], [15], we focus on a shared intermediate round where FedAvg like baselines have largely left the underfitting regime, and we also report best round (oracle) summaries. Overall, our contributions are: (i) a subpopulation view of fairness for federated PD speech using site by diagnosis cells and metrics such as minimum balanced accuracy, worst cell error, and error dispersion; (ii) the FedSafe aggregation family, which keeps the Wav2Vec 2.0 encoder and local training unchanged while modulating aggregation weights via bounded functions of PD and HC error statistics at each site; and (iii) an empirical study on three PD speech corpora (Italian clinical, MDVR KCL mobile, and UAMS telephone vowels) under a fixed communication budget and an oracle view, showing improved early worst cell performance and dispersion relative to FedAvg like baselines, together with the tradeoff in later rounds avgerage accuracy.

2. Federated Subpopulation Aware Fairness Encoder aggregation

We consider cross site federated learning (FL) for speech based Parkinson's disease (PD) detection, where each client is a clinical site. Privacy and regulatory constraints make FL preferable to centralizing audio data, and prior work shows that speech based PD detection is feasible [2] and that FL can train PD models across institutions and languages [4]–[6]. We introduce a subpopulation aware fairness objective for a site by diagnosis cell based aggregation rule. Our main configuration, **FedSafe** (Federated Subpopulation Aware Fairness Encoder aggregation), applies moderate q style emphasis on high error cells together with a variance style regularizer across sites. Variants with different q values (for example q=0.0 or q=0.1) or alternative penalty schedules are treated as ablations of the same family rather than distinct algorithms.

2.1. Federated setting and subpopulation structure

We consider S federated clients (sites) indexed by $s \in \{1, \ldots, S\}$, each with local data

$$\mathcal{D}_s = \{(x_i, y_i, \ell_i)\}_{i=1}^{n_s}, \tag{1}$$

where x_i is a speech segment, $y_i \in \{0, 1\}$ is the diagnosis label (PD vs healthy control, with y=1 for PD), and ℓ_i denotes site specific attributes (for example language or recording protocol). A central server runs T communication rounds; at round t it broadcasts $f_{\theta^{(t)}}$ to sampled sites, which fine tune on $\mathcal{D}_s$ and return updates aggregated into $\theta^{(t+1)}$. Local training uses the cross entropy loss

$$\ell_s(\theta) = \frac{1}{n_s} \sum_{(x,y)\in\mathcal{D}_s} \mathrm{CE}\big(f_\theta(x), y\big), \tag{2}$$

and the global FL objective

$$\min_\theta \sum_{s=1}^{S} p_s \, \ell_s(\theta), \tag{3}$$

with $p_s \propto n_s$ recovering FedAvg weighting [1].

Client level averages can hide clinically important disparities. Motivated by worst group and equal performance viewpoints [10]–[12], [27], we treat each *subpopulation cell* (s, g), where s is a site and $g \in \{\mathrm{PD}, \mathrm{HC}\}$, as a basic unit and track its validation recall (or error) $R_{s,g}$. If some $R_{s,g}$ is consistently lower, patients in that cell receive systematically worse performance even when global metrics look acceptable. Our goal is to introduce a server side fairness objective that keeps global performance comparable to a strong FedAvg baseline while explicitly pushing site by diagnosis performance toward a more balanced profile, especially in worst cell accuracy and dispersion. Rather than assuming perfect equalization, we empirically show that members of our aggregation family can improve early round worst cell metrics under a fixed communication budget.

2.2. Client model and local fine tuning

Each client uses the same speech encoder architecture: a pretrained self supervised model (Wav2Vec 2.0 [24]) plus a lightweight classification head, following FL work on speech and ASR [7]. Each model is $f_\theta(x) = h_\phi(g_\psi(x))$, with encoder g_ψ and task head h_ϕ. In all experiments we fine tune only the last L transformer layers of g_ψ together with h_ϕ using AdamW and mixed precision, consistent with standard Wav2Vec transfer learning [7], [24]. On client s, local training runs for E epochs per round, minimizing

$\ell_s(\theta)$ with minibatch SGD. This procedure and architecture are identical across all configurations (FedAvg like baseline and subpopulation aware variants), so differences in fairness metrics stem solely from the aggregation rule. After local training, each client evaluates $f_{\theta^{(t)}}$ on its validation split, computing accuracy, macro F1, and class specific recalls for PD and HC. Let $\mathrm{TPR}_{\mathrm{PD},s}$ and $\mathrm{TPR}_{\mathrm{HC},s}$ denote PD and HC recall on site s; the site level balanced accuracy $\mathrm{BA}_s = \frac{1}{2}\big(\mathrm{TPR}_{\mathrm{PD},s} + \mathrm{TPR}_{\mathrm{HC},s}\big)$ is less sensitive to label imbalance than raw accuracy [14]. These site level metrics are returned to the server along with model updates; the aggregation family uses only the per site PD and HC recalls to build site by diagnosis error statistics, without accessing raw audio or per example data.

2.3. Subpopulation error statistics

For each site s we define PD and HC error rates $e_{\mathrm{PD},s} = 1 - \mathrm{TPR}_{\mathrm{PD},s}$ and $e_{\mathrm{HC},s} = 1 - \mathrm{TPR}_{\mathrm{HC},s}$. At the end of round t the server computes cross site means $\bar{e}_{\mathrm{PD}} = \frac{1}{S}\sum_s e_{\mathrm{PD},s}$ and $\bar{e}_{\mathrm{HC}} = \frac{1}{S}\sum_s e_{\mathrm{HC},s}$, and empirical standard deviations σ_{PD}, σ_{HC}. These statistics support fairness metrics such as mean and minimum balanced accuracy over site by diagnosis cells, maximum subpopulation error, and simple dispersion summaries, in line with worst group and dispersion based criteria in healthcare [11], [12], [27].

For the aggregation weights we focus on how much each site's PD and HC error exceeds the cross site mean. We define positive normalized deviations

$$z^{+}_{g,s} = \max\left\{0,\ \frac{e_{g,s} - \bar{e}_g}{\sigma_g + \delta}\right\}, \quad g \in \{\mathrm{PD}, \mathrm{HC}\}, \tag{4}$$

where $\delta > 0$ is a small stabilizing constant. This follows group DRO style objectives that emphasize groups with above average loss [11] and federated group DRO formulations that shift weight toward high loss clients [9], with groups here defined as site by diagnosis cells.

2.4. Subpopulation aware q fair aggregation family

In FedAvg, server side aggregation uses sample size weights $w_s^{\mathrm{FedAvg}} \propto n_s$ [1]. This is communication efficient and widely used [1], [26], but can favor large or well performing clients and exacerbate disparities when some sites are under represented or harder to classify [27].

q Fair Federated Learning (q FFL) mitigates this by upweighting high loss clients [16]:

$$w_s^{\mathrm{qFFL}} \propto n_s\big(\ell_s(\theta^{(t)}) + \varepsilon\big)^q, \tag{5}$$

where $q > 0$ controls reweighting strength and $\varepsilon > 0$ smooths the loss. Larger q increases the influence of high loss clients and can reduce accuracy spread across devices at a given average performance [16], but ℓ_s is a coarse client level summary that does not indicate which subpopulations are poorly served.

Our subpopulation aware q fair aggregation family keeps the q FFL base weights and multiplies them by a fairness factor γ_s tied to PD and HC errors at each site. Intuitively, sites whose PD or HC subpopulations have above average error receive a larger γ_s, and hence more influence in aggregation. Concretely,

$$\gamma_s = \mathrm{clip}\big(1 + \tau\big(\alpha_{\mathrm{PD}} z^{+}_{\mathrm{PD},s} + \alpha_{\mathrm{HC}} z^{+}_{\mathrm{HC},s}\big),\ \gamma_{\min},\ \gamma_{\max}\big), \tag{6}$$

where $z^{+}_{g,s}$ is the positive part of the standardized error for group $g \in \{\mathrm{PD}, \mathrm{HC}\}$ at site s, $\alpha_{\mathrm{PD}}, \alpha_{\mathrm{HC}} \geq 0$ control emphasis on PD vs HC disparities, τ sets the overall strength, and $\gamma_{\min}, \gamma_{\max}$ bound the correction to avoid extreme weights, in line with group DRO style observations about overfitting small groups [11], [12], [27].

The aggregation weight in this family is

$$w_s^{\mathrm{subq}} \propto n_s\big(\ell_s(\theta^{(t)}) + \varepsilon\big)^q \gamma_s. \tag{7}$$

It recovers standard FedAvg when $q{=}0$ and $\gamma_s{\equiv}1$, and client level q FedAvg when $q{>}0$ and $\gamma_s{\equiv}1$ [16]. When $q{\geq}0$ and $\tau{>}0$, the fairness factor γ_s increases the weight of sites whose PD or HC error is above the cross site mean. Intuitively, groups (site by diagnosis cells) with unusually high error receive extra emphasis, similar in spirit to group DRO and federated group DRO [9], [11], while clipping keeps the method closer to an average case optimizer than to pure worst group optimization.

2.5. FedSafe configuration and ablations

All configurations studied in our experiments share the same encoder, local optimizer, and communication protocol. They differ only in how they instantiate the subpopulation aware q fair weighting. We summarize the main settings here; the exact hyperparameters for FedSafe are made explicit so that our configuration is reproducible.

- **Subpop FedAvg (**$q{=}0.0$**).** Sets $q = 0$ and $\tau = 0$, so $\gamma_s \equiv 1$ and $w_s^{\mathrm{subq}} \propto n_s$. This is the closest variant to standard FedAvg, while still computing and logging subpopulation statistics.
- **Subpop q FedAvg (**$q{=}0.1$**).** Uses a small $q > 0$ with $\tau = 0$, giving a q FFL style client level reweighting without explicit site by diagnosis emphasis, analogous to [16].
- **UP Pen (**$q{=}0.2$**).** Uses $q > 0$ together with a tighter upper clipping to moderate the contribution of already strong cells, probing whether downweighting high performing cells alone can improve fairness.
- **FedSafe (**$q{=}0.2$ **with fairness factor).** Our main configuration fixes $q = 0.2$ and $\varepsilon = 10^{-3}$, defines $u_s = \alpha_{\mathrm{PD}} z^{+}_{\mathrm{PD},s} + \alpha_{\mathrm{HC}} z^{+}_{\mathrm{HC},s}$ with $(\alpha_{\mathrm{PD}}, \alpha_{\mathrm{HC}}) = (1.0, 0.5)$, and mixes worst cell and mean deviations with coefficient 0.7 (70 percent worst, 30 percent mean). The overall push toward high error cells is controlled by $\tau = 0.3$ and a small extra bump on the worst cell, and the fairness factor $\gamma_s = \mathrm{clip}\big(1 + \tau\, b_s, \gamma_{\min}, \gamma_{\max}\big)$ is bounded with $\gamma_{\min} = 0.7$ and $\gamma_{\max} = 1.4$, so no site is downweighted by more than 30 percent or upweighted by more than 40 percent. These bounds keep FedSafe closer to an average case optimizer than a pure worst group procedure [11], [12], [27].

We emphasize that these four settings are not four independent algorithms. They are operating points within the same subpopulation aware q fair aggregation family, chosen to probe the effect of (i) activating q style client reweighting and (ii) activating site by diagnosis fairness factors. In the experiments we examine whether and where it can improve early round worst cell performance and dispersion metrics while keeping global accuracy in a comparable range to the most relevant baselines, in line with recent work that studies performance fairness trade offs under realistic communication budgets [12], [15].

2.6. Fairness metrics and analysis

To assess whether FedSafe meets its subpopulation oriented fairness goal, we follow fairness evaluation practices from medical ML and group robust optimization [4], [10], [11], [27]. At each round we track: (i) *global utility* via overall accuracy and macro F1 on all validation speech, to check clinically acceptable discrimination [13]; (ii) *mean and worst balanced accuracy*, computing BA_s per site and reporting its mean and minimum to summarize typical and worst site level performance under label imbalance [14], [27]; and (iii) a *site × diagnosis dispersion* metric given by the worst cell error across all site by diagnosis subgroups, analogous to worst group risk in group DRO [11] and highlighting under served subpopulations. FedSafe is a lightweight, encoder preserving reweighting layer that steers optimization toward under performing site by diagnosis cells while remaining compatible with strong speech FL baselines and common healthcare FL practices [4], [7], [8].

3. Experimental Setup

We evaluate our subpopulation aware aggregation family on a binary PD vs healthy control (HC) task using three PD speech corpora that are widely used or recently introduced in voice based PD studies [28]. ***Italian Parkinson's Voice and Speech (ItalianPVS).*** This corpus comes from a study at the Università degli Studi di Bari on speech intelligibility in PD [17]. We use 50 participants (22 elderly HC, 28 PD), each providing fixed read text recordings at 44.1 kHz, for HC vs PD classification.

MDVR-KCL mobile speech. The English MDVR-KCL corpus contains clinical recordings from King's College Hospital, London [18]. Speech was captured with a Moto G4 smartphone at 44.1 kHz while participants read two fixed passages and produced brief spontaneous dialogue; recordings are annotated with PD vs HC diagnosis and severity scores (Hoehn–Yahr, UPDRS-II, UPDRS-III).

UAMS telephone vowel dataset (PD-Voice figshare). The "Voice Samples for Patients with Parkinson's Disease and Healthy Controls" dataset from UAMS provides telephone collected sustained /a/ vowels from 50 PD and 50 HC participants [19]. Callers sustained /a/ for about 3 seconds into a voicemail line, recorded at 8 kHz with 16 bit resolution. The corpus was introduced by Iyer et al. for telephone based PD detection [22] and has since been reused in deep learning studies on spectrograms and multimodal voice features [20], [21].

Federated sites and subpopulations. Following cross site FL studies on PD speech [4], [5], we treat each corpus as a federated client (site). ItalianPVS, MDVR-KCL, and UAMS thus form three sites with both PD and HC speech but different languages, microphones, and recording conditions. Within each corpus we create speaker disjoint train, validation, and test splits so that no speaker appears in more than one split, mirroring prior PD speech protocols [23]. Subpopulation cells for fairness analysis are defined as site by diagnosis (PD vs HC), so each site contributes two cells.

All audio is converted to mono 16 kHz to match the pretrained Wav2Vec 2.0 encoder [24]. Following recent PD speech work with self supervised encoders [22], [23], [25], we apply per recording amplitude normalization, energy based VAD trimming at the beginning and end, discard sustained vowels shorter than 1.5 s after trimming [22], and use task specific cropping or zero padding so that sustained vowels use a central 1.5 s window while read speech has a fixed maximum duration. UAMS telephone recordings are first upsampled from 8 kHz to 16 kHz by band limited interpolation and then processed with the same pipeline, without attempting to restore frequencies beyond the telephone bandwidth [20], [22], so that performance differences mainly reflect federated aggregation rather than corpus specific preprocessing.

3.1. Wav2Vec 2.0 model and training setup

Base encoder and head.. All sites share the same pretrained encoder and classifier head. We use a Wav2Vec 2.0 base model pretrained on large scale read speech [24], which has shown strong performance on pathological and clinical speech, including PD voice tasks [7], [23], [25]. On top of the encoder we attach a lightweight two layer feedforward head h_ϕ that maps the pooled representation to a PD vs HC logit, so each client implements

$$f_\theta(x) = h_\phi\big(g_\psi(x)\big), \tag{8}$$

with Wav2Vec 2.0 encoder g_ψ and task specific head h_ϕ. We fine tune only the last L transformer blocks of g_ψ together with h_ϕ, keeping earlier blocks frozen, following recent clinical and cross site adapters for Wav2Vec 2.0 [7], [23], [24]. This provides enough capacity to adapt to language and channel differences while keeping the trainable parameter count moderate for FL.

3.1.1. Local training and evaluation protocol.

Local optimization uses AdamW with weight decay and mixed precision, as in standard Wav2Vec 2.0 fine tuning [7], [24]. Unless otherwise stated, we use a learning rate in the range 10^{-4} to $5 \cdot 10^{-5}$ with cosine decay, minibatch sizes chosen to fit a single GPU, and one or two local epochs per round, in line with prior PD speech and cross site FL studies with Wav2Vec style encoders [4], [5], [23], [25]. All configurations (FedAvg like baselines and FedSafe variants) share the same local settings so that differences arise only

TABLE 1: Fixed communication budget comparison at round $R_{\text{budget}} = 33$. All methods are trained for the same number of global rounds; we report overall accuracy, mean and minimum balanced accuracy over site-by-diagnosis cells, the maximum subpopulation error, and dispersion measures of PD and HC error across sites.

Method	Acc	Mean BA	Min BA	Max subpop err	SBER	$\mathrm{Var}(e_{\mathrm{PD}})$	$\mathrm{Var}(e_{\mathrm{HC}})$
FedSafe (q=0.2)	0.783	0.785	0.528	0.500	0.215	0.043	0.045
Subpop-FedAvg (q=0.0)	0.756	0.735	0.347	0.750	0.265	0.049	0.058
UP-Pen (q=0.2)	0.677	0.664	0.500	1.000	0.336	0.076	0.112
Subpop-qFedAvg (q=0.1)	0.515	0.550	0.267	1.000	0.450	0.076	0.088

TABLE 2: Comparing the metrics over number of rounds it takes. For each method we report the round at which its mean balanced accuracy is maximized, together with the corresponding test metrics. These oracle early stopping points complement the fixed budget comparison in Table 1.

Method	Round	Acc	Mean BA	Min BA	Max subpop err	SBER	$\mathrm{Var}(e_{\mathrm{PD}})$	$\mathrm{Var}(e_{\mathrm{HC}})$
FedSafe (q=0.2)	33	0.783	0.785	0.528	0.500	0.215	0.043	0.045
Subpop-FedAvg (q=0.0)	85	0.855	0.855	0.472	0.556	0.145	0.047	0.053
UP-Pen (q=0.2)	59	0.797	0.798	0.493	0.889	0.202	0.019	0.125
Subpop-qFedAvg (q=0.1)	11	0.732	0.694	0.486	0.778	0.306	0.071	0.110

from aggregation. We run synchronous FL for $T{=}100$ global rounds with a fixed fraction of sites participating each round, following PD speech FL protocols [4], [5]. After every round we evaluate the global model on each site's held out test set, compute per cell recalls and balanced accuracies, and log the fairness metrics from Section 2. In line with fairness aware FL under limited communication [12], [15], our main comparison focuses on an intermediate round $R_{\text{budget}}{=}33$, where a FedAvg like baseline has largely left the underfitting regime, complemented by an oracle best round summary based on mean balanced accuracy per method.

4. Evaluation and Discussion

We hypothesize that subpopulation-aware reweighting is most useful under realistic, limited communication budgets. Table 1 therefore reports test performance at a shared budget of $R_{\text{budget}} = 33$ global rounds, which we treat as a practically relevant regime for cross-site PD speech FL (Section 3). At this budget, FedSafe attains the best overall and mean balanced accuracy ($\mathrm{Acc} = 0.783$, $\mathrm{Mean\ BA} = 0.785$) and, from a fairness perspective, the highest minimum balanced accuracy (0.528) and lowest maximum subpopulation error (0.500). Relative to Subpop-FedAvg ($q{=}0.0$), FedSafe improves mean balanced accuracy (0.785 vs 0.735), raises the minimum cell BA (0.528 vs 0.347), and reduces the worst-cell error (0.500 vs 0.750), with slightly lower dispersion (SBER and error variances). UP-Pen and Subpop-qFedAvg perform worse at this budget: both show lower mean BA and higher maximum subpopulation error, and Subpop-qFedAvg is weakest overall, consistent with the idea that client-level q reweighting without explicit subpopulation structure is a poor proxy for subgroup fairness [12], [16]. Overall, the fixed-budget view suggests that incorporating simple site-by-diagnosis error statistics into the aggregation weight can improve the balance of accuracies across cells while keeping global performance competitive with a strong FedAvg-like baseline. Table 2 reports a best-round oracle analysis, where for each method we select the training round with highest *test* mean balanced accuracy, following best-epoch summaries in fair FL benchmarks [12], [15]. Subpop-FedAvg ($q{=}0.0$) peaks late and achieves the strongest average performance, while FedSafe's oracle round (33) attains the highest minimum balanced accuracy and lowest maximum subpopulation error but lower averages than late Subpop-FedAvg. UP-Pen and Subpop-qFedAvg peak earlier and never dominate on both average and worst-cell metrics, indicating that mild client-level q reweighting without explicit subpopulation structure does not equalize site-by-diagnosis cells. Overall, the oracle view suggests that average-focused objectives can eventually narrow fairness gaps, whereas FedSafe mainly reshapes the early to mid training trajectory in favor of underperforming cells rather than improving asymptotic averages.

5. Conclusion

We studied fairness for site by diagnosis subpopulations in multilingual cross site federated PD speech classification. Our federated system design, keeps the Wav2Vec 2.0 encoder and local training unchanged and adds a bounded reweighting layer based on per-site PD and HC error statistics. On a three-site FL setting (Italian clinical recordings, MDVR-KCL mobile speech, and UAMS telephone vowels), FedSafe achieved higher mean and minimum balanced accuracy and lower worst-cell error than a closely matched FedAvg-like baseline at a moderate communication budget, while maintaining comparable overall accuracy. An oracle early-stopping view showed that a plain subpopulation FedAvg objective can eventually recover stronger average performance if training continues, whereas FedSafe mainly reshapes the early–mid training trajectory in favor of underperforming site-by-diagnosis cells. These results suggest that even lightweight, encoder-preserving subpopulation reweighting can be a practical tool for improving early-round subgroup behavior in federated PD speech, but also highlight that it trades off with late-stage average performance. In the future, we will extend this idea to richer subgroup structures, larger cohorts, and broader fairness objectives.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156287). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00256629) grant funded by the Korea government(MSIT). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

[1] HMcMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.

[2] Vásquez-Correa, Juan Camilo, et al. ”Convolutional neural networks and a transfer learning strategy to classify Parkinson’s disease from speech in three different languages.” Iberoamerican Congress on Pattern Recognition. Cham: Springer International Publishing, 2019.

[3] Sarlas, Athanasios, Alexandros Kalafatelis, Georgios Alexandridis, Michail-Alexandros Kourtis, and Panagiotis Trakadas. ”Exploring federated learning for speech-based parkinson’s disease detection.” In Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1-6. 2023.

[4] Quan, Changqin, et al. ”FedOcw: optimized federated learning for cross-lingual speech-based Parkinson’s disease detection.” npj Digital Medicine 8.1 (2025): 357.

[5] Arasteh, Soroosh Tayebi, et al. ”Federated learning for secure development of AI models for Parkinson’s disease detection using speech from different languages.” arXiv preprint arXiv:2305.11284 (2023).

[6] Danek, Benjamin P., et al. ”Federated learning for multi-omics: A performance evaluation in Parkinson’s disease.” Patterns 5.3 (2024).

[7] Nguyen, Tuan, et al. ”Federated learning for asr based on wav2vec 2.0.” ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[8] Xu, Jie, et al. ”Federated learning for healthcare informatics.” Journal of healthcare informatics research 5.1 (2021): 1-19.

[9] Guo, Z., & Yang, T. (2024). Communication-efficient federated group distributionally robust optimization. Advances in Neural Information Processing Systems, 37, 23040-23077.

[10] Jin, Ruinan, et al. ”Fairmedfm: fairness benchmarking for medical imaging foundation models.” Advances in Neural Information Processing Systems 37 (2024): 111318-111357.

[11] Sagawa, Shiori, et al. ”Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.” arXiv preprint arXiv:1911.08731 (2019).

[12] Zhang, Lin, et al. ”Federated learning for non-iid data via unified feature learning and optimization objective alignment.” Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[13] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

[14] Brodersen, Kay Henning, et al. ”The balanced accuracy and its posterior distribution.” 2010 20th international conference on pattern recognition. IEEE, 2010.

[15] Ye, Rongguang, and Ming Tang. ”Learning Heterogeneous Performance-Fairness Trade-offs in Federated Learning.” arXiv preprint arXiv:2504.21775 (2025).

[16] Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2019). Fair resource allocation in federated learning. arXiv preprint arXiv:1905.10497.

[17] Dimauro, Giovanni, and Francesco Girardi. ”Italian parkinson’s voice and speech.” (2019).

[18] H. Jaeger, D. Trivedi, M. Stadtschnitzer, Mobile Device Voice Recordings at King’s College London (MDVR-KCL) from both early and advanced Parkinson’s disease patients and healthy controls, Zenodo (Cern European Organization For Nuclear Research). (2019). https://doi.org/10.5281/zenodo.2867216.

[19] Prior, F., et al. ”Voice Samples for Patients with Parkinson’s Disease and Healthy Controls.” Dataset, figshare (2023).

[20] Rahmatallah, Yasir, et al. ”Pre-trained convolutional neural networks identify Parkinson’s disease from spectrogram images of voice samples.” Scientific Reports 15.1 (2025): 7337.

[21] Chen, Wenna, et al. ”Parkinson’s disease detection using spectrogram-based multi-model feature fusion networks.” Frontiers in Neurology 16 (2025): 1706317.

[22] Iyer, Anu, et al. ”A machine learning method to process voice samples for identification of Parkinson’s disease.” Scientific reports 13.1 (2023): 20615.

[23] Klempíř, O., Příhoda, D., & Krupička, R. (2023). Evaluating the performance of wav2vec embedding for parkinson’s disease detection. Measurement Science Review.

[24] Baevski, Alexei, et al. ”wav2vec 2.0: A framework for self-supervised learning of speech representations.” Advances in neural information processing systems 33 (2020): 12449-12460.

[25] Sedigh Malekroodi, Hadi, et al. ”Speech-Based Parkinson’s Detection Using Pre-Trained Self-Supervised Automatic Speech Recognition (ASR) Models and Supervised Contrastive Learning.” Bioengineering 12.7 (2025): 728.

[26] Liu, Bingyan, Nuoyan Lv, Yuanchun Guo, and Yawen Li. ”Recent advances on federated learning: A systematic survey.” Neurocomputing 597 (2024): 128019.

[27] Petersen, Eike, Sune Holm, Melanie Ganz, and Aasa Feragen. ”The path toward equal performance in medical machine learning.” Patterns 4, no. 7 (2023).

[28] Xavier, Daniela, et al. ”Voice analysis in Parkinson’s disease-a systematic literature review.” Artificial Intelligence in Medicine (2025): 103109.