npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01763-3

FedOcw: optimized federated learning for cross-lingual speech-based Parkinson's disease detection

Changqin Quan[1] ✉, Zhonglue Chen[2], Kang Ren[2] & Zhiwei Luo[1]

Accurate detection of Parkinson's disease (PD) through speech analysis holds great promise for early diagnosis and improved patient management. However, developing robust machine learning models is challenging due to the decentralized nature of medical data and the substantial heterogeneity in multilingual PD speech datasets. Conventional federated learning (FL) methods struggle in these heterogeneous, non-independent and identically distributed (non-IID) environments, where differences in data distributions arise from variations in language, speech content, recording conditions, medical measurement techniques, and dataset sizes. To address these challenges, we propose FedOcw, an optimized FL framework designed to enhance cross-lingual knowledge transfer and improve convergence stability. Through extensive multilingual experiments, we demonstrate that FedOcw consistently outperforms traditional FL models by achieving superior diagnostic accuracy while ensuring adaptive and equitable weight distribution across clients. These findings highlight FedOcw as an effective FL solution for privacy-preserving, speech-based PD detection across diverse linguistic and institutional settings.

Parkinson's disease (PD), one of the most prevalent neurodegenerative disorders, is projected to affect over 12 million individuals globally by 2040, driven by an aging population[1]. Speech impairments are among the earliest and most prevalent symptoms of PD, with 89% of patients exhibiting vocal disorders, 45% experiencing articulatory impairments, and 20% suffering from fluency issues, as reported by Logeman et al.[2]. The perceptual, acoustic, and kinematic characteristics of PD-related speech deterioration have been extensively documented[3–6], underscoring the potential for speech-based diagnostic tools to enhance early detection and disease management.

Recent advances in machine learning (ML) have significantly contributed to the development of automated PD diagnosis from speech signals. Traditional ML approaches, such as Support Vector Machines[7–9], K-Nearest Neighbors[9,10], Decision Trees[11], Naïve Bayes (NB)[11], Genetic Algorithms[12], and Gaussian Process Classification[13], typically rely on hand-engineered speech features, including Mel-frequency cepstral coefficients (MFCC), pitch, jitter, and shimmer, to distinguish PD patients from healthy controls. More recently, deep learning (DL) models, such as multi-layer perceptrons (MLPs), convolutional neural networks (CNNs), and long short-term memory (LSTM) networks, have shown improved performance by automatically extracting salient patterns from input features. Building on this, end-to-end architectures based on CNNs and recurrent neural networks have demonstrated the ability to capture complex temporal and spectral characteristics directly from raw audio[14,15]. Current research has increasingly focused on more robust and generalizable approaches for speech-based Parkinson's disease detection, particularly self-supervised speech encoders and Transformer-based architectures. These foundational models have demonstrated superior performance compared to traditional methods[16]. In parallel, model interpretability has become a critical focus, with efforts to elucidate the inner workings of these deep models and their alignment with clinical speech markers[17].

Despite these advancements, the performance of ML models is highly dependent on the availability of large and diverse training datasets. However, medical speech data is often decentralized across institutions, with significant variations in measurement techniques, dataset sizes, and linguistic content. The privacy-sensitive nature of medical data further complicates data sharing, limiting the potential for robust model training. Federated learning (FL) presents a compelling solution by enabling collaborative model training across institutions without centralizing patient data. FL has demonstrated success in medical applications, including brain anomaly detection[18], COVID-19 diagnosis[19,20], breast tumor classification[21–23], and predicting high-risk gastric cancer recurrence[24], as well as in biomedical natural language processing[25]. Specifically for PD detection, FL has been explored in functional MRI-based studies[26], and its

[1]Graduate School of System Informatics, Kobe University, Kobe, Japan. [2]GYENNO SCIENCE CO. LTD., Shenzhen, China.
✉e-mail: quanchqin@gold.kobe-u.ac.jp

feasibility has been validated for speech-based FL models across multiple institutions while preserving patient privacy[27].

However, conventional FL approaches face substantial challenges in heterogeneous (non-IID) data environments, which are particularly pronounced in multilingual PD speech datasets. Existing FL methods, such as Federated Averaging (FedAvg)[28], aggregate local models by averaging client updates, often leading to suboptimal generalization when data distributions vary significantly. To address statistical heterogeneity, FedProx[29] introduces a proximal term that constrains local updates, but it does not provide personalized solutions tailored to individual clients[30]. Besides, Local customization methods offer an alternative solution by customizing a well-trained global model. The global model undergoes local fine-tuning by incorporating the private data of each client to create personalized models for those clients[31,32]. Alternative strategies, including Scaffold[33] (variance reduction for gradient stabilization) and FedNova[34] (adaptive model selection), have sought to improve FL robustness, yet they remain limited in their ability to handle both statistical and linguistic diversity.

Some recent studies have explored pairwise collaboration strategies in FL. Huang et al.[35] introduced federated attentive message passing to facilitate collaboration among clients with similar data, while Smith et al.[36] modeled pairwise collaboration by extending distributed multi-task learning to FL. However, these methods struggle when data exhibits both statistical heterogeneity and variations in linguistic features, making it difficult to form effective collaboration groups.

Real-world multilingual PD speech datasets exhibit high variability in medical measurement techniques, speech content, and language structure. Several studies have demonstrated that PD detection performance can vary significantly depending on speech input. For instance, an LSTM model achieved 88.08% classification accuracy with sentence-based speech data but only 73.52% with the sustained vowel sound '/a/'[37]. Similarly, an end-to-end deep learning model trained on a dataset of Chinese short sentences achieved only 49.4% accuracy when tested on a Spanish dataset[38], highlighting the critical impact of cross-lingual generalization on model performance. Moreover, Botelho et al.[39] emphasized that performance discrepancies may arise not only from linguistic differences but also from technical factors such as variations in recording conditions or equipment.

While FL offers a promising solution for privacy-preserving PD diagnosis, many existing studies fail to adequately address the challenges introduced by heterogeneous and cross-lingual data distributions, particularly prevalent in multilingual speech datasets. To address this gap, this study introduces FedOcw, a dynamic optimization-based aggregation framework that enables client nodes to develop customized models that adapt to their local datasets. This targeted optimization enables more effective knowledge transfer across linguistically and clinically diverse datasets, enhancing the robustness and accuracy of cross-lingual PD detection.

Additionally, we integrate an end-to-end deep learning model that combines time-distributed 2D convolutional neural networks (2D-CNNs) and 1D convolutional neural networks (1D-CNNs). This architecture is designed to capture both temporal and spatial features from speech data, enhancing model robustness for PD detection across linguistic and institutional variations.

Our study seeks to address the challenges posed by data heterogeneity in federated learning for speech-based Parkinson's disease detection through the proposed FedOcw framework. Specifically, we aim to (1) evaluate FedOcw's effectiveness in enhancing Parkinson's disease detection across diverse, multilingual, and institutionally heterogeneous datasets; (2) investigate the impact of dynamically optimized client weighting on the stability and efficiency of the global federated learning process; and (3) analyze the statistical properties and linguistic diversity of client datasets that influence their aggregation weights and contributions to global learning. By achieving these objectives, our work aims to advance the scalability, personalization, and adaptability of federated learning, supporting the development of privacy-preserving AI-driven diagnostic tools for Parkinson's disease.

Results

To evaluate the effectiveness of our proposed federated learning framework for speech-based Parkinson's disease detection, we utilized five multilingual datasets, incorporating Spanish, Italian, Chinese, Czech, and English speech samples. These datasets vary in recording conditions, linguistic structure, and phonetic tasks, providing a diverse and heterogeneous training environment that closely resembles real-world clinical scenarios.

Dataset-1 (Spanish), sourced from the PC-GITA repository[40], comprises speech recordings from 100 individuals, including 50 Parkinson's disease (PD) patients and 50 healthy controls (HCs). All recordings were conducted in professional soundproof booths at 44.1 kHz sampling frequency with 16-bit resolution. The PD participants, aged 33 to 81 years, were evaluated in the ON state by three expert phoneticians. Speech samples included:

(1) Sustained vowels: Three repetitions of /a/, /i/, /e/, /o/, and /u/.
(2) Isolated words: /blusa/, /petaka/, /apto/, /campana/, /llueve/, /reina/, /braso/, and /viaje/.
(3) Sentence reading: Simple (/laura/, /loslibros/, /luisa/, etc.) and complex (/preocupado/, /juan/, etc.) structures.
(4) Spontaneous speech: Monologues (~44.86 s on average).

These tasks were designed to capture phonation, articulation, and prosody impairments, which are critical for detecting Parkinsonian dysarthria.

Dataset-2 (Italian) was originally developed to assess speech intelligibility in PD patients using automatic speech recognition systems[41]. This dataset[42] includes 28 PD patients and 37 HCs, featuring recordings of:

(1) Phonemically balanced text reading (twice, with a 30-s pause).
(2) Repetitions of syllables (/pa/ and /ta/ for 5 s each).
(3) Sustained vowels (/a/, /i/, /e/, /o/, /u/).
(4) Phonemically balanced word and phrase reading.

Recordings were conducted in low-noise, echo-free environments, with microphones placed 15–25 cm from the speaker's lips. All PD participants were receiving antiparkinsonian treatment.

Dataset-3 (Chinese), obtained from the GYENNO SCIENCE Parkinson's Disease Research Center[37], consists of 30 PD patients and 15 HCs, aged 37 to 75 years. Speech tasks included:

(1) Sustained vowels (/a/ and /e/).
(2) Short sentence reading (e.g., /si shi si zhi shi shi zi/).

Speech samples were recorded using smartphones positioned 10 cm from the speaker's mouth. All PD participants were assessed by two neurologists and recorded in the ON state.

Dataset-4 (Czech) was designed to differentiate idiopathic Parkinson's disease from other parkinsonian syndromes via prolonged vowel analysis[43]. The dataset includes 22 PD patients, alongside 21 patients with multiple system atrophy, 18 with progressive supranuclear palsy, and 22 HCs. For this study, we utilized data from PD patients and HCs only.

Recordings were performed using a headset condenser microphone (5 cm from the lips) at 48 kHz sampling frequency with 16-bit resolution. Participants were instructed to sustain vowels (/A/ and /I/) in a modal voice for as long and steadily as possible.

Dataset-5 (English), the MDVR-KCL dataset (Mobile Device Voice Recordings at King's College London)[44], was developed to explore non-invasive Parkinson's disease monitoring through smartphone-based voice analysis[45]. This dataset includes 16 PD patients and 21 HCs, recorded using a Motorola Moto G4 smartphone at 44.1 kHz sampling frequency with 16-bit resolution.

Table 1 provides a summary of the demographic, clinical, and recording characteristics of participants, including the distribution of PD and HC groups by gender, age, disease severity, recording conditions, and speech tasks across Dataset-1 to Dataset-5. These datasets provide a comprehensive and multilingual foundation for evaluating federated learning models in Parkinson's disease detection.

Table 1 | Demographic, clinical, and recording characteristics of participants across Dataset-1 to Dataset-5

Dataset (Language)	Gender Male/ Female	Age Range (Mean ± SD)	Disease Severity (Mean ± SD)	Recording Conditions	Speech Tasks
Dataset-1 (Spanish)[40]	PD: 25/25 HC: 25/25	PD: 33–81 (61 ± 9.4) HC: 31–86 (61 ± 9.5)	UPDRS speech score: 6–93 (37.7 ± 18.3)	Soundproof booth; 44.1 kHz, 16-bit; professional recording setup	Sustained vowels, isolated words, sentence reading, spontaneous speech
Dataset-2 (Italian)[41]	PD: 19/9 HC: 23/14	PD: 40–80 (67.2 ± 8.7) HC: 19–77 (48.3 ± 23.4)	UPDRS II speech score: 0–4 (1.1 ± 1.2)	Echo-free environment; 15–25 cm mic distance	Text and phrase reading, syllable repetition (/pa/, /ta/), sustained vowels
Dataset-3 (Chinese)[37]	PD: 16/14 HC: 7/8	PD: 36–86 (60 ± 13.6) HC: 23–72 (51.9 ± 14.1)	Hoehn and Yahr: 1–5 (2.5 ± 0.8)	Smartphone; 10 cm from mouth	Sustained vowels (/a/, /e/), short sentence reading
Dataset-4 (Czech)[43]	PD: 10/ 12 HC: 11/ 11	PD: 48–82 (64.4 ± 9.6) HC: 41–79 (63.6 ± 10.0)	UPDRS III: 6–34 (15.9 ± 7.6)	Headset mic; 5 cm distance; 48 kHz, 16-bit	Sustained vowels (/A/, /I/)
Dataset-5 (English)[44]	PD: 9/7 HC: 19/2	–	UPDRS II Part 5: 0–3 (0.8 ± 0.9)	Smartphone (Moto G4); 44.1 kHz, 16-bit	Not specified in full; smartphone-based voice tasks

"–" denotes unavailable data.

To evaluate the effectiveness of the proposed federated learning approach on multilingual speech data, we define five experimental scenarios that integrate five datasets with varying language distributions and client allocations:

Scenario A (Fig. 1): Speech data from Dataset-1 (Spanish) and Dataset-2 (Italian) is distributed across eight clients (C0–C7), with an uneven distribution of Parkinson's disease (PD) cases and healthy controls (HCs). Clients C0–C3 are assigned data from Dataset-1 (Spanish), while Clients C4–C7 receive data from Dataset-2 (Italian).

Scenario B (Fig. 1): Speech data from Dataset-1 (Spanish) and Dataset-3 (Chinese) is allocated to seven clients (C0–C6). Clients C0–C3 are assigned data from Dataset-1 (Spanish), while Clients C4–C6 are assigned data from Dataset-3 (Chinese).

Scenario C (Fig. 1): Speech data from Dataset-2 (Italian) and Dataset-3 (Chinese) is used, with seven clients (C0–C6). Clients C0–C3 receive data from Dataset-2 (Italian), and Clients C4–C6 are assigned data from Dataset-3 (Chinese).

Scenario D (Fig. 1): Speech data from Dataset-1 (Spanish), Dataset-2 (Italian), and Dataset-4 (Czech) is used. Clients C0–C3 are assigned Dataset-1 (Spanish), C4–C7 receive data from Dataset-2 (Italian), C8 is allocated Dataset-4 (Czech).

Scenario E (Fig. 1): All five datasets are incorporated for a comprehensive multilingual evaluation. Clients C0–C3 are assigned Dataset-1 (Spanish), C4–C7 receive Dataset-2 (Italian), C8–C10 are allocated Dataset-3 (Chinese), C11 is assigned Dataset-4 (Czech), and C12 receives Dataset-5 (English).

This experimental setup enables a comprehensive evaluation of the federated model's generalization across linguistically diverse datasets. Each client was assigned speech samples from its respective dataset, which included a variety of task types such as sustained vowels, sentence reading, and spontaneous speech. A single model was trained per client using the entire local training dataset, without further partitioning based on individual speech tasks. This approach reflects real-world deployment conditions in federated learning, where heterogeneity in assessment protocols and data characteristics is common across different clinical sites.

To promote robust generalization, all speech samples were partitioned into non-overlapping training and test sets. Training data remained strictly localized on each edge client, while evaluation was independently performed on each client's separate test node. Importantly, no speaker overlap existed across clients' training sets, strengthening the model's ability to generalize across languages and participants. For final evaluation, each client was tested on its corresponding test node using local testing samples, providing a comprehensive assessment of the model's cross-lingual performance.

Regarding the choice of languages for the bilingual experiments, Spanish, Italian, and Chinese were prioritized due to the availability of well-balanced datasets with large sample sizes and a diverse set of speech tasks. These characteristics provided a robust and heterogeneous foundation for evaluating cross-lingual generalization. In contrast, the English and Czech datasets, while valuable, had comparatively smaller sample sizes and fewer speech tasks, limiting their suitability for the bilingual scenarios. Instead, English and Czech were incorporated in Scenarios D and E to further explore the impact of increasing language diversity on model performance.

Figure 1 provides a circular visualization of client distributions, including sample sizes, case-control ratios (shown as bar plots), and the number of participants (indicated in brackets). Percentages represent each client's relative contribution to the overall training dataset.

In Fig. 1, the box plots present the evaluation results over 100 rounds of federated aggregation, capturing performance across accuracy, F1-score, and Matthews correlation coefficient (Mcc). The mathematical formulations for these metrics are detailed in Eqs. (1)–(5).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Fig. 1 | FedOcw for Parkinson's disease detection from speech across five scenarios, showing data distributions and model performance. Different combinations of multilingual datasets assigned to clients: (Scenarios A) Spanish–Italian, (Scenarios B) Spanish–Chinese, (Scenarios C) Italian–Chinese, (Scenarios D) Spanish–Italian–Czech, and (Scenarios E) Spanish–Italian–Chinese–Czech–English. Each sub-panel shows data allocation across clients (e.g., C0–C3: Spanish, C4–C7: Italian, etc.) and box plots comparing the performance of five federated learning methods—FedAvg, FedProx, Scaffold, FedNova, and the proposed FedOcw—on client test data. Box plots indicate performance distributions, where the center line marks the median, the circle denotes the mean, box limits correspond to the 1st and 3rd quartiles, whiskers span 1.5 times the interquartile range, and outliers are shown individually.

$$F1-score = \frac{2 \times specifity \times sensitivity}{specificity + sensitivity} \tag{2}$$

$$specifity = \frac{TP}{TP + FP} \tag{3}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Here, TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Sensitivity and specifity quantify the model's ability to correctly identify positive and negative cases. The F1-score represents the harmonic mean of sensitivity and specifity, providing a balanced assessment of classification performance. Mcc measures the overall quality of binary classifications, ranging from −1 to +1, where +1 indicates perfect prediction, −1 signifies total disagreement between predictions and actual labels, and 0 reflects performance equivalent to random guessing.

Tables 2–6 summarize the average performance across 100 aggregation rounds for Scenarios A, B, C, D, and E with the best-performing federated learning methods highlighted in bold. Alongside federated learning approaches, the tables also report the average results for clients trained and evaluated on their isolated local datasets (Local) and the outcomes of centralized learning for comparison.

To ensure a fair and meaningful comparison, all baseline methods were carefully tuned and evaluated under consistent experimental conditions. For FedProx, we explored the proximal term coefficient $\mu \in \{0.01, 0.1, 1.0\}$ and selected the value $\mu = 0.1$ that achieved the best performance in each experimental setting. For SCAFFOLD, we followed the standard configuration, with control variates updated at the end of every local training

Table 2 | Mean performance metrics for Scenario A using Dataset-1 (Spanish) and Dataset-2 (Italian) over 100 federated aggregation rounds

Scenario A	Local	Federated learning methods					Centralized learning
		FedAvg	FedProx	Scaffold	FedNova	FedOcw	
Accuracy	65.23 (61.04–68.99)	72.11 (61.88–76.38)	71.84 (57.23–74.99)	69.68 (59.24–73.95)	59.72 (44.13–75.89)	74.81 (65.5–77.78)	69.51 (66.8–70.4)
F1-score	61.5 (57.1–66.67)	68.92 (53.3–74.49)	68.49 (51.53–73)	66.83 (47.76–72.69)	52.34 (30.5–74.47)	73.35 (64.64–76.79)	68.23 (65.01–69.21)
Specifity	65.95 (62.37–69.71)	69.73 (58.88–74.47)	69.58 (56.24–73.03)	67.74 (54.92–73.17)	61.92 (50–75.28)	74.16 (66.35–77.44)	68.48 (65.63–69.4)
Sensitivity	67.17 (61.85–72.85)	75.66 (62.45 79.46)	74.81 (58.6–77.52)	71.76 (59.8–75.24)	50.58 (22.06–78.15)	76.23 (67.05–79.21)	70.64 (68.03–71.55)
Mcc	0.331 (0.252–0.403)	0.446 (0.21–0.535)	0.438 (0.133–0.499)	0.39 (0.135–0.481)	0.252 (0–0.526)	0.502 (0.333–0.565)	0.39 (0.335–0.409)

Table 3 | Mean performance metrics for Scenario B using Dataset-1 (Spanish) and Dataset-3 (Chinese) over 100 federated aggregation rounds

Scenario B	Local	Federated learning methods					Centralized learning
		FedAvg	FedProx	Scaffold	FedNova	FedOcw	
Accuracy	62.89 (58.25–67.39)	62.23 (57.42–68.35)	63.37 (57.16–68.84)	64.84 (59.48–69.37)	55.84 (42.86–68.82)	67.85 (59.69–72.06)	62.3 (61.5–63.33)
F1-score	56.65 (52.12–62.94)	54.22 (39.11–63.5)	58.22 (39.41–65.67)	59.33 (41.13–66.41)	48.52 (29.76–65.73)	61.8 (43.57–68.11)	60.65 (59.83–61.89)
Specifity	60.85 (55.87–65.95)	57.63 (51.55–64.68)	60.06 (50.41–65.76)	61.56 (52.34–66.89)	57.06 (50–66.14)	63.24 (52.8–68.58)	61.34 (60.57–62.46)
Sensitivity	62.81 (56.38–69.64)	60.86 (36.15–72.28)	63.33 (39.7–70.06)	65.88 (41.16–73.91)	47.03 (21.43–69.72)	66.44 (50.5–74.09)	62.69 (61.71–63.73)
Mcc	0.232 (0.122–0.346)	0.184 (0.05–0.345)	0.229 (0.023–0.349)	0.266 (0.058–0.36)	0.153 (0–0.35)	0.288 (0.082–0.4)	0.24 (0.222–0.261)

Table 4 | Mean performance metrics for Scenario C using Dataset-2 (Italian) and Dataset-3 (Chinese) over 100 federated aggregation rounds

Scenario C	Local	Federated learning methods					Centralized learning
		FedAvg	FedProx	Scaffold	FedNova	FedOcw	
Accuracy	73.75 (68.18–77.72)	81.42 (76.49–84.22)	82.08 (67.59–84.97)	79.99 (70.56–82.99)	72.57 (36.15–85.14)	82.78 (52.76–87.08)	81.77 (80.9–82.68)
F1-score	70.37 (63.93–74.49)	79.08 (74.47–82.68)	80.66 (63.9–83.87)	76.79 (68.72–80.35)	68.82 (26.52–83.35)	81.93 (51.55–86.06)	80.19 (79.35–81.76)
Specifity	72.52 (66.77–76.42)	79.54 (74.47–84)	82.2 (64.3–85.46)	77.48 (68.69–81.06)	73.83 (50–85.35)	84.19 (61.11–88.44)	80.1 (79.45–81.03)
Sensitivity	72.76 (65.44–80.81)	81.89 (75.69–84.97)	83.11 (66.79–85.97)	79.89 (71.39–84.32)	68.58 (18.07–86.41)	83.44 (63.3–87.31)	80.5 (79.51–81.49)
Mcc	0.452 (0.327–0.533)	0.61 (0.505–0.683)	0.652 (0.309–0.713)	0.567 (0.398–0.644)	0.484 (0–0.707)	0.676 (0.241–0.757)	0.606 (0.59–0.625)

Table 5 | Mean performance metrics for scenario D using Dataset-1 (Spanish), Dataset-2 (Italian), and Dataset-4 (Czech) over 100 aggregation rounds

Scenario D	Local	Federated learning methods					Centralized learning
		FedAvg	FedProx	Scaffold	FedNova	FedOcw	
Accuracy	64.45 (60.38–68.16)	70.03 (60.32–74.55)	68.76 (60.27–72.83)	69.49 (60.32–73.94)	44.52 (44.52–44.52)	72.53 (62.04–75.69	68.42 (67.54–69.1)
F1-score	60.63 (55.86–64.21)	65.43 (47.33–71.41)	63.45 (52.54–69.34)	64.99 (48.81–71.05)	30.69 (30.69–30.69)	69.8 (51.54–73.68)	67.39 (66.36–68.23)
Specifity	64.81 (60.3–68.39)	67.58 (56~ 72.94)	66.12 (56.44–71.08)	67.33 (55.83–72.94)	50 (50–50)	71 (58.3–74.68)	67.66 (66.73–68.43)
Sensitivity	66.33 (58.46–72.31)	73.15 (56.12–78.55)	70.02 (60.96–77.2)	70.35 (58.96–76.62)	22.26 (22.26–22.26)	76.27 (67.58–79.39)	69.09 (68.21–69.63)
Mcc	0.309 (0.203–0.384)	0.408 (0.15–0.489)	0.371 (0.166–0.457)	0.384 (0.152–0.467)	0 (0–0)	0.465 (0.244–0.535	0.367 (0.349–0.38)

Table 6 | Average results for scenario E on all five datasets over 100 aggregation rounds

Scenario E	Local	Federated learning methods					Centralized learning
		FedAvg	FedProx	Scaffold	FedNova	FedOcw	
Accuracy	65.68 (61.07–70.58)	69.58 (57.41–73.12)	69.35 (60.35–73.85)	67.37 (59.34–70.72)	46.09 (43.04–75.64)	72.63 (62.5–75.58	68.31 (67.76–68.87)
F1-score	61.55 (56.46–67.84)	63.15 (37.09–68.61)	62.72 (43.83–69.33)	62.39 (41.24–66.82)	33.9 (29.87–72.09)	68.16 (48.74–72.46)	66.83 (65.53–67.6)
Specifity	65.2 (60.8–70.24)	65.4 (50.47–69.41)	65.05 (53.08–70.36)	64.59 (52.71–68.31)	52.14 (50–73.73)	69.41 (56.37–73.4)	67 (66.05–67.68)
Sensitivity	66.06 (61.3–72.22)	70.3 (36.3–76.88)	67.81 (52.04–75.81)	66.64 (47.51–72.15)	26.96 (21.52–77.27)	75.1 (55.89–79.25)	68.79 (68.3–69.42)
Mcc	0.312 (0.221–0.422)	0.357 (0.026–0.433)	0.339 (0.093–0.45)	0.325 (0.01–0.392)	0.045 (0–0.504)	0.435 (0.173–0.508	0.357 (0.348–0.369)

round. The standard implementation of FedNova was used without modification. To maintain comparability across methods, all experiments used the same optimization settings: Adam optimizer with a learning rate of 0.001, 10 local epochs per round, and a training batch size of 8.

As shown in Tables 2–6, FedOcw consistently outperforms conventional FL methods across all evaluated scenarios (A–E) in accuracy, F1-score, specifity, sensitivity, and Mcc, demonstrating superior stability and training effectiveness. It not only surpasses FedAvg, FedProx, Scaffold, and FedNova, but also outperforms centralized learning across all key metrics. These findings highlight the advantages of federated models in privacy-preserving and heterogeneous learning environments.

FedOcw's adaptability to linguistic diversity is evident across all scenarios. In the Spanish–Italian setting (Scenario A), it achieves the highest accuracy (74.81%) and Mcc (0.502), demonstrating effective knowledge transfer between related languages. In the Spanish–Chinese scenario (Scenario B), the model maintains strong performance with 67.85% accuracy and an Mcc of 0.288, though the increased linguistic divergence presents convergence challenges. In the Italian–Chinese setting (Scenario C), FedOcw achieves high specifity (84.19%) and sensitivity (83.44%), indicating a balanced classification approach. In the trilingual scenario (Scenario D), it maintains top performance with the highest accuracy (72.53%), F1-score (69.8%), and Mcc (0.465). Even in the most heterogeneous multilingual scenario (Scenario E), the model sustains robust performance, achieving 72.63%accuracy and an Mcc of 0.435, highlighting its robustness and ability to generalize across linguistic domains.

Table 7 reports the p values for the mean Accuracy, F1-score, and Mcc metrics across all five scenarios, evaluating the statistical significance of differences between our proposed federated model (FedOcw) and alternative methods.

Table 7 shows that FedOcw achieves statistically significant improvements over individual learning (Local), alternative federated learning methods, and centralized learning in key performance metrics, including accuracy, F1-score, specifity, and Mcc. However, sensitivity is an exception, where it performs comparably to FedProx ($p = 0.0525$), indicating no significant difference. Compared to centralized learning, FedOcw demonstrates significant advantages across all evaluation metrics, reinforcing its effectiveness in diverse settings. Notably, FedOcw outperforms FedAvg with strong statistical significance in specifity ($p = 0.0003$) and Mcc ($p = 0.0011$). However, the lack of statistical significance in some comparisons ($p > 0.05$) for sensitivity with FedProx) suggests that certain methods may still be competitive in specific aspects. These findings highlight FedOcw's robustness in handling heterogeneous and multilingual datasets, reinforcing its potential for broader cross-linguistic and clinical applications.

To better understand the model's behavior across multilingual settings, we conducted a language-wise accuracy analysis of client models in Scenarios A–E. Figure 2 presents the individual accuracy scores for each language (Spanish, Italian, Chinese, Czech, and English), highlighting the specific contributions of each client group to the overall federated learning performance.

As shown in Fig. 2, the Italian client consistently achieves the highest accuracy across scenarios in which it is present, reaching up to 94% in Scenario C and 91.6% in Scenario A. This suggests that the Italian dataset may contain more consistent or discriminative speech features for Parkinson's detection, possibly due to better recording conditions, more clearly defined task protocols, or less intra-class variability. In contrast, Spanish and Chinese clients show more variable performance, with Chinese accuracy rising from 63.14% in Scenario B to 67.8% in Scenario C, depending on the pairing.

The performance gap between Scenario B (Spanish–Chinese) and Scenario C (Italian–Chinese) is particularly informative. While both involve cross-lingual collaboration with Chinese data, Scenario C significantly outperforms Scenario B. This may be attributed to greater similarity in task structure or feature distribution between Italian and Chinese datasets, leading to more effective model generalization. Alternatively, the Spanish

Table 7 | Statistical significance (*p* values) of comparisons between the proposed FedOcw model and other learning methods across all evaluation metrics

p-value of metric	Local	Federated learning methods				Centralized learning
		FedAvg	FedProx	Scaffold	FedNova	
Accuracy	0.0007	0.0123	0.0089	0.0022	0.0079	0.0074
F1-score	0.0026	0.0033	0.0082	0.0019	0.0108	0.032
Specifity	0.0154	0.0003	0.0021	0.0073	0.0067	0.0064
Sensitivity	0.0025	0.0297	0.0525	0.0246	0.0145	0.0029
Mcc	0.0064	0.0011	0.0069	0.0035	0.0096	0.0019

P values less than 0.05 indicate statistical significance.

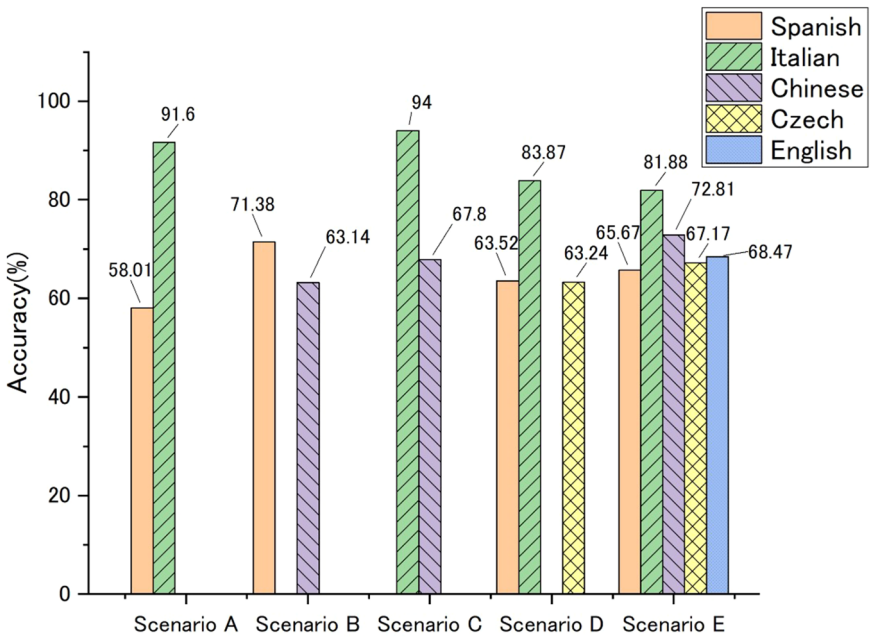

Fig. 2 | Accuracy by language group scross Scenarios A–E. This figure presents individual accuracy scores for clients using Spanish, Italian, Chinese, Czech, and English datasets, illustrating each language group's contribution to the overall federated learning performance. The results provide insight into cross-lingual generalization capabilities across different scenarios.

dataset may differ more substantially in prosody, phonetic structure, or participant characteristics, making knowledge transfer more challenging.

A similar trend is observed when comparing Scenario A (Spanish–Italian) and Scenario D (Spanish–Italian–Czech). In Scenario A, a large performance gap exists between Italian (91.6%) and Spanish (58.01%), suggesting unbalanced contributions and potential dominance of the Italian dataset during model aggregation. However, when Czech is added in Scenario D, the gap narrows: Italian performance drops slightly to 83.87%, while Spanish improves to 63.52%, and Czech reaches 63.24%. This shift indicates that adding a third, linguistically distinct client introduces more diversity into the training process, which likely promotes better generalization across heterogeneous clients.

In Scenario E, where five languages are present, performance becomes more balanced across clients with different languages, though Italian still maintain relatively strong accuracy. This suggests that FedOcw is able to preserve generalization even under high linguistic and distributional heterogeneity.

To evaluate the global stability and efficiency of the federated learning framework, Fig. 3 presents the training loss convergence of various federated learning models across five evaluation scenarios (A, B, C, D, and E). The models compared include FedAvg, FedProx, Scaffold, FedNova, and the proposed FedOcw. The x-axis denotes the number of communication rounds, while the y-axis represents the average training loss across local clients. A lower training loss over time indicates improved convergence and model stability.

As shown in Fig. 3, FedOcw consistently achieves the lowest training loss across all scenarios, demonstrating superior convergence stability and effectiveness. In Scenario A, it stabilizes at a loss of ~0.3, while other models exhibit significant fluctuations, indicating sensitivity to data heterogeneity. Scenario B follows a similar pattern, with FedOcw maintaining low and stable training loss, whereas FedAvg and FedNova experience sharp oscillations, leading to poor convergence. In Scenario C, FedOcw again outperforms all models, stabilizing around 0.2, while the other methods struggle to converge, with increasing training loss over rounds, reflecting poor adaptation to the scenario. Similar trends are observed in Scenario D and E, where FedOcw demonstrates the best stability, while FedAvg, FedProx, and FedNova continue to show erratic loss patterns. These findings underscore FedOcw's robustness in addressing non-IID data challenges, offering enhanced convergence stability and adaptability across diverse multilingual datasets. The observed training loss trends further highlight its resilience in handling complex learning environments, making it a promising candidate for real-world federated learning applications.

To examine the impact of the weighting strategy on individual clients during the federated learning process, we analyze client model C0 across five scenarios (A, B, C, D, and E) as case studies, focusing on the optimized client weights assigned by FedOcw. Table 8 presents the sample standard deviation (STDEV.S) over 100 rounds for the optimized weights of local clients when updating client model C0 in the five scenarios, considering various layer parameters of the deep learning model.

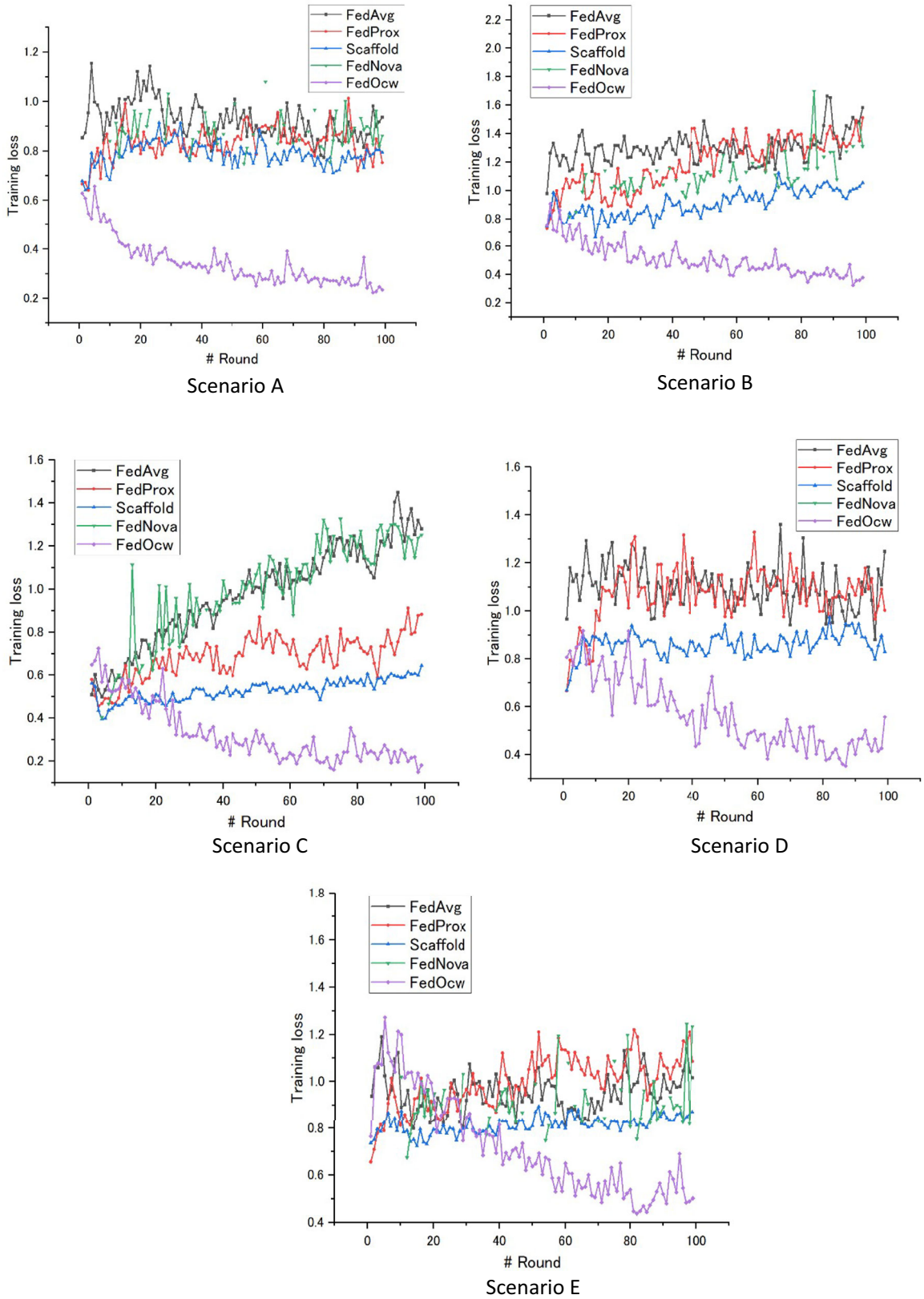

Fig. 3 | Training loss convergence across federated learning models. This figure compares the training loss convergence of five federated learning models—FedAvg, FedProx, Scaffold, FedNova, and the proposed FedOcw—across Scenarios A–E. Lower loss values over communication rounds indicate better convergence and stability. Missing lines for FedNova indicate instances where the training loss was undefined (NaN).

Table 8 | Sample standard deviation (STDEV.S) of optimized client weights over 100 rounds when updating client model C0 across five scenarios (A, B, C, D, and E), considering various layer parameters of the deep learning model

Deep learning model layer	STDEV.S of optimized weights of local clients for updating client model C0				
	Scenario A	Scenario B	Scenario C	Scenario D	Scenario E
Time-distributed 2D-CNNs weight	**0.0336**	**0.0263**	**0.0228**	**0.1111**	**0.0181**
Time-distributed 2D-CNNs bias	0.0111	0.0124	0.0018	0.0028	0.0029
Time-distributed Batch normalization weight	0.0037	0.0067	0.0043	0.0065	0.0053
Time-distributed Batch normalization bias	0.0005	0.0007	0.0004	0.0004	0.0004
1D-CNN weight	0.0136	0.0134	0.0065	0.013	0.0107
1D-CNN bias	0	1.9E-16	1.9E-16	1.6E-17	7.4E-17
Fully connected	0.0017	0.0077	0.0066	0.0091	0.0083

As presented in Table 8, the weights assigned to the Time-Distributed 2D-CNN layer exhibit the highest variability across aggregation rounds, underscoring their critical role in shaping the deep learning model's performance. A similar trend is observed across other client models, indicating the central influence of this layer in the federated learning process. Given this, we focus on the Time-Distributed 2D-CNN layer for a more in-depth analysis of how the weighting strategy impacts individual clients during training.

Figure 4 shows the adjacency matrix of the weights assigned to the Time-Distributed 2D-CNN layer across five scenarios (A, B, C, D, and E). The y-axis represents the clients receiving updates, with each row corresponding to the aggregate weights assigned to the local clients. The weights are averaged over 100 rounds. The color bar visually indicates the weight values, emphasizing the relative importance of each client's input space to the target client receiving updates.

As shown in Fig. 4, FedOcw does not confine weight assignment to clients within the same language group across all scenarios. Instead, updates are exchanged between clients from different linguistic backgrounds, demonstrating that the model enables cross-lingual knowledge transfer without imposing language-based isolation. The weight distribution remains relatively balanced, ensuring that model updates are equitably shared, allowing each client to both contribute to and benefit from diverse sources. Additionally, certain clients receive higher-weighted updates, suggesting that the personalization strategy enhances model performance by dynamically prioritizing influential clients. Importantly, these higher-weighted assignments do not consistently correspond to a specific language group, reinforcing the model's adaptability.

To better understand the dynamics behind these weight assignments, we examined Scenario E to determine whether the most influential clients, defined as those consistently receiving higher weights, correlate with dataset-specific attributes such as training sample size, class distribution, or speech task diversity. Table 9 presents hypothetical examples illustrating this analysis.

As shown in Table 9, The analysis of FedOcw's weight assignment strategy in Scenario E reveals that client influence is not determined solely by dataset size or task diversity. While one might expect larger or more diverse datasets to receive higher weights, FedOcw instead appears to prioritize clients with balanced class distributions, as these tend to contribute more reliable and generalizable updates. Notably, the Czech client (C11) with relatively small dataset and limited task diversity receives one of the highest weight assignments, suggesting that FedOcw values the informativeness and alignment of updates over raw data quantity. This indicates that FedOcw adopts a nuanced aggregation strategy that promotes fairness and generalization by emphasizing the quality and complementary value of each client's contribution rather than relying on size or frequency alone.

Discussion

Our findings highlight the advantages of federated learning (FL) in multilingual settings, with FedOcw enabling cross-lingual knowledge transfer while preserving privacy. Among FL approaches, FedOcw excels in handling heterogeneous data distributions, particularly in linguistically diverse scenarios.

FedOcw consistently outperforms FedAvg, FedProx, Scaffold, and FedNova across key metrics, with statistically significant improvements in specifity ($p = 0.0003$) and Mcc ($p = 0.0011$) compared to FedAvg. While its sensitivity is comparable to that of FedProx ($p = 0.0525$), FedOcw surpasses centralized learning across all evaluated metrics. Statistical validation confirms its significant ($p < 0.05$) improvements in most metrics, reinforcing FL's effectiveness in privacy-preserving, heterogeneous environments.

FedOcw's adaptability is evident across different language pairings. In Spanish–Italian (Scenario A), it achieves the highest accuracy (74.81%) and Mcc (0.502), demonstrating effective transfer between related languages. In Spanish–Chinese (Scenario B), greater linguistic divergence introduces challenges, yet the model maintains strong performance (67.85% accuracy, Mcc = 0.288). In Italian–Chinese (Scenario C), FedOcw achieves high specifity (84.19%) and sensitivity (83.44%), reflecting balanced classification. In the trilingual scenario (Scenario D), it maintains top performance with the highest accuracy (72.53%), F1-score (69.8%), and Mcc (0.465). Even in the most heterogeneous multilingual setting (Scenario E), it maintains robust accuracy (72.63%) and Mcc (0.435), demonstrating its strong generalization ability across diverse linguistic domains.

Language-Wise performance analysis show that Italian clients consistently achieve the highest accuracy, reaching up to 94% in Scenario C, likely due to more consistent speech features or better data quality. Spanish and Chinese clients exhibit more variable performance, influenced by differences in task similarity and language characteristics. The better performance of Scenario C (Italian–Chinese) compared to Scenario B (Spanish–Chinese) suggests greater alignment between Italian and Chinese datasets. Introducing Czech in Scenario D narrows the accuracy gap between Italian and Spanish clients, indicating that increased linguistic diversity enhances generalization. In Scenario E, which includes all five languages, performance becomes more balanced, underscoring FedOcw's ability to generalize effectively amid high linguistic and distributional heterogeneity.

Convergence analysis further highlights FedOcw's stability and efficiency. Unlike competing FL models that struggle with non-IID distributions, FedOcw consistently achieves the lowest and most stable training loss across all scenarios, demonstrating its resilience in handling complex multilingual data.

Examining client weight distributions reveals that FedOcw enables effective cross-lingual knowledge sharing without restricting weight assignments to specific language groups. The model ensures balanced contributions across clients while integrating personalization, prioritizing influential clients without linguistic bias. In Scenario E, the weight distribution remains stable, further affirming FedOcw's capacity for multilingual generalization.

An in-depth analysis of client weights in Scenario E reveals that FedOcw does not favor clients solely based on dataset size or task diversity. Instead, higher weights are assigned to clients with balanced class distributions, as they provide more reliable and generalizable updates. For

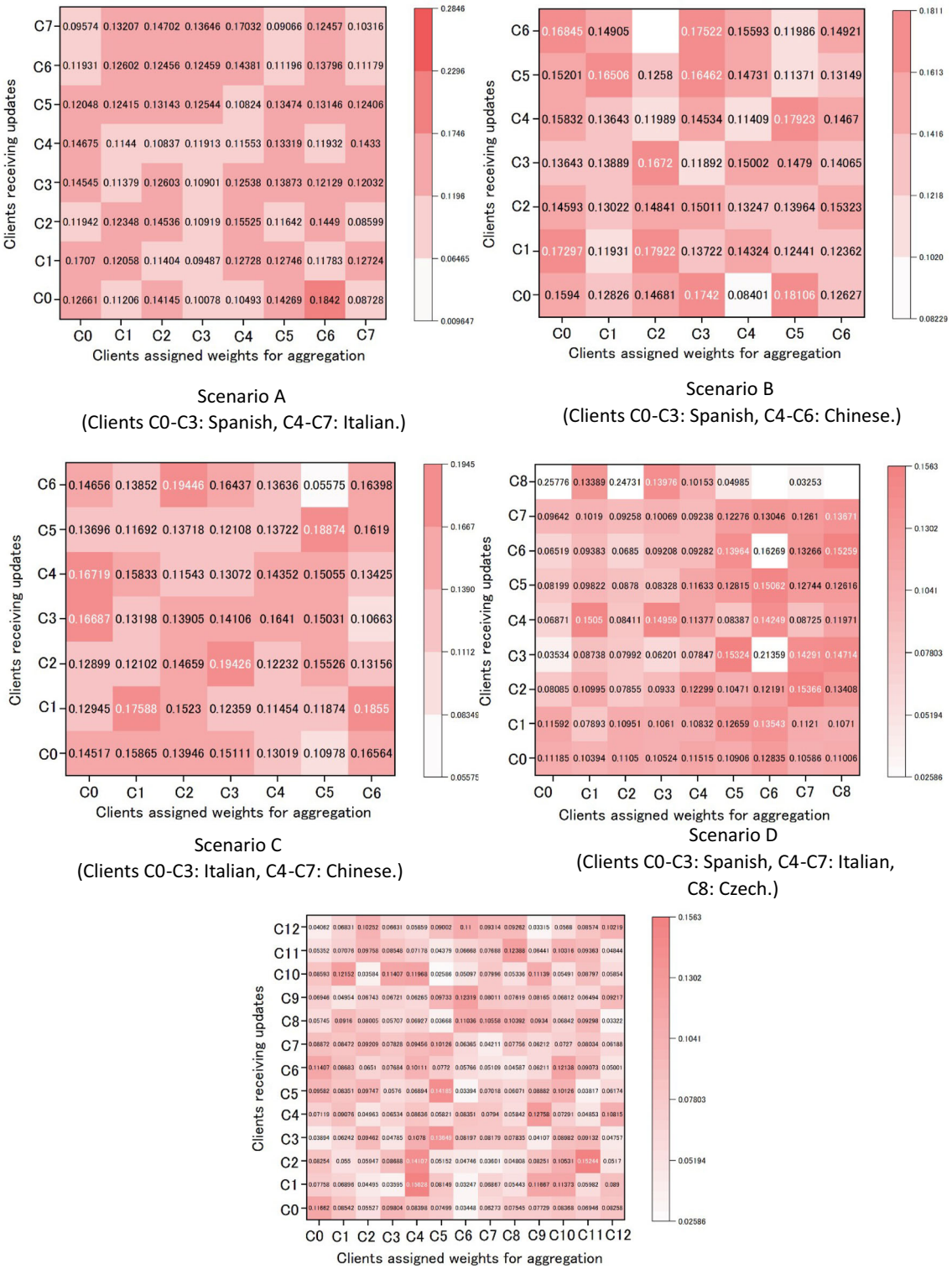

Fig. 4 | Aggregation weights in Time-Distributed 2D-CNN across scenarios. This figure displays the adjacency matrix of client-to-client aggregation weights assigned to the time-distributed 2D-CNN layer, averaged over 100 communication rounds for each scenario (A–E). The *y*-axis represents receiving clients, and each row shows the weights assigned to local clients. Color intensity reflects the relative importance of each client's input in the aggregation process.

example, the Czech client, despite the relatively small dataset and limited task diversity, receives one of the highest weight assignments. This indicates that FedOcw's aggregation strategy prioritizes update quality and alignment over raw data quantity, promoting fairness and improved generalization.

Despite its advantages, the weighting strategy in FedOcw may inadvertently give disproportionate influence to clients with poor convergence, which can degrade overall performance in highly heterogeneous environments. To address this limitation, future work will explore adaptive weighting mechanisms that account for both convergence dynamics and local model quality. Reinforcement learning–based aggregation strategies also present a promising direction for optimizing weight assignments and enhancing robustness in practical deployments.

In our current experimental setup (Scenarios A–E), cross-lingual heterogeneity is simulated by assigning distinct language-specific datasets to different groups of clients. However, we recognize that a more stringent setting, where each client is assigned a fully unique dataset with no overlap, would more closely reflect the diversity encountered in real-world federated learning scenarios. To address this, we plan to extend our framework in future work to support a strict one-to-one mapping between clients and datasets. This may involve assigning each client a different language, recording condition, or assessment protocol, enabling a more comprehensive evaluation of the proposed model's generalizability and robustness in highly heterogeneous environments.

Methods

Federated learning framework

In this study, we propose a novel method within the federated learning framework for determining the weights of client models. This approach, termed FedOcw (Optimized Client Weights for Federated Learning), enables client nodes to develop customized models that adapt to their local datasets.

Figure 5 provides an overview of FedOcw. In the federated learning setup with M clients (as illustrated in Fig. 5), at the onset of a federated aggregation round ($t = 0$), the central server initiates the process by dispatching the initial global model with parameters θ^0 to each local client $k \in \{1, \ldots, M\}$. Each client then performs local training on its private dataset, producing an updated local model θ_k^t and the gradient $\nabla l_k(\theta_k^t)$ of the local loss function $l_k(\theta_k^t)$ with respect to θ_k^t.

These client-specific updates $\{\theta_k^t, \nabla l_k(\theta_k^t)\}$ are sent back to the server. Upon receiving all updates, the server computes a personalized aggregation for each client. Specifically, for client k, the server derives a weighted combination of all client models based on an optimized vector of weights $\boldsymbol{w}_k^t = [w_{k(1)}^t, \cdots, w_{k(M)}^t]^{\mathrm{T}}$, where $w_{k(m)}^t$ denotes the contribution of client m's model to the updated model for client k. This design allows the central server to tailor each client's updated model θ_k^{t+1} using a dynamically weighted fusion of all available models, as shown in the equation at the top of Fig. 5. The updated models $\{\theta_1^{t+1}, \cdots, \theta_M^{t+1}\}$ are then sent back to their corresponding clients, and the process repeats over a certain number of aggregation rounds.

Please note that in the context of federated learning, the client-specific weight vector $\boldsymbol{w}_k^t$ is dynamically optimized for each participating client k based on the local trained model θ_k^t and the gradient $\nabla l_k(\theta_k^t)$ of the local loss function. Consequently, the updated model θ_k^{t+1} is tailored to better fit the local data distribution of client k.

The pseudo code for FedOcw is provided in Algorithm 1.

Table 9 | Preliminary Observations (Hypothetical Example Based on Scenario E)

Client	Language	Samples	PD/HC Ratio	Task Diversity	Avg. Weight
C0	Spanish	High	Imbalanced	High	Medium
C5	Italian	Medium	Balanced	High	High
C8	Chinese	Medium	Imbalanced	Medium	Low
C11	Czech	Low	Balanced	Low	High
C12	English	Low	Slightly imbalanced	Low	Low

Algorithm 1. Optimizing Client Weights for Federated Learning (FedOcw)

Input:

- Each client is indexed by k
- Each communication round is indexed by t
- M: number of clients participating in round t
- n_k: number of data samples on client k
- θ_k^t: model parameters for client k at round t
- $\nabla l_k(\theta_k^t)$: gradient of the local loss function $l_k(\theta_k^t)$ with respect to θ_k^t.

1. Initialize model parameters θ^0 and distribute them to all clients.
2. For each aggregation round t:
   a. Client-side (executed in parallel on all clients):

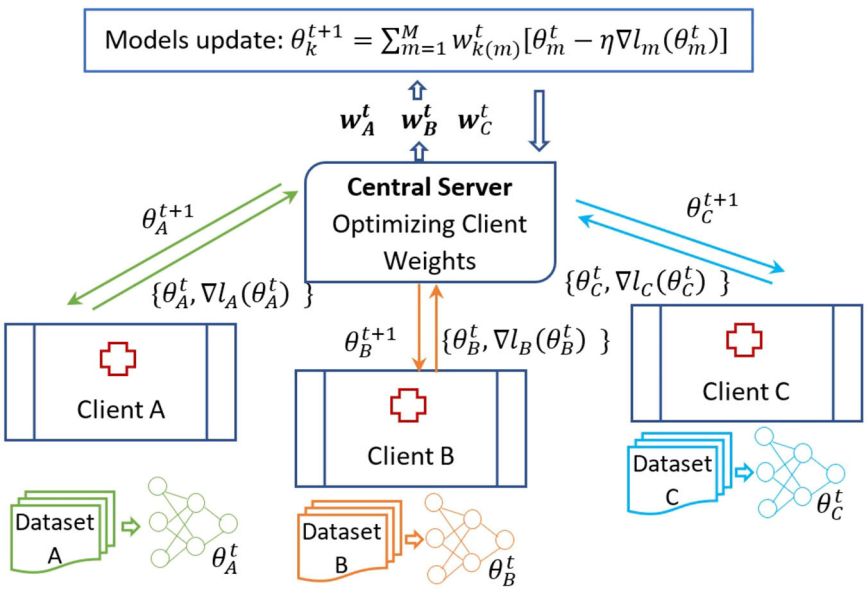

Fig. 5 | Overview of the training process with FedOcw. The figure illustrates the federated learning workflow in which a central server distributes the initial global model to M clients, each of which performs local training on private data to update model parameters and compute local gradients. These updates are then used to optimize client-specific aggregation weights, enabling personalized and stable global model convergence.

- Train the local model using end-to-end deep learning on the client's private dataset.
- Compute and send the updated local parameters θ_k^t and gradient $\nabla l_k(\theta_k^t)$ to the server.

b. Server-side:

- Upon receiving θ_k^t and $\nabla l_k(\theta_k^t)$ from all clients:
  - For each client k (in parallel):
    - Compute the optimal aggregation weights $\boldsymbol{w}_k^t$ using the proposed optimization method.
    - Update the personalized model for client k as:

$$\theta_k^{t+1} = \sum_{m=1}^{M} w_{k(m)}^t \left[\theta_m^t - \eta \nabla l_m(\theta_m^t)\right]$$

As depicted in Algorithm 1, in each aggregation round, every client computes a personalized weight determined by its local empirical loss and the gradient of the local loss function concerning the parameters. This mechanism empowers clients to develop tailored models that aptly adjust to their local datasets.

Previous studies have generally minimized the global loss across all clients by using uniform or static aggregation weights. In contrast, this study adopts a dynamic and personalized approach: at each aggregation round t, the optimal client weights are calculated to maximize the expected reduction in local loss for each client. That is, the weight vector $\boldsymbol{w}_k^t = [w_{k(1)}^t, \cdots, w_{k(M)}^t]^{\mathrm{T}}$ for client k defines the relative contributions of each client $m \in \{1, \ldots, M\}$ to client k's next-round model.

Let $l_k(\theta_k^t)$ denote the empirical loss of client k at round t. The goal is to compute $\boldsymbol{w}_k^t$ such that the expected loss reduction for client k is maximized:

$$\max_{\boldsymbol{w}_k^t} \{l_k(\theta_k^t) - l_k(\theta_k^{t+1})\} \tag{6}$$

where θ_k^t represents current model of client k and θ_k^{t+1} is the updated model.

The updated model for client k is computed as:

$$\theta_k^{t+1} = \sum_{m=1}^{M} w_{k(m)}^t \left[\theta_m^t - \eta \nabla l_m(\theta_m^t)\right] \tag{7}$$

Here, M is the total number of participating clients, $w_{k(m)}^t$ denotes the contribution of client m's update to the model of client k, and η is the learning rate. $\nabla l_m(\theta_m^t)$ is the gradient of the local loss function $l_m(\theta_m^t)$ with respect to θ_m^t

To simplify the optimization, we apply a first-order Taylor approximation to the loss function:

$$l_k(\theta_k^{t+1}) \approx l_k(\theta_k^t) + \nabla l_k(\theta_k^t) \cdot (\theta_k^{t+1} - \theta_k^t) \tag{8}$$

Replacing this approximation and using the aggregation function (Eq. (7)) for θ_k^{t+1}, we obtain Eq. (9):

$$\begin{aligned} \max_{\boldsymbol{w}_k^t} \{l_k(\theta_k^t) - l_k(\theta_k^{t+1})\} &\approx \max_{\boldsymbol{w}_k^t} \{l_k(\theta_k^t) - l_k(\theta_k^t) - \nabla l_k(\theta_k^t) \cdot (\theta_k^{t+1} - \theta_k^t)\} \\ &\approx \max_{\boldsymbol{w}_k^t} \{-\nabla l_k(\theta_k^t) \cdot (\theta_k^{t+1} - \theta_k^t)\} \\ &\approx \max_{\boldsymbol{w}_k^t} \{-\nabla l_k(\theta_k^t) \cdot (\sum_{m=1}^{M} w_{k(m)}^t \left[\theta_m^t - \eta \nabla l(\theta_m^t)\right] - \theta_k^t)\} \end{aligned} \tag{9}$$

The objective function (6) can be reformulated into a minimization problem by negating Eq. (9), yielding Eq. (10):

$$\operatorname*{argmin}_{\boldsymbol{w}_k^t} \left\{ (\boldsymbol{w}_k^t - \boldsymbol{\alpha}_k^t) \cdot \nabla l_k(\theta_k^t) \cdot (\boldsymbol{\theta}^t - \eta \nabla^t) - \eta \|\nabla l_k(\theta_k^t)\|^2 \right\} \tag{10}$$

Since $\eta > 0$, this minimization problem can be further represented by Eq. (11)

$$\operatorname*{argmin}_{\boldsymbol{w}_k^t} (\boldsymbol{w}_k^t - \boldsymbol{\alpha}_k^t) \cdot \nabla l_k(\theta_k^t) \cdot (\boldsymbol{\theta}^t - \eta \nabla^t) \tag{11}$$

subject to

$$\boldsymbol{w}_k^t = \begin{bmatrix} w_{k(1)}^t \\ \vdots \\ w_{k(M)}^t \end{bmatrix} \geq \mathbf{0} \tag{12}$$

(where M is the number of clients participating at round t)and

$$\sum_{i=1}^{M} w_{k(i)}^t = 1 \tag{13}$$

$$\boldsymbol{\alpha}_k^t = \begin{bmatrix} \alpha_{k(1)}^t \\ \vdots \\ \alpha_{k(M)}^t \end{bmatrix}, \qquad \alpha_{k(i)}^t \triangleq \begin{cases} 1, & if\, i = k \\ 0, & if\, i \neq k \end{cases} \tag{14}$$

$$\boldsymbol{\theta}^t = \begin{bmatrix} \theta_1^t \\ \vdots \\ \theta_M^t \end{bmatrix} \tag{15}$$

$$\nabla^t = \begin{bmatrix} \nabla l_1(\theta_1^t) \\ \vdots \\ \nabla l_M(\theta_M^t) \end{bmatrix} \tag{16}$$

The first constraint in equation (12) ensures all client weights are non-negative, while the second constraint in Eq. (13) enforces the sum of weights for all clients to be equal to one, thereby maintaining a balanced contribution to the model parameters θ_k^t and an effective learning rate. However, this optimization problem may have a trivial solution where all weights $w_{k(i)}^t$ converge to $\alpha_{k(i)}^t$, except for the client with the smallest value of $\nabla l_k(\theta_k^t) \cdot (\boldsymbol{\theta}^t - \eta \nabla^t)$. In such a scenario, the system would rely solely on one client to update the parameters, severely hampering aggregation efficiency. To prevent this, an additional regularization term is required, penalizing small values of $\nabla l_k(\theta_k^t) \cdot (\boldsymbol{\theta}^t - \eta \nabla^t)$ from a standard weight $w_k^* = \frac{n_k}{N}$ for client k (where n_k is the number of data samples on client k, $N = \sum_j n_j$ is the total number of data samples).

Let $\boldsymbol{v}_k^t = \begin{bmatrix} v_{k(1)}^t \\ \vdots \\ v_{k(M)}^t \end{bmatrix} = \boldsymbol{w}_k^t - \boldsymbol{\alpha}_k^t$. By adding the regularization term, this minimization problem can be expressed as Eq. (17):

$$\operatorname*{argmin}_{\boldsymbol{V}_k^t} \left\{ \boldsymbol{v}_k^t \cdot \nabla l_k(\theta_k^t) \cdot (\boldsymbol{\theta}^t - \eta \nabla^t) + \mu \left|\left| \boldsymbol{v}_k^t - \boldsymbol{\beta}_k^t \right|\right|^2 \right\} \tag{17}$$

subject to

$$\boldsymbol{v}_k^t \geq -\boldsymbol{\alpha}_k^t \tag{18}$$

Table 10 | Hyperparameter settings for the client-end deep learning model and the control parameters for the CVXOPT solver

Layer/Component	Parameter Settings
Time-distributed 2D-CNNs	Filters = 16, Kernel size = (3,3), Padding = (1,1), Activation = ReLU
Time-distributed Batch Normalization	Momentum = 0.1
Time-distributed 2D AveragePooling	Pool size = (3,3)
Time-distributed Dropout	Dropout rate = 0.3
Time-distributed Flatten	—
1D-CNN	Filters = 8, Kernel size = 3, Padding = 1, Activation = Sigmoid
1D AveragePooling	Pool size = 3, Stride = 1
Flatten	—
Fully Connected Layer	384 hidden units, Activation = ReLU
Dropout	Dropout rate = 0.3
Output Layer	Activation = Sigmoid
Common Settings	Optimizer = Adam, Learning rate = 0.001, Epochs per round = 10, Batch size = 8
CVXOPT Solver Parameters	Maxiters = 20, Abstol = 1e-3, Reltol = 1e-3, Feastol = 1e-3

and

$$\sum_{i=1}^{M} v_{k(i)}^{t} = 0 \tag{19}$$

$$\boldsymbol{\beta}_k^t = \begin{bmatrix} \beta_{k(1)}^t \\ \vdots \\ \beta_{k(M)}^t \end{bmatrix} = w_k^* - \boldsymbol{\alpha}_k^t \tag{20}$$

where

$$w_k^* = \frac{n_k}{N} \tag{21}$$

This quadratic problem can be effectively solved using suitable optimization algorithms, with the regularization term controlled by parameter $\mu > 0$. The value of μ is empirically set and consistently fixed at 0.05 in all experiments.

Once we have obtained $\boldsymbol{v}_k^t$ by solving this quadratic problem, the weight $\boldsymbol{w}_k^t$ can be obtained by:

$$\boldsymbol{w}_k^t = \boldsymbol{v}_k^t + \boldsymbol{\alpha}_k^t \tag{22}$$

The computation complexity of the optimization problem (Eq. (17)) primarily depends on the precomputation of the values of vectors ∇^t and $\boldsymbol{\theta}^t$. As these values can be obtained by locally training the client model k, they can be saved and transmitted directly to the central server.

Client-end deep learning model and hyperparameter settings

This study adopts an end-to-end deep learning architecture for client-side training, which combines time-distributed 2D Convolutional Neural Networks (2D-CNNs) with a 1D Convolutional Neural Network (1D-CNN). The architecture is composed of two key modules: (1) Time-distributed 2D-CNN Module: This module applies a series of 2D convolutional operations independently to each time step, extracting time-series dynamic features from the input log Mel-spectrogram. (2) 1D-CNN Module: The resulting time-series features are then passed through a 1D-CNN block, which captures temporal dependencies between segments. Further details about this model can be found in[38].

The input to the model consists of log Mel-spectrograms derived from speech signals. Each input sample is represented as a sequence of overlapping segments extracted from the spectrogram of a speech recording. Audio recordings were resampled to 22,050 Hz and processed using Librosa[46], with a hop length of 512 and 55 Mel-frequency bands. To handle varying lengths of input recordings, zero-padding was applied to ensure uniform tensor dimensions compatible with PyTorch. The spectrogram frame count and time-series length were fixed at 40 and 50, respectively, to maintain consistent input shapes across batches.

The architecture and parameter settings were guided by prior work[38], where this configuration achieved a favorable balance between model complexity and performance. It is particularly well-suited to resource-constrained and privacy-sensitive environments such as federated learning.

All experiments were implemented in PyTorch and conducted in a consistent computing environment featuring an NVIDIA RTX A6000 GPU, an Intel Xeon Silver 4210 CPU (2.20 GHz), Ubuntu 20.04.6 LTS (64-bit), and 64 GiB of memory. The quadratic optimization problem (Eq. (17)) was solved using CVXOPT[47].

The same set of hyperparameters was used across all experiments. Details of the client-end model architecture and the control parameters used in the CVXOPT solver are summarized in Table 10.

Code availability

The code utilized for training and evaluating the federated learning models in this study is available upon reasonable request. To ensure reproducibility and facilitate further research, the implementation details, including data preprocessing scripts, model architecture, and training configurations, can be accessed by contacting the corresponding author. Additionally, we are committed to open science and intend to release the full codebase on a public repository upon acceptance of this manuscript, ensuring accessibility and transparency in our research methods.

Received: 2 April 2025; Accepted: 1 June 2025;
Published online: 13 June 2025

References

1. Dorsey, E. R. & Bloem, B. R. The Parkinson pandemic—A call to action. *JAMA Neurol.* **75**, 9–10 (2018).
2. Logemann, J. A., Fisher, H. B., Boshes, B. & Blonsky, E. R. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *J. Speech Hear. Disord.* **43**, 47–57 (1978).
3. Sapir, S. et al. Voice and speech abnormalities in Parkinson disease: relationship to severity of motor impairment, duration of disease, medication, depression, gender, and age. *J. Med. Speech-Lang Pathol.* **9**, 213–226 (2001).
4. Forrest, K., Weismer, G. & Turner, G. S. Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults. *J. Acoust Soc. Am.* **85**, 2608–2622 (1989).
5. Ackermann, H. & Ziegler, W. Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *J. Neurol. Neurosurg. Psychiatr.* **5**, 1093–1098 (1991).
6. Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L. & Gates, S. Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **11**, 131–137 (1998).
7. Narendra, N. P., Schuller, B. & Alku, P. The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1925–1936 (2021).
8. Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J. & Rami, L. O. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **59**, 1264–1271 (2012).

9. Sakar, B. E. et al. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J. Biomed. Health Inf.* **17**, 828–834 (2013).
10. Tuncer, T., Dogan, S. & Acharya, U. R. Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. *Biocybern. Biomed. Eng.* **40**, 211–220 (2020).
11. Lahmiri, S., Dawson, D. A. & Shmuel, A. Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures. *Biomed. Eng. Lett.* **8**, 29–39 (2017).
12. Shahbakhi, M., Far, D. & Tahami, E. Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine. *J. Biomed. Sci. Eng.* **7**, 147–156 (2014).
13. Despotovic, V., Skovranek, T. & Schommer, C. Speech based estimation of Parkinson's disease using Gaussian processes and automatic relevance determination. *Neurocomputing* **401**, 173–181 (2020).
14. Vásquez-Correa, J. C. et al. Multimodal assessment of Parkinson's disease: a deep learning approach. *IEEE J. Biomed. Health Inf.* **23**, 1618–1630 (2019).
15. Fujita, T., Luo, Z., Quan, C., Mori, K. & Cao, S. Performance evaluation of RNN with hyperbolic secant in gate structure through application of Parkinson's disease detection. *Appl. Sci*. **11**, https://doi.org/10.3390/app11104361 (2021).
16. La Quatra, M. et al. Exploiting foundation models and speech enhancement for Parkinson's disease detection from speech in real-world operative conditions. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, https://doi.org/10.21437/Interspeech.2024-522 (2024).
17. Gimeno-Gómez, D. et al. Unveiling interpretability in self-supervised speech representations for Parkinson's diagnosis. *IEEE J. Sel. Top. Signal Process.* **99**, 1–14 (2025).
18. Bercea, C. I. et al. Federated disentangled representation learning for unsupervised brain anomaly detection. *Nat. Mach. Intell.* **4**, 685–695 (2022).
19. Dou, Q. et al. Author Correction: Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *npj Digit. Med.* **5**, 56 (2022).
20. Dayan, I. et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat. Med.* **27**, 1735–1743 (2021).
21. Ogier du Terrail, J. et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat. Med.* **29**, 135–146 (2023).
22. Kumar, A., Purohit, V., Bharti, V., Singh, R. & Singh, S. K. Medisecfed: private and secure medical image classification in the presence of malicious clients. *IEEE Trans. Ind. Inf.* **18**, 5648–5657 (2021).
23. Agbley, B. L. Y. et al. Federated fusion of magnified histopathological images for breast tumor classification in the internet of medical things. *IEEE J. Biomed. Health Inf*. **28**, 3389–3400 (2023).
24. Feng, B. et al. Robustly federated learning model for identifying high-risk patients with postoperative gastric cancer recurrence. *Nat. Commun.* **15**, 742 (2024).
25. Peng, L. et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *npj Digit. Med.* **7**, 127 (2024).
26. Dipro, S. H., Islam, M., Nahian, M., Al, A. & Azad, M. S. *A federated learning approach for detecting Parkinson's Disease through privacy preserving by blockchain*. Ph.D. Thesis, Brac University (2022).
27. Sarlas, A., Kalafatelis, A., Alexandridis, G., Kourtis, M.-A. & Trakadas, P. Exploring Federated learning for speech-based Parkinson's disease detection. In *The 18th International Conference on Availability, Reliability and Security (ARES 2023)*, August 29-September 01, 2023, Benevento, Italy. https://doi.org/10.1145/3600160.3605088 (ACM, 2023).
28. McMahan, H. B. et al. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics* (PMLR, 2017).
29. Li, T. et al. Federated optimization in heterogeneous networks. In *Proc. MLSys Conference*, Austin, TX, USA (PMLR, 2020).
30. Huang, W. et al. Fairness and accuracy in horizontal federated learning. *Inf. Sci.* **589**, 170–185 (2022).
31. Mansour, Y., Mohri, M., Ro, J. & Suresh, A. T. Three approaches for personalization with applications to federated learning. *arXiv* preprint arXiv:2002.10619 (2020).
32. Wang, K. et al. Federated evaluation of on-device personalization. *arXiv* preprint arXiv:1910.10252 (2019).
33. Karimireddy, S. P. et al. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143 (PMLR, 2020).
34. Wang, J., Liu, Q., Liang, H., Joshi, G. & Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Adv. Neural Inf. Process. Syst*. **33**, 7611–7623 (2020).
35. Huang, Y. et al. Personalized cross-silo federated learning on non-iid data. In *Proc. AAAI Conference on Artificial Intelligence*, Vol. **35**, 7865–7873 (ACM, 2021).
36. Smith, V., Chiang, C.-K., Sanjabi, M., Talwalkar, A. S. Federated multi-task learning. *Adv. Neural Inform. Process. Syst*. **30**, 4424–4434 (2017).
37. Quan, C., Ren, K. & Luo, Z. A deep learning-based method for Parkinson's disease detection using dynamic features of speech. *IEEE Access* **9**, 10239–10252 (2021).
38. Quan, C., Ren, K., Luo, Z., Chen, Z. & Ling, Y. End-to-end deep learning approach for Parkinson's disease detection from speech signals. *Biocybern. Biomed. Eng.* **42**, 556–574 (2022).
39. Botelho, C., Schultz, T., Abad, A. & Trancoso, I. Challenges of using longitudinal and cross-domain corpora on studies of pathological speech. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 1921–1925. https://doi.org/10.21437/Interspeech.2022-10995 (2022).
40. Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., González-Rátiva, M. C. & Nöth, E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Proc. LREC 2014*, 342–347 (ELRA, 2014).
41. Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D. & Girardi, F. Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access* **5**, 22199–22208 (2017).
42. Dimauro, G. & Girardi, F. Italian Parkinson's voice and speech. *IEEE Dataport*, June 11, https://doi.org/10.21227/aw6b-tg17 (2019).
43. Hlavnička, J., Čmejla, R., Tykalová, T., Šťastná, M. & Rektorová, I. Acoustic tracking of pitch, modal, and subharmonic vibrations of vocal folds in Parkinson's disease and parkinsonism. *IEEE Access* **7**, 150339–150354 (2019).
44. Trivedi, D., Jaeger, H. & Stadtschnitzer, M. Mobile device voice recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. Zenodo https://doi.org/10.5281/zenodo.2867216 (2019).
45. Huang, F., Xu, H., Shen, T. & Jin, L. Recognition of Parkinson's disease based on residual neural network and voice diagnosis. In *Proc. 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 381–386 https://doi.org/10.1109/ITNEC52019.2021.9586915 (2021).
46. McFee, B. et al. Librosa: v0.5.0 (2021). https://doi.org/10.5281/zenodo.293021. Accessed 1 March 2023.
47. Andersen, M. S., Dahl, J. & Vandenberghe, L. CVXOPT: A Python package for convex optimization (2019). http://cvxopt.org. Accessed 10 July 2023.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP25K15078.

Author contributions

C.Q. conceptualized the study, developed the federated learning framework, implemented the FedOcw algorithm, and drafted the

manuscript. Z.C. performed experimental analysis and performance evaluations. K.R. managed data collection and preprocessing. Z.L. provided critical feedback, contributing to research design, analysis, and manuscript revision. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Changqin Quan.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.