FedMedICL: Towards Holistic Evaluation of Distribution Shifts in Federated Medical Imaging

Multimedia Content Analysis Report (CS6880)

Sanskriti Agarwal, Ishu Priya, Prof C. Krishna Mohan

***Abstract*—The authors highlight that achieving generalization is essential for medical imaging AI models to be clinically reliable. However, this goal is challenged by the presence of multiple, co-existing distribution shifts—such as label imbalance, demographic variation, and temporal change—as well as the limited diversity of data siloed within individual medical institutions. While federated learning offers a promising solution to the issue of data decentralization, current benchmarks fail to evaluate these distribution shifts jointly. To address this gap, the authors propose FedMedICL, a unified framework and benchmark that captures the combined effects of label, demographic, and temporal shifts in federated medical imaging scenarios. Through extensive experiments spanning six diverse datasets (amounting to 550 GPU hours), including simulations of COVID-19 spread across hospitals, they assess the adaptability of various federated and continual learning methods. Interestingly, their results show that a simple batch balancing method consistently outperforms more advanced techniques, raising concerns about the reliability of insights drawn from earlier benchmarks that focus on isolated shift types.**

***Index Terms*—Federated learning, continual learning, medical imaging, domain shift, generalization, benchmark**

I. Introduction

Medical imaging has benefited tremendously from advances in machine learning, especially in automating diagnosis and enhancing clinical decision-making. However, a persistent challenge is to ensure that models trained on available data generalize well across diverse clinical scenarios. A primary obstacle is the presence of *distribution shifts*, wherein differences in data characteristics (such as label prevalence, patient demographics, and temporal changes) between training datasets and real-world clinical environments can lead to significant performance degradation.

Data pooling could be a potential solution, but in practice, institutional policies and delayed access to large-scale data prevent its widespread use. Federated learning offers a promising alternative by allowing collaborative model training while preserving data privacy. Yet, existing evaluations tend to overlook the combined effects of different types of distribution shifts. **FedMedICL** addresses this gap by providing a framework that integrates label, demographic, and temporal shifts into a unified benchmark for federated medical imaging.

II. Motivation

The impetus for developing **FedMedICL** stems from several real-world challenges commonly encountered in medical imaging and AI deployment:

A. Label Imbalance

Certain diseases occur more frequently in specific hospitals than in others. As a result, AI models trained on data from one region may underperform when exposed to environments where different, less-represented diseases are more prevalent. This imbalance can hinder the model's ability to recognize rare but critical conditions.

B. Demographic Imbalance

The patient population varies significantly across institutions. For instance, some hospitals may predominantly serve elderly patients, while others cater to a younger demographic. These variations lead to differences in imaging data distribution, affecting model accuracy and generalization across age groups.

C. Temporal Shifts

Disease prevalence is subject to change over time, driven by seasonal trends or emerging health crises such as pandemics. Models that are not continuously updated may fail to recognize these shifts, leading to a drop in diagnostic performance. Temporal adaptability is therefore essential for clinically reliable AI systems.

III. Problem Statement

In the federated learning framework, multiple hospitals collaborate to train a shared global model while maintaining the privacy of their local data. Traditional evaluation strategies tend to focus on isolated distribution shifts or employ benchmarks that do not adequately reflect the siloed and heterogeneous nature of real-world medical data.

FedMedICL introduces a dual-function problem formulation that aims to holistically capture the challenges faced in federated medical imaging:

1) Unified Modeling: The framework systematically models the combined effects of label imbalance, demographic variation, and temporal shifts. By addressing these three types of distribution shifts simultaneously, FedMedICL better reflects the dynamic, multi-faceted challenges observed in clinical deployment.

2) Comprehensive Benchmark: FedMedICL leverages available demographic metadata to automatically generate federated and continual learning tasks. This enables the creation of a realistic testbed for evaluating federated learning algorithms under diverse and evolving healthcare settings.

This unified problem formulation is designed not only to assess how well models perform within individual hospitals but also to measure their ability to generalize across unseen client environments with distinct patient populations and temporal dynamics.

IV. Related Works

Two notable efforts provide foundational context for the development of FedMedICL:

1) SubpopBench [1]: SubpopBench is a benchmark designed to evaluate the robustness of models to subpopulation shifts. However, it does not incorporate temporal shifts and does not assess the combined effect of multiple shift types. Its scope is therefore limited when applied to real-world clinical settings where such shifts frequently co-occur.

2) MEDFAIR [2]: MEDFAIR is another benchmark that emphasizes fairness across demographic groups in medical imaging AI. While it effectively captures demographic disparities, it does not address label imbalance or temporal changes. Moreover, it overlooks the federated nature of medical data, limiting its applicability in privacy-preserving training scenarios.

These gaps in existing literature highlight the need for a benchmark like FedMedICL—one that integrates multiple distribution shifts and aligns with the decentralized, privacy-sensitive nature of modern healthcare data infrastructure.

V. Background: Federated and Continual Learning

A. Federated Learning (FL)

In FL, the dataset is split across K hospitals:

$$D = \{D_1, D_2, \ldots, D_K\}$$

where D_k represents the dataset local to hospital k. Each hospital serves a distinct patient population, resulting in a different demographic distribution. For instance, Hospital A may predominantly have elderly patients, while Hospital B may cater to younger individuals.

The probability of observing a specific attribute a_i (e.g., "elderly" or "male") at hospital k is defined as:

$$p_i^k = P(x, y) \sim D_k(a = a_i)$$

Here, p_i^k represents the likelihood of seeing attribute a_i in data sampled from hospital k.

B. Continual Learning (CL)

In continual learning, instead of splitting data across hospitals, a single hospital's data is divided into time-based segments:

$$D_i = \{D_{i1}, D_{i2}, \ldots, D_{iT}\}$$

where D_{it} denotes the dataset from hospital i at time t. This temporal partitioning allows tracking of shifts over time.

C. Learning Setup

- **Input Image and Label (x, y):** A medical image x is associated with a diagnosis label $y \in Y = \{1, 2, \ldots, L\}$.
- **Classifier Function (f_θ):** The model f_θ maps input images to a probability distribution over disease labels:

$$f_\theta : X \to P(Y)$$

- **Patient Attributes:** Additional metadata (e.g., age, sex) are represented as:

$$A = \{a_1, a_2, \ldots, a_m\}$$

VI. Proposed Methodology

We define a federated learning system with K clients (hospitals), each possessing a private dataset D_k. The objective is to simulate realistic distribution shifts across both space (between hospitals) and time (evolving over time).

Client Splitting: Balanced vs. Skewed Clients

Each client (hospital k) holds a dataset of medical records represented as:

$$(x, y) \sim P_k(x, y)$$

where x is the patient data (e.g., images), y is the diagnosis label, and P_k is the local data distribution at hospital k.

- **Balanced Clients:** A hospital is considered balanced if all demographic groups are equally represented:

$$\forall i, j \in \{1, \ldots, m\}, \quad p_k^i \approx p_k^j$$

- **Skewed Clients:** A hospital is skewed if at least one demographic group dominates:

$$p_k^i \gg \frac{1}{m}$$

  Example: A pediatric hospital will have most samples from the "children" group.

By mixing both types of clients, FedMedICL replicates real-world demographic heterogeneity across institutions.

Temporal Task Splitting: Modeling Distribution Over Time

To incorporate temporal dynamics, each client's dataset is further split into T time-based segments $\{D_{k1}, D_{k2}, \ldots, D_{kT}\}$.

Localized Split: Seasonal Demographic Changes

Hospitals experience time-dependent shifts in patient attributes. For any two hospitals k and l at time t:

$$p_k^i(t) \neq p_l^i(t)$$

Example: Hospital k in a cold region may see more flu cases in winter, unlike hospital l in a tropical region.

Novel Disease Split: Emergence of New Conditions

A new disease label y_{new} is introduced at time T:

$$\text{At } T = 1 : \quad P(y = y_{\text{new}}) = 0 \quad \forall D_{i1}$$

$$\text{At } T > 1: \quad \exists D_{ij} \text{ such that } P(y = y_{\text{new}}) > 0$$

This setting simulates the emergence of novel diseases (e.g., COVID-19) that appear in some hospitals but not others, mimicking real-world outbreak propagation patterns.

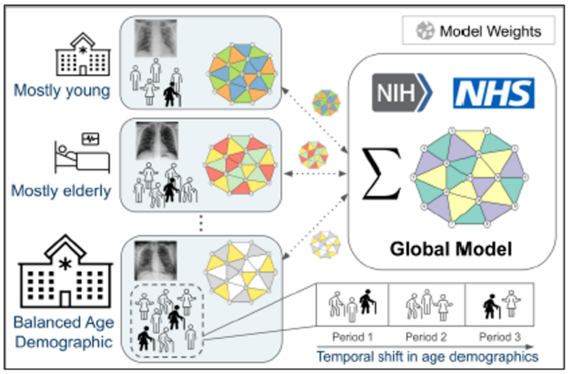


Fig. 1: **Problem Setup.** We model a federated medical imaging scenario, in which siloed hospitals experience demographic imbalances and temporal shifts.

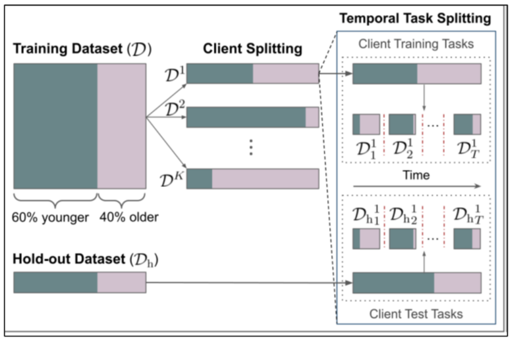


Fig. 2: **FedMedICL Benchmark Construction.** We construct client datasets (D_1 to D_K), each representing a hospital with unique demographic characteristics and temporal training tasks (D_{i1} to D_{iT}). We evaluate models on temporally aligned test tasks for adaptability to local demographic shifts, and on a hold-out set (D_h) for generalization to unseen demographics.

VII. Baseline Methods

To benchmark the robustness of FedMedICL under distribution shifts, several existing learning strategies were adapted to the federated learning setting:

- **Empirical Risk Minimization (ERM):** Standard training baseline that learns directly from available local data without adjustment for imbalance or shift.
- **Data Augmentation:** MixUp [3] generates new training samples by linear interpolations of inputs and labels to improve generalization.
- **Domain Generalization:** SWAD [4] uses stochastic weight averaging with early stopping to prevent overfitting to local domains.
- **Continual Learning:** Experience Replay (ER) [5] replays a buffer of past samples to retain knowledge across tasks.
- **Handling Imbalanced Data:**
  - *Group Balancing (GB)* – Reweights losses across demographic groups.
  - *Class Balancing (CB)* – Reweights based on class frequency.
  - *Class Rebalancing with Two-stage Training (CRT)* [6] – First trains representations, then retrains classifiers on balanced data.

All methods were adapted for federated learning by integrating with FedAvg and are denoted with an "F-" prefix (e.g., F-CRT, F-MixUp).

VIII. Datasets

FedMedICL uses six diverse public medical imaging datasets covering multiple modalities, disease types, and demographics:

- **CheXpert** [7]: Chest X-ray images with multi-label disease annotations. Split among $K = 50$ clients due to dataset size.
- **Fitzpatrick17k** [8]: Skin images with Fitzpatrick skin-type metadata. Emphasizes fairness across skin tones.
- **HAM10000** [9]: Dermatoscopic images for skin lesion diagnosis.
- **OL3I** [10]: Abdominopelvic CT scans with ischemic heart disease risk labels.
- **PAPILA** [11]: Fundus images for glaucoma diagnosis.
- **CheXCOVID (Novel Dataset)**: Combination of CheXpert with COVID-19 patient images to simulate real-time emergence of a novel disease.

IX. Metrics

Many of the datasets used in FedMedICL suffer from significant class imbalance, where certain disease labels are overrepresented. In such settings, naively predicting the most frequent class can result in deceptively high overall accuracy, failing to reflect true model robustness.

To mitigate this, FedMedICL follows the principles of **Long-Tailed Recognition (LTR)** and employs:

- **LTR Accuracy:** Ensures fair evaluation by considering class-wise performance, rather than being dominated by large classes. It calculates accuracy across all classes equally, making it suitable for imbalanced medical data.
- **Imbalance Factor (IF):** Quantifies the degree of class imbalance in a dataset. It is computed as:

$$\text{IF} = \frac{\text{Size of the largest class}}{\text{Size of the smallest class}}$$

A higher IF value indicates a greater imbalance, meaning that the most common class heavily outnumbers the rarest one.

By incorporating LTR metrics, FedMedICL provides a more realistic and fair assessment of model generalization across both frequent and rare disease categories.

X. Experimental Setup

The experimental design in FedMedICL reflects real-world federated learning scenarios with multiple clients (hospitals), each experiencing demographic and temporal variations. The following setup was used across all experiments:

Client and Task Splitting

- Each dataset is split among $K = 10$ clients, where each client represents an individual hospital.
- Every client is assigned $T = 4$ training tasks, designed to simulate temporal changes such as seasonal variations in disease prevalence and hospital admissions.
- **Example:** Flu season may significantly shift the age distribution of hospital patients.
- Each client also receives $T = 4$ temporally aligned test sets, with demographic distributions matching their respective training tasks.

Training Process

- Training proceeds across multiple communication rounds (**150 rounds**).
- Each communication round includes:
  - **5 local iterations** per client.
  - **Batch size** of 10 for local updates.
  - **Federated Averaging (FedAvg)** is performed after each communication round to aggregate model weights.
- After completing each training task, the model is evaluated on all previous tasks to measure **continual learning performance** and the ability to retain past knowledge.

Final Evaluation

- Once all $T = 4$ training tasks are completed, the final model from each client is evaluated on:
  - Their respective local test sets (for task-specific adaptation).
  - A **shared global hold-out set** (D_h), which reflects the overall demographic distribution across the entire population.
- This setup enables assessment of both **local adaptation** and **global generalization** capabilities of the models.

TABLE I: Experimental configuration details

Architecture	ResNet18 (also tried ResNet34, ResNet50)
Batch size	10
Imbalance ratios	“balanced”: 0.2, “spare”: 0.2
GPU	A100, L4 GPU
PyTorch version	2.6.0 + cu124
Number of clients	10
Learning rate	0.0001
Federated learning (use_FL)	True for all methods, except ERM
Number of communication rounds	150
Number of local iterations	5
Number of tasks	4
Optimizer	SGD, Adam

XI. Reproduced Results

The F-CB (class-balancing) method (green curve in the plots) performed best on 5 out of 6 datasets, outperforming more advanced methods like F-SWAD and F-CRT.

Although these methods worked well in previous benchmarks (SubpopBench, MEDFAIR), their performance degraded in the federated learning setup.

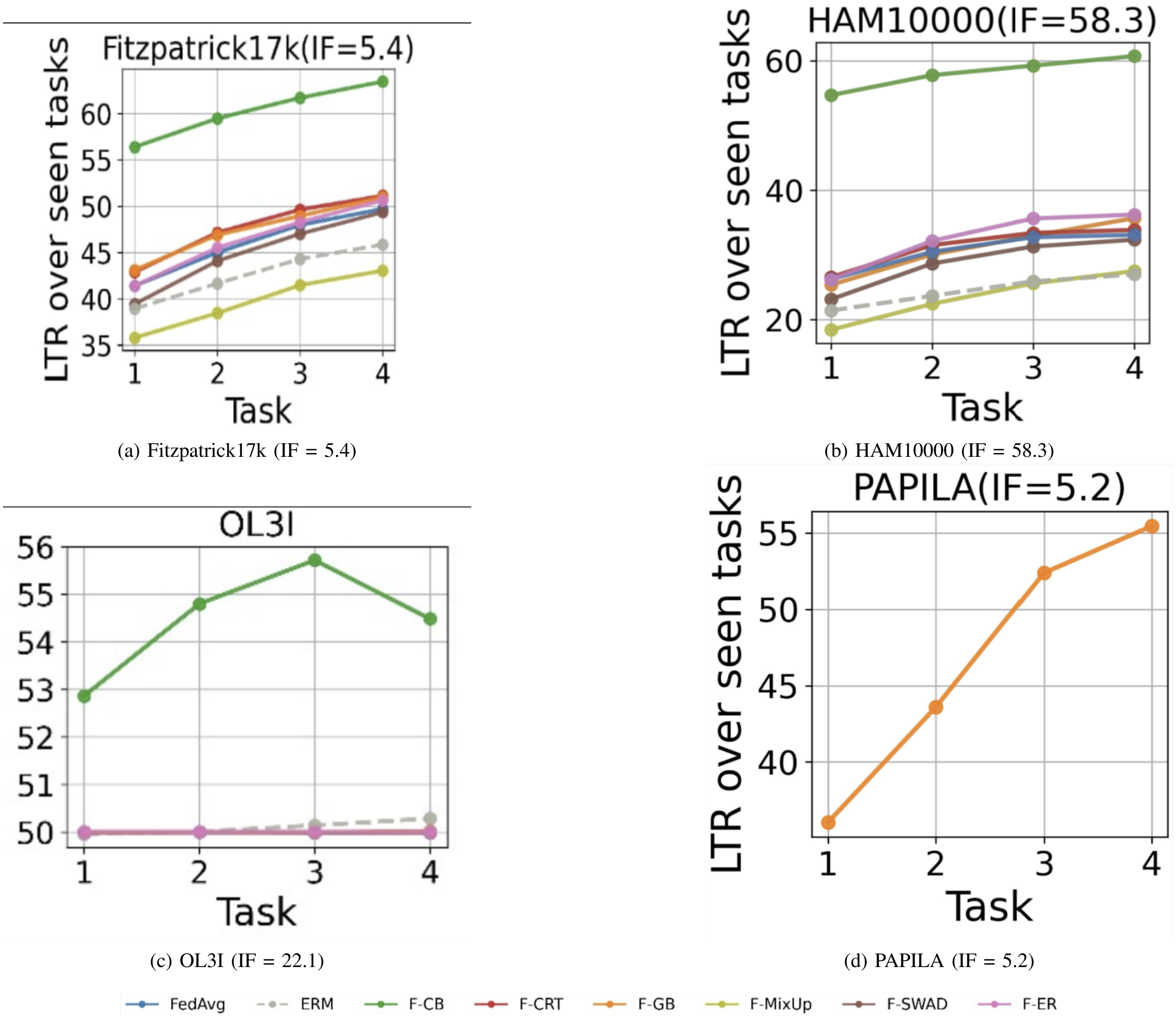
Why Did Advanced Methods Fail?

- **F-SWAD** relies on dense stochastic weight averaging, which takes too long to adapt. This makes it unsuitable for federated learning, where models must quickly adjust to new data distributions.

- **F-CRT** uses a two-stage training process, which is inefficient for continual and federated learning where frequent updates are required.

Key Takeaway

The results show that models performing well on local client data also generalize better to unseen demographic distributions. This confirms that handling just one type of data shift is not enough—federated learning requires adaptability across multiple shifts at the same time.



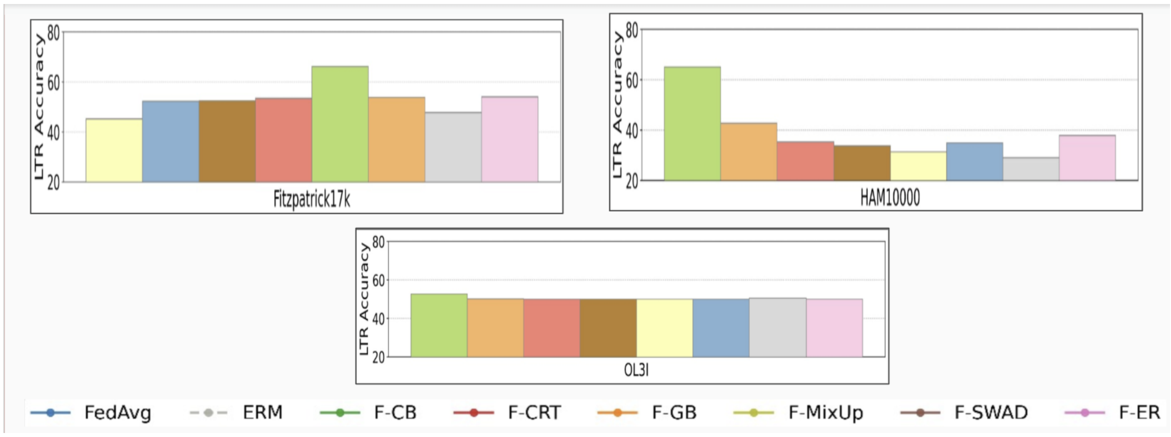
(a) Fitzpatrick17k (IF = 5.4)

(b) HAM10000 (IF = 58.3)

(c) OL3I (IF = 22.1)

(d) PAPILA (IF = 5.2)

Fig. 3: Top: Test accuracy across 4 sequential tasks, averaged over 10 clients and 5 runs. Each curve shows the mean `test_per_category_acc` per method, highlighting generalization and retention performance over time.



XII. Novelty: Methods & Results

Abstract : Over successive iterations, our work began by identifying a critical gap in FedMedICL: although it simulated label, demographic, and temporal shifts, it lacked mechanisms to reconcile client-specific and global feature distributions. To address this, we introduced FedDG, which employs dual batch normalization and a lightweight adapter network to learn an optimal interpolation between local and global statistics—an approach essential for medical imaging, where acquisition protocols and patient populations vary widely. Recognizing that domain adaptation alone could not overcome severe label imbalance, we then developed FedDGHybrid by integrating focal loss with class-balanced weighting into the dual-BN framework, ensuring that rare but clinically important conditions receive sufficient emphasis during optimization. Building on this, FedAIM advances the paradigm by meta-learning loss weights in response to global performance feedback—targeting minority-class recall—while enforcing consistency between local and global BN outputs and leveraging knowledge distillation to transfer domain-invariant knowledge. This progression—from basic domain alignment to simultaneous imbalance mitigation and finally to adaptive, performance-guided optimization—reflects a systematic strategy tailored to the non-IID and imbalanced realities of federated healthcare applications.

A. Data & Preprocessing

We use the CDC *COVID-19 Case Surveillance Public Use* dataset (8.4 M records, 5 death prevalence). After filtering to retain only records with `death_yn` $\in$ {Yes, No}, we shuffle (seed = 0) and split into 5 IID shards ($\approx$ 1.68 M each). Each shard is further split 80/20 into local train/test.

- **Features**: Categorical: {`age_group`, `sex`, `race_ethnicity`, `medcond_yn`}. Target: binary `death_yn`.
- **Encoding**: Fill missing → `Unknown`, cast to `category`, map to integer codes $[0, V_i - 1]$.
- **Batching**: DataLoader(batch_size = 4096, shuffle=True, drop_last=True).

B. Model: TabTransformer

We adopt a categorical-only TabTransformer with:

- **Embeddings**: C features, vocab sizes V_i, embedding dim $d = 256$.
- **Transformer Encoder**: 4 layers, 8 heads, feed-forward dim 1024, layer-norm replaced by BatchNorm1d for domain adaptation.
- **Classifier**: mean-pool transformer outputs → BN → Linear(256→1) $\to \sigma$.

Total parameters $\approx$ 3.2 M.

C. Federated Algorithms

All methods run 20 global rounds; local update: SGD (LR = 0.01, momentum 0.9, weight decay = 1e-4).

a) FedAvg: Standard FedAvg: each client trains 1 epoch, sends weights for simple averaging.

b) FedDG: Episodic domain-generalization:

$$\hat{\theta}_k = \theta_k - \beta \nabla_\theta L(D^k_{\text{src}}; \theta_k),$$
$$\theta_k \leftarrow \theta_k - \alpha \nabla_\theta L(D^k_{\text{aug}}; \hat{\theta}_k),$$

with $\beta = 5 \times 10^{-3}$, $\alpha = 2 \times 10^{-3}$, and Fourier-amplitude augmentation to simulate unseen domains.

c) FedDGHybrid: Two-stage:

1) $T_1 = 10$ rounds FedAvg pre-training.
2) $T_2 = 10$ rounds FedDG fine-tuning on the checkpoint.

d) FedAIM: Adaptive imbalance mitigation:

1) Server updates positive weight $w_+ \leftarrow w_+ + \eta(R^* - r)$ ($\eta = 0.1$, $R^* = 0.7$).
2) SMOTE-style oversampling (neighbors=5).
3) Teacher–student KL distillation with hard-mining.
4) BN-consistency $L_{\text{consist}} = D_{KL}(p_{\text{globalBN}} \parallel p_{\text{localBN}})$ (weight $\lambda = 0.5$).

D. Results

Performance on death prediction (averaged over 4 outcome tasks):

Method	Acc.	Prec.	Rec.	F1	ROC AUC	AP
FedCB	0.909	0.302	0.706	0.423	0.923	0.384
FedDG	0.940	0.381	0.584	0.461	0.924	0.396
Hybrid	0.917	0.299	0.675	0.413	0.912	0.327
FedAIM	0.953	0.401	0.514	0.450	0.922	0.383

TABLE II: 20-round federated performance (death prediction).

Key Insight. FedDG best balances precision–recall via domain-shift meta-learning; FedAIM meets recall targets with controlled trade-offs; Hybrid accelerates recall gains; FedCB maximizes recall at precision cost.

XIII. Conclusion

We introduced **FedMedICL**, a benchmark for evaluating federated continual learning under realistic medical imaging scenarios with demographic imbalances and temporal shifts. Our experiments across four healthcare datasets show that simple class-balancing methods like **F-CB** outperform more complex approaches in both retaining knowledge and generalizing to new distributions. These results highlight the need for fast-adapting, imbalance-aware methods in federated healthcare settings. **FedMedICL** provides a foundation for future research toward robust and fair federated learning in medical applications.

References

[1] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, “Change is hard: A closer look at subpopulation shift,” in *International Conference on Machine Learning*, 2023.

[2] Y. Zong, Y. Yang, and T. Hospedales, “Medfair: Benchmarking fairness for medical imaging,” in *International Conference on Learning Representations*, 2023.

[3] H. Zhang *et al.*, “Mixup: Beyond empirical risk minimization,” in *International Conference on Learning Representations*, 2018.

[4] J. Cha *et al.*, “Swad: Domain generalization by seeking flat minima,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[5] A. Chaudhry *et al.*, “On tiny episodic memories in continual learning,” *arXiv preprint arXiv:1902.10486*, 2019.

[6] B. Kang *et al.*, “Decoupling representation and classifier for long-tailed recognition,” in *International Conference on Learning Representations (ICLR)*, 2020.

[7] J. Irvin *et al.*, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” *AAAI*, 2019.

[8] M. Groh *et al.*, “Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] P. Tschandl *et al.*, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, 2018.

[10] J. Zambrano Chaves *et al.*, “Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data,” *medRxiv*, 2021.

[11] O. Kovalyk *et al.*, “Papila: Dataset with fundus images and clinical data for glaucoma assessment,” *Scientific Data*, 2022.