
Optimising Clinical Federated Learning through Mode Connectivity-based Model Aggregation

Anshul Thakur^{1 ψ}

Soheila Molaei^{1 ψ}

Patrick Schwab²

Danielle Belgrave²

Kim Branson²

David A. Clifton^{1,3}

¹Institute of Biomedical Engineering, University of Oxford, UK

²GlaxoSmithKline, London, UK

³Oxford-Suzhou Institute of Advanced Research (OSCAR), Suzhou, China

^{ψ} Equal contribution

Abstract

Federated Learning (FL) involves a server aggregating local models from clients to compute a global model. However, this process can struggle to position the global model in low-loss regions of the parameter space for all clients, resulting in subpar convergence and inequitable performance across clients. This issue is particularly pronounced in non-IID settings, common in clinical contexts, where variations in data distribution, class imbalance, and training sample sizes result in client heterogeneity. To address this issue, we propose a mode connectivity-based FL framework that ensures the global model resides within the overlapping low-loss regions of all clients in the parameter space. This framework models the low-loss regions as *non-linear mode connections* between the current global and local models, and optimises to identify an intersection among these mode connections to define the new global model. This approach enhances training stability and convergence, yielding better and more equitable performance compared to standard FL frameworks like federated averaging. Empirical evaluations across multiple healthcare datasets demonstrate the benefits of the proposed framework.

1 INTRODUCTION

Federated Learning (FL) has become a vital approach for developing clinical models across distributed healthcare datasets while addressing privacy concerns and regulatory constraints (Soltan et al., 2023; Molaei et al., 2024; Rieke et al., 2020). Instead of centralising data, FL transmits model updates, enabling collaborative model development while maintaining data privacy, making it particularly suitable for healthcare (McMahan et al., 2017; Thakur et al., 2021).

Despite its potential, FL faces significant optimisation challenges, especially due to client heterogeneity, where data across clients often follows non-IID (non-Independent and Identically Distributed) distributions (Li et al., 2020; Kairouz et al., 2021). This heterogeneity—marked by differences in class distribution, data characteristics, and sample sizes—causes substantial divergence in client-specific models. Most FL frameworks use Federated Averaging (FedAvg), which aggregates client-specific models into a global model during each training round (McMahan et al., 2017). However, this divergence, combined with the non-convex and highly irregular nature of neural network loss landscapes (Li et al., 2018), often results in global models that settle in high-loss regions of the parameter space during training (Zhou et al., 2023), as illustrated in Figure 1. Consequently, this leads to poor convergence and inconsistent performance across clients (Karimireddy et al., 2020).

These challenges are particularly pronounced in clinical applications, where FL is deployed across a diverse range of medical institutions that vary in size, geographic location, and patient demographics (Rieke et al., 2020; Lu et al., 2022). This diversity exacer-

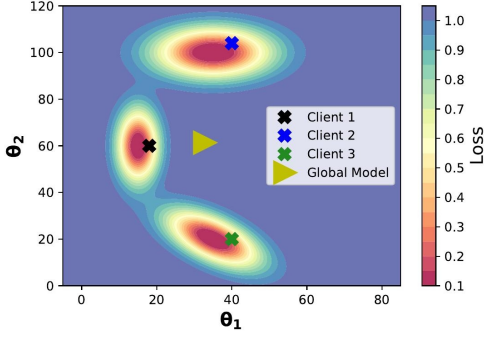


Figure 1: Illustration of a hypothetical scenario in parameter space where the divergence of client models results in a global model being situated in a high-loss region within the loss landscape.

bates the heterogeneity of data characteristics among clients, complicating the optimisation process and increasing the likelihood of the global model being suboptimal or exhibiting uneven performance across institutions. While many clinical FL frameworks still rely on Federated Averaging (FedAvg), methods like FedProx (Li et al., 2020) and SCAFFOLD (Karimireddy et al., 2020) have been proposed to alleviate the impact of client heterogeneity. These methods primarily aim to stabilise convergence and improve model performance by addressing client divergence. However, these methods do not explicitly address the issue of client fairness, which is crucial to ensuring that the global model performs equitably across all clients, irrespective of disparities in data distribution. This is particularly important in clinical contexts, where subpar performance on data from smaller or underrepresented institutions could have severe consequences. Therefore, there is a pressing need for FL frameworks that not only overcome data heterogeneity but also enhance client fairness, ensuring robust and equitable model performance across all participants.

In response to this need, this paper presents a novel FL framework that directly addresses client model divergence and the resulting suboptimal global model aggregation, thereby tackling the challenges of client heterogeneity more effectively. The framework leverages the concept of *mode connectivity* (Garipov et al., 2018) to ensure that the global model is positioned at or near the intersection of the low-loss regions corresponding to each client’s model. In this approach, each client learns a non-linear low-loss path, or manifold, called a mode connection, which links the global model with its local model within the parameter space. Every point on this path represents a viable optimal solution for the client’s task. These paths, along with their associated loss values, are transmitted to the server, which optimises to find the intersection of these paths. This

intersection is used as the updated global model, ensuring it remains within a shared low-loss region. By maintaining the global model in an optimal region of the loss landscape throughout training, the framework improves convergence and ensures more consistent performance across all clients. Figure 2 illustrates the difference in the process of computing local models in FedAvg and the proposed framework.

2 EARLIER STUDIES

FEDERATED LEARNING: FedAvg is a foundational framework for FL, widely used in various domains, including clinical applications. While it typically converges well with similar clients (Karimireddy et al., 2020), challenges arise due to client heterogeneity and irregular neural network loss landscapes, leading to suboptimal global model. This issue is often overlooked in clinical FL studies, such as CURIAL-FL (Soltan et al., 2023), which trains COVID-19 prediction models across multiple sites, as well as in approaches like neighbourhood regularisation (Thakur et al., 2021) and augmented graph attention network-based FL (Molaei et al., 2024) in medical informatics.

To address FedAvg’s limitations in non-IID scenarios, several variants have been developed. FedProx (Li et al., 2020) introduces a regularisation term that mitigates the impact of client update divergence during aggregation, improving stability in heterogeneous environments. SCAFFOLD (Karimireddy et al., 2020) addresses data heterogeneity by using control variates to reduce variance in client updates, leading to more consistent global model updates. FedMarl (Zhang et al., 2022) leverages multi-agent reinforcement learning to optimise client selection, improving the training process under client heterogeneity. More recently, Federated Neural Propagation (FedNP) (Wu et al., 2023) has been proposed, which introduces an auxiliary task of estimating a latent global data distribution to regularise local models, thereby enhancing performance in non-IID settings. Building on these advancements, FedIMA (Zhou et al., 2024) also aims to alleviate the optimisation issue of global model aggregation in non-IID scenarios by introducing iterative moving averaging of past global models, which smooths out fluctuations across training rounds and enhances the convergence and robustness of FL in heterogeneous environments. Furthermore, hypernetwork-based FL (Shamshian et al., 2021) leverages hypernetworks to generate personalised models for each client, facilitating parameter sharing while retaining the ability to create unique models that adapt to client-specific data distributions.

More recent efforts have focused on improving local and global optimization strategies. Sharpness-Aware

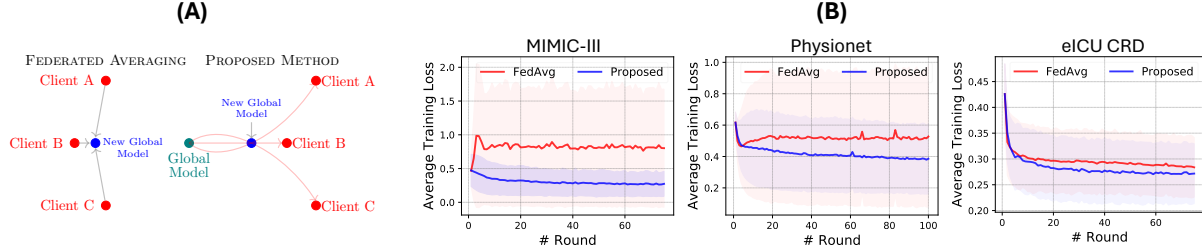


Figure 2: **(a)** FedAvg vs. proposed method: FedAvg averages client models, often placing the global model outside low-loss regions. The proposed method intersects low-loss paths (red) between the global and client models, maintaining optimal positioning. **(b)** Training dynamics on 15 clients (MIMIC-III, Physionet) and 100 clients (eICU-CRD). MIMIC-III and Physionet exhibit higher heterogeneity. The proposed method achieves better convergence in all cases.

Minimization (SAM) improves local generalization by encouraging flatter minima, mitigating sharp optima caused by non-IID data (Qu et al., 2022). This results in more stable local updates, leading to a better aggregated global model. In contrast, Multi-Model Collaborative Optimization (MMCO) enables models to collaborate by jointly optimizing different layers or parameter subsets before aggregation, enhancing generalization in heterogeneous settings where traditional FL struggles with local adaptation (Hu et al., 2024). However, its impact may be limited in shallow models with fewer layers available for collaborative optimization. Additionally, FedGuCci+ (Li et al., 2024) explores model connectivity in FL by improving group-wise connectivity among local models using fixed anchor models and transitivity properties in the loss landscape. This approach enhances generalization by stabilizing model fusion, particularly in non-IID environments.

MODE CONNECTIVITY: Mode connectivity refers to the idea that different local minima found by independently trained models are often connected by a low-loss path in the parameter space, suggesting a broader and more interconnected space of good solutions (Garipov et al., 2018; Lubana et al., 2023). Early studies modelled non-linear mode connections using Bezier curves and polygon chains, leveraging these for ensemble-based uncertainty estimation. In contrast, linear mode connectivity has been used in regularisation for multi-tasking and continual learning, ensuring new tasks stay close to previous ones, similar to the regularisation constraints in FedProx (Mirzadeh et al., 2021).

Apart from that, Zhou *et al.* studied mode connectivity to analyse the impact of data heterogeneity on FL (Zhou et al., 2023). They found that linear connections between client-specific and global models, as well as between iterations of the global model in FedAvg, often lead to a loss barrier under heterogeneity. How-

ever, low-loss paths can be achieved with non-linear connections like polygonal chains. Additionally, reducing data heterogeneity and using wider models improves mode connectivity, enhancing alignment and performance in FL.

COMPARISON WITH PROPOSED APPROACH: While the FL methods discussed above improve upon FedAvg in non-IID settings, they do not ensure that the global model consistently remains in the low-loss regions for all clients. In contrast, the proposed framework actively keeps the global model near the intersection of these low-loss regions, represented as mode connections, in the loss landscape. This approach guarantees robust performance across the entire client base. Additionally, unlike Zhou et al. (2023), which studies mode connections under data heterogeneity, the proposed framework directly exploits these connections to address optimisation challenges.

3 BACKGROUND

PROBLEM STATEMENT: We consider a FL framework with K clients, each participating in every training round. A central server initialises the global model and coordinates the training across clients to minimise a shared global objective:

$$\min_{\theta_G} \ell(\theta) = \sum_{k=1}^K p_k \ell_k(\theta) = \mathbb{E}_k[\ell_k(\theta)], \quad (1)$$

where p_k is the weight of client k , with $p_k > 0$ and $\sum_{k=1}^K p_k = 1$. The goal is to optimise the global model θ_G to perform well across all clients, despite the heterogeneity across their local datasets \mathcal{D}_k . Each client optimises its local objective ℓ_k on its own data, contributing to the improvement of the global model while maintaining data privacy.

MODE CONNECTIVITY: Consider two optima, or modes, $\theta_A \in \mathbb{R}^N$ and $\theta_B \in \mathbb{R}^N$, in the parameter space

for a given task. Mode connectivity refers to the low-loss path, $\gamma_{\theta_A \rightarrow \theta_B}$, between these two optima, where every point along the path represents an effective solution with minimal or no increase in loss.

While a linear path (a straight line between θ_A and θ_B) may encounter loss barriers, non-linear paths provide a more reliable way to connect these modes without hitting such barriers (Zhou et al., 2023). In this work, we focus exclusively on non-linear mode connections, modelled using a parameterised quadratic Bézier curve, $\Phi_\phi(t)$, as introduced by (Garipov et al., 2018):

$$\Phi_\phi(t) = (1-t)^2\theta_A + 2t(1-t)\phi + t^2\theta_B, \quad (2)$$

where $\phi \in \mathbb{R}^N$ are the trainable parameters defining the curve. The variable t controls the position along the curve, with $t = 0$ corresponding to θ_A and $t = 1$ to θ_B , for $t \in [0, 1]$.

4 PROPOSED METHOD

4.1 FedMode : Proposed Framework

The proposed FL framework, termed FedMode, enhances FedAvg by incorporating mode connectivity-based global model computation and modifying both server-side and client-side operations. The detailed steps of these operations in a single training round of FedMode are as follows:

CLIENT-SIDE OPERATIONS: Each client k receives the global model θ_g from the server and initialises its local model θ_k using the global model, such that $\theta_k = \theta_g$. The client then performs local updates using its local dataset \mathcal{D}_k by minimising the local loss function $\ell_k(\theta_k)$ through the following update rule:

$$\theta_k = \theta_k - \eta \nabla_{\theta_k} \ell_k(\theta_k), \quad (3)$$

where η is the learning rate.

After updating the local model θ_k , the client aims to learn a low-loss path or mode connection $\gamma_{\theta_g \rightarrow \theta_k}$ between the global model θ_g and the local model θ_k modelled as parameterised Bézier curve, $\Phi_{\phi_k}(t)$, defined in Equation 8. This Φ_{ϕ_k} is trained by updating ϕ_k to achieve minimum loss at different values of t , thereby, obtaining the desired mode connection $\gamma_{\theta_g \rightarrow \theta_k}$:

$$\phi_k = \phi_k - \eta \nabla_{\phi_k} \ell_k(\Phi_{\phi_k}(t)), t \sim \mathcal{U}(0, 1). \quad (4)$$

Once $\gamma_{\theta_g \rightarrow \theta_k}$ is obtained, models are sampled along this path using a set τ of predefined t values provided by the server. The sampled models are then used to compute the loss ℓ_k for the training samples along the Bézier curve: $\mathcal{L}^k = \{\ell_k(\Phi_{\phi_k}(t)) \mid \forall t \in \tau\}$. The client

then transmits the local model θ_k , the Bézier curve parameters ϕ_k , and the losses \mathcal{L}^k along $\gamma_{\theta_g \rightarrow \theta_k}$.

SERVER-SIDE OPERATIONS: The server, similar to FedAvg, begins by initializing the global model θ_g and distributing it to all clients. Along with the model, the server also transmits τ to each client, which is used for sampling models along the mode connections. From each client k , the server receives the client-specific model θ_k , the Bézier curve parameters ϕ_k , and the losses \mathcal{L}^k corresponding to points along these Bézier curves (mode connections).

The server then samples models along the mode connection for each client k as: $\mathcal{M}^k = \{\Phi_{\phi_k}(t) \mid t \in \tau\}$, where \mathcal{M}_i^k denotes the i^{th} sampled model on client k 's mode connection, and \mathcal{L}_i^k is its corresponding loss. The objective is to determine a global model Θ that resides in a region of the parameter space with consistently low loss across all clients. To achieve this, the server optimizes the following objective function:

$$\mathcal{J}(\Theta) = \sum_{k=1}^K \sum_{i=1}^{|\tau|} \frac{1}{\mathcal{L}_i^k + \epsilon} \|\Theta - \mathcal{M}_i^k\|_F^2 - \lambda \|\Theta - \theta_g\|_F^2, \quad (5)$$

where θ_g is the current global model, λ is a coefficient controlling how far to move from θ_g , and ϵ is a small positive constant (e.g., 10^{-6}) to avoid division by zero.

The first term in $\mathcal{J}(\Theta)$ encourages Θ to stay close to points on each client's mode connection where the loss is relatively small, by assigning larger weights $\frac{1}{\mathcal{L}_i^k + \epsilon}$ to those low-loss models. Meanwhile, the second term, due to its negative sign, pushes Θ away from the current global model θ_g . The strength of this push is determined by the coefficient λ . Higher λ means the new global model will wander further from θ_g , allowing more radical updates. Lower λ keeps the update closer to θ_g , leading to more conservative movement.

In practice, the server iteratively minimizes $\mathcal{J}(\Theta)$ via gradient-based steps: $\Theta \leftarrow \Theta - \alpha \nabla_{\Theta} \mathcal{J}(\Theta)$. After convergence, the server sets $\theta_g \leftarrow \Theta$ as the new global model, distributing it to the clients for the next communication round.

4.2 Theoretical Insights

We study the proposed FedMode method to gain deeper insights into its behaviour, with a focus on how it handles data heterogeneity, ensures convergence, and promotes equitable performance across clients. The theoretical analysis is based on standard assumptions regarding the smoothness and convexity of the local loss functions, bounded variance of stochastic gradients, and the level of data heterogeneity across clients. For clarity, we describe these assumptions in

the supplementary material.

CONVERGENCE ANALYSIS: Given these assumptions, we now examine the convergence of FedMode and its sensitivity to data heterogeneity.

Theorem 4.1. *Under the assumptions of smoothness and convexity of local loss functions, bounded variance of stochastic gradients, and bounded heterogeneity, meaning the difference between the local gradients and the global gradient is bounded by δ , i.e., $\|\nabla \ell_k(\theta) - \nabla \ell(\theta)\| \leq \delta$ for all k , the FedMode algorithm converges to a stationary point of the global loss function. Specifically, for FedMode, after T training rounds or iterations, the following holds:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \delta, \quad (6)$$

where θ_g^t is the global model at the t -th iteration.

Proof. Section C of the supplementary document provides the complete proof. \square

From Li et al. (2019), the convergence rate of FedAvg is given by: $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \delta^2$. This result shows that FedAvg’s performance degrades significantly in the presence of data heterogeneity, as the residual error—measuring the model’s deviation from a stationary point or zero gradient norm—scales quadratically with the heterogeneity parameter δ . In contrast, as shown in Theorem 4.1, FedMode exhibits improved resilience to non-IID data, achieving a convergence rate where the residual error depends linearly on δ . This reflects FedMode’s ability to better handle variations in client data distributions. While both methods perform similarly under homogeneous data conditions, FedMode’s superior robustness in heterogeneous environments makes it more suitable for real-world FL scenarios, where client data often varies significantly.

CLIENT FAIRNESS IN FEDMODE: Since the global model in FedMode is explicitly forced to lie at or near the intersection of mode connections, it resides near the low-loss regions of local models associated with every client. On the other hand, there are no such constraints on the global model computed by FedAvg, and it is quite likely that in non-IID settings, the performance of global model gets skewed towards some clients. As a result, the variance of local losses computed for the global model generated by FedMode is expected to be lesser than FedAvg:

$$\text{Var}(\{\ell_k(\theta_g^{\text{Mod}})\}_{k=1}^K) \leq \text{Var}(\{\ell_k(\theta_g^{\text{Avg}})\}_{k=1}^K), \quad (7)$$

where θ_g^{Mod} and θ_g^{Avg} represent the global models obtained by FedMode and FedAvg, respectively, and

$\ell_k(\theta)$ is the local loss function of client k evaluated at the model θ .

The improved client fairness, reflected in reduced variance among client losses in FedMode, results from its approach to constructing mode connections and weighting models based on local losses in the global objective. By assigning greater weight to models with lower local losses, FedMode ensures that clients with better local performance have a stronger influence on the global model, thereby decreasing the overall variance in client losses. This is especially valuable in non-IID settings, such as clinical applications, where client data distributions are inherently heterogeneous. By reducing disparities in client performance, FedMode ensures more equitable outcomes.

5 EXPERIMENTS

5.1 Datasets used

The proposed framework is evaluated on the following healthcare datasets:

- **PHYSIONET 2012 CHALLENGE** (Silva et al., 2012): This dataset is for in-hospital mortality (IHM) prediction, using the first 48 hours of ICU stay. It includes 8,000 time-series examples, each with 48 hourly steps and 44 features. The data is partitioned into 10 simulated clients in a non-IID manner, suitable for federated learning.
- **MIMIC-III** (Johnson et al., 2016): This extensive dataset from critical care units is processed for IHM and phenotype classification. For IHM, 21,156 time-series (48 time-steps, 76 features each) are sampled. Phenotyping involves 41,902 ICU stays, classified into 25 categories with variable-length time-series. The data is partitioned into 10 simulated clients in a non-IID manner for federated learning.
- **eICU COLLABORATIVE RESEARCH DATABASE: eICU** (Pollard et al., 2018; Tang et al., 2020) is a large multi-hospital dataset comprising 164,333 ICU stays. We utilise a pre-processed version (Tang et al., 2020) of eICU for predicting shock and acute respiratory failure (ARF) within the first 4 hours of ICU stay. Each stay is represented as a time series with 4 time steps and 375 features for shock prediction, and 345 features for ARF prediction (vital signs and demographics). For federated settings, we select the 100 hospitals with the most positive examples as clients.

5.2 Experimental design

Comparative Methods: We evaluate the performance of the proposed FedMod against several base-

Table 1: Performance of comparative methods for in-hospital mortality prediction across 10 clients sampled from (a) the MIMIC-III dataset and (b) the Physionet 2012 Challenge dataset.

(a) MIMIC-III			(b) PHYSIONET		
Methods	AUROC	AUPRC	Methods	AUROC	AUPRC
FEDAVG	0.781 ± 0.023	0.634 ± 0.019	FEDAVG	0.739 ± 0.038	0.409 ± 0.031
HNET-FL	0.805 ± 0.025	0.647 ± 0.024	HNET-FL	0.768 ± 0.031	0.413 ± 0.028
FEDPROX	0.834 ± 0.013	0.677 ± 0.017	FEDPROX	0.79 ± 0.027	0.421 ± 0.026
SCAFFOLD	0.821 ± 0.011	0.659 ± 0.021	SCAFFOLD	0.791 ± 0.021	0.418 ± 0.019
FEDMARL	0.835 ± 0.022	0.679 ± 0.016	FEDMARL	0.793 ± 0.026	0.425 ± 0.021
FEDIMA	0.832 ± 0.022	0.657 ± 0.019	FEDIMA	0.778 ± 0.027	0.415 ± 0.016
FEDNP	0.837 ± 0.021	0.676 ± 0.012	FEDNP	0.792 ± 0.023	0.426 ± 0.026
MOFEDSAM	0.841 ± 0.012	0.673 ± 0.011	MOFEDSAM	0.798 ± 0.011	0.429 ± 0.018
FEDGuCCI+	0.845 ± 0.014	0.677 ± 0.013	FEDGuCCI+	0.804 ± 0.016	0.43 ± 0.018
PROPOSED	0.854 ± 0.019	0.688 ± 0.015	PROPOSED	0.811 ± 0.018	0.442 ± 0.019

lines, including (1) *FedAvg* (McMahan et al., 2017), (2) *Hypernetwork-based FL* (Shamsian et al., 2021), (3) *SCAFFOLD* (Karimireddy et al., 2020), (4) *FedProx* (Li et al., 2020), (5) *FedMARL* (Zhang et al., 2022), (6) *FedIMA* (Zhou et al., 2024), (7) *FedNP* (Wu et al., 2023), (8) *FedGuCCI+* Li et al. (2024) and (9) *moFedSAM* (Qu et al., 2022).

For each client, the dataset is split into 60% for training, 15% for validation, and 25% for testing. Testing is performed on each client using the global model, except for the hypernetwork-based FL (HNet-FL). In HNet-FL, the global model acts as a hypernetwork that generates weights for the prediction models, which are then used for evaluation.

We use the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPRC) as performance metrics in all experiments.

Models & Parameter Settings: For all prediction tasks, we used LSTM-based models with a consistent structure: an LSTM layer with N hidden nodes, followed by a dense output layer with either 1 node (for binary tasks) or 25 nodes (for phenotype classification). The output layer is followed by a sigmoid activation function. The value of N is set to 64 for the MIMIC-III and PhysioNet datasets, and 128 for the eICU dataset.

For all methods, local training at each client is conducted using the Adam optimiser with a fixed learning rate of 0.001. During each training round, local training is performed for a single epoch, meaning the global model is updated once using every available training batch. In FedMod, we sample 10 equidistant points along each mode connection, represented as $\tau = \{\frac{i}{9}\}_{i=0}^9$. At the server, optimisation is carried out using the Adam optimiser with a learning rate of 0.001 for 200 iterations to determine the intersection or update the global model.

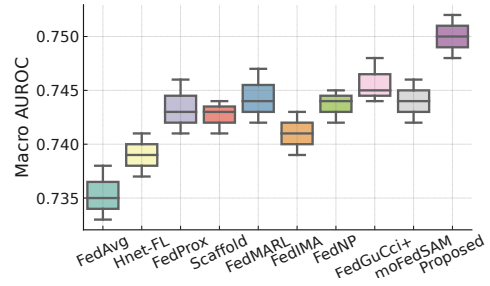


Figure 3: Comparative performance of various FL Methods for phenotype prediction using the MIMIC-III dataset.

More details about the parameter settings are provided in the supplementary document.

6 RESULTS & DISCUSSION

6.1 MIMIC-III and Physionet results

Table 1 presents the performance of the proposed FL framework compared to baseline methods for in-hospital mortality prediction across 10 clients simulated using the MIMIC-III and Physionet datasets. These clients are simulated with a non-IID data partitioning strategy, resulting in varying sample sizes and differing degrees of class imbalance, as previously discussed. The analysis of this table highlights the following:

- FedMod consistently outperforms FedAvg across both datasets, showing an average relative improvement of 9.29% in AUROC and 9.61% in AUPRC on the MIMIC-III and Physionet datasets, respectively.
- As expected, all baselines designed to handle client heterogeneity show noticeable improvement over FedAvg across both datasets. On the MIMIC-

Table 2: Performance of comparative methods for predicting (a) shock and (b) acute respiratory failure across 100 clients (hospitals) in the eICU-CRD dataset.

(a) SHOCK			(b) ACUTE RESPIRATORY FAILURE		
Methods	AUROC	AUPRC	Methods	AUROC	AUPRC
FEDAVG	0.725 ± 0.017	0.278 ± 0.012	FEDAVG	0.648 ± 0.012	0.177 ± 0.031
HNET-FL	0.726 ± 0.021	0.279 ± 0.018	HNET-FL	0.656 ± 0.022	0.181 ± 0.019
FEDPROX	0.731 ± 0.017	0.287 ± 0.016	FEDPROX	0.662 ± 0.017	0.184 ± 0.014
SCAFFOLD	0.732 ± 0.015	0.305 ± 0.017	SCAFFOLD	0.667 ± 0.021	0.186 ± 0.017
FEDMARL	0.73 ± 0.012	0.285 ± 0.016	FEDMARL	0.671 ± 0.01	0.189 ± 0.016
FEDIMA	0.728 ± 0.013	0.28 ± 0.014	FEDIMA	0.659 ± 0.015	0.182 ± 0.012
FEDNP	0.728 ± 0.018	0.281 ± 0.019	FEDNP	0.669 ± 0.017	0.185 ± 0.011
MOFEDSAM	0.734 ± 0.015	0.293 ± 0.017	MOFEDSAM	0.672 ± 0.017	0.189 ± 0.013
FEDGUCCI+	0.733 ± 0.016	0.292 ± 0.015	FEDGUCCI+	0.673 ± 0.013	0.191 ± 0.014
PROPOSED	0.739 ± 0.013	0.3 ± 0.011	PROPOSED	0.68 ± 0.018	0.193 ± 0.012

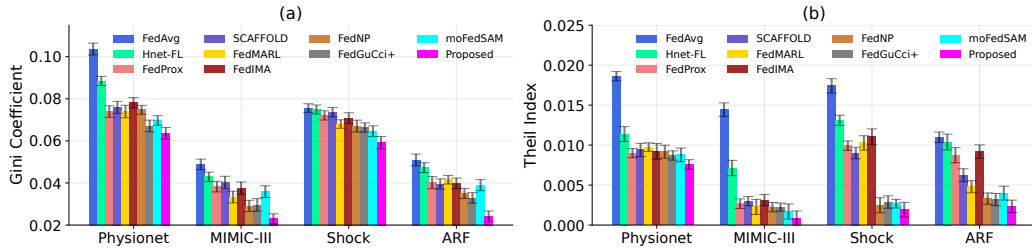


Figure 4: Client fairness across different FL methods, measured by the Gini Coefficient and Theil Index, for in-hospital mortality prediction using the Physionet and MIMIC-III datasets, and for shock and ARF prediction using the eICU dataset. Lower values indicate better fairness.

III dataset, all baselines, except for Hnet-FL and SCAFFOLD, demonstrate comparable performances. Notably, moFedSAM and FedGuCci+ outperform prior methods like FedProx and FedNP, with FedGuCci+ achieving the highest performance among baselines. On the Physionet dataset, SCAFFOLD, FedNP, FedMARL, moFedSAM, and FedGuCci+ exhibit similar levels of performance, highlighting their effectiveness in mitigating heterogeneity-related performance degradation.

- The proposed FedMode consistently outperforms all baselines. On the MIMIC-III dataset, it surpasses FedGuCci+, the best-performing baseline, by 1.07% in AUROC and 1.62% in AUPRC. On the Physionet dataset, it exceeds FedGuCci+ by 0.87% in AUROC and 2.79% in AUPRC, further improving over moFedSAM by 1.63% in AUROC and 3.03% in AUPRC. This superior performance is attributed to FedMode’s ability to ensure the global model resides near low-loss regions for all clients.

Figure 3 illustrates the performance of various methods for phenotype prediction. The trends are similar to those observed in in-hospital mortality prediction. moFedSAM and FedGuCci+ again exhibit strong performance, often outperforming prior baselines like Fed-

Prox and FedMARL. However, the proposed method still achieves the best results, with relative improvements of up to 2% over FedGuCci+ and FedMARL, and around 5% over FedAvg. This consistent improvement across different prediction tasks underscores the robustness and effectiveness of the proposed method in handling client heterogeneity and ensuring optimal model performance.

6.2 Performance on the eICU dataset

Table 2 presents the performance of different methods for predicting shock and acute respiratory failure on the eICU-CRD dataset. The results show that the proposed method achieves either comparable or better performance across both tasks.

- **SHOCK PREDICTION:** The proposed method shows a relative improvement of 0.95% in AUROC over moFedSAM, the best-performing baseline. While Scaffold achieves the highest AUPRC, the difference with the proposed method is minimal. Notably, moFedSAM and FedGuCci+ outperform earlier baselines like FedProx and FedMARL.
- **ACUTE RESPIRATORY FAILURE PREDICTION:** The proposed method outperforms all baselines, with a relative improvement of 1.04% in AUROC over

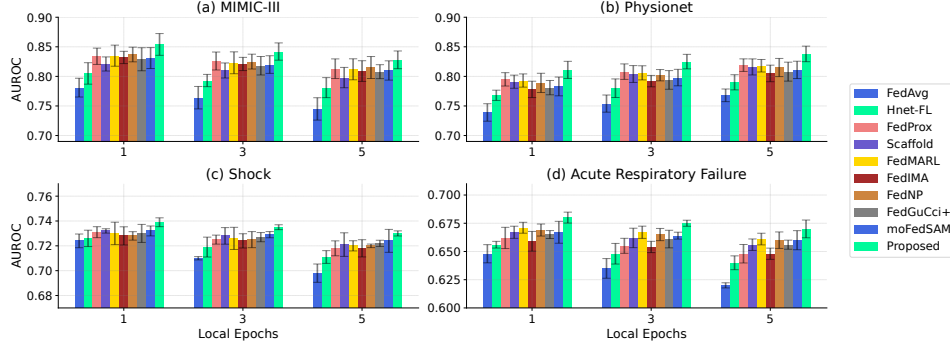


Figure 5: Impact of the number of local epochs on the performance of different FL methods on MIMIC-III dataset.

FedGuCci+ and 1.19% over moFedSAM. Additionally, it achieves a 1.05% improvement in AUPRC over FedGuCci+, reinforcing its effectiveness in optimising predictive performance.

The performance improvements observed with FedGuCci+, moFedSAM, and the proposed method over FedAvg on the eICU dataset are relatively smaller compared to those on the Physionet and MIMIC-III datasets. This may be due to the preprocessing applied to the eICU dataset, such as binarization or binning of clinical features, which can reduce the impact of distribution shifts and consequently lessen client heterogeneity. Nonetheless, this also highlights that the proposed framework maintains its effectiveness under varying degrees of client heterogeneity.

6.3 Evaluating client fairness

Client fairness refers to the equitable treatment and consistent performance of the global model across all participating clients, regardless of variations in data distribution, computational resources, or other factors (Mohri et al., 2019). In clinical FL, it is essential not only to enhance the average performance across clients but also to ensure that the global model serves all clients effectively. In this work, we assess client fairness using the Gini Coefficient and Theil Index to measure performance inequality (in terms of AUROC) across clients in each setting, as illustrated in Figure 4. These results demonstrate that the proposed method consistently achieves lower Gini and Theil values across datasets, indicating improved fairness and more equitable performance distribution compared to other FL methods. Hence, the proposed framework results in better performance as well as improve client fairness making it suitable for clinical applications.

6.4 Impact of local update frequency

In typical FL scenarios with similar client data, increasing the number of local epochs in each training round can lead to faster and more effective convergence, as local models become more refined and their updates align closely with the global objective. However, with data heterogeneity, increasing the number of local epochs can cause local models to diverge, leading to a more suboptimal global model. This behaviour is illustrated in Figure 5, which shows the impact of increasing local epochs on the performance of different methods.

As the number of epochs increases, FedAvg experiences a significant performance drop on the MIMIC-III and eICU (Shock & ARF) tasks, reflecting its inability to handle data heterogeneity. In contrast, the proposed framework (FedMode) and other baselines maintain stable performance, with some methods—including FedGuCci+ and moFedSAM—showing more robustness to client drift. This can be attributed to their ability to mitigate divergence across clients, a capability that FedAvg lacks. However, despite these improvements, FedMode consistently outperforms all baselines across all settings, demonstrating its superior handling of non-IID client distributions.

On the Physionet dataset, however, FedAvg, along with all other methods, shows performance improvement despite client heterogeneity. This may be due to the smaller number of training samples per client compared to other datasets, meaning that increasing local epochs to 3 or 5 does not cause significant deviation, unlike in larger datasets.

6.5 Regularisation in Global Objective

The regularisation term in the server’s objective function (Equation 5) guides the new global model along mode connections between the current global model

Table 3: Comparison of FedMode and FedMode-SAM across all datasets in terms of the AUROC.

Method	MIMIC-III	Phenotyping	Physionet	Shock	Respiratory Failure
FEDMODE	0.854 ± 0.019	0.750 ± 0.003	0.811 ± 0.018	0.739 ± 0.013	0.680 ± 0.018
FEDMODE + SAM	0.855 ± 0.017	0.752 ± 0.002	0.815 ± 0.014	0.739 ± 0.012	0.683 ± 0.014

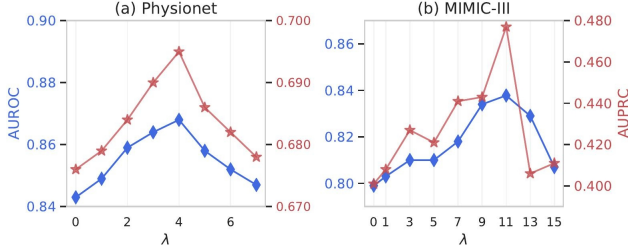


Figure 6: Impact of λ on the performance of the proposed FedMod framework for in-hospital mortality prediction using (a) the Physionet dataset and (b) the MIMIC-III dataset.

and clients’ local models. The regularisation coefficient λ controls this movement, balancing exploration and stability. A high λ promotes exploration by encouraging the global model to move further from its previous state, allowing it to shift toward an intersection of client-specific low-loss regions. However, this exploration is constrained within the clients’ low-loss regions, ensuring a safer pathway. If λ is too large, the global model may deviate excessively, increasing the risk of instability—particularly in highly non-IID settings where client optima are far apart.

Conversely, a low λ maintains stability by limiting movement away from the previous global model, leading to more conservative updates. While this can help prevent instability, it may result in a suboptimal model due to restricted exploration of the parameter space.

The impact of λ on the performance of the proposed framework on the MIMIC-III and PhysioNet datasets is shown in Figure 6. This analysis reveals that performance improves with increasing λ up to a certain point, after which it begins to decline, validating the exploration rationale discussed above. It is noteworthy that, across all parameter settings, the proposed framework consistently outperforms FedAvg and is either superior or comparable to other baselines.

6.6 Augmenting FedMode with SAM

Sharpness-Aware Minimization (SAM) has been widely utilized to improve generalization by encouraging flatter minima during optimization. To further enhance the effectiveness of FedMode, we integrate SAM

into local training, forming a new baseline, FedMode-SAM.

Since FedMode primarily optimizes the global model, its aggregation strategy remains unchanged, while local updates leverage SAM to enhance generalization. Notably, mode connection training does not interfere with local learning, making SAM a natural complement to FedMode. We adopt the same local training procedure as FedSAM, ensuring consistency while improving local optima.

To assess its impact, we compare FedMode-SAM against FedMode across all four datasets. Table 3 presents the results, demonstrating that FedMode-SAM achieves moderate improvements over FedMode, highlighting complementary nature of SAM and FedMode.

7 CONCLUSION & LIMITATIONS

In this paper, we introduced FedMod, a novel Federated Learning framework designed to address optimisation issues stemming from client heterogeneity. By utilising mode connectivity, FedMode ensures the global model remains in low-loss regions for all clients, resulting in better convergence and performance compared to traditional methods like FedAvg, particularly in diverse and heterogeneous datasets.

While FedMode offers improvements in accuracy, convergence, and fairness, it comes with increased computational and communication overhead. Each client must compute a Bézier curve, and the server solves for intersections, which may limit its use in resource-constrained environments. Despite these drawbacks, the performance and fairness advantages of FedMode makes it well-suited for critical applications such as healthcare, where accuracy and client fairness is paramount.

Future work will focus on reducing the computational complexity of learning mode connections, making FedMode more feasible in resource-limited settings. We also plan to deploy FedMode in real-world applications, such as across NHS Trusts, and explore integrating differential privacy to induce privacy preservation while maintaining performance.

Code

The implementation of the proposed method is available at <https://github.com/AnshThakur/FedMode>.

Acknowledgements

David A. Clifton is supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Center (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Wellcome Trust; the UKRI; and the InnoHK Hong Kong Center for Center for Cerebrocardiovascular Engineering (COCHE).

References

- Garipov, T., Izmailov, P., Podoprikin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in Neural Information Processing Systems*, 31.
- Hu, M., Yue, Z., Xie, X., Chen, C., Huang, Y., Wei, X., Lian, X., Liu, Y., and Chen, M. (2024). Is aggregation the only choice? federated learning via layer-wise model recombination. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, pages 1096–1107.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.
- Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.
- Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR.
- Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31.
- Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450.
- Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.
- Li, Z., Lin, J., Li, Z., Zhu, D., Ye, R., Shen, T., Lin, T., and Wu, C. (2024). Improving group connectivity for generalization of federated deep learning. *arXiv preprint arXiv:2402.18949*.
- Lu, M. Y., Chen, R. J., Kong, D., Lipkova, J., Singh, R., Williamson, D. F., Chen, T. Y., and Mahmood, F. (2022). Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76:102298.
- Lubana, E. S., Bigelow, E. J., Dick, R. P., Krueger, D., and Tanaka, H. (2023). Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.
- Mirzadeh, S. I., Farajtabar, M., Gorur, D., Pascanu, R., and Ghasemzadeh, H. (2021). Linear mode connectivity in multitask and continual learning. In *9th International Conference on Learning Representations, ICLR 2021*.
- Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR.
- Molaei, S., Thakur, A., Niknam, G., Soltan, A., Zare, H., and Clifton, D. A. (2024). Federated learning for heterogeneous electronic health records utilising augmented temporal graph attention networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1342–1350. PMLR.
- Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13.
- Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. (2022). Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning*, pages 18250–18280. PMLR.
- Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119.
- Shamsian, A., Navon, A., Fetaya, E., and Chechik, G. (2021). Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR.

- Silva, I., Moody, G., Scott, D. J., Celi, L. A., and Mark, R. G. (2012). Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology*, pages 245–248. IEEE.
- Soltan, A. A., Thakur, A., Yang, J., Chauhan, A., D’Cruz, L. G., Dickson, P., Soltan, M. A., Thickett, D. R., Eyre, D. W., Zhu, T., et al. (2023). Scalable federated learning for emergency care using low cost microcomputing: Real-world, privacy preserving development and evaluation of a covid-19 screening test in uk hospitals. *medRxiv*, pages 2023–05.
- Tang, S., Davarmanesh, P., Song, Y., Koutra, D., Sjoding, M. W., and Wiens, J. (2020). Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934.
- Thakur, A., Sharma, P., and Clifton, D. A. (2021). Dynamic neural graphs based federated reptile for semi-supervised multi-tasking in healthcare applications. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1761–1772.
- Thakur, A., Wang, C., Ceritli, T., Clifton, D., and Eyre, D. (2023). Mode connections for clinical incremental learning: Lessons from the covid-19 pandemic. *medRxiv*, pages 2023–05.
- Wu, X., Huang, H., Ding, Y., Wang, H., Wang, Y., and Xu, Q. (2023). Fednp: Towards non-iid federated learning via federated neural propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10399–10407.
- Zhang, S. Q., Lin, J., and Zhang, Q. (2022). A multi-agent reinforcement learning approach for efficient client selection in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9091–9099.
- Zhou, T., Lin, Z., Zhang, J., and Tsang, D. H. (2024). Understanding and improving model averaging in federated learning on heterogeneous data. *IEEE Transactions on Mobile Computing*.
- Zhou, T., Zhang, J., and Tsang, D. H. (2023). Mode connectivity and data heterogeneity of federated learning. *arXiv preprint arXiv:2309.16923*.

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
 - (b) Complete proofs of all theoretical results. [Yes]
 - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Yes]
 - (b) The license information of the assets, if applicable. [Yes]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
 - (d) Information about consent from data providers/curators. [Yes]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

- (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable]

A Dataset Details

Table A1 documents the details of the datasets used in this study.

As discussed in paper, for eICU-CRD dataset, we used available hospital information to simulate clients (each hospital as one client). However, for MIMIC-III (mortality) and Physionet datasets, we manually divided each dataset into 10 clients using a non-IID data partitioning process. First, a minimum number of samples per class m were assigned to each client. After this, the remaining data was distributed across clients using a Dirichlet distribution $\text{Dir}(\alpha \mathbf{1}_k)$, where α controls the degree of heterogeneity. This ensures that each client receives a non-uniform and heterogeneous subset of the data, reflecting real-world federated learning scenarios where clients possess different data distributions.

For both datasets, we use $m = 10$ (minimum number of samples per class) and $\alpha = 0.2$ (lower values induce heterogeneity).

The characteristics of resultant clients for both datasets are documented in Tables A2 and A3.

For Phenotyping experiment, we randomly divided the samples among 10 clients. These splits are inherently non-IID, given the nature of phenotypes and multi-label multi-class classification.

Table A1: Nature of healthcare datasets used in the experimental analysis.

Dataset	Task	#Examples	#Clients	Nature of clients
PHYSIONET	MORTALITY PREDICTION	8000	10	SIMULATED NON-IID
eICU-CRD	SHOCK PREDICTION	164,333	100	MULTI-CENTER, HOSPITALS
	ARF PREDICTION	138,840		
MIMIC-III	MORTALITY PREDICTION	21,156	10	SIMULATED NON-IID
	PHENOTYPING	41,902	10	

Table A2: Heterogeneity of clients in MIMIC-III (mortality).

	Clients									
	1	2	3	4	5	6	7	8	9	10
SAMPLES	558	5605	1141	2564	355	3382	349	3909	2187	1706
POSITIVES (%)	35.5	30.9	12.3	13.4	13.8	11.6	68.1	1.5	4.5	11.6

Table A3: Heterogeneity of clients in Physionet dataset.

	Clients									
	1	2	3	4	5	6	7	8	9	10
SAMPLES	1077	1186	37	565	2120	723	2189	233	44	26
POSITIVES (%)	0.75	1.85	65.5	65.5	0.47	1.1	31.5	19.7	61.3	38.5

B Implementation details

To learn the mode connections between the global model θ_g and local model θ_k , we use the same strategy used in earlier works (Garipov et al., 2018; Thakur et al., 2023) on non-linear mode connections. In each training round, we randomly sample t from a pre-defined set τ (containing values between 0 and 1). We use this value to generate an interpolated model $\Phi_\phi(t)$ using Bezier curve:

$$\Phi_\phi(t) = (1-t)^2\theta_g + 2t(1-t)\phi + t^2\theta_k, \quad (8)$$

where ϕ are the control points or trainable parameters of this curve. Then, $\Phi_\phi(t)$ is used to compute loss for a

training batch. This loss is then used to compute gradients and update ϕ :

$$\phi_k = \phi_k - \eta \nabla_{\phi_k} \ell_k(\Phi_{\phi_k}(t)). \quad (9)$$

This process is repeated for N epochs to achieve the trained Bezier curve.

C Theorem 4.1 Proof

C.1 Assumptions

We make the following assumptions to facilitate the analysis:

SMOOTHNESS OF LOSS FUNCTIONS: Each local loss function $\ell_k(\theta)$ is L -smooth:

$$\|\nabla \ell_k(\theta) - \nabla \ell_k(\theta')\| \leq L \|\theta - \theta'\|, \quad \forall \theta, \theta' \in \mathbb{R}^d. \quad (10)$$

CONVEXITY OF LOSS FUNCTIONS: Each $\ell_k(\theta)$ is convex.

BOUNDED VARIANCE OF STOCHASTIC GRADIENTS: The stochastic gradients have bounded variance:

$$\mathbb{E}_{\xi_k \sim \mathcal{D}_k} [\|\nabla f(\theta; \xi_k) - \nabla \ell_k(\theta)\|^2] \leq \sigma_k^2, \quad \forall \theta \in \mathbb{R}^d. \quad (11)$$

BOUNDED HETEROGENEITY (GRADIENT DISSIMILARITY): There exists $\delta \geq 0$ such that:

$$\|\nabla \ell_k(\theta) - \nabla \ell(\theta)\| \leq \delta, \quad \forall \theta \in \mathbb{R}^d, \forall k, \quad (12)$$

where $\ell(\theta) = \sum_{k=1}^K p_k \ell_k(\theta)$.

LIPSCHITZ CONTINUITY OF MODE CONNECTIONS: The Bézier curves $\Phi_{\phi_k}(t)$ are Lipschitz continuous with constant L_Φ :

$$\|\Phi_{\phi_k}(t) - \Phi_{\phi_k}(t')\| \leq L_\Phi |t - t'|, \quad \forall t, t' \in [0, 1]. \quad (13)$$

GRADIENT LIPSCHITZNESS ALONG MODE CONNECTIONS: The gradients along the mode connections satisfy:

$$\|\nabla \ell_k(\Phi_{\phi_k}(t)) - \nabla \ell_k(\Phi_{\phi_k}(t'))\| \leq L'_\Phi |t - t'|, \quad \forall t, t' \in [0, 1]. \quad (14)$$

C.2 Convergence Proof

We will start with analysing the expected decrease in the global loss function $\ell(\theta)$ over one communication round of FedMode. The global objective $J(\Theta)$ minimised at the server is defined as:

$$\mathcal{J}(\Theta) = \sum_{k=1}^K \sum_{i=1}^S w_{k,i} \|\Theta - \mathcal{M}_i^k\|^2 - \lambda \|\Theta - \theta_g^t\|^2. \quad (15)$$

We first compute the closed-form solution of this objective:

Compute the gradient of $\mathcal{J}(\Theta)$ with respect to Θ and set it to zero:

$$\nabla_{\Theta} \mathcal{J}(\Theta) = 2 \sum_{k=1}^K \sum_{i=1}^S w_{k,i} (\Theta - \mathcal{M}_i^k) - 2\lambda (\Theta - \theta_g^t) = 0, \quad (16)$$

$$\sum_{k=1}^K \sum_{i=1}^S w_{k,i} (\Theta - \mathcal{M}_i^k) - \lambda (\Theta - \theta_g^t) = 0, \quad (17)$$

$$\left(\sum_{k=1}^K \sum_{i=1}^S w_{k,i} + \lambda \right) \Theta = \sum_{k=1}^K \sum_{i=1}^S w_{k,i} \mathcal{M}_i^k + \lambda \theta_g^t, \quad (18)$$

Let us assume $W = \sum_{k=1}^K \sum_{i=1}^S w_{k,i}$ and $\bar{\mathcal{M}} = \frac{1}{W} \sum_{k=1}^K \sum_{i=1}^S w_{k,i} \mathcal{M}_i^k$. Then, we arrive at:

$$(W + \lambda)\Theta = W\bar{\mathcal{M}} + \lambda\theta_g^t, \quad (19)$$

and by solving for Θ , we get:

$$\Theta = \frac{W}{W + \lambda} \bar{\mathcal{M}} + \frac{\lambda}{W + \lambda} \theta_g^t. \quad (20)$$

Hence, the global model update can be expressed as:

$$\theta_g^{t+1} = \Theta = \gamma \bar{\mathcal{M}} + (1 - \gamma) \theta_g^t, \quad (21)$$

where $\gamma = \frac{W}{W + \lambda}$.

Now, we will establish a relation between $\bar{\mathcal{M}} - \theta_g^t$ to the global gradient $\nabla \ell(\theta_g^t)$.

We know that each client performs a local update: $\theta_k^t = \theta_g^t - \eta \nabla \ell_k(\theta_g^t)$. Then, it samples models \mathcal{M}_i^k along the mode connection between θ_g^t and θ_k^t . For small η and under Lipschitz continuity, we can approximate:

$$\mathcal{M}_i^k \approx \theta_g^t - \eta_i \nabla \ell_k(\theta_g^t), \quad (22)$$

where η_i is a small step size corresponding to the position along the mode connection. Then, the weighted average $\bar{\mathcal{M}}$ becomes:

$$\bar{\mathcal{M}} = \theta_g^t - \frac{1}{W} \sum_{k=1}^K \sum_{i=1}^S w_{k,i} \eta_i \nabla \ell_k(\theta_g^t). \quad (23)$$

Now, let us compute the change in global model i.e. $\theta_g^{t+1} - \theta_g^t$:

$$\theta_g^{t+1} - \theta_g^t = \gamma (\bar{\mathcal{M}} - \theta_g^t) = -\gamma \left(\frac{1}{W} \sum_{k=1}^K \sum_{i=1}^S w_{k,i} \eta_i \nabla \ell_k(\theta_g^t) \right). \quad (24)$$

Assuming $\eta_i = \eta$ for all i , we have: $\eta_{\text{eff}} = \eta$.

Then, Equation 24 becomes:

$$\theta_g^{t+1} - \theta_g^t = -\gamma \eta_{\text{eff}} \sum_{k=1}^K \tilde{p}_k \nabla \ell_k(\theta_g^t), \quad (25)$$

where $\tilde{p}_k = \frac{\sum_{i=1}^S w_{k,i}}{W}$.

Under the smoothness assumption, the expected decrease in global loss is given as:

$$\ell(\theta_g^{t+1}) \leq \ell(\theta_g^t) + \nabla \ell(\theta_g^t)^\top (\theta_g^{t+1} - \theta_g^t) + \frac{L}{2} \|\theta_g^{t+1} - \theta_g^t\|^2. \quad (26)$$

Also,

$$\nabla \ell(\theta_g^t)^\top (\theta_g^{t+1} - \theta_g^t) = -\gamma \eta_{\text{eff}} \sum_{k=1}^K \tilde{p}_k \nabla \ell(\theta_g^t)^\top \nabla \ell_k(\theta_g^t). \quad (27)$$

Using the bounded heterogeneity assumption, we get:

$$\nabla \ell(\theta_g^t)^\top \nabla \ell_k(\theta_g^t) \geq \|\nabla \ell(\theta_g^t)\|^2 - \|\nabla \ell(\theta_g^t)\| \delta. \quad (28)$$

$$\nabla \ell(\theta_g^t)^\top (\theta_g^{t+1} - \theta_g^t) \leq -\gamma \eta_{\text{eff}} (\|\nabla \ell(\theta_g^t)\|^2 - \|\nabla \ell(\theta_g^t)\| \delta). \quad (29)$$

By bounding the gradient norm, we get:

$$\|\theta_g^{t+1} - \theta_g^t\|^2 \leq (\gamma\eta_{\text{eff}}G)^2, \quad (30)$$

where $G = \max_k \|\nabla \ell_k(\theta_g^t)\|$.

Substituting everything back in Equation 26, we get:

$$\ell(\theta_g^{t+1}) \leq \ell(\theta_g^t) - \gamma\eta_{\text{eff}} (\|\nabla \ell(\theta_g^t)\|^2 - \|\nabla \ell(\theta_g^t)\|\delta) + \frac{L}{2}(\gamma\eta_{\text{eff}}G)^2. \quad (31)$$

$$\ell(\theta_g^{t+1}) \leq \ell(\theta_g^t) - \gamma\eta_{\text{eff}}\|\nabla \ell(\theta_g^t)\|^2 + \gamma\eta_{\text{eff}}\|\nabla \ell(\theta_g^t)\|\delta + \frac{L}{2}(\gamma\eta_{\text{eff}}G)^2. \quad (32)$$

Then, the expected decrease in global loss over T iterations is given as:

$$\sum_{t=0}^{T-1} \gamma\eta_{\text{eff}}\|\nabla \ell(\theta_g^t)\|^2 \leq \ell(\theta_g^0) - \ell(\theta_g^T) + \sum_{t=0}^{T-1} \left(\gamma\eta_{\text{eff}}\|\nabla \ell(\theta_g^t)\|\delta + \frac{L}{2}(\gamma\eta_{\text{eff}}G)^2 \right). \quad (33)$$

Divide both sides by $T\gamma\eta_{\text{eff}}$:

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\|^2 \leq \frac{\ell(\theta_g^0) - \ell^*}{T\gamma\eta_{\text{eff}}} + \delta \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\| + \frac{L\gamma\eta_{\text{eff}}G^2}{2\gamma\eta_{\text{eff}}}. \quad (34)$$

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\|^2 \leq \frac{\ell(\theta_g^0) - \ell^*}{T\gamma\eta_{\text{eff}}} + \delta a + \frac{LG^2}{2}, \quad (35)$$

where $a = \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\|$.

Rewriting the inequality and solving for a :

$$a^2 - \delta a - \left(\frac{\ell(\theta_g^0) - \ell^*}{T\gamma\eta_{\text{eff}}} + \frac{LG^2}{2} \right) \leq 0. \quad (36)$$

$$a \leq \frac{\delta}{2} + \sqrt{\left(\frac{\delta}{2} \right)^2 + \frac{\ell(\theta_g^0) - \ell^*}{T\gamma\eta_{\text{eff}}} + \frac{LG^2}{2}}. \quad (37)$$

For large T , the dominant term is $\frac{1}{T}$, leading to:

$$a \leq \frac{\delta}{2} + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \quad (38)$$

Therefore, the convergence rate is: $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla \ell(\theta_g^t)\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \delta$.

D Parameter Settings

MODEL ARCHITECTURE:

In this work, we evaluate have used LSTM-based prediction models for all datasets. The architecture of these models is as follows:

LSTM WITH N NODES \rightarrow RELU ACTIVATION \rightarrow DROPOUT WITH 0.1 RATE \rightarrow

DENSE LAYER WITH 1 NODE \rightarrow SIGMOID ACTIVATION

The N is set to be 64 nodes for MIMIC-III and Physionet tasks, whereas N is set to be 128 nodes for eICU tasks.

GENERIC PARAMETERS USED IN FEDERATED BASELINES AND THE PROPOSED METHOD:

For all experiments, both the baseline methods and FedMode were executed for 200 communication rounds. Client-side training was conducted for 1 epoch using the Adam optimiser with a fixed learning rate of 0.001 and a batch size of 64 examples, ensuring consistent model updates across all datasets and tasks.

PARAMETERS IN FEDMODE:

To learn the mode connections, Bezier curve control points were trained for 1 epoch using an Adam optimiser with a fixed learning of 0.001. Note that the training data to train local model and Bezier curve is same. For server-side optimisation, as discussed in the main text, Adam optimiser with learning rate of 0.001 is used to solve the global objective defined in Equation 5 for 200 iterations. Across all experiments, we use $\tau = \left\{\frac{i}{9}\right\}_{i=0}^9$.

PARAMETERS IN OTHER BASELINES:

Hypernetwork is fully-connected DNN that is used to generate weights for DNN and LSTM models. Hypernet used by (Shamsian et al., 2021) is also used in this work. Their implementation is available at <https://github.com/AvivSham/pFedHN>. The input embedding of 32 dimensions, 2 shared hidden layers with 128 nodes and spectral norm on initialised weights is used across all datasets in the hypernet.

In FedProx, the proximal regularisation coefficient (μ) is tuned to provide the best performance on the validation examples across all clients. We defined the sample set of μ to be $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1\}$. Based on the validation performance, we selected 0.1, 0.25 and 0.1 for MIMIC-III, PhysioNet and eICU tasks.

For FedMARL, each MARL agent uses a two-layer fully-connected model with a hidden layer of 256 neurons. The size of the historical information window used by the MARL agents is set to be 3 for training latencies and 5 for communication latencies. The reward weights $w_1 = 1$, $w_2 = 0.2$ and $w_3 = 0.1$ were used in all experiments. Also, to maintain consistency other baselines, all clients are involved in every training round.

For FedIMA, the Time Window (P), which determines how many global models from previous rounds are averaged to produce the IMA model, is set to be 3. Also, IMA begins at around 75% of the total training rounds, as suggested by the authors.

For FedNP, the search space for λ and ϵ were fixed to be $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1\}$ and $\{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$. The values of 0.2, 0.25 and 0.05 for λ were used for MIMIC-III, Physionet and eICU tasks. For all experiments, ϵ was set to be 10^{-5} . These values provided the best validation performance.

For moFedSAM, we follow the implementation of (Qu et al., 2022), where the sharpness-aware minimization (SAM) step is applied to local training. The perturbation magnitude ρ is selected from $\{0.01, 0.05, 0.1, 0.2\}$ based on validation performance, with $\rho = 0.1$ chosen for MIMIC-III and PhysioNet, and $\rho = 0.05$ for eICU tasks. The learning rate and batch size remain consistent with other baselines.

For FedGuCci+, we adopt the group-wise connectivity strategy as proposed in (Li et al., 2024). The number of anchor models per group is set to 3, and the transitivity-based group merging occurs every 5 communication rounds. The weight regularization parameter γ is tuned from $\{0.1, 0.25, 0.5, 1\}$, with $\gamma = 0.25$ selected for MIMIC-III and PhysioNet, and $\gamma = 0.5$ for eICU tasks.

Note that all experiments are run with 10 random seeds to obtain the results presented in the main text.