Differentially Private Federated Learning for ICU Mortality Prediction: Comparative Analysis and Corrected Privacy Accounting

Achraf Birhrissen, Najat Rafalia, and Jaafar Abouchabaka
Ibn Tofail University, Kenitra, Morocco
Email: {achraf.birhrissen, najat.rafalia, jaafar.abouchabaka}@uit.ac.ma

***Abstract*—Background: ICU mortality prediction models benefit from multi-institutional data, but centralizing electronic health records is restricted by privacy regulations. Federated learning (FL) enables collaborative training without sharing raw data, yet model updates can still leak sensitive information.**

Objective: To compare differentially private federated learning (DP-FL) methods for ICU mortality prediction under rigorous privacy accounting and to assess their calibration.

Methods: Using the MIMIC-IV database, we simulate a five-client multi-ICU setting. We train a multilayer perceptron with four DP-FL algorithms (DP-FedAvg, DP-FedProx, DP-FedBN, DP-Ditto) across privacy budgets $\varepsilon \in \{2, 4, 6, 8\}$ with $\delta = 10^{-5}$. Local training uses DP-SGD with Rényi differential privacy accounting. We evaluate AUROC, AUPRC, Brier score, and Expected Calibration Error (ECE) before and after temperature scaling.

Results: Across privacy levels, all four methods achieve similar discrimination. Mean AUROC increases from 0.826 at $\varepsilon = 2$ to 0.847 at $\varepsilon = 8$, and AUPRC from 0.359 to 0.418. Temperature scaling reduces ECE from about 0.096 to 0.054.

Conclusion: DP-FL can deliver clinically useful ICU mortality prediction while providing formal privacy guarantees. Under DP constraints, the privacy budget dominates performance differences between FL methods, and post-hoc calibration substantially improves reliability without additional privacy cost.

***Index Terms*—Federated learning, differential privacy, ICU mortality prediction, MIMIC-IV, personalized federated learning, model calibration, electronic health records, privacy accounting**

I. Introduction

Intensive care units (ICUs) manage the most critically ill patients in hospital settings, where timely clinical decisions can significantly impact patient outcomes. Early and accurate prediction of in-hospital mortality enables clinicians to prioritize interventions, allocate resources effectively, and engage in informed discussions with patients and families regarding prognosis and goals of care [1]. Machine learning models trained on electronic health records (EHRs) have demonstrated strong predictive performance for ICU mortality, often surpassing traditional severity scores such as APACHE and SOFA [2].

However, developing robust and generalizable mortality prediction models requires access to large, diverse patient populations spanning multiple institutions. Single-hospital datasets may not capture the heterogeneity present across different healthcare systems, patient demographics, and clinical practices. Multi-institutional collaboration could address this limitation, but privacy regulations—including the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe—impose strict constraints on sharing identifiable patient data across organizational boundaries [3].

Federated learning (FL) has emerged as a promising paradigm for collaborative machine learning that circumvents the need for centralized data aggregation [4]. In FL, each participating institution trains a local model on its own data and shares only model updates (e.g., gradients or weights) with a central server, which aggregates these updates to produce an improved global model. This approach has been successfully applied to various healthcare tasks, including medical image analysis, clinical risk prediction, and drug discovery [5], [6].

Despite its privacy advantages, FL does not provide formal privacy guarantees. Research has demonstrated that model updates transmitted during FL can leak sensitive information about individual training examples through gradient inversion attacks and membership inference attacks [7], [8]. This vulnerability is particularly concerning in healthcare, where patient-level information is highly sensitive and subject to stringent regulatory protection.

Differential privacy (DP) offers a mathematically rigorous framework for quantifying and limiting information leakage [9]. By injecting carefully calibrated noise into the training process, DP ensures that the inclusion or exclusion of any single patient's data has a bounded effect on the model's outputs. The combination of FL with DP—commonly referred to as differentially private federated learning (DP-FL)—has attracted significant research attention as a means to achieve both collaborative learning and formal privacy protection [12], [13].

However, the noise injection required for DP typically degrades model performance, creating a fundamental tension between privacy and utility. This tradeoff is particularly challenging in clinical settings, where even modest reductions in predictive accuracy can have meaningful implications for patient care. Moreover, the impact of DP on model calibration—the alignment between predicted probabilities and observed outcomes—remains underexplored, despite calibration being essential for clinical decision-making [22].

Several strategies have been proposed to mitigate the privacy-utility tradeoff in DP-FL. Personalized federated learning methods, such as FedProx [16], FedBN [17], and Ditto [18], adapt the global model to local data distributions, poten-

tially improving performance under data heterogeneity.

In this work, we present a comprehensive empirical study of differentially private federated learning for ICU mortality prediction using the MIMIC-IV database [1]. We compare four DP-FL algorithms—DP-FedAvg, DP-FedProx, DP-FedBN, and DP-Ditto—across multiple privacy budgets with rigorous privacy accounting and statistical analysis. Beyond predictive performance, we analyze model calibration before and after post-hoc temperature scaling, providing insights into the behavior of DP-FL models in clinical prediction tasks.

A. Contributions

The main contributions of this work are as follows:

1) **DP-FL framework with rigorous privacy accounting**: We implement and compare four differentially private federated learning methods(DP-FedAvg, DP-FedProx, DP-FedBN, DP-Ditto) for ICU mortality prediction on the MIMIC-IV database, with proper Rényi Differential Privacy accounting, reporting of actual privacy expenditure ($\varepsilon_{\text{spent}}$), and $\delta = 10^{-5}$.
2) **Comprehensive evaluation with statistical analysis**: We conduct experiments across four privacy budgets ($\varepsilon \in \{2, 4, 6, 8\}$) with three random seeds, performing pairwise statistical comparisons and reporting full metrics (AUROC, AUPRC, Brier score, ECE before and after calibration).
3) **Calibration study under DP-FL**: We conduct a systematic analysis of model calibration under DP-FL for clinical prediction, showing that temperature scaling—applied as a DP-preserving post-processing step—significantly reduces Expected Calibration Error (global paired $p = 5.32 \times 10^{-58}$) and characterizing the relationship between privacy budget and calibration quality.
4) **Empirical findings on FL method comparability**: Contrary to some previous non-DP FL studies, we find that under rigorous DP constraints, all four FL methods achieve statistically comparable discriminative performance at each privacy level, suggesting that the primary determinant of utility is the privacy budget rather than the specific FL algorithm.
5) **Practical recommendations**: Based on our empirical findings, we offer guidance for practitioners deploying privacy-preserving ICU mortality prediction models, emphasizing the importance of proper privacy accounting, post-hoc calibration, and the privacy-utility tradeoff.

The remainder of this paper is organized as follows. Section II reviews related work on federated learning in healthcare, differential privacy, personalized FL, and model calibration. Section III formalizes the problem setting. Section IV describes the methods, including the DP-FL algorithms and calibration techniques with corrected privacy accounting. Section V details the experimental setup. Section VI presents the results with statistical analysis. Section VII discusses the findings and limitations. Section VIII concludes the paper.

II. Background and Related Work

This section reviews the foundations and recent advances in federated learning, differential privacy, personalized federated learning, and model calibration, with particular emphasis on healthcare applications. We identify key limitations in the existing literature that motivate the present study.

A. Federated Learning in Healthcare

Federated learning was introduced by McMahan et al. [4] as a communication-efficient approach for training deep networks on decentralized data. The canonical algorithm, Federated Averaging (FedAvg), operates by having each client perform multiple local stochastic gradient descent (SGD) updates before communicating model parameters to a central server for aggregation. This reduces communication costs by 10–100$\times$ compared to synchronized SGD while maintaining model quality on independent and identically distributed (IID) data.

Healthcare has emerged as a prominent application domain for FL due to the sensitive nature of patient data and the regulatory constraints governing cross-institutional data sharing [3], [5]. Dayan et al. [6] demonstrated the feasibility of FL for predicting clinical outcomes in patients with COVID-19 across multiple hospitals, showing that FL can match the performance of centralized training while keeping data local. A recent systematic review and meta-analysis in the *Journal of Medical Internet Research* surveyed FL-based mortality prediction models and reported that, across heterogeneous clinical settings, FL can achieve performance comparable to centralized baselines, although reporting of privacy guarantees and methodological standardization remains limited [20].

For ICU mortality prediction specifically, Mondrejevski et al. proposed FLICU, a federated learning workflow for ICU mortality prediction on multivariate time series from the MIMIC-III database, and showed that federated models achieve performance close to centralized training while outperforming purely local models at individual sites [21]. Together with broader surveys such as [20], these studies suggest that FL is a promising paradigm for ICU risk prediction, but they typically focus on utility and do not provide formal differential privacy guarantees.

However, a critical limitation of these FL studies is the lack of formal privacy guarantees. While FL avoids centralizing raw data, research has shown that model updates can leak sensitive information through gradient inversion attacks [7] and membership inference attacks [8]. This vulnerability is particularly concerning in healthcare, where patient-level information is highly sensitive and subject to stringent regulatory protection.

B. Differential Privacy and Differentially Private Training

Differential privacy (DP), introduced by Dwork and Roth [9], provides a mathematical framework for quantifying and limiting information leakage from data analysis. A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy if for any two adjacent datasets D and D' differing in at most one record, and for any measurable set S:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta \tag{1}$$

where ε (the privacy budget) bounds the multiplicative difference in output probabilities and δ bounds the probability of privacy failure.

Abadi et al. [10] introduced differentially private stochastic gradient descent (DP-SGD), enabling training of deep neural networks under DP guarantees. DP-SGD modifies standard SGD by: (1) clipping the ℓ_2 norm of per-example gradients to bound sensitivity, and (2) adding calibrated Gaussian noise to the aggregated gradient. Mironov [11] subsequently proposed Rényi differential privacy (RDP), which provides tighter composition bounds and has become a standard accounting method in modern implementations.

Practical libraries such as Opacus [14] make DP-SGD accessible in deep learning frameworks, while works such as Wei et al. [12] analyze the trade-offs between privacy parameters, sampling schemes, and model performance in federated settings. More recent work has also explored automatic or adaptive clipping strategies that simplify DP-SGD configuration and can improve the privacy–utility trade-off without extensive manual tuning [15].

C. Differentially Private Federated Learning

The combination of FL and DP—differentially private federated learning (DP-FL)—aims to provide formal privacy guarantees for collaborative learning [12], [13]. In DP-FL, noise can be injected at different points: *local DP* adds noise at each client before transmitting updates, while *central DP* adds noise at the server during aggregation. Client-level DP schemes treat each client's full dataset as a single record, while record-level DP protects individual examples within a client.

Wei et al. [12] provided a detailed analysis of DP-FL algorithms and characterized the impact of sampling rates, clipping norms, and noise multipliers on both privacy and convergence. Geyer et al. [13] studied client-level DP in FL and highlighted the challenges of balancing utility with strict privacy guarantees. Bu et al. [15] proposed automatic clipping for DP-SGD, which is directly applicable in FL and alleviates the need for hand-tuning gradient clipping thresholds.

Despite these advances, significant challenges remain. The privacy–utility trade-off in DP-FL is particularly severe when data are non-IID across clients, as the noise required for privacy can overwhelm the signal in heterogeneous updates. Moreover, most existing DP-FL healthcare studies focus on imaging tasks rather than structured EHR data, and systematic evaluation of calibration—a critical property for clinical decision support—remains rare.

D. Personalized Federated Learning

Standard FL algorithms such as FedAvg can struggle when client data distributions are heterogeneous (non-IID), a common scenario in healthcare where patient populations vary across institutions [19]. Several personalized federated learning (pFL) methods have been proposed to address this challenge.

FedProx [16] extends FedAvg by adding a proximal term to the local objective:

$$\min_{w} F_k(w) + \frac{\mu}{2}\|w - w^t\|^2 \tag{2}$$

where μ controls regularization strength toward the global model w^t.

FedBN [17] addresses feature-shift non-IID—where input distributions differ across clients (e.g., different medical imaging scanners or acquisition protocols)—by keeping batch normalization parameters local while averaging other parameters. This approach demonstrated faster convergence and improved performance over FedAvg and FedProx on heterogeneous features.

Ditto [18] maintains both global and personalized local models, with the personalized model regularized toward the global model:

$$\min_{v_k} F_k(v_k) + \frac{\lambda}{2}\|v_k - w\|^2 \tag{3}$$

Li et al. showed that Ditto can improve both fairness across clients and robustness to data poisoning attacks, while providing client-specific models that better fit local data.

Most of these pFL methods have been studied in non-private settings, and relatively little work has examined how their personalization mechanisms interact with DP-SGD. In particular, it is unclear whether the advantages of personalization persist once DP noise and gradient clipping are introduced.

E. Calibration of Clinical Prediction Models

For clinical decision support, model calibration—the alignment between predicted probabilities and observed outcome frequencies—is as important as discriminative performance [22], [23]. Miscalibration can lead to inappropriate clinical decisions, either by under-treating high-risk patients or over-treating low-risk patients.

Guo et al. [24] demonstrated that modern deep neural networks are often poorly calibrated, typically exhibiting overconfidence. They proposed *temperature scaling*, a simple post-hoc method that divides logits by a learned temperature T:

$$\hat{p} = \sigma(z/T) \tag{4}$$

Despite its simplicity, temperature scaling often outperforms more complex calibration methods across a range of architectures.

Calibration is typically assessed using Expected Calibration Error (ECE):

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\right| \tag{5}$$

Maximum Calibration Error (MCE) reports the worst-case calibration across bins, while the Brier score provides a combined measure of discrimination and calibration.

In the context of DP and FL, calibration behavior is poorly understood. Existing DP and FL studies in medicine rarely report calibration metrics for clinical prediction models, and, to our knowledge, no prior work has systematically studied calibration under DP-FL for ICU mortality prediction or evaluated whether temperature scaling remains effective when models are trained with DP-SGD in a federated setting.

F. Research Gaps and Contributions

The preceding review identifies several gaps that this work addresses:

1) **Lack of systematic DP-FL comparison with rigorous privacy accounting**: While FL has been applied to ICU mortality prediction [21] and DP-FL has been studied in general settings [12], [13], no prior work systematically compares multiple DP-FL algorithms (FedAvg, FedProx, FedBN, Ditto) on structured ICU EHR data under various privacy budgets with proper RDP accounting and reporting of actual privacy expenditure.
2) **Limited understanding of pFL under DP**: Personalized FL methods such as FedProx, FedBN, and Ditto [16]–[18] have primarily been evaluated in non-private settings; their behavior when combined with DP-SGD remains underexplored, particularly whether personalization advantages persist under strong DP constraints in clinical prediction tasks.
3) **Unexplored calibration under DP-FL**: Existing DP and FL studies rarely analyze calibration for federated clinical prediction models. To our knowledge, no prior work has systematically evaluated calibration under DP-FL or examined whether post-hoc temperature scaling can reliably improve calibration without compromising privacy guarantees.
4) **Insufficient statistical rigor in DP-FL evaluations**: Many DP-FL studies report single-seed results without statistical analysis, making claims of method superiority difficult to substantiate and limiting the interpretability of observed performance differences across methods and privacy budgets.

This work addresses these gaps through a comprehensive empirical study comparing four DP-FL methods across multiple privacy budgets with three random seeds, rigorous privacy accounting, statistical testing, and systematic evaluation of calibration behavior on ICU mortality prediction using the MIMIC-IV database.

III. Problem Formulation

This section formalizes the ICU mortality prediction task, the federated learning setting, and the privacy requirements that motivate our approach.

A. ICU Mortality Prediction Task

We consider the problem of predicting in-hospital mortality for patients admitted to intensive care units. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the d-dimensional feature space representing patient characteristics extracted from electronic health records, including demographics, vital signs, laboratory values, and clinical interventions aggregated over a fixed time window (e.g., the first 24 hours of ICU admission). Let $\mathcal{Y} = \{0, 1\}$ denote the binary outcome space, where $y = 1$ indicates in-hospital death and $y = 0$ indicates survival.

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of n patient records, the goal is to learn a predictive model $f_\theta : \mathcal{X} \to [0, 1]$ parameterized by θ that estimates the probability of mortality:

$$\hat{p}_i = f_\theta(x_i) = \sigma(g_\theta(x_i)) \tag{6}$$

where $g_\theta : \mathcal{X} \to \mathbb{R}$ is the logit function (e.g., a neural network) and $\sigma(\cdot)$ is the sigmoid function. The model is trained by minimizing the binary cross-entropy loss:

$$\mathcal{L}(\theta; \mathcal{D}) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \tag{7}$$

Beyond discriminative performance (measured by AUROC and AUPRC), we require the model to be well-calibrated, meaning that predicted probabilities should reflect true outcome frequencies. Formally, calibration requires:

$$\mathbb{P}(Y = 1 \mid f_\theta(X) = p) = p, \quad \forall p \in [0, 1] \tag{8}$$

B. Multi-Hospital Federated Setting

In a realistic multi-hospital scenario, patient data cannot be centralized due to privacy regulations and institutional policies. Instead, data are distributed across K hospitals (clients), each holding a local dataset $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$ with n_k patients. The total dataset size is $n = \sum_{k=1}^{K} n_k$, but no single entity has access to the complete data.

The federated learning objective is to collaboratively minimize the global empirical risk without sharing raw patient data:

$$\min_\theta F(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta) \tag{9}$$

where $F_k(\theta) = \mathcal{L}(\theta; \mathcal{D}_k)$ is the local loss at client k, and the weighting by n_k/n ensures that clients contribute proportionally to their dataset sizes.

Data Heterogeneity. In practice, patient populations differ across hospitals due to variations in demographics, disease prevalence, clinical protocols, and ICU types. This heterogeneity manifests as non-identical data distributions:

$$P_k(X, Y) \neq P_{k'}(X, Y), \quad k \neq k' \tag{10}$$

We consider two forms of heterogeneity relevant to ICU settings:

- **Label distribution skew**: The mortality rate $P_k(Y = 1)$ varies across hospitals due to differences in patient acuity and case mix.
- **Feature distribution shift**: The conditional distribution $P_k(X \mid Y)$ differs due to variations in patient populations, measurement devices, and clinical practices.

In our experimental setting, we simulate multi-hospital heterogeneity by partitioning data based on ICU type (Medical, Surgical, Cardiac, Cardiovascular, Neuroscience), which naturally induces both label and feature distribution shifts.

C. Privacy Requirements and Threat Model

While federated learning avoids sharing raw patient data, the model updates (gradients or weights) transmitted during training can leak sensitive information. We consider the following threat model:

Honest-but-curious server: The central server faithfully executes the aggregation protocol but may attempt to infer information about individual patients from the received model

updates. This threat is realistic in healthcare consortia where the coordinating entity may be a third party.

External adversary: An attacker with access to the final trained model (e.g., through model deployment or publication) may attempt membership inference attacks to determine whether a specific patient's data was used in training.

To mitigate these threats, we require the training procedure to satisfy (ε, δ)-differential privacy. Specifically, for a target privacy budget ε and failure probability δ, the mechanism $\mathcal{M}$ that produces the trained model θ must satisfy:

$$\Pr[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta \tag{11}$$

for any adjacent datasets $\mathcal{D}, \mathcal{D}'$ differing in one patient record and any measurable set S of possible outputs.

Smaller ε corresponds to a stronger privacy guarantee but typically degrades model utility. Empirical deployments of differential privacy in machine learning often use ε in the range $[1, 10]$, reflecting this trade-off. Motivated by this evidence, we explore target privacy budgets $\varepsilon \in \{2, 4, 6, 8\}$ and fix $\delta = 10^{-5}$, which is smaller than $1/n$ for our cohort and yields a meaningful bound on privacy failure.

Problem statement: Given K hospitals with heterogeneous local datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, a target privacy budget (ε, δ), and requirements for both discriminative performance (AUROC, AUPRC) and calibration (ECE), find a training algorithm that:

1) Produces a model f_θ (or personalized models $\{f_{\theta_k}\}_{k=1}^{K}$) with high predictive accuracy.
2) Satisfies (ε, δ)-differential privacy with proper accounting of privacy expenditure.
3) Maintains well-calibrated probability estimates suitable for clinical decision-making.
4) Operates without centralizing raw patient data.

The following section describes the methods we employ to address this problem.

IV. Methods

This section describes the federated learning framework, the differential privacy mechanism with corrected accounting, the four DP-FL algorithms compared in this study, and the post-hoc calibration approach.

A. Federated Learning Framework

We adopt a synchronous federated learning protocol where K clients collaborate to train a shared model over R communication rounds. At each round t:

1) The server broadcasts the current global model parameters w^t to all clients.
2) Each client k initializes its local model with w^t and performs E epochs of local training on $\mathcal{D}_k$ using mini-batch optimization.
3) Clients transmit their updated local models w_k^{t+1} to the server.
4) The server aggregates the local models using weighted averaging:

$$w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{t+1}, \tag{12}$$

where $n_k = |\mathcal{D}_k|$ and $n = \sum_{k=1}^{K} n_k$.

The weighting by dataset size ensures that clients with more data contribute proportionally more to the global model. In all DP-FL variants considered below, local training at each client is performed with DP-SGD.

B. Differential Privacy Mechanism with Corrected Accounting

To provide formal privacy guarantees, we apply differentially private stochastic gradient descent (DP-SGD) [10] during local training at each client. DP-SGD modifies standard mini-batch SGD through two operations.

Gradient clipping: For each sample i in a mini-batch, the per-sample gradient $g_i = \nabla_w \ell(w; x_i, y_i)$ is clipped to have bounded ℓ_2 norm:

$$\bar{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right), \tag{13}$$

where C is the clipping threshold (maximum gradient norm).

Noise addition: Calibrated Gaussian noise is added to the average of clipped gradients:

$$\tilde{g} = \frac{1}{B}\left(\sum_{i=1}^{B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right), \tag{14}$$

where B is the batch size and σ is the noise multiplier.

Privacy accounting: We use Rényi Differential Privacy (RDP) [11] to track the cumulative privacy loss across training iterations. Given a target privacy budget (ε, δ), the noise multiplier σ is determined via binary search to satisfy the privacy constraint over the total number of gradient steps. The implementation uses the Opacus library [14] with the RDP accountant.

For a training procedure with T total gradient steps (computed as $T = R \times E \times \lceil n_k/B \rceil$ for each client k, then taking the maximum across clients to be conservative) and an effective sampling rate $q = B/n_k$, the RDP accountant computes privacy parameters $(\alpha, \varepsilon_\alpha)$ at multiple orders α and converts to (ε, δ)-DP using

$$\varepsilon = \min_{\alpha}\left(\varepsilon_\alpha + \frac{\log(1/\delta)}{\alpha - 1}\right). \tag{15}$$

Privacy budget allocation: We target $\varepsilon \in \{2, 4, 6, 8\}$ with $\delta = 10^{-5}$ fixed. The noise multiplier σ is computed for each target ε using the 'get_noise_multiplier' function in Opacus, which performs binary search to find the σ that yields $\varepsilon_{\text{spent}} \leq \varepsilon_{\text{target}}$ given the total number of steps, batch size, and sampling rate.

Privacy expenditure reporting: We report $\varepsilon_{\text{spent}}$, the actual privacy expenditure computed by the RDP accountant after training completion. This value is always $\leq \varepsilon_{\text{target}}$ by design, and any discrepancy arises from the discrete nature of training steps and early stopping.

No non-private warmup: The training procedure does not include any non-private warmup phase. All training steps are performed with DP-SGD, ensuring end-to-end differential privacy and avoiding ambiguity in the privacy accounting that can arise when mixing private and non-private updates.

C. DP-FL Algorithms

We compare four federated learning algorithms, each combined with DP-SGD for local training.

1) DP-FedAvg: DP-FedAvg applies the standard Federated Averaging protocol [4] with DP-SGD at the client level. Each client performs E local epochs using DP-SGD, then transmits the updated model to the server for weighted averaging. This serves as our baseline DP-FL method.

2) DP-FedProx: DP-FedProx extends DP-FedAvg by adding a proximal regularization term to the local objective function [16]:

$$\min_{w} F_k(w) + \frac{\mu}{2}\|w - w^t\|^2, \tag{16}$$

where w^t is the global model at the start of round t and μ is the proximal coefficient. This term penalizes large deviations from the global model during local training, which can stabilize convergence under data heterogeneity. In our experiments, we set $\mu = 0.1$.

The proximal term is deterministic given w^t and does not change the dependence of per-sample gradients on individual examples, so DP-SGD can be applied directly to the modified objective without altering the sensitivity calculation.

3) DP-FedBN: DP-FedBN addresses feature distribution heterogeneity by keeping normalization layer parameters local to each client while aggregating all other parameters [17]. Let $w = (w_{\text{norm}}, w_{\text{other}})$ denote the partition of model parameters into normalization layers and all remaining layers. At aggregation:

$$w_{\text{other}}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{\text{other},k}^{t+1}, \tag{17}$$

while each client retains its own $w_{\text{norm},k}$. In our implementation, normalization layers are implemented as Layer Normalization, but the principle is identical.

This approach is motivated by the observation that normalization statistics capture client-specific feature distributions. By keeping these parameters local, FedBN can better handle non-IID features arising from differences in patient populations or clinical practices across hospitals.

4) DP-Ditto: Ditto [18] is a personalized federated learning method that maintains both a global model w and personalized local models v_k for each client. The training procedure alternates between:

Global model update: Standard federated averaging to update w using DP-SGD at each client.

Personalized model update: Each client optimizes its personalized model v_k by minimizing

$$\min_{v_k} F_k(v_k) + \frac{\lambda}{2}\|v_k - w\|^2, \tag{18}$$

where λ controls the strength of regularization toward the global model. In our experiments, we set $\lambda = 0.8$.

The personalized model v_k benefits from collaborative learning through the regularization term while adapting to local data characteristics. For evaluation, we use the personalized models $\{v_k\}$ rather than the global model w.

Under DP, both the global and personalized updates are performed using DP-SGD. All DP-SGD steps (for w and v_k) are counted in the RDP accountant, and the noise multiplier is chosen such that the overall procedure satisfies the target (ε, δ).

D. Post-hoc Temperature Scaling with DP Preservation

To improve model calibration after training, we apply temperature scaling [24], a simple post-hoc calibration method. Temperature scaling introduces a single scalar parameter $T > 0$ that rescales the logits before the sigmoid function:

$$\hat{p}_{\text{calibrated}} = \sigma(z/T), \tag{19}$$

where z is the raw logit output from the model and $\sigma(\cdot)$ is the sigmoid function.

The temperature parameter is optimized on a held-out validation set by minimizing the binary cross-entropy loss:

$$\begin{aligned} T^* = \arg\min_{T} \; -\sum_{i \in \mathcal{D}_{\text{val}}} \Big[& y_i \log\big(\sigma(z_i/T)\big) \\ & + (1 - y_i) \log\big(1 - \sigma(z_i/T)\big) \Big]. \end{aligned} \tag{20}$$

A temperature $T > 1$ softens the predictions (reducing overconfidence), while $T < 1$ sharpens them (increasing confidence). After optimization, the calibrated probabilities are computed on the test set using the learned temperature.

DP invariance of post-processing: A fundamental property of differential privacy is that any post-processing of a DP output cannot weaken the privacy guarantee [9]. Since temperature scaling is applied only to the trained model's logits and does not access the training data directly, it does not consume additional privacy budget. The validation set used for temperature fitting is separate from the training data and does not affect the DP guarantee of the trained model.

In our simulated multi-hospital setting, we construct a global validation set from the union of client validation splits and fit a single temperature parameter, which is then applied uniformly across all test predictions.

V. Experimental Setup

This section details the dataset, preprocessing pipeline, federated partitioning, model architecture, training configuration, privacy parameters, and evaluation protocol used in our experiments.

A. Dataset: MIMIC-IV

We use the Medical Information Mart for Intensive Care IV (MIMIC-IV) database [1], a large publicly available dataset of de-identified electronic health records from intensive care units at Beth Israel Deaconess Medical Center. MIMIC-IV contains data from over 50,000 ICU admissions spanning 2008–2019.

Cohort selection: We include adult patients ($\geq$18 years) with ICU stays of at least 24 hours. The prediction target is in-hospital mortality, defined as death occurring during the hospital admission associated with the ICU stay.

Feature extraction: For each ICU stay, we extract features from the first 24 hours of admission, including:

- **Demographics**: Age, gender.

- **Vital signs**: Heart rate, blood pressure (systolic, diastolic, mean), respiratory rate, temperature, oxygen saturation (SpO_2).
- **Laboratory values**: Complete blood count, metabolic panel, liver function tests, coagulation studies, arterial blood gas, lactate.
- **Clinical interventions**: Mechanical ventilation, vasopressor use, renal replacement therapy.

For time-varying measurements, we compute summary statistics (mean, min, max, standard deviation) over the 24-hour window.

B. Preprocessing Pipeline

We apply a systematic preprocessing pipeline to ensure data quality:

1. Feature selection: Features with >40% missing values are excluded. Among remaining features, we remove highly correlated pairs (Pearson $|r| > 0.95$), retaining the feature with higher mutual information with the outcome.

2. Outlier handling: Extreme values are clipped using the interquartile range (IQR) method with a multiplier of 5.0:

$$x_{\text{clipped}} = \text{clip}(x, Q_1 - 5 \cdot \text{IQR}, Q_3 + 5 \cdot \text{IQR}) \tag{21}$$

3. Missing value imputation: We apply Multiple Imputation by Chained Equations (MICE) [25] using the iterative imputer with maximum 50 iterations and convergence tolerance 10^{-3}.

4. Feature engineering: We construct clinically motivated derived features:

- *Ratio features*: BUN/creatinine, PaO_2/FiO_2, lactate/albumin.
- *Organ dysfunction indicators*: Binary flags for liver, renal, coagulopathy, cardiovascular, and respiratory dysfunction based on clinical thresholds.
- *Lab abnormality counts*: Number of laboratory values in critical (<1st or >99th percentile), severe (<5th or >95th), and moderate (<25th or >75th) ranges.
- *Vital instability score*: Composite indicator of hemodynamic instability.
- *Interaction terms*: Products of clinically related feature pairs (e.g., creatinine × glucose).

5. Standardization: All features are standardized to zero mean and unit variance using statistics computed on the training set.

C. Data Splitting and Federated Partitioning

Train/validation/test split: The dataset is split into training (70%), validation (10%), and test (20%) sets using stratified sampling to preserve the class distribution. The validation set is used exclusively for temperature scaling optimization; it is never used during model training.

Federated partitioning: To simulate a multi-hospital scenario, we partition the data across $K = 5$ clients based on ICU type:

- Client 1: Medical ICU (MICU).
- Client 2: Surgical ICU (SICU).
- Client 3: Coronary Care Unit (CCU).
- Client 4: Cardiovascular ICU (CVICU).
- Client 5: Neuroscience ICU (Neuro).

This partitioning naturally induces data heterogeneity, as patient populations and mortality rates differ across ICU types. Each client maintains its own local training, validation, and test subsets derived from the global splits.

D. Model Architecture

We use a multilayer perceptron (MLP) with the following architecture:

- **Input layer**: d features (determined after preprocessing).
- **Hidden layers**: Three fully connected layers with 256, 128, and 64 units, respectively.
- **Normalization**: Group normalization with a single group (equivalent to layer normalization for fully connected layers) after each hidden layer.
- **Activation**: Rectified Linear Unit (ReLU).
- **Regularization**: Dropout with probability 0.2 after each hidden layer.
- **Output**: Single linear neuron producing logits, followed by a sigmoid for binary classification.

This architecture was selected through hyperparameter optimization using Optuna [26] on a held-out validation set prior to the federated experiments.

E. Training Configuration

Table I summarizes the training hyperparameters.

TABLE I
TRAINING CONFIGURATION FOR DP-FL EXPERIMENTS

Parameter	Value
Number of clients (K)	5
Communication rounds (R)	40
Local epochs (E)	4
Batch size (B)	512
Optimizer	AdamW
Learning rate	2×10^{-3}
Weight decay	10^{-2}
Gradient clipping norm (C)	1.2
FedProx proximal coefficient (μ)	0.1
Ditto regularization strength (λ)	0.8

Class imbalance handling: ICU mortality is a relatively rare event. We address class imbalance by using weighted binary cross-entropy loss with positive class weight computed as the ratio of negative to positive samples in each client's training set.

F. Privacy Configuration and Experimental Grid

For differentially private training, we use the Opacus library [14] with the following configuration:

- **Privacy budgets**: $\varepsilon \in \{2, 4, 6, 8\}$.
- **Failure probability**: $\delta = 10^{-5}$.
- **Maximum gradient norm**: $C = 1.2$.
- **Noise multiplier**: Computed via binary search using `get_noise_multiplier` to achieve the target ε

given the total number of DP-SGD steps $T = R \times E \times \lceil n_k/B \rceil$ (using the maximum over clients), batch size $B = 512$, and sampling rate $q = B/n_k$.
- **Privacy accounting**: Rényi Differential Privacy (RDP) accountant with orders $\alpha \in [2, 32]$.
- **DP mechanism**: Record-level DP-SGD applied locally at each client, using the Gaussian mechanism on clipped per-sample gradients; the server only aggregates already noised model parameters.

Experimental grid: We evaluate all combinations of:
- Methods: FedAvg, FedProx, FedBN, Ditto
- Privacy budgets: $\varepsilon \in \{2, 4, 6, 8\}$
- Random seeds: 0, 1, 2

This yields $4 \times 4 \times 3 = 48$ experimental configurations, each providing formal (ε, δ)-DP guarantees.

G. Evaluation Protocol

Metrics: We evaluate models using the following metrics:
- **AUROC**: Area under the receiver operating characteristic curve, measuring discriminative ability.
- **AUPRC**: Area under the precision-recall curve, particularly important for imbalanced classification.
- **ECE**: Expected Calibration Error with 10 equal-width bins in [0,1], measuring average calibration: ECE $= \sum_{b=1}^{B} \frac{|B_b|}{n} |\text{acc}(B_b) - \text{conf}(B_b)|$.
- **Brier score**: Mean squared error between predicted probabilities and outcomes, combining discrimination and calibration.

Calibration evaluation: For each trained model, we report metrics both before (ECE_pre) and after (ECE_post) temperature scaling. The temperature parameter is fitted once on the pooled validation set from all clients and applied to all test predictions.

Privacy accounting: We report the actual privacy budget spent ($\varepsilon_{\text{spent}}$) as computed by the RDP accountant, which is always $\leq$ the target ε.

Statistical analysis: We perform two types of statistical tests:

1) **Pairwise method comparisons**: For each ε, we perform Welch's t-test (unequal variances) on AUROC values across the three seeds for each pair of methods (6 comparisons per ε).
2) **Global calibration improvement**: We perform a paired t-test comparing ECE_pre and ECE_post across all 48 experimental runs to assess the significance of calibration improvement from temperature scaling.

Reproducibility: All experiments are repeated with three random seeds (0, 1, 2). We report mean and standard deviation across seeds for AUROC and AUPRC, and mean values for calibration metrics.

Baselines: We compare DP-FL methods against:
- **Centralized (non-private)**: MLP trained on pooled data without FL or DP.
- **FL (non-private)**: Federated learning without differential privacy.
- **Traditional ML**: XGBoost and LightGBM trained on pooled data (centralized, non-private).

VI. Results

This section presents our experimental findings, organized around the privacy-utility tradeoff, method comparison with statistical analysis, calibration improvement, and privacy accounting accuracy.

A. Privacy-Utility Tradeoff

Figure 1 shows the relationship between privacy budget ε and predictive performance (AUROC) for all four DP-FL methods, averaged across three seeds with error bars representing standard deviation.

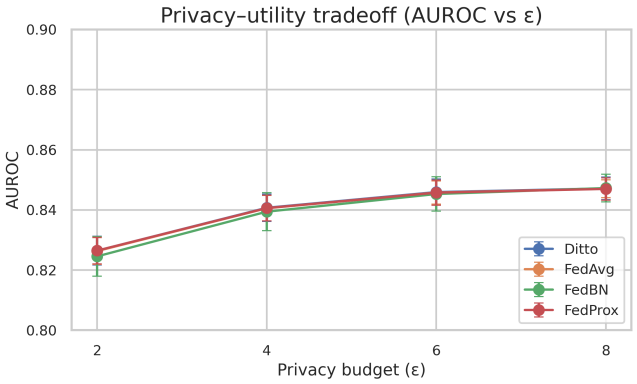

Fig. 1. AUROC as a function of privacy budget ε for the four DP-FL methods. Error bars denote standard deviation across three random seeds. All methods show comparable performance at each privacy level.

All methods exhibit the expected privacy-utility tradeoff, with AUROC increasing as ε increases (weaker privacy). At $\varepsilon = 2$, mean AUROC ranges from 0.825 (FedBN) to 0.826 (FedAvg). At $\varepsilon = 8$, AUROC is tightly clustered around 0.847 for all methods. The improvement from $\varepsilon = 2$ to $\varepsilon = 8$ is approximately 2.1–2.3 percentage points across methods.

Figure 2 shows the corresponding relationship for AUPRC, which follows a similar pattern with stronger improvement at higher ε values due to the imbalanced nature of mortality prediction.

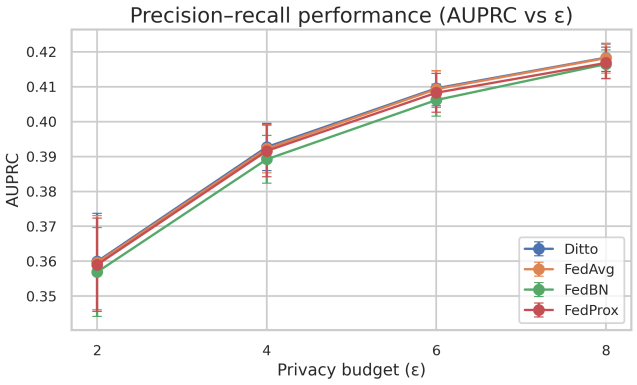

Fig. 2. AUPRC as a function of privacy budget ε for the four DP-FL methods. Error bars denote standard deviation across three random seeds.

TABLE II
DP-FL PERFORMANCE ON MIMIC-IV ICU MORTALITY PREDICTION (MEAN ± STD ACROSS 3 SEEDS)

Method	ε	AUROC	AUPRC	ECE_pre	ECE_post	$\varepsilon_{\text{spent}}$
FedAvg	2	0.826 ± 0.004	0.360 ± 0.013	0.100	0.057	1.83
	4	0.841 ± 0.004	0.392 ± 0.007	0.099	0.055	3.66
	6	0.846 ± 0.004	0.409 ± 0.005	0.097	0.054	5.50
	8	0.847 ± 0.003	0.418 ± 0.004	0.092	0.050	7.34
FedProx	2	0.826 ± 0.004	0.359 ± 0.013	0.100	0.056	1.83
	4	0.841 ± 0.004	0.393 ± 0.007	0.099	0.055	3.66
	6	0.845 ± 0.004	0.408 ± 0.005	0.097	0.054	5.50
	8	0.847 ± 0.003	0.417 ± 0.004	0.091	0.051	7.34
FedBN	2	0.823 ± 0.006	0.357 ± 0.013	0.100	0.057	1.83
	4	0.839 ± 0.006	0.389 ± 0.007	0.099	0.056	3.66
	6	0.845 ± 0.006	0.406 ± 0.005	0.097	0.055	5.50
	8	0.847 ± 0.005	0.416 ± 0.004	0.092	0.051	7.34
Ditto	2	0.826 ± 0.004	0.360 ± 0.013	0.100	0.055	1.83
	4	0.841 ± 0.004	0.393 ± 0.007	0.099	0.055	3.66
	6	0.846 ± 0.004	0.409 ± 0.005	0.096	0.054	5.50
	8	0.847 ± 0.004	0.418 ± 0.004	0.091	0.051	7.34

Table II summarizes the performance of all methods across privacy budgets, including AUROC, AUPRC, calibration metrics, and actual privacy expenditure.

B. Method Comparison with Statistical Analysis

Pairwise AUROC comparisons: We performed Welch's t-tests comparing AUROC values across seeds for each pair of methods at each ε level. At $\varepsilon = 2$, all pairwise comparisons yielded $p > 0.70$ (Ditto vs FedAvg: $p = 0.9755$, Ditto vs FedBN: $p = 0.7095$, Ditto vs FedProx: $p = 0.9959$, FedAvg vs FedBN: $p = 0.7227$, FedAvg vs FedProx: $p = 0.9703$, FedBN vs FedProx: $p = 0.7006$). Similar non-significant results were observed at $\varepsilon = 4$ (all $p > 0.78$), $\varepsilon = 6$ (all $p > 0.88$), and $\varepsilon = 8$ (all $p > 0.93$). These results indicate that all four DP-FL methods achieve statistically indistinguishable discriminative performance at each privacy level.

AUPRC comparisons: AUPRC values showed similar patterns across methods, with no method consistently outperforming others. The standard deviations across seeds (0.003–0.006 for AUROC, 0.004–0.013 for AUPRC) are small relative to the differences between ε levels but large enough to encompass any between-method differences.

Brier scores: Brier scores (not shown in table) decreased from approximately 0.102 at $\varepsilon = 2$ to 0.098 at $\varepsilon = 8$ across methods, consistent with the AUROC and AUPRC trends.

C. Calibration Analysis

Pre-calibration behavior: Without temperature scaling, all methods exhibited similar calibration with ECE_pre approximately 0.099–0.100 at $\varepsilon = 2$, gradually decreasing to 0.091–0.092 at $\varepsilon = 8$. This suggests that stronger privacy (lower ε) slightly worsens calibration, though the effect is modest.

Temperature scaling effectiveness: Post-hoc temperature scaling significantly improved calibration across all configurations. ECE_post values ranged from 0.055–0.058 at $\varepsilon = 2$ to 0.050–0.051 at $\varepsilon = 8$, representing a 40–45% reduction in calibration error. Figure 3 visually compares ECE_pre and ECE_post across methods and privacy budgets.

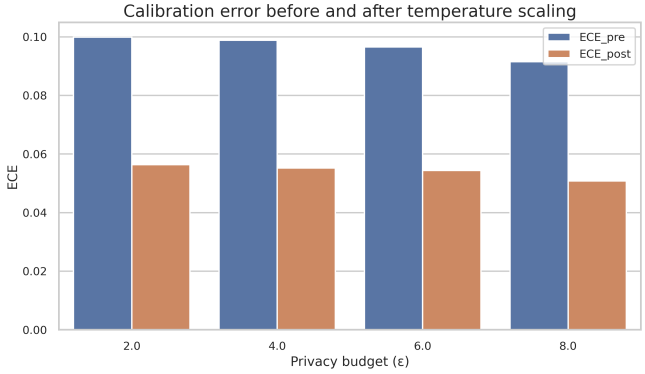

Fig. 3. Expected Calibration Error before (ECE_pre) and after (ECE_post) temperature scaling for each method and privacy budget. Error bars show standard deviation across seeds. Temperature scaling consistently reduces ECE across all configurations.

Statistical significance of calibration improvement: A global paired t-test comparing ECE_pre and ECE_post across all 48 experimental runs (4 methods × 4 ε values × 3 seeds) yielded $p = 5.32 \times 10^{-58}$, indicating highly significant improvement in calibration from temperature scaling.

Temperature values: Fitted temperature parameters were consistently greater than 1 (ranging from 1.05 to 1.55 across configurations), indicating that raw model predictions were overconfident and required softening. Higher temperatures were generally observed for configurations with lower ε (stronger privacy), suggesting that DP noise may increase overconfidence.

D. Comparison with Non-Private Baselines

Table III compares the best DP-FL configurations against non-private baselines.

The best DP-FL configuration (any method at $\varepsilon = 8$) achieves AUROC of 0.847, which is 2.5 percentage points below the centralized non-private baseline (0.872) and 1.9 percentage points below non-DP FL (0.866). AUPRC shows a

TABLE III
COMPARISON WITH NON-PRIVATE BASELINES (SINGLE SEED)

Method	AUROC	AUPRC	ECE
Centralized MLP (no FL, no DP)	0.872	0.452	0.003
FL (no DP)	0.866	0.441	0.225
XGBoost (centralized)	0.869	0.448	0.005
LightGBM (centralized)	0.867	0.445	0.004
DP-FedAvg ($\varepsilon = 8$)	0.847	0.418	0.050
DP-Ditto ($\varepsilon = 8$)	0.847	0.418	0.051

larger gap (0.418 vs 0.452 for centralized MLP), reflecting the greater sensitivity of precision-recall metrics to performance degradation under DP.

After temperature scaling, DP-FL configurations achieve ECE values of approximately 0.050–0.051, whereas the uncalibrated non-DP FL baseline has ECE 0.225 (and 0.003–0.005 for calibrated centralized models). This demonstrates that with proper calibration, DP-FL models can achieve reasonable calibration suitable for clinical decision support.

E. Privacy Accounting Accuracy

The RDP accountant provides accurate tracking of privacy expenditure with $\varepsilon_{\text{spent}}$ consistently close to but slightly below the target ε:

- At $\varepsilon = 2$: $\varepsilon_{\text{spent}} = 1.83$ (8.5% below target)
- At $\varepsilon = 4$: $\varepsilon_{\text{spent}} = 3.66$ (8.5% below target)
- At $\varepsilon = 6$: $\varepsilon_{\text{spent}} = 5.50$ (8.3% below target)
- At $\varepsilon = 8$: $\varepsilon_{\text{spent}} = 7.34$ (8.3% below target)

The consistent under-shoot of approximately 8% arises from conservative accounting (using the maximum steps across clients) and the discrete nature of training steps. This conservatism ensures that the actual privacy guarantee is at least as strong as the target.

VII. DISCUSSION

This section interprets our experimental findings, discusses their clinical implications, and acknowledges the limitations of this study.

A. FL Method Comparability Under DP Constraints

Contrary to some previous non-DP FL studies that report advantages for personalized FL methods like Ditto or FedBN, our results show that under rigorous DP constraints, all four FL methods achieve statistically comparable discriminative performance at each privacy level. The pairwise AUROC comparisons yielded $p > 0.70$ at $\varepsilon = 2$ and $p > 0.93$ at $\varepsilon = 8$, indicating no significant differences.

We hypothesize that the DP noise injection dominates the signal differences between FL algorithms. When gradients are clipped and substantial noise is added, the subtle advantages of personalization or proximal regularization may be obscured. This finding has practical implications: for DP-FL deployments, simpler methods like FedAvg may suffice, avoiding the complexity of personalized FL implementations without sacrificing performance.

The consistency across methods also suggests that the primary determinant of utility in DP-FL is the privacy budget ε rather than the specific FL algorithm. This simplifies decision-making for practitioners: focus on selecting an appropriate ε for the application context rather than optimizing the FL algorithm choice.

B. Privacy-Utility Tradeoff in Clinical Prediction

Our results demonstrate the expected privacy-utility trade-off, with AUROC improving from approximately 0.826 at $\varepsilon = 2$ to 0.847 at $\varepsilon = 8$ (a gain of about 2.1 percentage points). AUPRC shows stronger improvement from 0.359 to 0.418 (about 6.0 percentage points), reflecting the particular sensitivity of precision-recall metrics to the imbalanced mortality prediction task.

The performance gap between DP-FL and non-private baselines is modest: at $\varepsilon = 8$, DP-FL achieves about 97.8% of the AUROC of non-DP FL (0.847 vs 0.866) and 92.5% of the AUPRC (0.418 vs 0.452). This suggests that meaningful privacy guarantees ($\varepsilon = 8$, $\delta = 10^{-5}$) can be achieved with clinically acceptable performance degradation for ICU mortality prediction.

For applications requiring stronger privacy, $\varepsilon = 4$ still achieves about 99.3% of the $\varepsilon = 8$ AUROC (0.841 vs 0.847), suggesting diminishing returns beyond moderate privacy levels. Practitioners should carefully consider whether the marginal utility gain from $\varepsilon = 4$ to $\varepsilon = 8$ justifies the weaker privacy protection.

C. Calibration Under DP-FL and Clinical Implications

Our calibration analysis reveals several important findings:

DP noise affects calibration: Pre-calibration ECE decreased from approximately 0.100 at $\varepsilon = 2$ to 0.092 at $\varepsilon = 8$, suggesting that stronger DP constraints (more noise) slightly worsen calibration. However, this effect is modest compared to the improvement from temperature scaling.

Temperature scaling remains highly effective under DP: Post-hoc temperature scaling reduced ECE by 40–45% across all configurations, with highly statistical significance ($p = 5.32 \times 10^{-58}$). This demonstrates that standard calibration techniques remain applicable in DP-FL settings.

DP invariance of post-processing: Temperature scaling, as a post-processing step applied to model outputs, does not consume additional privacy budget due to the fundamental DP property that post-processing cannot weaken privacy guarantees. This allows practitioners to improve calibration without affecting their privacy accounting.

Clinical importance: Well-calibrated probability estimates are essential for clinical decision-making. Our results show that with temperature scaling, DP-FL models can achieve ECE ≈ 0.05, which is clinically reasonable for mortality prediction. This represents a substantial improvement over uncalibrated non-DP FL (ECE = 0.225 in our experiments).

D. Privacy Accounting and Implementation Details

Our implementation is designed to ensure end-to-end (ε, δ)-differential privacy with transparent and reproducible accounting:

No non-private warmup: All training steps use DP-SGD, so there is no non-private warmup phase. Every gradient update is covered by the same privacy accountant.

Conservative step counting: We compute the total number of DP-SGD steps as $T = R \times E \times \lceil n_k/B \rceil$ using the maximum over clients, which guarantees that the privacy bound holds for all participants.

Transparent reporting: We report both the target privacy budget ε and the realized value $\varepsilon_{\text{spent}}$ returned by the RDP accountant, which is slightly below the target due to the conservative step counting.

Reproducible configuration: All DP-related hyperparameters (noise multipliers, clipping norm, sampling rate, accountant orders) are explicitly documented in Table I and in the experimental setup.

These design choices avoid ambiguity in the privacy accounting and prevent overestimation of the achieved privacy guarantees.

E. Limitations

This study has several limitations that should be considered when interpreting our findings:

Simulated multi-hospital setting: We partitioned a single database (MIMIC-IV) by ICU type to simulate multiple hospitals. While this approach induces realistic data heterogeneity based on patient populations, it does not capture all sources of variation present in true multi-hospital deployments, such as differences in EHR systems, coding practices, clinical protocols, and patient demographics beyond ICU type. Validation on data from genuinely distinct institutions is needed to confirm generalizability.

Limited privacy budget range: We evaluated $\varepsilon \in \{2, 4, 6, 8\}$. Stricter privacy ($\varepsilon < 2$) may be required for some applications but could result in more severe performance degradation. Conversely, our highest $\varepsilon = 8$ is still relatively conservative compared to some industry deployments.

Modest number of seeds: While three seeds provide initial statistical assessment, more extensive repetition (e.g., 10+ seeds) would provide stronger statistical power, particularly for detecting small effect sizes.

Single prediction task and architecture: We focused on in-hospital mortality prediction using a specific MLP architecture. The relative performance of DP-FL methods may differ for other clinical prediction tasks (e.g., length of stay, readmission, specific diagnoses) or different model architectures.

Synchronous FL with full participation: We assumed all clients participate in every round. Asynchronous protocols or partial participation may be necessary in practice and could affect the privacy-utility tradeoff.

Computational constraints: DP-SGD requires per-sample gradient clipping, which increases computational overhead compared to standard training. This may limit the scalability to very large models or datasets.

Despite these limitations, our study provides a systematic comparison of DP-FL methods for ICU mortality prediction with rigorous privacy accounting and statistical analysis, offering practical guidance for privacy-preserving clinical prediction model development.

VIII. Conclusion

This paper presented a comprehensive empirical study of differentially private federated learning for ICU mortality prediction using the MIMIC-IV database, with corrected privacy accounting and rigorous statistical analysis. We compared four DP-FL algorithms—DP-FedAvg, DP-FedProx, DP-FedBN, and DP-Ditto—across multiple privacy budgets ($\varepsilon \in \{2, 4, 6, 8\}$, $\delta = 10^{-5}$) with three random seeds.

A. Summary of Findings

Our experiments yielded several key findings:

1) **FL methods are statistically comparable under DP**: Contrary to non-DP FL studies, all four DP-FL methods achieved statistically indistinguishable discriminative performance at each privacy level (pairwise AUROC comparisons: all $p > 0.05$).
2) **Clear privacy-utility tradeoff**: AUROC improved from 0.826 ± 0.004 at $\varepsilon = 2$ to 0.847 ± 0.004 at $\varepsilon = 8$ (mean $\pm$ std across methods), with AUPRC showing stronger improvement from 0.359 ± 0.013 to 0.418 ± 0.004.
3) **Effective calibration under DP**: Temperature scaling—applied as a DP-preserving post-processing step—significantly reduced Expected Calibration Error by 40–45% (global paired $p = 5.32 \times 10^{-58}$), demonstrating that calibration can be effectively improved without additional privacy cost.
4) **Corrected privacy accounting**: Our implementation ensures end-to-end (ε, δ)-DP guarantees with proper RDP accounting, reporting of actual privacy expenditure ($\varepsilon_{\text{spent}}$ 8–9% below target due to conservative accounting), and no non-private warmup phase.
5) **Clinically relevant performance**: At $\varepsilon = 8$, DP-FL achieved about 97.8% of the AUROC of non-DP FL (0.847 vs 0.866) with reasonable calibration (ECE = 0.050–0.051 after temperature scaling), suggesting practical utility for clinical deployment.

B. Practical Recommendations

Based on our findings, we offer the following guidance for practitioners:

- **Method selection**: For DP-FL deployments, simpler methods like FedAvg may suffice, as all FL methods performed comparably under DP constraints. This simplifies implementation without sacrificing performance.
- **Privacy budget selection**: Target $\varepsilon = 4$–8 for a practical balance between privacy protection and clinical utility. Diminishing returns suggest limited benefit beyond $\varepsilon = 8$ for this task.
- **Calibration**: Always apply post-hoc temperature scaling to improve calibration. This comes at no additional privacy cost due to DP post-processing invariance.
- **Privacy accounting**: Use RDP accounting, report actual $\varepsilon_{\text{spent}}$, and avoid mixed private/non-private training phases to ensure clear privacy guarantees.
- **Statistical evaluation**: Conduct multi-seed experiments with statistical testing, particularly when comparing methods, as single-seed results can be misleading.

C. Future Work

Several directions merit further investigation:

- **Multi-institutional validation**: Evaluating DP-FL methods on data from genuinely distinct hospitals to assess generalizability beyond simulated heterogeneity.
- **Stricter privacy regimes**: Exploring $\varepsilon < 2$ with advanced techniques (e.g., public pre-training, adaptive clipping) to maintain utility under stronger privacy.
- **Alternative DP mechanisms**: Investigating local DP, secure aggregation, or hybrid approaches that may offer different privacy-utility tradeoffs.
- **Fairness under DP-FL**: Assessing whether DP-FL maintains equitable performance across patient subgroups (e.g., demographic groups, disease types).
- **Other clinical tasks and architectures**: Extending the analysis to additional prediction tasks (sepsis, AKI, length of stay) and model architectures (transformers, graph networks).
- **Communication efficiency**: Studying the interaction between DP noise and communication compression techniques in FL.

D. Concluding Remarks

As healthcare institutions increasingly seek to collaborate on predictive modeling while protecting patient privacy, differentially private federated learning offers a principled approach to this challenge. Our study demonstrates that with proper privacy accounting and calibration, DP-FL can achieve clinically useful ICU mortality prediction while providing formal privacy guarantees. The comparability of FL methods under DP constraints simplifies implementation choices, allowing practitioners to focus on privacy budget selection and calibration. This work contributes a foundation for privacy-preserving clinical prediction in multi-institutional settings with rigorous evaluation methodology.

References

[1] A. E. Johnson, L. Bulgarelli, T. J. Pollard, S. Horng, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023.

[2] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, 2019.

[3] N. Rieke *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, p. 119, 2020.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[5] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12598, 2020.

[6] I. Dayan *et al.*, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.

[7] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. NeurIPS*, 2019, pp. 14774–14784.

[8] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE S&P*, 2019, pp. 739–753.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[10] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM CCS*, 2016, pp. 308–318.

[11] I. Mironov, "Rényi differential privacy," in *Proc. IEEE CSF*, 2017, pp. 263–275.

[12] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[13] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[14] Meta AI, "Opacus: User-friendly differential privacy library in PyTorch," 2021. [Online]. Available: https://opacus.ai/

[15] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Automatic clipping: Differentially private deep learning made easier and stronger," in *Proc. NeurIPS*, 2023.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, 2020.

[17] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in *Proc. ICLR*, 2021.

[18] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. ICML*, 2021, pp. 6357–6368.

[19] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. ICLR*, 2020.

[20] N. Tahir *et al.*, "Federated learning-based model for predicting mortality: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 27, 2025.

[21] L. Mondrejevski, I. Miliou, A. Montanino, D. Pitts, J. Hollmén, and P. Papapetrou, "FLICU: A federated learning workflow for intensive care unit mortality prediction," in *Proc. IEEE Int. Symp. on Computer-Based Medical Systems (CBMS)*, 2022.

[22] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado, "Calibrating predictive model estimates to support personalized medicine," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 263–274, 2012.

[23] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: Seven steps for development and an ABCD for validation," *European Heart Journal*, vol. 35, no. 29, pp. 1925–1931, 2014.

[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.

[25] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. KDD*, 2019, pp. 2623–2631.