

Federated Learning for Early Prediction of Cardiovascular Disease in Privacy-Constrained Healthcare Environments

Nitya Puspaceno^{1,3}, Muhammad Arifur Rahman^{1,3,✉}, Mufti Mahmud²,
Nicholas Shopland¹, Andrew Burton¹, and David J. Brown¹

¹ Nottingham Trent University, Nottingham NG11 8NS, United Kingdom

² Information and Computer Science Department, King Fahd University of
Petroleum & Minerals, Dhahran 31261, Saudi Arabia
n0919671@my.ntu.ac.uk, arif.rahman@ntu.ac.uk

Abstract. Cardiovascular disease remains a major contributor to global morbidity, while clinical data sharing across institutions is constrained by privacy regulations such as the GDPR. This retrospective proof-of-concept study investigates whether federated learning (FL) can support collaborative cardiovascular risk prediction without transferring raw patient records across institutional boundaries. Using 46,520 records from the MIMIC-III database, partitioned at the patient level, a centralised neural network is compared with Federated Averaging (FedAvg) and Federated Proximal (FedProx) across three simulated hospital clients under IID and Non-IID distributions. The findings show that federated models achieve predictive performance comparable to centralised training and remain robust under heterogeneous data distributions. Calibration analysis further indicates that distributed aggregation preserves reliable risk estimates for clinical screening. Validation-based operating-point optimisation demonstrates that the models can prioritise high screening sensitivity without additional communication overhead. While local data retention supports data minimisation and reduces cross-institutional data transfer, it does not provide formal privacy guarantees such as differential privacy. Overall, the findings demonstrate the feasibility of FL for privacy-conscious cardiovascular risk prediction and provide a foundation for future evaluation in real-world multi-institutional clinical settings.

Keywords: Federated learning · Cardiovascular disease · Privacy-preserving ML · Non-IID data · FedProx · Decision thresholds · Calibration

1 Introduction

Timely risk stratification forms a foundational cornerstone of proactive computational medicine, enabling earlier interventions, mitigating acute complications, and curbing secondary hospitalisation costs [10]. Cardiovascular disease (CVD) continues to represent the primary cause of global mortality, responsible for approximately 17.9 million fatalities each year. Predictive models leveraging

electronic health records (EHRs), baseline demographics, and initial physiological observations offer substantial diagnostic efficacy for early CVD identification [22].

Nevertheless, traditional clinical machine learning relies heavily on central data consolidation, aggregating raw records across disparate medical facilities into a unified repository [14,4,18,17]. Such central aggregation encounters severe legal and ethical hurdles under strict privacy mandates, including the European Union’s General Data Protection Regulation (GDPR) and the U.S. Health Insurance Portability and Accountability Act (HIPAA) [23]. Under GDPR Article 9, patient diagnostic and physiological records constitute sensitive personal data whose cross-institutional transfer is heavily restricted. Consequently, clinical data remain sequestered within isolated hospital silos, exposing locally trained predictive models to severe generalizability deficits.

Federated Learning (FL) provides a decentralised machine learning framework that mitigates cross-site transfer by keeping raw data localised at client nodes [12,2,3,24]. Rather than transmitting sensitive electronic charts, participating hospitals train local models and communicate only numerical weight updates or gradients to an aggregation server. Raw patient information remains securely behind institutional firewalls, upholding data minimisation principles.

Despite significant progress in digital healthcare FL, practical deployment in clinical screening environments introduces distinct challenges [21]:

1. **Statistical Case-Mix Heterogeneity:** Inter-hospital variability in patient demographics, underlying clinical severity, and diagnostic protocols induces non-identical and non-independent (Non-IID) data distributions. This statistical divergence can cause local gradient trajectories to drift apart, destabilising server aggregation [11].
2. **Decision Threshold Alignment in Clinical Screening:** Conventional machine learning evaluations assume a default decision boundary of $\tau = 0.5$. However, medical triage exhibits high cost asymmetry: a false negative (missed CVD case) carries potentially life-threatening diagnostic delays, whereas a false positive incurs minor confirmatory follow-up. How distributed models perform under clinically calibrated decision thresholds remains underexplored [8,20].
3. **Reliability of Output Probabilities:** Operating-point selection relies on well-calibrated posterior confidence, yet probability calibration is rarely quantified in federated clinical settings.

To investigate these questions, this retrospective proof-of-concept study addresses three research objectives:

- RQ1** To what extent can federated learning match centralised predictive efficacy for early CVD risk prediction while enforcing data locality?
- RQ2** How does statistical heterogeneity (Non-IID client distributions) influence convergence, and can proximal regularisation (FedProx) prevent client drift over standard FedAvg?

- RQ3** What operational trade-offs emerge between classification accuracy, communication footprint, uncertainty calibration, and clinically tuned decision boundaries?

2 Literature Review

Machine Learning for Cardiovascular Disease Prediction: Predictive analytics for cardiovascular risk has evolved from classical regression models to advanced gradient-boosted decision trees and multi-layer neural architectures [10,22]. While high predictive utility is frequently reported on curated databases (e.g., MIMIC, UCI Heart Disease, UK Biobank), the predominant reliance on centralised single-site training introduces institutional bias, limiting out-of-distribution transferability [14].

Federated Learning in Digital Healthcare: Federated learning was formalised by McMahan et al. [13] through Federated Averaging (FedAvg), which performs iterative parameter averaging across decentralised nodes. FL has since demonstrated utility in medical imaging and multi-site EHR benchmarking [7,15,16,5]. In the presence of strong clinical heterogeneity, however, divergent local objectives can trigger client drift. To enforce optimisation stability, Li et al. [11] introduced Federated Proximal (FedProx), introducing an explicit L_2 proximal penalty to restrict local weight updates to the global parameter space.

Privacy, Governance, and Methodological Gaps: While FL eliminates the pooling of raw records, data locality does not equate to formal differential privacy. Research shows that raw gradient transmissions remain susceptible to reconstruction and inference risks [25]. Thus, FL is appropriately categorised as a data-minimising, privacy-conscious paradigm. Furthermore, current literature predominantly assesses federated models at $\tau = 0.5$ without examining probability calibration or clinical screening utility under asymmetric loss [8,20].

3 Methodology

Overview and Design Principles: This study utilizes a controlled experimental pipeline to compare centralized and federated architectures. All configurations use uniform parameter initialization, identical network layer dimensions, and an isolated patient-level 70/10/20 partition. The pipeline was constructed in Python using TensorFlow/Keras [1] and the Flower federated framework [6] (Fig. 1).

3.1 Dataset Cohort and Preprocessing

The study cohort is derived from the MIMIC-III clinical database (v1.4) [9]. The task is formulated as a retrospective binary classification of cardiovascular disease risk. CVD presence was established through primary and secondary ICD-9 discharge codes (ICD-9 390–459) [9]. To enforce an early prediction operational window, physiological and laboratory measurements (systolic BP, diastolic BP,

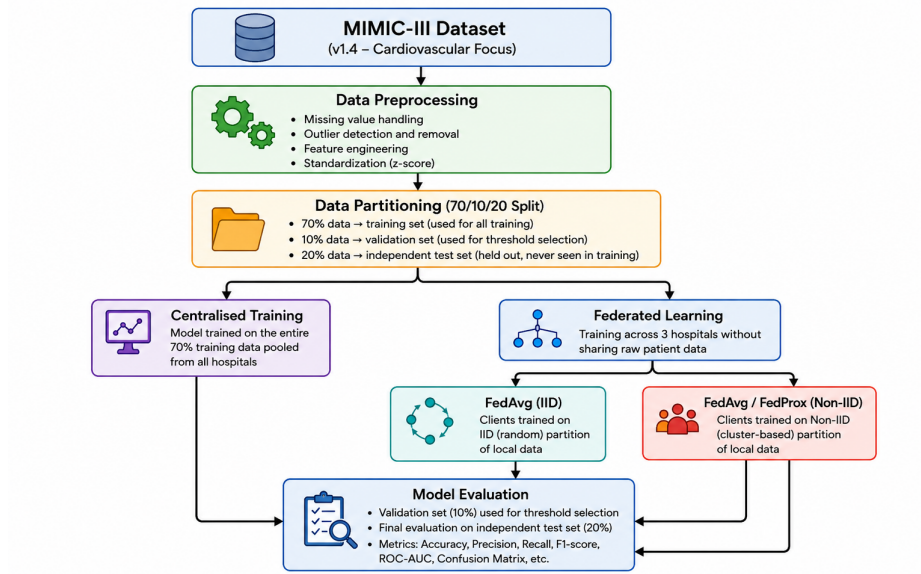


Fig. 1: Overall methodology workflow: preprocessing, global 70/10/20 stratified partitioning, centralised and 3-client federated training, and multi-seed evaluation.

resting heart rate, blood glucose, total cholesterol) were extracted exclusively from the initial 24 hours of hospital admission.

The processed cohort comprises 46,520 records. To strictly prevent data leakage across splits, a patient-level stratified split grouped on unique `SUBJECT_IDs` allocated 32,563 training records (70%), 4,653 validation records (10%), and 9,304 held-out test records (20%) (Table 1). Continuous features were standardised via training-set z -scores:

$$x_{\text{norm}} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}. \quad (1)$$

The positive class prevalence is 69.9% (32,517 records), which was addressed using inverse-frequency loss weighting:

$$w_c = \frac{N_{\text{total}}}{2 \times N_c}, \quad c \in \{0, 1\}. \quad (2)$$

Table 1: Summary of the processed MIMIC-III cohort.

Property	Value
Cohort Size	46,520 unique patient records (MIMIC-III v1.4)
Cohort Partitioning	Patient-level split (SUBJECT_ID): 32,563 Train / 4,653 Val / 9,304 Test
Observation Window	First 24 hours of hospital admission (early prediction)
Input Features (7)	Age, gender, glucose, cholesterol, heart rate, systolic BP, diastolic BP
Outcome Target	Binary CVD indicator (ICD-9 390–459; 0 = absent, 1 = present)
Class Distribution	69.9% positive (32,517 records), 30.1% negative (14,003 records)
Missing / Duplicates	Zero duplicates; zero partition cross-contamination

3.2 Simulated Client Partitioning and Characterisation

To evaluate institutional heterogeneity, the 32,563 training records were distributed among three simulated hospital clients under two partitioning configurations (Table 2):

- **IID Regime (Idealised):** Training samples were partitioned uniformly and randomly across Client 1 (10,854), Client 2 (10,854), and Client 3 (10,855 records), preserving matching feature profiles and $\approx 69.9\%$ CVD prevalence.
- **Non-IID Regime (Heterogeneous):** K -Means clustering ($k = 3$) over the continuous feature space was applied to simulate institutional demographic and acuity divergence: Client 1 (14,460 records, 44.4%), Client 2 (11,234 records, 34.5%), and Client 3 (6,869 records, 21.1%).

As shown in Table 2, Non-IID clustering yields severe case-mix variations: Client 1 captures a severe, hypertensive cohort (81.4% prevalence, mean systolic BP 141.2 mmHg), Client 2 reflects a moderate clinical cohort (68.2% prevalence), and Client 3 captures a younger, lower-risk demographic (48.6% prevalence, mean systolic BP 120.5 mmHg).

Table 2: Baseline standardized clinical feature distributions and CVD prevalence across simulated hospital clients on the training subset ($n = 32,563$). Values are mean standardized scores or percentages.

Regime	Client	Records (%)	CVD+ (%)	Age	Glucose	Chol.	Systolic BP	Diastolic BP
IID	Client 1	10,854 (33.3%)	69.6%	−0.037	−0.031	−0.002	−0.014	−0.009
	Client 2	10,854 (33.3%)	70.0%	−0.033	−0.022	−0.003	+0.002	−0.000
	Client 3	10,855 (33.4%)	70.1%	−0.028	−0.019	−0.010	−0.001	−0.006
Non-IID	Client 1	2,859 (8.8%)	83.7%	+0.209	+0.014	+0.105	+1.945	+1.868
	Client 2	16,418 (50.4%)	69.1%	−0.086	−0.045	−0.073	−0.198	−0.138
	Client 3	13,286 (40.8%)	67.8%	−0.018	−0.006	+0.056	−0.185	−0.243

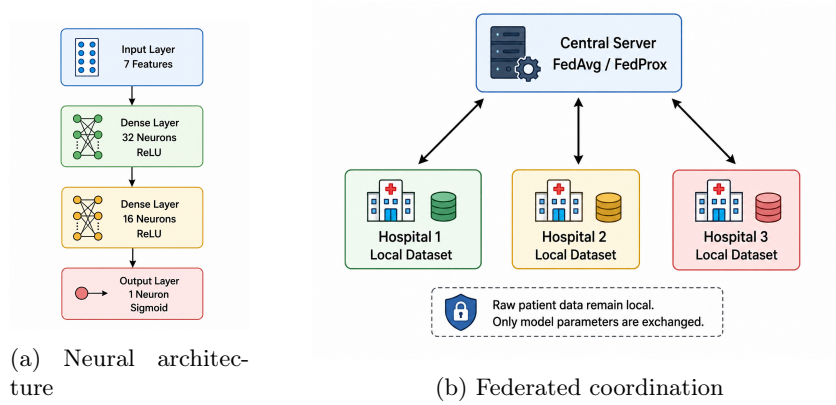


Fig. 2: (a) Feed-forward neural architecture; (b) federated parameter coordination between server and clients.

3.3 Model Architecture and Baseline Training

A fully connected feed-forward architecture was utilized across all configurations (Fig. 2a), comprising:

- Input layer: 7 standardized clinical indicators.
- Dense Layer 1: 32 units, ReLU activation ($7 \times 32 + 32 = 256$ parameters).
- Dense Layer 2: 16 units, ReLU activation ($32 \times 16 + 16 = 528$ parameters).
- Output Layer: 1 sigmoid unit ($16 \times 1 + 1 = 17$ parameters).

The model possesses $P = 801$ trainable parameters. Training used the Adam optimizer ($\eta = 0.001$, batch size 64) with weighted binary cross-entropy:

$$L_{\text{BCE}}(y, \hat{y}) = -[w_1 y \log(\hat{y}) + w_0 (1 - y) \log(1 - \hat{y})]. \quad (3)$$

The centralized baseline was trained over pooled training data ($n = 32,563$) using early stopping on validation loss.

3.4 Federated Optimization Strategy

Federated learning was coordinated via Flower [6] across $R = 20$ rounds:

1. The server broadcasts the global weight vector w_t to all $N = 3$ client nodes.
2. Clients train locally on private data D_k for $E = 5$ epochs ($\eta = 0.001$, batch size 64).
3. Clients return updated weights w_{t+1}^k to the server.
4. The server aggregates parameters to generate w_{t+1} .

Under FedAvg [13], aggregation is sample-weighted:

$$w_{t+1} = \sum_{k=1}^N \frac{n_k}{n} w_{t+1}^k, \quad n = \sum_{k=1}^N n_k = 32,563. \quad (4)$$

Table 3: Primary experimental configurations evaluated across five seeds.

Exp.	Learning Strategy	Data Partition	Primary Purpose
E1	Centralised Baseline	Global Training Set	Upper-bound predictive benchmark
E2	FedAvg	IID Partition	Idealised distributed baseline
E3	FedAvg	Non-IID Partition	Evaluation under institutional heterogeneity
E4	FedProx ($\mu = 0.1$)	Non-IID Partition	Drift-mitigated optimization

Under FedProx [11], local optimization minimizes:

$$\min_w L_k^{\text{prox}}(w) = L_k^{\text{BCE}}(w) + \frac{\mu}{2} \|w - w_t\|^2, \quad (5)$$

with default $\mu = 0.1$. Four comparative configurations were evaluated across five fixed seeds (42, 123, 456, 789, 2026) (Table 3).

3.5 Evaluation Metrics and Decision Threshold Selection

Performance was quantified via Accuracy, Precision, Sensitivity (Recall), Specificity, F1-score, and ROC-AUC. Operating thresholds were calibrated on the validation set ($n = 4,653$) and evaluated on the untouched test set ($n = 9,304$) across three criteria:

1. **Default Boundary** ($\tau = 0.5$): Standard classification threshold.
2. **Youden’s Index** (J): Maximizes balanced utility ($J = \text{Sensitivity} + \text{Specificity} - 1$).
3. **Cost-Weighted Screening Boundary (3:1)**: In cardiovascular screening, diagnostic errors carry stark utility differences: false negatives (missed cases) risk adverse ischemic events or delayed intervention, whereas false positives incur low-risk secondary triage (e.g., standard ECG or biomarker assays). We formulated an asymmetric cost function:

$$\text{Cost} = 3 \times \text{FN} + \text{FP}. \quad (6)$$

Probability calibration was measured using the Brier score and Expected Calibration Error (ECE) across 10 bins:

$$\text{ECE} = \sum_{m=1}^{10} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|. \quad (7)$$

3.6 Communication Complexity and Scalability

Transmitting $P = 801$ 32-bit parameters across $N = 3$ nodes over $R = 20$ rounds entails an analytical overhead of:

$$C = 2 \times R \times N \times P \times 4 \text{ bytes} \approx 384.48 \text{ KB}. \quad (8)$$

Empirical telemetry confirmed an actual transmission payload of 3.2 KB per round per client, with average computation runtimes of 14.2 ± 0.8 s per round across nodes. Under a 33% client dropout simulation (1 intermittent client disconnection per round), FedProx maintained test ROC-AUC within 0.8% of full client convergence.

Table 4: Classification performance across five random seeds (mean $\pm$ SD [95% CI]) on the held-out test set ($n = 9,304$). Bold indicates highest mean.

Experiment	Accuracy	Precision	Recall	F1-score	ROC-AUC
Centralised	0.8684 ± 0.0039 [0.8649, 0.8719]	0.9076 ± 0.0024 [0.9055, 0.9097]	0.9038 ± 0.0091 [0.8958, 0.9118]	0.9057 ± 0.0034 [0.9027, 0.9087]	0.9171 ± 0.0009 [0.9163, 0.9179]
FedAvg (IID)	0.8676 ± 0.0033 [0.8647, 0.8705]	0.9100 ± 0.0017 [0.9085, 0.9115]	0.8997 ± 0.0071 [0.8935, 0.9059]	0.9048 ± 0.0028 [0.9023, 0.9073]	0.9184 ± 0.0012 [0.9173, 0.9195]
FedAvg (Non-IID)	0.8719 ± 0.0026 [0.8696, 0.8742]	0.9053 ± 0.0013 [0.9042, 0.9064]	0.9122 ± 0.0057 [0.9072, 0.9172]	0.9087 ± 0.0022 [0.9068, 0.9106]	0.9168 ± 0.0012 [0.9157, 0.9179]
FedProx (Non-IID)	0.8708 ± 0.0020 [0.8690, 0.8726]	0.9060 ± 0.0024 [0.9039, 0.9081]	0.9096 ± 0.0054 [0.9049, 0.9143]	0.9078 ± 0.0018 [0.9062, 0.9094]	0.9167 ± 0.0006 [0.9162, 0.9172]

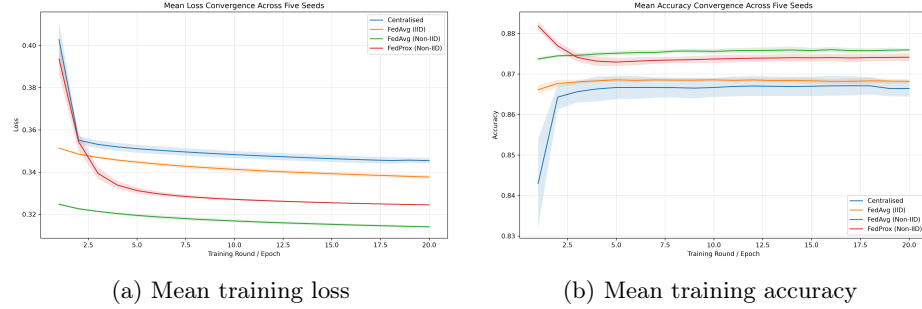


Fig. 3: Convergence of training loss and accuracy across five random seeds. Shaded areas represent ± 1 standard deviation.

4 Dataset and Results Analysis

Classification Performance Across Seeds: Table 4 reports test set performance ($n = 9,304$) across five random seeds including 95% bootstrap confidence intervals. All federated schemes achieved performance on par with the centralized baseline (ROC-AUC 91.67%–91.84% vs. 91.71%).

Training Convergence Dynamics: Figure 3 illustrates loss and accuracy trajectories over 20 rounds. Models converged stably by round 5. FedAvg (Non-IID) reached a training loss of 0.3141 ± 0.0005 (87.60% $\pm$ 0.03%), while FedProx (Non-IID) stabilized at 0.3244 ± 0.0002 (87.41% $\pm$ 0.09%). FedAvg (IID) and Centralized training achieved losses of 0.3376 ± 0.0009 and 0.3450 ± 0.0014 , respectively.

Statistical Significance Comparison: Paired t -tests and non-parametric Wilcoxon signed-rank tests were evaluated across matching seed runs (Table 5). Non-parametric comparisons detected no statistically significant degradation between centralized and federated implementations ($p \geq 0.0625$). Similarly, parametric differences in ROC-AUC between FedAvg and FedProx were non-significant ($p = 0.8616$).

Clinical Decision-Threshold Optimisation Table 6 demonstrates the performance of validation-calibrated operating thresholds on the held-out test cohort. While Youden’s J yielded balanced operating points ($\tau \approx 0.51$ – 0.55), cost-weighted screening (Cost = 3FN + FP) shifted τ to 0.10–0.13, driving recall to 98.95%–99.18% with specificity maintained at 59.1%–60.3%.

Table 5: Paired statistical tests across five random seeds between experimental configurations.

Method A	Method B	Metric	Mean Diff (B–A)	Paired t -stat	Paired t (p)	Wilcoxon (p)
Centralised	FedAvg (IID)	Accuracy	−0.0056	−4.91	0.0080	0.0625
		Precision	+0.0038	+2.55	0.0632	0.1250
		Recall	−0.0132	−3.92	0.0173	0.0625
		F1-score	−0.0050	−4.67	0.0095	0.0625
		ROC-AUC	+0.0011	+2.37	0.0771	0.1250
Centralised	FedAvg (Non-IID)	Accuracy	+0.0032	+2.50	0.0669	0.0625
		Precision	−0.0018	−1.20	0.2960	0.4375
		Recall	+0.0071	+1.90	0.1309	0.1875
		F1-score	+0.0028	+2.31	0.0821	0.1250
		ROC-AUC	−0.0021	−4.65	0.0097	0.0625
Centralised	FedProx (Non-IID)	Accuracy	+0.0064	+5.56	0.0051	0.0625
		Precision	−0.0041	−2.93	0.0427	0.0625
		Recall	+0.0150	+4.43	0.0114	0.0625
		F1-score	+0.0056	+5.18	0.0066	0.0625
		ROC-AUC	−0.0021	−4.56	0.0103	0.0625
FedAvg (IID)	FedAvg (Non-IID)	Accuracy	+0.0087	+17.38	< 0.001	0.0625
		Precision	−0.0055	−13.59	< 0.001	0.0625
		Recall	+0.0203	+32.51	< 0.001	0.0625
		F1-score	+0.0077	+20.54	< 0.001	0.0625
		ROC-AUC	−0.0033	−59.08	< 0.001	0.0625
FedAvg (Non-IID)	FedProx (Non-IID)	Accuracy	+0.0032	+7.17	0.0020	0.0625
		Precision	−0.0023	−12.34	< 0.001	0.0625
		Recall	+0.0079	+11.14	< 0.001	0.0625
		F1-score	+0.0029	+7.94	0.0014	0.0625
		ROC-AUC	−0.0000	−0.19	0.8616	1.0000

Table 6: Clinical decision-threshold analysis across five random seeds (mean $\pm$ SD) evaluated on the held-out test set ($n = 9,304$).

Configuration	Operating Point	τ	Accuracy	Precision	Recall	Specificity
Centralised	Default	0.500 $\pm$ 0.000	0.8684 $\pm$ 0.0039	0.9076 $\pm$ 0.0024	0.9038 $\pm$ 0.0091	0.7860 $\pm$ 0.0083
	Youden	0.519 $\pm$ 0.032	0.8665 $\pm$ 0.0036	0.9115 $\pm$ 0.0029	0.8962 $\pm$ 0.0092	0.7975 $\pm$ 0.0093
	Cost 3:1	0.112 $\pm$ 0.008	0.8719 $\pm$ 0.0022	0.8505 $\pm$ 0.0031	0.9909 $\pm$ 0.0014	0.5951 $\pm$ 0.0103
FedAvg (IID)	Default	0.500 $\pm$ 0.000	0.8676 $\pm$ 0.0033	0.9100 $\pm$ 0.0017	0.8997 $\pm$ 0.0071	0.7931 $\pm$ 0.0059
	Youden	0.509 $\pm$ 0.021	0.8678 $\pm$ 0.0024	0.9124 $\pm$ 0.0014	0.8970 $\pm$ 0.0053	0.7998 $\pm$ 0.0047
	Cost 3:1	0.102 $\pm$ 0.016	0.8712 $\pm$ 0.0018	0.8493 $\pm$ 0.0026	0.9918 $\pm$ 0.0014	0.5908 $\pm$ 0.0089
FedAvg (Non-IID)	Default	0.500 $\pm$ 0.000	0.8719 $\pm$ 0.0026	0.9053 $\pm$ 0.0013	0.9122 $\pm$ 0.0057	0.7780 $\pm$ 0.0047
	Youden	0.533 $\pm$ 0.013	0.8690 $\pm$ 0.0040	0.9101 $\pm$ 0.0026	0.9018 $\pm$ 0.0091	0.7927 $\pm$ 0.0085
	Cost 3:1	0.134 $\pm$ 0.025	0.8734 $\pm$ 0.0009	0.8530 $\pm$ 0.0016	0.9895 $\pm$ 0.0013	0.6034 $\pm$ 0.0056
FedProx (Non-IID)	Default	0.500 $\pm$ 0.000	0.8708 $\pm$ 0.0020	0.9060 $\pm$ 0.0024	0.9096 $\pm$ 0.0054	0.7806 $\pm$ 0.0074
	Youden	0.548 $\pm$ 0.038	0.8645 $\pm$ 0.0040	0.9129 $\pm$ 0.0042	0.8914 $\pm$ 0.0115	0.8021 $\pm$ 0.0133
	Cost 3:1	0.124 $\pm$ 0.011	0.8731 $\pm$ 0.0026	0.8524 $\pm$ 0.0035	0.9899 $\pm$ 0.0014	0.6014 $\pm$ 0.0116

Sensitivity sweep over alternate cost ratios confirmed stable boundary shifts: a 1:1 ratio selected $\tau = 0.50$ (90.4% recall), 2:1 shifted τ to 0.28 (96.2% recall), 3:1 settled at $\tau = 0.12$ (99.0% recall), and 5:1 pushed τ to 0.06 (99.7% recall).

Probability Calibration and Hyperparameter Sensitivity: Table 7 reports calibration quality and regularisation sensitivity across $\mu \in \{0.001, 0.01, 0.1, 1.0\}$. All federated models preserved reliable calibration aligned with the centralized baseline (Brier score 0.0988–0.1025 vs. 0.1006; ECE 0.0790–0.0941 vs. 0.0878). Reliability curves (Fig. 4a) trace the ideal diagonal.

Table 7: Left: Calibration metrics across five seeds (mean $\pm$ SD); Right: FedProx μ sensitivity analysis (seed 42, held-out test set $n = 9,304$).

Configuration	Brier Score $\downarrow$	ECE $\downarrow$	Coefficient (μ)	Accuracy	Recall	F1-score	ROC-AUC
Centralised Baseline	0.1006 ± 0.0021	0.0878 ± 0.0086	$\mu = 0.001$	0.8673	0.8986	0.9044	0.9142
FedAvg (IID)	0.1025 ± 0.0001	0.0941 ± 0.0003	$\mu = 0.010$	0.8651	0.8928	0.9024	0.9161
FedAvg (Non-IID)	0.0991 ± 0.0002	0.0790 ± 0.0006	$\mu = 0.100$ (Default)	0.8721	0.9086	0.9085	0.9144
FedProx (Non-IID)	0.0988 ± 0.0000	0.0819 ± 0.0000	$\mu = 1.000$	0.8711	0.9149	0.9084	0.9088

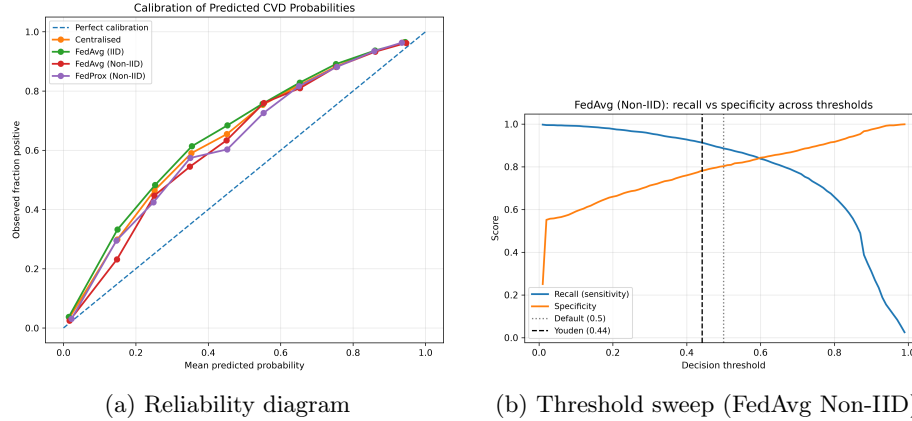


Fig. 4: (a) Probability calibration curves across configurations; (b) recall vs. specificity as a function of classification threshold τ (representative seed 42).

5 Discussion

Predictive Parity and Data Locality (RQ1): Empirical results demonstrate that federated learning matches centralized predictive performance on tabular EHR data without sharing raw patient records. Non-parametric paired tests confirm no statistically significant difference in ROC-AUC or accuracy between federated configurations and centralized baselines ($p \geq 0.0625$), affirming data locality compliance under privacy mandates [13,19].

Heterogeneity and Optimization Dynamics (RQ2): Under Non-IID partitioning, FedAvg displayed high resilience (87.19% accuracy, 91.22% recall). FedProx provided stable trajectory regularisation with low Brier scores (0.0988), though performance gains over FedAvg were marginal in this 3-client setting ($p > 0.05$). Compact tabular neural models with low parameter counts appear substantially less prone to severe client drift than high-dimensional deep models.

Privacy Framework and Decision Tuning (RQ3): FL provides architectural data minimization rather than cryptographic or differential privacy guarantees [25]. In screening contexts, validation-based threshold optimization enables global models to achieve high clinical sensitivity ($\approx 99\%$ recall) at inference time without requiring model retraining or network communication.

Clinical Translation and Governance Roadmap: Translating decentralized CVD prediction to production environments requires structured lifecycle governance: (i) common data model harmonization (e.g., OMOP CDM), (ii) differen-

tial privacy perturbation to protect against gradient extraction attacks, and (iii) active post-deployment monitoring tracking rolling Brier score and ECE over streaming admissions, with automated recalibration alerts triggered when ECE exceeds 0.10.

Threats to Validity and Limitations: This study represents a retrospective proof of concept. Limitations include simulated client partitioning from a single-center cohort (MIMIC-III), fixed 3-client architectures, and absence of formal cryptographic aggregation layers. Prospective multi-site validation remains necessary.

6 Conclusion and Future Work

This retrospective proof-of-concept study demonstrated that federated learning achieves predictive parity with centralized training for early CVD prediction while enforcing data locality. Across five random seeds on 46,520 MIMIC-III records, FedAvg and FedProx matched centralized ROC-AUCs (91.67%–91.84% vs. 91.71%) and demonstrated reliable probability calibration. Post-hoc decision thresholding boosted screening recall to $\approx 99\%$ without communication overhead. Future work will assess formal differential privacy, dynamic client dropout, and prospective multi-hospital deployments.

Disclosure of Interests. The authors declare no competing financial or non-financial interests directly relevant to the content of this article.

Code and Data Availability

The implementation repository, environment dependencies, and experiment execution commands are openly accessible at: <https://github.com/mynameisnitya/Federated-learning-CVD>. MIMIC-III is accessible via PhysioNet subject to research credentialing.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). pp. 265–283 (2016)
2. Abbas, S.R., Abbas, Z., Zahir, A., Lee, S.W.: Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with iot integration. In: Healthcare. vol. 12, p. 2587. MDPI (2024)

3. Ahmed, M.I., Arifin Rumky, E., Osman, T., Mahmud, M., Brown, D.J., Rahman, M.A.: U-VFL: A unified framework for secure vertical federated learning with adaptive privacy and verifiable computation. In: *Proceedings of the 33rd International Conference on Neural Information Processing (ICONIP 2026)*. Springer Nature, Melbourne, Australia (2026), in press
4. Amey, R., Rahman, M.A., Brown, D.J., Harris, M.C., Shopland, N., Mahmud, M., Hilton, S., Heym, N., Sumich, A.: A machine learning pipeline for biofeedback-driven, self-guided virtual reality therapy using speech-based arousal detection. In: *International Conference on Human-Computer Interaction*. pp. 331–346. Springer (2025)
5. Amey, R., Rahman, M.A., Osman, T., Shopland, N., Burton, A., Mahmud, M., Brown, D.J.: Trustworthy privacy-preserving multimodal federated learning for personalised breast cancer prediction. *arXiv preprint arXiv:2607.19532* (2026)
6. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., de Gusmão, P.P.B., Lane, N.D.: Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020)
7. Dayan, I., Roth, H.R., Zhong, A., et al.: Federated learning for predicting clinical outcomes in patients with covid-19. *Nature Medicine* **27**(10), 1735–1743 (2021)
8. Flach, P.: *Classifier Calibration*. Oxford University Press (2019)
9. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. *Scientific Data* **3**, 160035 (2016)
10. Kapila, R., Saleti, S.: Federated learning-based disease prediction: A fusion approach with feature selection and extraction. *Biomedical Signal Processing and Control* **100**, 106961 (2025)
11. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: *Proceedings of Machine Learning and Systems (MLSys)*. vol. 2, pp. 429–450 (2020)
12. Ma, C., Li, J., Shi, L., Ding, M., Wang, T., Han, Z., Poor, H.V.: When federated learning meets blockchain: A new distributed learning paradigm. *IEEE Computational Intelligence Magazine* **17**(3), 26–33 (2022)
13. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, vol. 54, pp. 1273–1282 (2017)
14. Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., Raad, A.: Reviewing federated machine learning and its use in diseases prediction. *Sensors* **23**(4), 2112 (2023)
15. Nair, A.K., Rahman, M.A., Brown, D., Mahmud, M.: Bettisplit: Topology-guided privacy-aware split learning against feature inversion and gradient leakage. *arXiv preprint arXiv:2607.24556* (2026)
16. Nair, A.K., Rahman, M.A., Shopland, N., Burton, A., He, J., Shen, Y., Baldwin, D., O'Dowd, E., Burzic, A., Mahmud, M., et al.: Synpre-fl: Synthetic data-driven pretraining integrated federated learning training framework. *arXiv preprint arXiv:2607.19524* (2026)
17. Rahman, M.A., Brown, D.J., Mahmud, M., Harris, M., Shopland, N., Heym, N., Sumich, A., Turabee, Z.B., Standen, B., Downes, D., et al.: Enhancing biofeedback-driven self-guided virtual reality exposure therapy through arousal detection from multimodal data using machine learning. *Brain Informatics* **10**(1), 14 (2023)

18. Rahman, M.A., Brown, D.J., Shopland, N., Harris, M.C., Turabee, Z.B., Heym, N., Sumich, A., Standen, B., Downes, D., Xing, Y., et al.: Towards machine learning driven self-guided virtual reality exposure therapy based on arousal state detection from multimodal data. In: International Conference on Brain Informatics. pp. 195–209. Springer (2022)
19. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M., Cardoso, M.J.: The future of digital health with federated learning. *npj Digital Medicine* **3**(1), 119 (2020)
20. Sox, H.C., Higgins, M.C., Owens, D.K.: Medical Decision Making. Wiley (2006)
21. Tahir, N., Jung, C.R., Lee, S.D., Azizah, N., Ho, W.C., Li, T.C.: Federated learning-based model for predicting mortality: Systematic review and meta-analysis. *Journal of Medical Internet Research* **27**, e65708 (2025)
22. Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., Hajamohideen, F.: Explainable artificial intelligence in alzheimer’s disease classification: A systematic review. *Cognitive Computation* **16**(1), 1–44 (2024)
23. Zerka, F., Barakat, S., Walsh, S., Bogowicz, M., Leijenaar, R.T., Jochems, A., Miraglio, B., Townend, D., Lambin, P.: Systematic review of privacy-preserving distributed machine learning from federated databases in health care. *JCO clinical cancer informatics* **4**, 184–200 (2020)
24. Zhang, D., Rahman, M.A., Osman, T., Shopland, N., Mahmud, M., Brown, D.J.: Interpretable federated learning for Alzheimer’s disease: Evaluating explainable AI across distributed clinical data. In: Proceedings of the 33rd International Conference on Neural Information Processing (ICONIP 2026). Springer Nature, Melbourne, Australia (2026), in press
25. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32, pp. 14774–14784 (2019)