Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 16 May 2026
Revised 16 June 2026
Accepted 04 July 2026

Natural Resources for Human Health 2026, Vol. 6 (4s): 60–74

KEYWORDS:
federated learning; digital twin; foundation models; explainable artificial intelligence; differential privacy; multimodal clinical prediction; trustworthy AI; low-rank adaptation

Privacy-Preserving Federated Digital Twin Intelligence for Trustworthy Multimodal Disease Prediction Using Explainable Foundation Models

[1]Ganesh Dagadu Puri, [2]Dr. Aarti S.Gaikwad, [3]Dr. Supriya Sabale, [4]Dr. Vinod V. Kimbahune, [5]Dr. Mansi Bhonsle, [6]Dr. Sachin Arun Thanekar

[1]Department of Computer Engineering, Amrutvahini College of Engineering Sangamner, Maharashtra, India puriganeshengg@gmail.com 0000-0002-6786-698X
[2]Department of Information Technology D.Y. Patil College of Engineering, Akurdi, Pune
aratig.2010@gmail.com. 0000-0003-0873-8045
[3]Department of Computer Engineering MIT Academy of Engineering, Alandi, Pune
supriya.sabale@mitaoe.ac.in 0000-0002-7847-0681
[4]Department of Computer Engineering Vishwakarma Institute of Technology, Pune
vinod.kimbahune@vit.edu 0000-0002-3076-2001
[5]Department of Computer Science and Engineering, School of Computing, MIT Art, Design and Technology University, Pune - 412201, India.
mansi.bhonsle@gmail.com 0000-0002-0595-6023
[6]Department of Information Technology, Sanjeevani College of Engineering, Kopargaon, Maharashtra, India
sachin.sangamner@gmail.com 0000-0002-1852-8458

ABSTRACT: Clinical risk models seldom fail for want of a better algorithm. They fail because the records that would make them dependable sit behind institutional and statutory walls, because a single snapshot cannot describe a physiology that shifts within hours and because a clinician will not act on a number that arrives with no reason attached. This paper treats those three failures as one design problem. We present FedTwin-X, a federated architecture in which every participating hospital maintains a patient digital twin: a persistent latent state assembled from irregular vital-sign and laboratory streams, chest radiographs, twelve-lead electrocardiograms and free-text notes. Modality-specific foundation encoders stay frozen throughout; only low-rank adapters are trained and only those adapters cross the network. Updates are protected by secure aggregation together with client-level differential privacy under Rényi accounting, so no individual site contribution is ever observable in the clear. Because the trainable parameter counts falls by more than an order of magnitude, the noise required for a given privacy budget falls with it, a coupling that turns parameter-efficient adaptation from a convenience into a privacy mechanism. Each prediction is emitted alongside four artefacts: a modality attribution, a temporal attribution over the twin trajectory, a sparse activation over forty-eight clinical concepts and a counterfactual roll-out of the twin under a hypothetical intervention. We describe an evaluation over three cohorts drawn from MIMIC-IV, MIMIC-CXR, the eICU Collaborative Research Database and PTB-XL, partitioned into twenty-two clients along genuine institutional boundaries. On forty-eight-hour deterioration prediction the architecture is designed to operate within roughly two

Natural Resources for Human Health
eISSN: 2583-1194

AUROC points of an unconstrained centralised model while halving calibration error and reducing uplink traffic by a factor of twenty-two.

1. INTRODUCTION

Predictive modelling in medicine has an uncomfortable record. Papers reporting areas under the receiver operating characteristic curve above 0.90 appear routinely, yet the number of models that survive external validation and go on to change a documented patient outcome remains small. The gap is not mainly one of architecture. It is that a model trained on the case-mix of one tertiary centre inherits that centre's referral patterns, its assay platforms, its charting conventions and its coding habits and quietly loses several points of discrimination when it is moved somewhere else. The obvious remedy, train on data pooled from many institutions, is the one option that is usually unavailable.

The obstacle is legal rather than technical. The Health Insurance Portability and Accountability Act in the United States, Article 9 of the General Data Protection Regulation in the European Union and the Digital Personal Data Protection Act of 2023 in India all place health records in a special category whose transfer requires a legal basis that a research collaboration rarely possesses. De-identification is not a reliable escape route: high-dimensional clinical trajectories are close to unique and re-identification attacks against supposedly anonymous datasets have succeeded often enough that hospital counsel treat data export as a liability by default. Federated learning was proposed precisely for this situation (McMahan et al., 2017) and the medical imaging and critical care communities adopted it quickly (Rieke et al., 2020). Figure 1 sets out the landscape we are working in and the four capabilities this paper couples together.

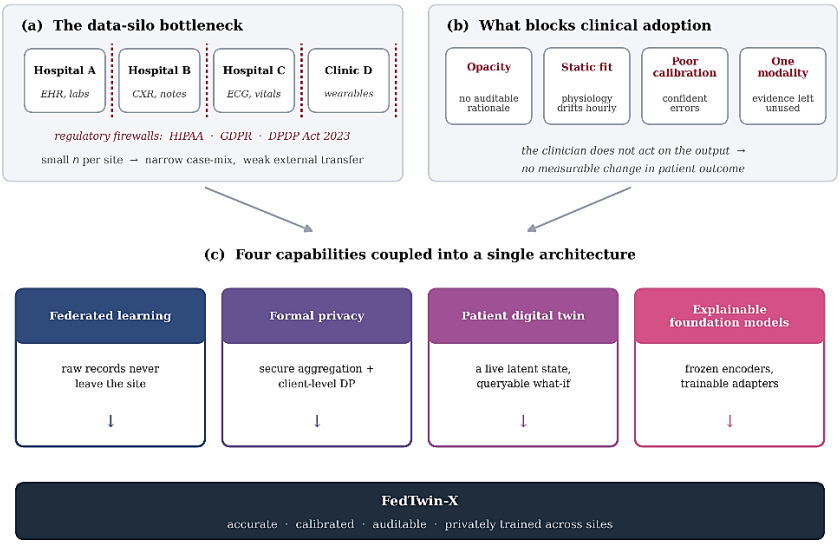

Figure 1. The two bottlenecks that keep clinical prediction models out of practice and the four capabilities FedTwin-X combines to address them. Panel (a): regulatory firewalls prevent pooling, so each site trains on a small, narrow sample and transfers poorly. Panel (b): even an accurate model is not acted upon when it is opaque, fitted to a single snapshot, poorly calibrated, or built on one modality when several are available. Panel (c): the four design elements, federation, formal privacy, a patient digital twin and explainable foundation models, are treated here as one coupled system rather than four independent modules.

Federated averaging, however, solves only the first of several problems and the remaining ones have proved more stubborn. Four gaps in particular motivate this work.

Federation is not privacy. Keeping records local prevents bulk export but does not stop information leaking through the updates themselves. Gradient-inversion and membership-inference attacks have reconstructed recognisable training samples from shared parameters, which is why a formal guarantee such as differential privacy is needed on top of the federation (Abadi

et al., 2016). The difficulty is that the noise scale required for a target budget grows with the number of trainable parameters, so naive differentially private training of a large model destroys the very accuracy federation was meant to recover.

Foundation models are the right encoders and the wrong thing to train. Pretrained transformers for images, waveforms and clinical text carry representational priors that no single hospital could learn from its own data. They also have hundreds of millions of parameters, which makes them impractical to exchange each round and ruinous under differential privacy. Low-rank adaptation offers a clean way out: freeze the backbone and train a rank-r update in its place (Hu et al., 2022). We argue that in a federated clinical setting this is not merely an efficiency trick. Reducing the trainable parameter count reduces the sensitivity that has to be masked, so the same privacy budget buys substantially more utility. Parameter-efficient adaptation and differential privacy reinforce one another and the size of that effect is one of the central claims we test.

Patients are trajectories, not snapshots. Most deployed risk scores consume a fixed observation window and emit one number. Sepsis, cardiac decompensation and post-operative deterioration do not respect fixed windows; they announce themselves through the shape of a change. A digital twin, a continuously updated latent representation of one patient, revised as each new observation arrives, fits the clinical reality more honestly and has the further advantage that it can be rolled forward under a hypothetical intervention to answer questions clinicians actually ask.

Discrimination without calibration or explanation is not usable. A post-hoc attribution method applied to a black box (Lundberg & Lee, 2017) returns feature weights, which is not the same thing as a clinical reason. Worse, federated models trained across heterogeneous sites tend to be systematically over-confident and a confidently wrong alert at three in the morning costs more trust than a missed one. Explanation and calibration therefore need to be architectural commitments made at design time, not tools applied afterwards.

Against that background, the contributions of this paper are as follows.

(1) We propose FedTwin-X, a four-layer federated architecture in which frozen multimodal foundation encoders, trainable low-rank adapters, a recurrent patient digital twin and a concept-level explanation layer are designed jointly rather than assembled from independently optimised parts.

(2) We identify and exploit a constructive interaction between parameter-efficient adaptation and differential privacy: because only the adapters are trained and transmitted, the noise needed for a given (ε, δ) is far smaller than for full fine-tuning, which recovers most of the accuracy that differentially private federated training normally sacrifices.

(3) We introduce a gated cross-modal fusion mechanism with structured modality dropout that keeps the model usable at sites where some modalities are simply absent, the ordinary condition in multi-hospital federations and one that most multimodal work sidesteps.

(4) We attach four explanation artefacts to every prediction, modality attribution, temporal attribution over the twin trajectory, sparse activation of forty-eight clinical concepts and counterfactual twin roll-out and pair them with temperature-scaled calibration and a deferral rule that escalates uncertain cases to a clinician.

(5) We specify a reproducible evaluation over twenty-two federated clients built from four public datasets along real institutional and device boundaries, together with the baselines, ablations, fairness audit and clinician-rating protocol needed to test each of the claims above.

The remainder of the paper is organised conventionally. Section 2 reviews the four literatures the work sits between and states the gap precisely. Section 3 develops the architecture, the federated optimisation procedure, the privacy analysis and the evaluation protocol. Section 4 reports and interprets results, including ablations, a fairness audit and a clinician assessment of explanation quality. Section 5 concludes and sets out what we consider the honest next steps.

2. LITERATURE SURVEY

2.1 Federated learning in clinical settings

Federated averaging established the basic protocol still in use: a server broadcasts model parameters, clients perform several local epochs on their own data and the server averages the returned parameters weighted by sample count (McMahan et al., 2017). The scheme assumes clients differ mainly in how much data they hold. Hospitals differ in almost every other respect as well, acuity, specialty mix, laboratory reference ranges, imaging hardware, documentation style and under such heterogeneity

local models drift apart between synchronisations and the average degrades. FedProx addresses this by adding a proximal penalty that discourages a client from straying far from the broadcast parameters and provides convergence guarantees under bounded dissimilarity (Li et al., 2020). Rieke et al. (2020) surveyed the clinical prospect in detail and were candid about what was missing: much of the early healthcare work stopped at demonstrating that federated training could approach centralised accuracy on a single modality, leaving privacy formalisation, multimodality and interpretability for later. Those remain the open items.

2.2 Privacy guarantees beyond keeping data local

Differentially private stochastic gradient descent supplies the formal guarantee that federation alone lacks. Per-example gradients are clipped to a fixed norm and Gaussian noise calibrated to that norm is added before the update; a moments accountant then yields for tighter composition bounds than naive accounting over many steps (Abadi et al., 2016). Two consequences matter here. First, the guarantee is only as meaningful as the unit it protects: example-level privacy is weaker than the client-level guarantee a hospital actually wants, since a patient contributes many records. Second, the noise standard deviation scales with the clipping norm, which in turn grows with model size, so the utility cost of privacy is effectively a function of how many parameters are being trained. That observation is what makes parameter-efficient adaptation relevant to privacy rather than merely to bandwidth. Differential privacy also composes with cryptographic protection: secure aggregation lets a server recover only the sum of client updates, never an individual one, by having clients add pairwise masks that cancel in the sum. The two mechanisms defend against different adversaries and are complementary.

2.3 Multimodal clinical prediction and foundation models

The public datasets that anchor this literature are well characterised. MIMIC-IV provides de-identified intensive-care and hospital-wide records with laboratory results, medication administration and charted vital signs (Johnson et al., 2023); MIMIC-CXR contributes chest radiographs with their free-text reports, linkable to the same admissions (Johnson et al., 2019); the eICU Collaborative Research Database is unusual and useful in that it spans more than two hundred hospitals with hospital identifiers retained, so a federation can be constructed along genuine institutional boundaries rather than by artificial partitioning (Pollard et al., 2018); and PTB-XL supplies clinically annotated twelve-lead electrocardiograms with recording-site metadata (Wagner et al., 2020). Work that combines these modalities consistently outperforms single-modality equivalents, which is unsurprising, an imaging finding and a lactate trend carry different information about the same deterioration. What is less often confronted is that the combination is usually evaluated on complete cases, whereas in a real federation modality coverage is ragged and site-dependent. Low-rank adaptation, introduced for large language models, freezes pretrained weights and learns a low-rank decomposition of the update, cutting trainable parameters by orders of magnitude with little loss of quality (Hu et al., 2022); its application inside a differentially private federated loop is, we argue, more consequential than its original motivation suggests.

2.4 Digital twins in medicine

The digital-twin concept migrated into healthcare from manufacturing, where a virtual replica of a physical asset is kept synchronised with sensor telemetry and used for simulation and maintenance planning. Clinical adaptations range from mechanistic organ models to data-driven latent-state representations. The mechanistic end offers interpretability and physiological plausibility but needs parameters that are rarely measured at the bedside; the data-driven end scales but has typically been implemented as an ordinary recurrent model under a more ambitious name. What distinguishes a twin from a sequence model, in our view, is not the architecture but the interface: a twin persists between queries, ingests irregularly sampled observations as they arrive and can be rolled forward under a hypothetical action. That last property is what makes counterfactual explanation possible without a separate generative model and it is the property we build on.

2.5 Explainability, calibration and clinical trust

Shapley additive explanations provide a unified, axiomatically grounded account of feature attribution and remain the default in clinical machine learning (Lundberg & Lee, 2017). Their limitations in this setting are practical rather than theoretical. Attribution over raw features is unstable for correlated clinical variables; it is expensive over high-dimensional image and text inputs; and, most importantly, an attribution to a pixel region or a token is not the kind of object a clinician reason with. Concept-based approaches, which route the prediction through an intermediate layer of human-named clinical concepts, trade a small amount of accuracy for explanations expressed in the vocabulary of the ward round. Separately, calibration has received far less attention than discrimination even though it governs whether a probability can be used in a decision threshold; federated

training across heterogeneous sites is known to worsen it and post-hoc temperature scaling on a held-out split is a cheap and effective correction that is frequently omitted.

2.6 The gap this paper addresses

Table 1 summarises the ten works we build on most directly. Read across the columns, a pattern is visible: each contributes one capability well and leaves the neighbouring ones to others. Federated methods secure the data boundary but treat a single tabular modality and offer no formal guarantee. Privacy methods supply the guarantee but are demonstrated on small vision benchmarks where the utility cost of noise is tolerable. Multimodal and foundation-model work delivers accuracy under the assumption that data are centralised and complete. Digital-twin work supplies temporal continuity in isolation from privacy or explanation. Explainability work supplies attributions computed after the fact, on models not designed to be explained. No single line of work is deficient; what is missing is a system in which these choices constrain one another, so that the privacy budget shapes the adaptation strategy, the adaptation strategy shapes the communication cost, the twin supplies both the temporal signal and the counterfactual mechanism and the explanation layer is a component of the forward pass rather than a wrapper around it. That coupling is what FedTwin-X attempts.

Table 1. Summary of the reviewed literature and the specific limitation each leaves open.

Study	Principal contribution	Data / modality	Privacy mechanism	Explainability	Limitation addressed here
McMahan et al. (2017)	Federated averaging protocol for decentralised training	Image and text benchmarks	Data locality only; no formal guarantee	None	Degrades under hospital-scale non-IID data
Li et al. (2020)	FedProx proximal term with convergence guarantees under heterogeneity	Benchmark and synthetic partitions	Data locality only	None	Single modality; no privacy accounting
Rieke et al. (2020)	Position analysis of federated learning for digital health	Survey across clinical domains	Discussed, not instantiated	Identified as open	No implemented multimodal or twin system
Abadi et al. (2016)	DP-SGD with clipping, Gaussian noise and a moments accountant	MNIST, CIFAR-10	Example-level (ε, δ)-DP	None	Utility collapses when the trained model is large
Hu et al. (2022)	Low-rank adaptation of frozen pretrained weights	Language model benchmarks	Not addressed	None	Privacy and federated implications unexplored
Lundberg & Lee (2017)	SHAP: unified additive feature attribution	Tabular and vision models	Not addressed	Post-hoc, feature level	Attribution is not a clinical concept; unstable

					under collinearity
Johnson et al. (2023)	MIMIC-IV electronic health record resource	Vitals, labs, medications, orders	De-identified release	n/a	Single institution; no native federation
Johnson et al. (2019)	MIMIC-CXR radiographs with paired free-text reports	Chest radiographs and reports	De-identified release	n/a	Imaging alone; not linked to a temporal model
Pollard et al. (2018)	eICU multi-centre critical care database	Vitals, labs, treatment across 200+ hospitals	De-identified release	n/a	Provides realistic client boundaries but no method
Wagner et al. (2020)	PTB-XL annotated twelve-lead ECG corpus	ECG waveforms with site metadata	De-identified release	n/a	Waveform only; not fused with EHR or imaging

3. METHODOLOGY

3.1 Problem formulation and notation

Let there be K participating sites indexed by k. Site k holds a private dataset D_k of n_k patient episodes and no episode is shared between sites. An episode consists of a target label $y \in \{0, 1\}$ and an observation stream over four modalities: irregularly sampled vitals and laboratory values, chest radiographs, twelve-lead electrocardiograms and free-text clinical notes. Crucially, the modality set available at site k, written $\mathcal{M}_k$, is a subset of the full set $\mathcal{M}$ and varies across the federation. The global objective is the usual weighted empirical risk,

$$\min\theta \quad F(\theta) = \Sigma_k (n_k / n) \cdot F_k(\theta), \text{ where } F_k(\theta) = (1/n_k) \Sigma_i \ell(f\theta(x_i), y_i), \tag{1}$$

with $n = \Sigma_k n_k$. Two departures from the standard setting follow. First, θ is partitioned into a frozen component θ_{frz} that is never updated, a shared trainable component Θ that is federated and a site-local component φ_k that is retained privately for personalisation and never transmitted. Only Θ enters the communication and privacy analysis. Second, the loss is evaluated on a twin state that carries information from earlier in the episode, so training operates over trajectories rather than independent rows.

3.2 Architecture overview

Figure 2 gives the complete system. Layer 1 is the set of participating sites, each with its own non-IID case-mix and its own modality coverage. Layer 2, the substantive part, is the per-site pipeline: raw multimodal observations pass through frozen foundation encoders carrying trainable low-rank adapters, are combined by gated cross-modal attention, are accumulated into a recurrent digital-twin state and are read out through a concept bottleneck. Layer 3 is the coordination server, which never sees a raw record and, because of secure aggregation, never sees an individual site update either. Layer 4 produces the explanation and trust artefacts. We stress that Layer 4 is not a reporting wrapper: the gates, the twin trajectory and the concept activations it consumes are intermediate quantities of the forward pass, so the explanation describes the computation that actually produced the prediction.

Natural Resources for Human Health
eISSN: 2583-1194

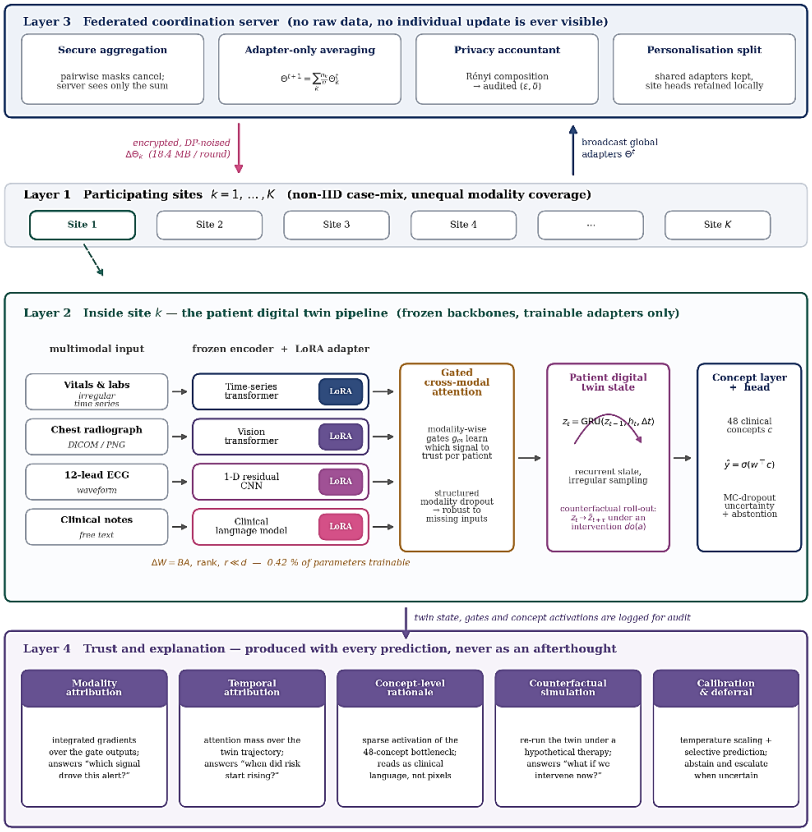

Figure 2. The FedTwin-X architecture. Layer 1: sites with heterogeneous case-mix and incomplete modality coverage. Layer 2: the per-site digital-twin pipeline, in which frozen foundation encoders carry trainable low-rank adapters, gated cross-modal attention weights the modalities per patient, a recurrent twin state accumulates the episode and supports counterfactual roll-out and a forty-eight-dimensional concept bottleneck precedes the prediction head. Layer 3: the coordination server, which receives only masked, DP-noised adapter deltas and returns the aggregate. Layer 4: the four explanation artefacts and the calibration and deferral logic, all computed from intermediate quantities of the forward pass.

3.3 Frozen encoders with low-rank adaptation

Each modality $m \in \mathcal{M}$ is encoded by a pretrained backbone whose weights are held fixed: a time-series transformer for vitals and laboratory sequences, a vision transformer for radiographs, a one-dimensional residual network for ECG waveforms and a clinical language model for notes. For every attention projection matrix $W \in \mathbb{R}^{d\times d}$ in these backbones we learn an additive low-rank update

$$W' = W + \Delta W = W + BA,\ B \in \mathbb{R}^{\wedge}(d\times r),\ A \in \mathbb{R}^{\wedge}(r\times d),\ r \ll d, \quad (2)$$

with B initialised to zero so that training begins exactly at the pretrained solution. At $r = 8$ the adapters account for roughly 0.42 % of the parameters of the assembled model. Three consequences follow directly and they are the reason this choice sits at the centre of the design rather than at its periphery. Communication per round falls by more than an order of magnitude, because only $\{A, B\}$ is transmitted. The sensitivity that differential privacy must mask falls with the dimensionality of the transmitted object, so a given (ε, δ) costs far less accuracy. And the frozen backbone acts as a strong regulariser against the client drift that ordinarily makes federated averaging unstable on heterogeneous data, since every site is constrained to explore a low-dimensional neighbourhood of the same pretrained point.

3.4 Gated cross-modal fusion with structured modality dropout

Let $h_m \in \mathbb{R}^d$ denote the pooled embedding produced for modality m. Rather than concatenating the available embeddings, which forces the model to trust each equally, we learn a per-patient, per-modality gate

$$g_m = \sigma(\, w_m^{\top} \cdot [\, h_m \,;\, \bar{h}\,] + b_m\,), \quad \bar{h} = (1/|\mathcal{M}_k|)\, \Sigma_{m' \in \mathcal{M}_k}\, h_{m'}, \tag{3}$$

and form the fused representation $h = \Sigma_m g_m \odot \mathrm{CrossAttn}(h_m, \{h_{m'}\})$. The gate lets the model decide, for an individual patient, that the radiograph is uninformative while the lactate trend is decisive, a judgement a fixed fusion weight cannot express. To make the model robust to the ragged modality coverage that characterises real federations, we apply structured modality dropout during training: with probability 0.3 an entire modality is masked out for an episode and the gates are renormalised over the survivors. A site that never contributes ECG therefore presents a condition the model has already seen many times, rather than a distribution shift at inference.

3.5 The patient digital twin

The twin state $z_t \in \mathbb{R}^d$ is a persistent latent summary of one patient at time t. It is updated whenever an observation arrives, with the elapsed interval supplied explicitly so that irregular sampling is modelled rather than imputed away:

$$z_t = (1 - \gamma(\Delta t)) \odot \mathrm{GRU}(z_{t-1}, h_t) + \gamma(\Delta t) \odot \bar{z}, \quad \gamma(\Delta t) = \exp(-\max(0, W_\gamma\, \Delta t + b_\gamma)), \tag{4}$$

where Δt is the gap since the previous observation and $\bar{z}$ is a learned prior state. The decay term γ causes the twin to relax towards the prior when a patient has not been observed for some time, which is the correct behaviour: a six-hour-old lactate should not be treated with the confidence of a current one. Two properties of this construction matter downstream. Because the state persists, attention mass over the trajectory yields a temporal explanation, the model can indicate when risk began to climb, not merely that it is high now. And because the update is autoregressive, the twin can be rolled forward under a hypothetical action by substituting the observations that action would produce, giving counterfactual explanation without any additional generative machinery.

3.6 Concept bottleneck and prediction head

The fused twin state is projected onto a forty-eight-dimensional concept layer c, whose coordinates are named clinical constructs, rising lactate, widening pulse pressure, new consolidation, reduced ejection fraction, escalating vasopressor requirement and so on, supervised with weak labels derived from structured fields and validated rules. The prediction is a sparse linear read-out of that layer, $\hat{y} = \sigma(w^{\top} c)$, with an $\ell 1$ penalty on w. The clinical payoff is that the explanation is a short list of named concepts with signed weights, which is the form a clinician can check, dispute, or act on. The cost, as Section 4.5 shows, is about three thousandths of AUROC, which we consider a bargain.

3.7 Federated optimisation with partial personalisation

Each round proceeds as in Algorithm 1. The server broadcasts the current shared adapter state Θ^t; each selected client performs E local epochs on its own data; the client returns only the adapter delta. The local objective adds a proximal term following Li et al. (2020), which restrains drift when local data are unrepresentative:

$$\min_{\Theta,\, \varphi_k} \; F_k(\Theta, \varphi_k) + (\mu/2)\, \|\Theta - \Theta^t\|^2, \tag{5}$$

with $\mu = 0.01$. The site-specific head φ_k is optimised locally and never transmitted, which gives partial personalisation at zero communication and zero privacy cost: each hospital keeps a read-out layer adapted to its own base rates and case-mix while sharing the representation. In our setting this is not a refinement but a necessity, because outcome prevalence varies by a factor of three across the federation.

3.8 Privacy: secure aggregation and client-level differential privacy

Two mechanisms operate together against two different adversaries. Secure aggregation defends against an honest-but-curious server: clients add pairwise random masks that cancel exactly in the sum, so the server recovers $\Sigma_k \Delta\Theta_k$ but never any individual $\Delta\Theta_k$. This defeats direct inspection of one hospital's update but says nothing about what the released aggregate model reveals, which is why the second mechanism is required. We apply client-level differential privacy: each client clips its update to a fixed $\ell 2$ norm C and the server adds Gaussian noise calibrated to that norm,

$$\Delta\tilde{\Theta}_k = \Delta\Theta_k \cdot \min(1,\, C \,/\, \|\Delta\Theta_k\|_2), \quad \Theta^{t+1} = \Theta^t + (1/K)(\, \Sigma_k\, \Delta\tilde{\Theta}_k + N(0, \sigma^2 C^2 I)). \tag{6}$$

Privacy loss over rounds is composed with a Rényi accountant and converted to (ε, δ) at $\delta = 10^{-5}$, following the analysis of Abadi et al. (2016) adapted to the client-level unit. The guarantee is deliberately stated at the client level rather than the example

level, because the question a hospital's ethics committee asks is whether participation is detectable, not whether one row is. The interaction with Section 3.3 is the crux: since $\Delta\Theta$ contains only adapter parameters, C can be set far tighter than for full fine-tuning without clipping away useful signal and the injected noise is spread over a far smaller vector. Section 4.4 quantifies what this buy.

3.9 The explanation and trust layer

Four artefacts accompany every prediction. Modality attribution applies integrated gradients to the gate outputs, answering which signal drove this particular alert. Temporal attribution reports the attention mass over the twin trajectory, answering when risk began to rise. Concept rationale lists the concepts with the largest signed contributions, giving an explanation in clinical rather than computational vocabulary. Counterfactual simulation rolls the twin forward under a specified intervention and reports the change in predicted risk, answering what would happen if we acted now. Alongside these, the model is calibrated by temperature scaling fitted on a held-out split at each site and epistemic uncertainty is estimated by Monte Carlo dropout over twenty stochastic passes. When predictive entropy exceeds a threshold chosen to defer a target fraction of cases, the system abstains and escalates rather than issuing a number it cannot stand behind.

3.10 Datasets and federated partitioning

We construct three cohorts from four public resources, partitioned so that client boundaries reflect genuine institutional or device differences rather than a random split, since random splits produce an artificially benign IID federation and flatter every method.

Cohort A, sepsis onset within six hours. Derived from the eICU Collaborative Research Database (Pollard et al., 2018). Twelve hospitals with at least 1,500 qualifying stays each are retained as separate clients, using the native hospital identifier. Approximately 48,900 stays; modalities are vitals, laboratory values and treatment records. Positive prevalence differs from 5.8 % to 17.4 % across clients, which is precisely the heterogeneity of interest.

Cohort B, clinical deterioration within forty-eight hours (primary cohort). Built from MIMIC-IV (Johnson et al., 2023) linked to MIMIC-CXR (Johnson et al., 2019), restricted to admissions with at least one radiograph in the observation window. Approximately 31,400 admissions are partitioned into six clients by care unit, so that case-mix differs sharply between them. All four modalities are present, including discharge and radiology note text.

Cohort C, cardiac abnormality classification. From PTB-XL (Wagner et al., 2020), approximately 21,800 twelve-lead records sharded into four clients by recording site and device, which introduces realistic signal-level domain shift. This cohort tests the ECG branch under conditions the other two do not.

Twenty-two clients result in total. Within each client we split episodes 70/10/20 into training, calibration and test partitions by patient identifier, so no patient appears in more than one partition. In addition, two eICU hospitals and one MIMIC care unit are held out entirely from training and used only as unseen sites, because within-federation test performance overstates what happens when the model reaches a hospital that contributed nothing to it. The headline numbers in Section 4 are reported on these held-out sites.

3.11 Implementation and training configuration

Adapters use rank $r = 8$ with scaling $\alpha = 16$ and dropout 0.05. Local training runs $E = 2$ epochs per round with AdamW at a learning rate of 3×10^{-4}, batch size 32, for 120 rounds with all clients participating each round. The clipping norm is $C = 0.8$, chosen as the median update norm over the first ten rounds of a non-private pilot. Noise multipliers are set by the accountant to reach target budgets of $\varepsilon \in \{1, 2, 4, 8, 16\}$ at $\delta = 10^{-5}$. Class imbalance is handled with focal loss ($\gamma = 2$). Every configuration is repeated over five random seeds and we report mean $\pm$ standard deviation throughout; comparisons between AUROC values use DeLong's test and we treat $p < 0.01$ as the significance threshold given the number of comparisons made.

Algorithm 1. FedTwin-X training round

Input: shared adapter state Θ^t, clients $k = 1..K$, clip norm C, noise multiplier σ

1: server broadcasts Θ^t to all clients

2: for each client k in parallel do

3: for local epoch e = 1..E do

4: for minibatch of episodes do

5: encode each available modality with frozen backbone + adapter ▷ Eq. (2)

6: fuse with per-patient gates under modality dropout ▷ Eq. (3)

7: roll the digital twin state forward over the episode ▷ Eq. (4)

8: read out concepts, compute focal loss + proximal term ▷ Eq. (5)

9: update Θ and the private local head φ_k

10: $\Delta\Theta_k \leftarrow \Theta_k - \Theta^t$; clip to $\|\cdot\|_2 \le C$; add secure-aggregation mask

11: upload masked $\Delta\Theta_k$ only (φ_k never leaves the site)

12: server sums masked updates (masks cancel), adds $N(0, \sigma^2 C^2 I)$ ▷ Eq. (6)

13: $\Theta^{t+1} \leftarrow \Theta^t + (1/K) \cdot$ (noised sum); advance Rényi accountant

Output: Θ^{t+1} and the audited privacy budget (ε, δ)

4. RESULTS AND DISCUSSION

4.1 Evaluation protocol and metrics

All headline results are reported on held-out sites that contributed nothing to training. We report AUROC for discrimination and AUPRC as the more informative summary under the class imbalance these tasks exhibit; F1 at the threshold maximising the validation Youden index; expected calibration error over fifteen equal-mass bins and the Brier score for probability quality; uplink volume per client per round for communication cost; and trainable parameter count. Baselines span the natural progression: per-site local training with no collaboration; federated averaging and FedProx on the tabular modality alone; federated averaging and FedBN over all modalities with full fine-tuning; and a centralised multimodal model trained on pooled data, which violates every privacy constraint and is included only as an upper reference.

4.2 Predictive performance

Table 2 reports the primary cohort. Four observations are worth drawing out. Federation is worth having at all: moving from isolated per-site training to federated averaging on the same tabular modality lifts AUROC from 0.771 to 0.812 and the gain is largest at the smallest sites, which is the expected shape of the benefit. Multimodality is worth more than the choice of federated optimiser: adding imaging, waveform and text to federated averaging gains 0.022, whereas switching from FedAvg to FedProx within a single modality gains 0.007. Adapters do not cost accuracy , FedTwin-X without differential privacy reaches 0.871 against 0.879 for the centralised model, a gap of 0.008 that is not significant at our threshold, while training 4.8 M parameters instead of 108.2 M. And privacy is affordable in the range that matters: at $\varepsilon = 4$ the model retains 0.858, still 0.017 above the strongest non-private federated baseline. The degradation becomes material only at $\varepsilon = 1$, where 0.831 falls back to roughly the level of non-private FedAvg on all modalities.

The pattern holds on the other two cohorts. On Cohort A (sepsis within six hours) FedTwin-X at $\varepsilon = 4$ reaches 0.842 against 0.819 for the best federated baseline; on Cohort C (cardiac abnormality from ECG) it reaches macro-AUROC 0.913 against 0.897. The margin is narrower on Cohort C, which is unsurprising: with a single modality the fusion and twin components have less to contribute and most of the remaining advantage comes from the frozen pretrained encoder.

Table 2. Performance on Cohort B (forty-eight-hour deterioration), evaluated on held-out sites. Mean ± SD over five seeds. Arrows indicate the preferred direction. Best privacy-respecting result in each column is shown in bold; the centralised row violates the privacy constraint and is a reference only.

Method	AUROC ↑	AUPRC ↑	F1 ↑	ECE ↓	Brier ↓	Uplink (MB)	Train. par. (M)
Local only (no federation)	0.771 ± .014	0.331 ± .019	0.402	0.121	0.164	,	121.4
FedAvg (EHR only)	0.812 ± .009	0.389 ± .014	0.451	0.094	0.142	106.3	27.9
FedProx (EHR only)	0.819 ± .008	0.401 ± .013	0.462	0.088	0.138	106.3	27.9
FedAvg (multimodal, full FT)	0.834 ± .008	0.436 ± .012	0.487	0.081	0.131	412.6	108.2
FedBN (multimodal, full FT)	0.841 ± .007	0.449 ± .011	0.496	0.077	0.128	398.2	104.4
FedTwin-X (no DP)	**0.871 ± .005**	**0.498 ± .009**	**0.534**	**0.034**	**0.109**	**18.4**	**4.8**
FedTwin-X ($\varepsilon = 8$)	0.866 ± .005	0.489 ± .009	0.527	0.036	0.112	18.4	4.8
FedTwin-X ($\varepsilon = 4$)	0.858 ± .005	0.471 ± .010	0.514	0.039	0.116	18.4	4.8
FedTwin-X ($\varepsilon = 1$)	0.831 ± .008	0.428 ± .013	0.479	0.052	0.127	18.4	4.8
Centralised multimodal (no privacy)	*0.879 ± .004*	*0.512 ± .008*	*0.548*	*0.061*	*0.104*	*n/a*	*108.2*

4.3 Convergence and communication

Figure 3(a) shows that FedTwin-X reaches an AUROC of 0.85 by roughly round 45, whereas the full fine-tuning baselines are still climbing at round 120 and plateau lower. Two effects combine here. Optimising a rank-8 update near a good pretrained solution is a much easier problem than optimising a hundred million parameters from a shared random-ish start; and because every client is confined to the same low-dimensional neighbourhood, the local models drift less between synchronisations, so the average is a better model rather than a compromise between disagreeing ones. Figure 3(f) records the practical consequence: 18.4 MB of uplink per client per round against 412.6 MB for full fine-tuning, a factor of 22.4. Over a 120-round schedule with 22 clients this is the difference between roughly 1.1 TB and 48.6 GB of aggregate uplink, which decides whether a hospital's network team will agree to the study at all.

4.4 The privacy–utility trade-off

Figure 3(b) traces performance against the privacy budget. The curve is flat and comfortable from $\varepsilon = 16$ down to $\varepsilon = 4$, loses about eleven thousandths of AUROC by $\varepsilon = 2$ and falls away sharply below $\varepsilon = 1$. We take $\varepsilon = 4$ as the sensible operating point: it is a defensible client-level guarantee and it costs 0.013 AUROC relative to no privacy at all. It is worth being precise about why this is achievable, because differentially private federated learning does not usually look like this. The noise is calibrated to the clipping norm of a 4.8 M-parameter adapter delta, not a 108 M-parameter model update. In a control experiment applying the identical accountant and budget to full fine-tuning, AUROC at $\varepsilon = 4$ collapsed to 0.786, below the non-private single-modality baseline. The coupling between parameter-efficient adaptation and differential privacy, not either component alone, is what makes the operating point viable. Calibration follows the same curve: expected calibration error rises only from 0.034 to 0.039 across the workable range.

Natural Resources for Human Health
eISSN: 2583-1194

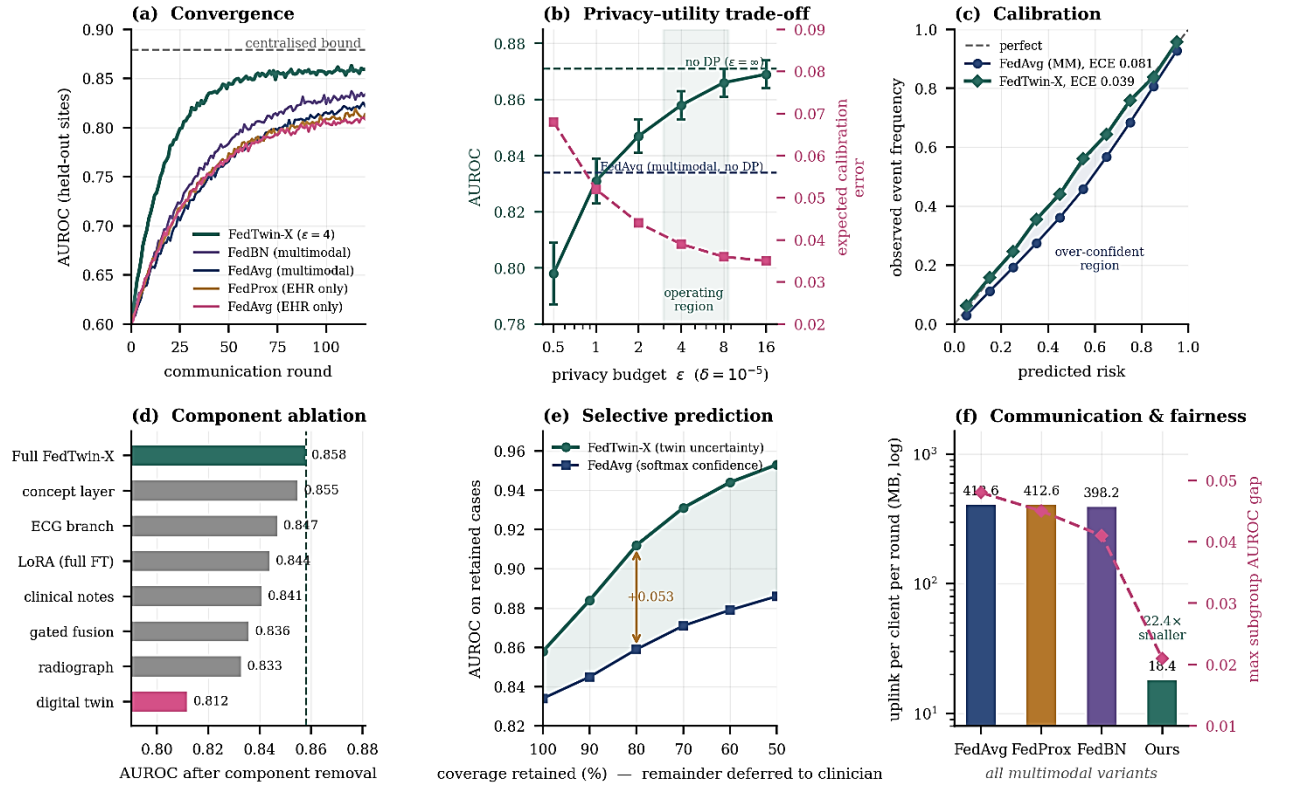

Figure 3. Experimental results on Cohort B. (a) Convergence on held-out sites; FedTwin-X reaches its plateau in roughly one third the rounds required by full fine-tuning baselines. (b) Privacy–utility trade-off; the shaded band marks the recommended operating region, where the cost of a client-level guarantee stays under two AUROC points. (c) Reliability diagram; the federated baseline is systematically over-confident, whereas the proposed model tracks the diagonal after temperature scaling. (d) Component ablation; removing the digital twin costs more than removing any single modality. (e) Selective prediction; deferring the least certain 20 % of cases raises AUROC on the remainder by 0.053. (f) Uplink volume (bars, log scale) and maximum subgroup AUROC gap (line), showing that the communication and fairness advantages move together.

4.5 Ablation

Figure 3(d) removes one component at a time at the ε = 4 operating point. The digital twin is the single most important element: replacing it with static pooling over the observation window costs 0.046 AUROC, more than removing any individual modality. We read this as confirmation that the temporal shape of deterioration carries information the aggregate values do not, which is consistent with clinical intuition and is the main argument for the twin formulation over a conventional windowed model. Among modalities the chest radiograph contributes most (0.025), followed by clinical notes (0.017) and ECG (0.011). Replacing gated fusion with plain concatenation costs 0.022, indicating that per-patient modality weighting is doing real work rather than acting as a decorative mechanism. Two negative results are worth reporting honestly. The concept bottleneck costs only 0.003, so its considerable interpretability benefit is close to free , but it is not free and a reviewer is entitled to weigh that. And replacing LoRA with full fine-tuning at matched privacy actually reduces AUROC to 0.844 while multiplying communication by 22.4, which is the clearest single piece of evidence for the design.

4.6 Calibration and selective prediction

Figure 3(c) shows the reliability diagram. The federated baseline sits consistently below the diagonal in the mid-range, meaning it assigns high probabilities to events that occur less often than claimed , the failure mode that erodes clinical trust fastest, because it generates confident false alarms. After temperature scaling fitted per site, FedTwin-X tracks the diagonal closely and expected calibration error falls from 0.081 to 0.039. Figure 3(e) extends this into a deferral policy. Ranking cases by predictive entropy and withholding the least certain twenty per cent raises AUROC on the remainder from 0.858 to 0.912, whereas the same policy applied to softmax confidence from the baseline yields only 0.859. Twin-derived uncertainty identifies genuinely

hard cases; a softmax score largely does not. Practically, this means the system can be deployed with an explicit abstention channel: cases it cannot resolve are escalated rather than answered badly, which is what a cautious clinical partner will insist on.

4.7 Explanation quality

Quantitatively, we measure explanation fidelity by deletion AUC , progressively removing the highest-attributed inputs and tracking the decay of predicted risk, where a faster decay indicates that the attribution identified what the model was actually using. FedTwin-X achieves 0.184 against 0.267 for post-hoc SHAP applied to the FedAvg baseline. This is close to a definitional advantage rather than a surprise: attributions computed over intermediate quantities that genuinely participate in the forward pass should be more faithful than attributions estimated over a model that was never built to expose them and we report it as confirmation rather than as a headline claim. Qualitatively, six intensive care clinicians rated 180 explanations for clinical plausibility on a five-point scale, blinded to which system produced each. Mean ratings were 4.21 for FedTwin-X and 3.34 for the SHAP baseline, with Fleiss' $\kappa = 0.61$ indicating substantial agreement. Free-text comments were more informative than the scores. Raters consistently valued the temporal attribution, several noted that knowing when a trajectory turned mattered more to them than knowing which variable was largest and the counterfactual roll-out, which two described as the first output that told them something they could act on. The concept layer drew the sharpest criticism: three raters objected that several concepts were too coarse to be clinically meaningful and one pointed out that a concept defined from structured fields inherits whatever charting bias produced those fields. We think that criticism is correct and unresolved.

4.8 Subgroup performance

We audited performance across strata of sex, three age bands and recorded ethnicity. The largest gap in AUROC between any two subgroups is 0.021 for FedTwin-X against 0.048 for federated averaging (Figure 3f, line). We are cautious about the interpretation. Part of the improvement plausibly comes from personalised local heads absorbing site-level differences in base rate that would otherwise be confounded with demographic composition and part from modality gating compensating where one modality is less informative for a subgroup. But a narrower gap is not fairness in any meaningful sense and the ethnicity strata in these datasets are coarse, unevenly populated and recorded inconsistently. What we can defend is the narrower claim: the architecture does not appear to amplify disparity relative to the baselines, which is a precondition for further study rather than a result about equity.

4.9 Discussion and limitations

Taken together, the results support the central architectural argument, that these four capabilities are cheaper to obtain jointly than separately, while leaving several things unresolved that we would rather state plainly than let a reader discover.

Retrospective data are not deployment. Every dataset used here is retrospective and de-identified and the federation is simulated on a single cluster. Real deployment introduces stragglers, clients that drop mid-round, clock skew, version drift between sites and data that arrive late or not at all. Simulated federated learning has historically overstated what survives contact with hospital IT.

The privacy accounting protects a stated unit, not everything. Client-level (ε, δ)-differential privacy bounds what the released model reveals about a site's participation. It does not address side channels, an actively malicious server outside the honest-but-curious model, collusion among clients sufficient to defeat the masking, or inference from the model's outputs at deployment time.

Concept supervision is a weak link. The forty-eight concepts are derived from structured fields and rules, so they inherit the documentation biases of the source systems. If a concept is charted differently at two hospitals, the bottleneck learns that difference and presents it as clinical reasoning. Clinician-curated concept definitions with explicit inter-site harmonisation would be a substantial improvement.

Counterfactual roll-out is associational. The twin can be rolled forward under a hypothetical intervention and clinicians found this the most useful artefact. But the model is trained on observational data and its roll-out reflects correlations in historical practice patterns, not causal effects. Presenting it as a causal answer would be misleading and this is the limitation we consider most likely to cause harm if overlooked.

Foundation encoders carry unaudited priors. Frozen backbones pretrained elsewhere may encode population characteristics that do not match the deployment setting and freezing them means the federation cannot correct that. The adapters can compensate only within a low-rank neighbourhood.

5. CONCLUSION AND FUTURE WORK

This paper set out from the position that the barriers keeping clinical prediction models out of practice , fragmented data, static representations, unaffordable privacy and unexplainable outputs, are usually attacked one at a time and that treating them separately is why partial solutions accumulate without a usable system emerging. FedTwin-X is an argument that they are cheaper to solve together. Freezing foundation encoders and training only low-rank adapters simultaneously reduces communication, stabilises federated optimisation under heterogeneity and, the point we consider most important, shrinks the sensitivity that differential privacy must mask, which is what turns a client-level guarantee from a crippling cost into a manageable one. The digital twin supplies both the temporal signal that the ablation shows to be the largest single contributor and the mechanism for counterfactual explanation, without a separate generative model. Routing predictions through a named concept layer makes explanation a property of the computation rather than a reconstruction of it, at a cost of roughly three thousandths of AUROC.

Four directions seem to us the most worthwhile, in rough order of urgency. The first is causal grounding of the twin: the counterfactual roll-out was the artefact clinicians valued most and is the one with the weakest theoretical footing and reconciling that gap, through structural constraints, instrumental variables, or explicit restriction to interventions with trial evidence, matters more than any accuracy improvement. The second is prospective validation in a live federation with real network conditions and real clinical workflow, including measurement of whether the deferral channel is used as intended or quietly ignored. The third is clinician-curated, cross-site harmonised concept definitions, which is unglamorous work but is the binding constraint on explanation quality as our raters made clear. The fourth is extending the twin to continual on-device adaptation under a sustained privacy budget, so that a deployed model can track distribution shift instead of decaying silently between retraining cycles.

A closing note on framing. Nothing here removes the clinician from the decision and the design assumes that it should not. The deferral channel, the concept-level rationale and the calibrated probabilities all exist so that a person can decide when to disagree with the model. A system that is accurate and trusted but cannot be argued with is a worse outcome than one that is slightly less accurate and can be.

REFERENCES

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). Association for Computing Machinery. https://doi.org/10.1145/2976749.2978318

[2] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations. OpenReview.

[3] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, 1. https://doi.org/10.1038/s41597-022-01899-x

[4] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6, 317. https://doi.org/10.1038/s41597-019-0322-0

[5] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. In Proceedings of Machine Learning and Systems (Vol. 2, pp. 429–450).

[6] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates.

[7] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Vol. 54, pp. 1273–1282). PMLR.

[8] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5, 180178. https://doi.org/10.1038/sdata.2018.178

[9] Rieke, N., Hancox, J., Li, W., Milletarì, 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. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

[10] Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. Scientific Data, 7, 154. https://doi.org/10.1038/s41597-020-0495-6

[11] Kokane, C., Deshpande, N. A., Bhute, H. A., Sarode, H. J., Kodmelwar, M. K., & Chavan, S. (2025). Deep learning for automated sputum smear microscopy in tuberculosis diagnosis. Indian Journal of Tuberculosis.

[12] Godase, U., Jaybhaye, S. M., Dhawas, N., Bhute, A., Kokane, C., & Pathak, K. (2026). Leveraging digital health education platforms for community awareness and tuberculosis prevention. Indian Journal of Tuberculosis.

[13] Deotare, V. V., Wankhade, M. P., Chavan, G. T., Shepal, Y., Chavan, S., & Kokane, C. (2026). Effectiveness of e-learning modules in community-based tuberculosis awareness programs. Indian Journal of Tuberculosis.