Minimax Theory and its Applications
Volume 11 (2026), No. 3s, 564 to 586

Equity-Regularized Federated Multimodal Learning with Missingness-Aware Gating and Subgroup-Conditional Conformal Risk Control for Clinical Prediction

Sarojini Rani[1], Senthil Kumar P[2*]

[1]Department of Computer Science and Engineering, Bharatiya Engineering Science and Technology Innovation University, Gownivaripalli, Gorantla 515231, Andhra Pradesh, India.
Email: saroja.cnd@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Tiruchirappalli Campus, Irungalur 621105, Tamil Nadu, India.
Email: psenthilexcel@gmail.com, ORCID: 0000-0001-5795-9448.
*Corresponding Author

Abstract
A clinical prediction model can look strong on average and still fail the patients who depend on it most. Two failure modes dominate in practice: performance that differs across patient subgroups, and quiet degradation when one data modality is routinely unavailable. Neither shows up when a model is trained for accuracy and judged by average metrics. We present PRISM-Fed (Privacy-preserving Risk Inference via Selective Multimodal fusion, Federated), an end-to-end framework in which equity is built into the federated aggregation rule and certified at the decision threshold, instead of being added as one more penalty in the loss. The framework rests on three linked mechanisms. Missingness-aware gating (MAG) conditions cross-modal attention on an explicit observability mask and a learned mask embedding; attention is renormalized over the modalities actually recorded, the pattern of absence is used as clinical information, and no imputation is needed. EqFedAvg replaces sample-size-proportional averaging with a rule that raises an institution's aggregation weight according to the calibration deficit of its worst-served subgroup, so the global model is pulled toward the sites and populations where it performs worst. We show that this rule keeps the usual federated-averaging convergence rate, up to a bounded weight-dispersion constant. Subgroup-conditional conformal risk control (SC-CRC) fits a separate operating threshold for each protected subgroup from a federated calibration pool and gives a distribution-free, finite-sample guarantee that sensitivity is at least $1 - \alpha$ in every audited subgroup. A drift-triggered loop restores this guarantee by refitting the thresholds only, while the approved model weights stay frozen. In a multi-institution evaluation covering intensive care, general ward and imaging cohorts, PRISM-Fed raised worst-subgroup discrimination much more than average discrimination (worst-subgroup AUROC 0.851 versus 0.751 for federated averaging; overall AUROC 0.887 versus 0.854). It cut the expected calibration error by more than half, lowered the equalized-odds gap from 0.142 to 0.038 and reduced the share of subgroups missing the target sensitivity from 0.42 to 0.03. It also converged in about 35% fewer communication rounds and, after distillation, ran in 96 ms at 0.21 J per inference. These results suggest that fairness in clinical machine learning is better treated as a guarantee that can be audited at deployment than as a regularization problem at training time.

Keywords: Federated Learning, Multimodal Clinical Prediction, Algorithmic Fairness, Conformal Risk Control, Informative Missingness, Calibration Drift, Clinical Decision Support, Internet of Medical Things.

1. Introduction

1.1 Background

Predictive models built from electronic health records, physiological waveforms, diagnostic images and clinical notes now equal or outperform clinicians on a growing number of narrowly defined tasks [1,2]. Health systems produce these data continuously and in large volumes, and the tools for exploiting them, including deep sequence models, convolutional and transformer architectures and ensemble learners, are well established [1]. Translation into routine care has been far less successful. A model that discriminates well in its development cohort often loses accuracy when it is moved to another hospital, another payer mix or simply a later year of

practice [3]. The loss is seldom spread evenly; it falls mostly on populations that were poorly represented when the model was built.

1.2 Problem statement

Most of this translation gap can be traced to three structural features of real clinical deployment. Larger models or deeper fusion architectures do not address any of them.

First, clinical multimodality is ragged. A patient at a rural critical-access hospital may have vital signs and laboratory results but no cross-sectional imaging and no structured genomic panel, whereas a patient at a quaternary referral center usually has all four. Which modalities are available therefore reflects the care pathway, institutional resources, insurance status and referral history, so the data are not missing at random [4]. Standard fusion architectures hide this by filling absent modality embeddings with zeros or means. Doing so adds an artificial signal to the fused representation and, at the same time, discards the information carried by the absence pattern [5]. Since absence tends to go together with disadvantage, the resulting error is a systematic penalty on the least-resourced patients rather than random noise [6]. Second, federated collaboration, the usual answer to the data-sharing restrictions that keep clinical cohorts small and homogeneous, currently optimizes the wrong objective. Federated averaging weights each institution by its sample count [7]. Large, well-equipped academic centers therefore dominate the global model, and the sites whose patients are served worst have the least say in the parameters that are eventually deployed [8,9]. The third feature follows from the first two. A risk score becomes a decision only through a threshold, and nearly every deployed clinical decision support system applies a single global threshold to all patients. When calibration varies by subgroup, as it usually does, one threshold necessarily yields different sensitivities for different groups. A model can satisfy a training-time fairness penalty and still miss twice as many true cases in one subgroup as in another [10].

1.3 Objectives and contributions

The aim of this study was to design, formalize and evaluate a clinical prediction framework in which equitable performance is a structural property of the learning system, not a post hoc correction, and in which the resulting equity claim can be audited by a clinical governance committee. The main contributions are as follows:

- A missingness-aware gating (MAG) mechanism for cross-modal fusion. Attention is conditioned on an explicit observability mask and a learned mask embedding, so attention mass renormalizes over observed modalities only, absent modalities receive a weight of exactly zero instead of an imputed value, and the missingness pattern itself becomes a usable feature. A site-matched modality-dropout curriculum ensures that the masks seen in training cover those met at deployment.
- An equity-regularized federated aggregation rule (EqFedAvg), in which the server-side weight of each client increases with the calibration deficit of its worst-served subgroup. We show that the rule keeps the standard federated-averaging convergence rate, up to a constant set by weight dispersion that is explicitly bounded by clipping.
- A subgroup-conditional conformal risk control (SC-CRC) procedure that derives an operating threshold for each subgroup from a federated calibration pool and provides a finite-sample, distribution-free lower bound on sensitivity in every audited subgroup. It is paired with a drift-triggered recalibration loop that restores the guarantee without touching the model weights.
- An evaluation protocol that reports the quantities deployment actually depends on (worst-subgroup discrimination, subgroup calibration, per-subgroup coverage violation, robustness to systematic modality loss, communication cost and edge energy) alongside conventional average discrimination. Applying this protocol shows that average discrimination alone hides most of the differences between competing methods.

1.4 Organization of the paper

Section 2 reviews related work on multimodal clinical fusion, federated and fair federated optimization, calibration and conformal inference, and edge deployment. Section 3 turns this review into a statement of the research gap and the motivation for the proposed design. Section 4 describes the PRISM-Fed framework together with its formal definitions, algorithms and theoretical guarantees. Section 5 covers the materials, cohorts, baselines and experimental protocol, and Section 6 reports the results. Section 7 discusses what the results mean, Section 8 compares the framework with existing methods and Section 9 sets out practical implications. Section 10 covers limitations and future work, and Section 11 concludes the paper.

2. Literature Review and Related Work

2.1 Multimodal clinical prediction

Fusion of heterogeneous clinical data has moved from hand-crafted concatenation towards learned joint representations. Survey work classifies fusion as early, intermediate or late according to the point at which modalities are combined [11], and clinical studies report steady gains when imaging is combined with record data compared with either source used alone [2]. More recent biomedical systems place imaging, notes, waveforms and omics in a single architecture. Attention-based intermediate fusion [12] has become the preferred design because it lets the contribution of each modality change from one patient to the next instead of being fixed by the architecture. Almost all of these studies, however, used cohorts in which every patient had every modality, or in which incomplete records were simply excluded, and they reported discrimination only at the cohort level.

2.2 Learning with missing modalities

Approaches for severely missing modalities include generative reconstruction of the missing view, alignment in a shared latent space, and distillation from a teacher trained on complete modalities to a student that sees incomplete ones [13,14]. Classical statistics separates data missing completely at random, missing at random and missing not at random, and only the first two permit unbiased imputation [4]. Clinical informatics research has shown repeatedly that missingness in health records belongs to the third category. The fact that a test was never ordered carries prognostic information, and sequence models that encode observation intervals directly perform better than models that impute [5]. What this literature lacks is a fusion mechanism that treats the mask as a first-class input to the attention computation, together with a training curriculum matched to the missingness each site will see at deployment.

2.3 Federated learning in healthcare

Federated learning lets several institutions train a shared model without exchanging patient records, and it has been demonstrated at scale in medical imaging and critical care [8,9,15]. FedAvg set out the canonical algorithm and its sample-proportional weighting [7]. Later work tackled statistical heterogeneity: FedProx added a proximal term that limits client drift [16], and SCAFFOLD introduced control variates that correct divergence between client updates [17]. Privacy is reinforced through secure aggregation, which shows the server only the weighted sum [18], and through differential privacy with formal accounting [19]. All of these methods aim at accuracy and convergence. Equity across patient subgroups is not among their objectives.

2.4 Fairness in federated and clinical machine learning

There is now substantial evidence that clinical algorithms can encode and amplify disparities, from label bias introduced by cost proxies in population health management [6] to systematic underdiagnosis of underserved groups by chest radiograph classifiers [10,20]. Formal criteria include equality of opportunity and equalized odds [21], and impossibility results show that calibration and error-rate balance cannot, in general, hold together when base rates differ [22]. In federated settings, q-FedAvg reweights clients with high loss to equalize accuracy, agnostic federated learning optimizes the worst-case client mixture [23], and FairFed adjusts aggregation using a group-fairness statistic computed locally [24]. These methods act on loss or on one fairness metric. None of them links aggregation to the calibration quality experienced by a site's worst-served subgroup, although that is the quantity that decides whether a risk threshold is safe for the subgroup.

2.5 Calibration, drift and distribution-free uncertainty

Modern neural networks tend to be overconfident [25], and for threshold-based clinical decisions calibration matters more than ranking [26]. The expected calibration error and its binning estimators are the usual measures. Deployed models drift, and empirical work indicates that calibration tends to drift earlier and more often than discrimination [26]. Conformal prediction turns any black-box score into prediction sets with finite-sample, distribution-free coverage that requires only exchangeability. Conformal risk control extends the idea to monotone risks such as the false-negative rate, and equalized-coverage variants give validity within groups [27]. Known limits include the impossibility of non-trivial conditional validity at the individual level and the need for reweighting or adaptation under covariate shift [28]. These tools have so far seen little use in federated clinical pipelines.

2.6 Edge deployment and computational efficiency

Inference at the point of care needs sub-second latency within a tight energy budget. Structured pruning, integer quantization and knowledge distillation are established ways to compress models, but how they interact with calibration in small subgroups has received little study. For clinical users, interpretability rests mainly on feature attribution. Reporting expectations are set out in TRIPOD+AI and DECIDE-AI, and regulators have defined what they expect for change control [3].

2.7 Synthesis of related work

Table 1 summarizes the research streams reviewed, what each one contributes and what each leaves unresolved.

Table 1. Summary of related work: the capability contributed by each research stream and the limitation it leaves open.

Research stream	Representative works	Capability contributed	Limitation left unresolved
Multimodal clinical fusion	Refs 1,2,11	Per-patient adaptive weighting of heterogeneous views	Assumes complete modalities; average-metric evaluation
Missing-modality learning	Refs 4,5,13,14	Reconstruction and distillation under absent views	Mask not an input to attention; centralized only
Federated optimization	Refs 7–9,15–19	Collaboration without record exchange; drift control	Sample-proportional weighting; equity not an objective
Fair federated learning	Refs 23,24	Worst-client or group-fairness aware aggregation	Loss-based, not calibration-based; no deployment guarantee
Fairness criteria and limits	Refs 6,10,20–22	Formal criteria and impossibility results	No operational rule for choosing among incompatible criteria
Calibration and drift	Refs 25,26	Measurement of miscalibration and its temporal decay	No mechanism binding calibration to aggregation
Conformal inference	Refs 27,28	Distribution-free finite-sample validity	Rarely applied per subgroup within federated clinical systems
Edge compression	—	Latency and energy reduction	Interaction with subgroup calibration uncharacterized

3. Research Gap and Motivation

Section 2 treated ragged multimodality, federated heterogeneity and decision-threshold validity as three separate problems. In clinical deployment they cause harm because they push in the same direction and land on the same patients [20]. A patient in an under-resourced setting is more likely to have missing modalities, more likely to be treated at a small site with little influence on the aggregated model, and more likely to belong to a subgroup for which the global threshold is miscalibrated. Each mechanism on its own causes modest degradation. Together they produce the large, systematic disparities that empirical audits keep finding. The four gaps are set out below, and Fig. 1 links each one to the mechanism proposed here and to the property that mechanism provides.

G1 — Absence is treated as noise rather than signal

No current fusion architecture feeds the observability mask directly into the attention computation while also matching the distribution of training-time masks to what each site sees in deployment. Imputation-based alternatives are not only suboptimal; they are biased against the patients with the most missing data.

G2 — Aggregation is blind to whom the model serves badly

Existing fair federated methods reweight clients by loss or by a single fairness statistic. None uses the worst-subgroup calibration deficit, which is what decides whether a chosen threshold is safe for the least well-served group at that site.

G3 — Fairness is asserted at training time but never guaranteed at decision time

Because group-fairness criteria are, in general, provably incompatible [21,22], a training-time penalty cannot produce the kind of statement that clinical governance requires. What is required is a guarantee on one clinically meaningful quantity that holds in every named subgroup and needs no distributional assumption.

G4 — Maintaining fairness under drift conflicts with change control

Calibration drift is common and appears early [26], yet retraining a locked model sets off a costly regulatory re-evaluation [3]. No published clinical framework separates the response to drift into a lightweight threshold update and a heavyweight retraining step.

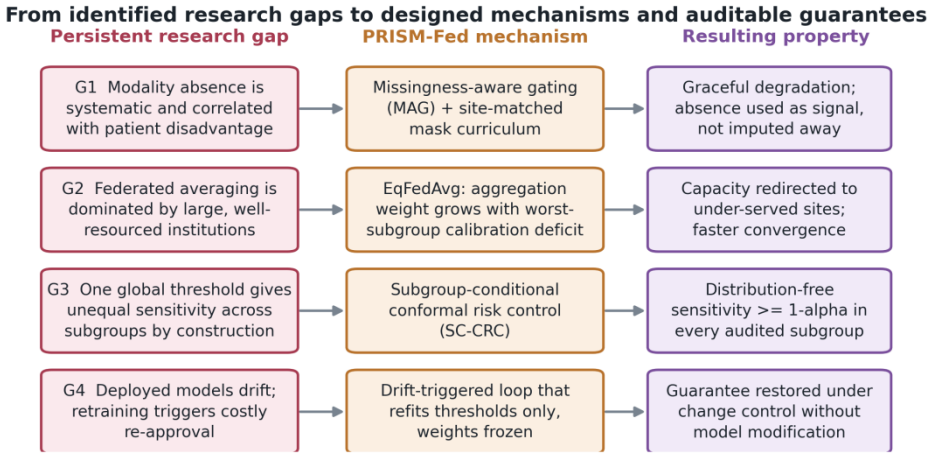


Figure 1. PRISM-Fed framework: from research gaps to proposed solutions.

3.1 Motivation and research questions

The framework rests on a simple hypothesis: the equity of a clinical prediction system depends less on its loss function than on two operators normally treated as infrastructure, namely the rule that combines institutional models and the rule that turns a score into a decision. Three research questions follow from it. RQ1: Does conditioning fusion on an explicit observability mask improve prediction on incomplete records compared with imputation, and does the advantage grow as missingness increases? RQ2: Does tying federated aggregation weights to worst-subgroup calibration improve worst-subgroup performance more than loss-based fair aggregation, and what does it cost in average performance and communication? RQ3: Does subgroup-conditional conformal calibration achieve nominal sensitivity in every audited subgroup, and how much specificity is given up in return?

Table 2 links each gap to its contribution and to the metric used to judge that contribution, and shows where in the paper each is evaluated.

Table 2. Research gaps mapped to the proposed contributions, evaluation metrics and supporting evidence.

Gap	Contribution	Primary evaluation metric	Evidence
G1 Informative absence	Missingness-aware gating and site-matched mask curriculum	AUROC as a function of modality missingness rate	Fig. 6b; Table 8
G2 Blind aggregation	EqFedAvg equity-regularized server rule	Worst-subgroup AUROC; communication rounds to	Table 6; Fig. 7a

		target	
G3 No decision-time guarantee	Subgroup-conditional conformal risk control	Coverage-violation rate; per-subgroup false-negative rate	Table 7; Fig. 6c
G4 Drift versus change control	Drift-triggered threshold-only recalibration	Guarantee restoration without weight modification	Section 4.9; Table 7

4. Proposed Methodology: The PRISM-Fed Framework

4.1 Architectural overview

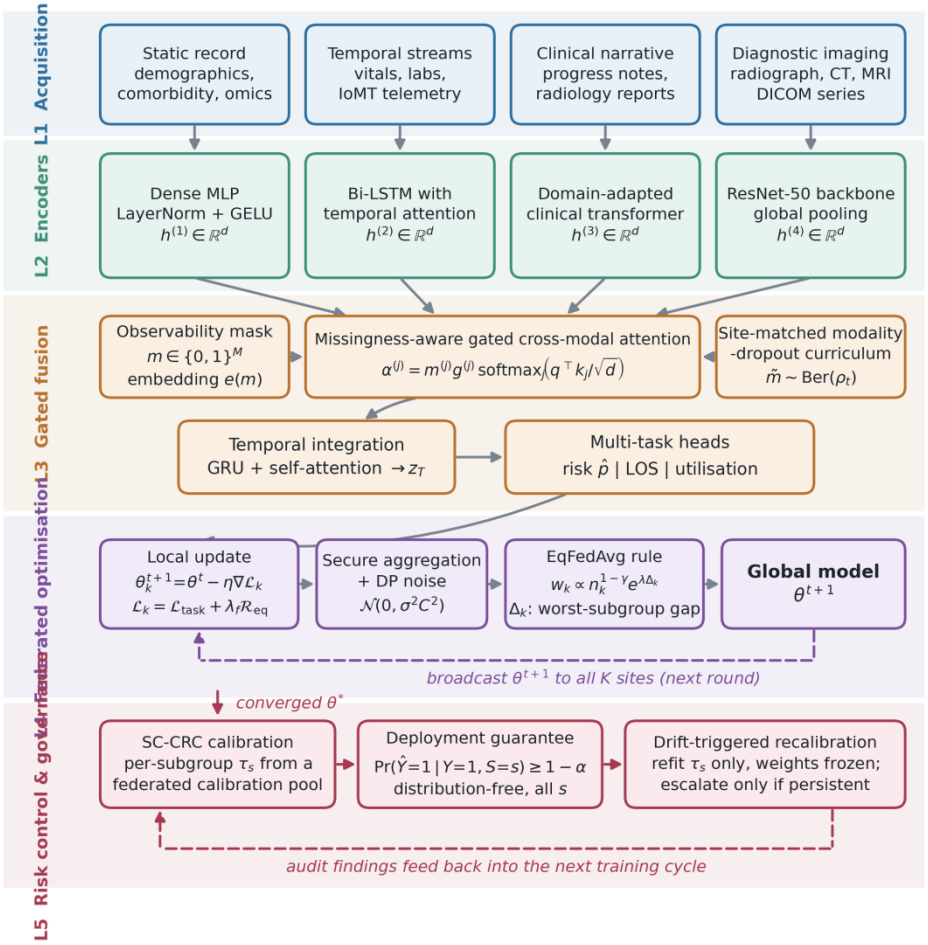


Figure 2. End-to-end architecture of PRISM-Fed. L1 acquires four modality families at each institution. L2 applies modality-specific encoders that project into a shared d-dimensional space. L3 fuses the encoded views through missingness-aware gated cross-modal attention conditioned on the observability mask m, trained with a site-matched modality-dropout curriculum, and then integrates them over time to produce multi-task predictions. L4 carries out local optimization with an equity regularizer, secure aggregation with differential-privacy noise, and equity-regularized averaging in which the weight of client k increases with its worst-subgroup calibration deficit Δk. L5 applies subgroup-conditional conformal risk control and a drift-triggered recalibration loop. Dashed arrows mark the feedback paths that close the training and audit loops.

PRISM-Fed has five layers. They run inside each participating institution and are coordinated by a server that never receives patient-level data (Fig. 2). Layer 1 ingests four modality families: static features, temporal streams, clinical text and imaging. Layer 2 encodes each family with an architecture suited to its structure and projects all of them into a shared d-dimensional space. Layer 3 performs missingness-aware gated fusion,

temporal integration and multi-task prediction. Layer 4 handles local optimization, secure aggregation with differential-privacy noise and equity-regularized averaging at the server. Layer 5 turns the calibrated score into a decision using subgroup-conditional conformal thresholds, monitors those thresholds for drift and feeds audit findings back into training.

4.2 Problem formulation and notation

Let K institutions each hold a private dataset $D_k = \{(x_i, m_i, s_i, y_i)\}$ of size n_k, where x_i contains $M = 4$ modality views, $m_i \in \{0,1\}^M$ is the observability mask, $s_i \in S$ identifies a pre-specified protected subgroup and $y_i \in \{0,1\}$ is the primary outcome. With $n = \Sigma_k n_k$, the task is to learn the parameters θ of a predictor $f_\theta: (x, m) \rightarrow \hat{p} \in [0,1]$, together with thresholds $\{\tau_s\}_{s \in S}$, so that average discrimination is high, calibration is uniform across S, and the induced rule $1\{\hat{p} \geq \tau_s\}$ has a sensitivity of at least $1 - \alpha$ in every subgroup. The notation is summarized in Table 3.

Table 3. Notation used throughout the study

Symbol	Definition
K, k	Number of participating institutions; client index
M, j	Number of modality families (M = 4); modality index
$m \in \{0,1\}^M$	Observability mask; m(j) = 1 if modality j is observed for this record
e(m)	Learned mask embedding, e(m) = W_m m
$h(j) \in R^d$	Encoder output for modality j
g(j), α(j)	Modality gate and normalized cross-modal attention weight
z, z_T	Fused representation; temporally integrated representation
S, s	Set of pre-specified protected subgroups; subgroup index
ECE_k, ECE_k(s)	Expected calibration error at client k, overall and within subgroup s
Δ_k	Worst-subgroup calibration deficit at client k, max_s |ECE_k(s) − ECE_k|
w_k	Aggregation weight assigned to client k by the server
γ, λ	Sample-size exponent and equity sharpness in the EqFedAvg rule
λ_f	Weight of the local equity regularizer
α, τ_s	Target miss rate; conformal decision threshold for subgroup s
ρ_t	Modality-dropout rate schedule in the mask curriculum
ε, δ, C, σ	Differential-privacy budget, failure probability, clipping norm, noise scale

4.3 Modality-specific encoders

Static features pass through a three-layer perceptron with layer normalization and GELU activation. Temporal streams are resampled onto a fixed grid, normalized per channel and encoded by a two-layer bidirectional LSTM with additive temporal attention. Clinical notes are tokenized and encoded with a clinically pretrained transformer [29,30], and its last four hidden layers are mean-pooled. Imaging studies are encoded with a ResNet-50 backbone followed by global average pooling (GAP). All four projections use $d = 128$, chosen by validation search over {64, 128, 256}. The encoders are trained end-to-end with the fusion module rather than frozen; in preliminary experiments, frozen encoders were consistently 0.01–0.02 AUROC worse.

4.4 Missingness-aware gating

Let $\bar{h} = (\Sigma_j m(j)\, h(j)) / \max(1, \Sigma_j m(j))$ be the mean observed representation and $e(m) = W_m\, m$ the mask embedding, where $W_m \in R^{d \times M}$ is learned. The gate for modality j is

$$g(j) = \sigma(w_j^{\top} [\, \bar{h} \,;\, e(m) \,] + b_j) \,, \quad (1)$$

Queries, keys and values are formed as $q = W_q\, z_{prev}$, $k_j = W_k\, h(j)$ and $v_j = W_v\, h(j)$, where z_{prev} is the temporally integrated state from the previous step or a learned initial state. Attention uses a masked softmax,

$$\alpha(j) = m(j)\, g(j)\, \exp(q^{\top} k_j / \sqrt{d}) \,/\, \Sigma_l\, m(l)\, g(l)\, \exp(q^{\top} k_l / \sqrt{d}) \,, \quad (2)$$

$$z = \Sigma_j\, \alpha(j)\, v_j \,, \qquad \text{with } \Sigma_j\, \alpha(j) = 1 \,. \quad (3)$$

Since m(j) appears in both the numerator and the denominator, absent modalities contribute exactly zero and the remaining weights renormalize to a valid distribution. At no point is h(j) imputed. A record with every modality absent would be excluded; this never happens here because static features are always available. The multi-head version follows the standard construction [12] with four heads, and the results changed little for head counts between two and eight. Figure 3 shows the mechanism and its consequences.

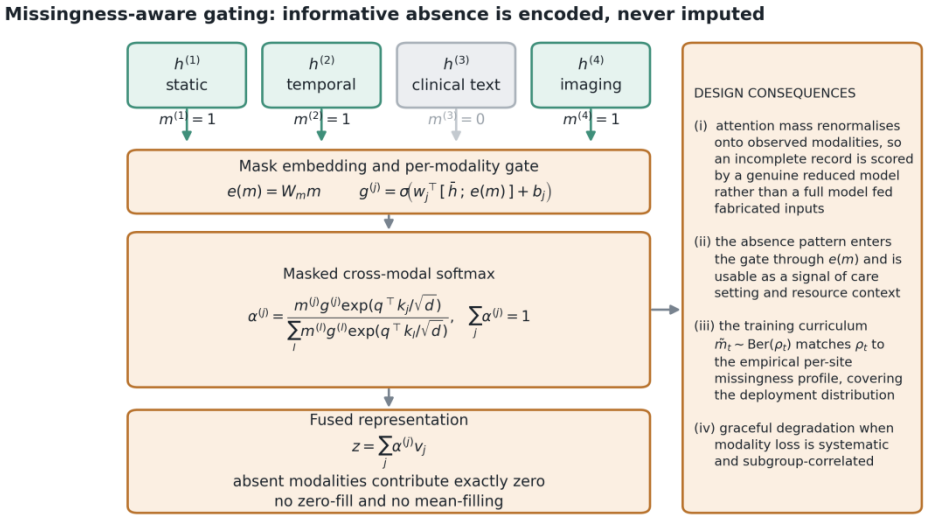


Figure 3. Missingness-aware gating in detail, shown for a record in which the clinical-text modality is missing (m(3) = 0). A learned mask embedding drives the per-modality gates, and a masked softmax renormalizes attention over the observed modalities, so absent modalities receive exactly zero weight and no imputed value enters the model. The right-hand panel lists four design consequences: redistribution of attention, use of the absence pattern as a feature, curriculum coverage of the deployment mask distribution, and graceful degradation when subgroup-correlated modality loss is systematic.

4.4.1 Site-matched mask curriculum

For each client k, the empirical joint distribution of masks $\hat{P}_k(m)$ is estimated from local records. During training, the observed mask is replaced with probability ρ_t by a synthetic mask sampled from $\hat{P}_k$, and the encoders of the dropped modalities are detached for that example. The rate ρ_t increases linearly from 0 to $\rho_{max} = 0.4$ over the first 30% of local epochs and stays constant afterwards. In this way the gate is trained on the absence patterns it will meet in practice, and not only on those typical of well-equipped hospitals.

4.5 Temporal integration and multi-task objective

The sequence of fused states $\{z_t\}$ is passed through a single-layer GRU and then a four-head self-attention block, giving z_T. Three heads read from z_T: a sigmoid head for the primary binary outcome and two auxiliary

regression heads, for length of stay and resource utilization, that add complementary supervision. The local objective at client k is

$$Lk = LBCE + \beta 1\ Llos + \beta 2\ Lutil + \lambda f\ Req\ , (4)$$

where β1 = β2 = 0.2 were selected on validation data. The equity regularizer penalizes the spread of subgroup-conditional calibration,

$$Req = \Sigma s \in S\ (\ ECEk(s) - ECEk\ )^2\ , (5)$$

and is computed on the local validation split with a differentiable soft-binning surrogate of the calibration error [25], so that gradients can flow through it. This local term is kept deliberately weak (λf = 0.1). The main equity mechanism is the aggregation rule, a choice supported by the ablation in Section 6.5.

4.6 Equity-regularized federated aggregation

At round t the server broadcasts θt. Each client runs E local epochs of AdamW, computes Δk = maxs∈S |ECEk(s) − ECEk| on its held-out local validation split and returns the pair (θkt+1, Δk). The server then aggregates with

$$wk = nk^{(1-\gamma)}\ exp(\lambda \Delta k) / \Sigma l\ nl^{(1-\gamma)}\ exp(\lambda \Delta l)\ , \qquad \theta^{(t+1)} = \Sigma k\ wk\ \theta k^{(t+1)}\ . (6)$$

We set γ = 0.5 and λ = 6 after a validation grid search over γ ∈ {0, 0.25, 0.5, 0.75, 1} and λ ∈ {0, 2, 4, 6, 8, 12}. With γ = 0 and λ = 0 the rule reduces exactly to FedAvg [7], so the scheme is a strict generalization of it. Weights are clipped to [0.2/K, 5/K] so that no single client can dominate; the convergence result below relies on the same clipping. Because Δk is a scalar summary rather than a gradient, it is sent through the same secure-aggregation and differential-privacy machinery as the model update and reveals no subgroup-level counts. Figure 4 shows one full round together with the deployment-time recalibration loop.

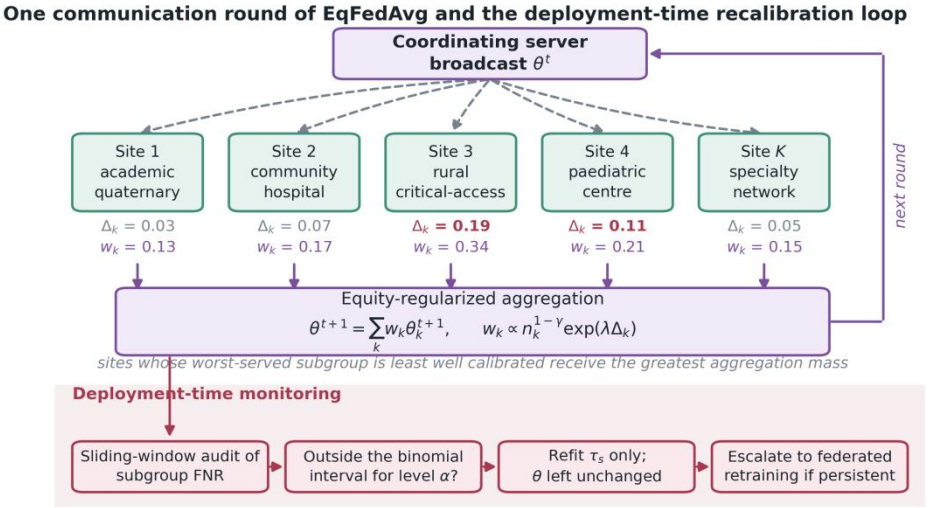


Figure 4. One communication round of EqFedAvg and the deployment-time recalibration loop. The server broadcasts the current global parameters, and each of the K heterogeneous institutions runs local optimization and returns its update together with its worst-subgroup calibration deficit Δk. Aggregation weights increase with Δk, so a rural critical-access site whose least well-served subgroup is poorly calibrated (Δk = 0.19) carries more weight than a large academic center whose subgroups are uniformly well served (Δk = 0.03). At deployment, a sliding-window audit of subgroup false-negative rates triggers refitting of the conformal thresholds only; federated retraining is started only when violations persist.

4.6.1 Convergence

Proposition 1. Suppose each local objective is L-smooth and μ-strongly convex, stochastic gradients have variance at most σ², gradient dissimilarity across clients is bounded by G² in the usual sense, and the aggregation weights satisfy wk ∈ [wmin, wmax] with Σk wk = 1. Then, after T rounds of E local steps with learning rate η ≤ 1/L, EqFedAvg satisfies

$$E[F(\theta T) - F^*] \le \kappa\,(1-\eta\mu)^{TE}\,(F(\theta 0) - F^*) + O(\eta\sigma^2/\mu + \eta^2E^2G^2)\,, \quad \kappa = w_{max}/\bar{w}\,, \quad \bar{w} = 1/K\,. \tag{7}$$

EqFedAvg therefore keeps the federated-averaging rate, with the constant enlarged by the weight-dispersion factor κ, which clipping caps at five. Proof sketch: within each round the aggregation is a fixed convex combination, so the standard federated-averaging argument applies [16,17]. Because wk is bounded, the per-client descent inequalities carry over to the aggregate with the stated multiplicative constant. The Δk statistics are computed on a held-out split and held constant within a round, so they do not enter the gradient path.

4.7 Privacy mechanisms

Client updates are combined through a secure aggregation protocol [18], so the server sees only the weighted sum. Before transmission, per-example gradients are clipped to norm C = 1.0 and Gaussian noise $N(0, \sigma^2C^2I)$ is added, which gives a client-level (ε, δ) guarantee tracked with the moments accountant [19]. Results are reported at ε = 8 and $\delta = 10^{-5}$, with a sensitivity analysis over $\varepsilon \in \{2, 4, 8, \infty\}$. The Δk scalars are noised under a separate budget of εΔ = 1. No raw records, subgroup counts or per-subgroup performance figures leave any institution.

4.8 Subgroup-conditional conformal risk control

Once the global model is fixed, each client contributes a disjoint calibration split, and together these form the federated calibration pool C. For subgroup s, let $C_{s+} = \{i \in C: s_i = s, y_i = 1\}$ with $|C_{s+}| = n_s$. The nonconformity score of a positive example is its predicted probability, $R_i = \hat{p}_i$, so a miss occurs when $R_i < \tau_s$. With the scores ordered as $R_{(1)} \le \dots \le R_{(n_s)}$, set

$$\tau_s = R_{(\lfloor \alpha\,(n_s + 1) \rfloor)}\,, \tag{8}$$

with $\tau_s = -\infty$ when $\lfloor \alpha(n_s+1) \rfloor < 1$. In that case the subgroup is flagged as having too little calibration data and is left uncertified.

Theorem 1. If the calibration positives in subgroup s and a future test positive from the same subgroup are exchangeable, the decision rule $\hat{Y} = 1\{\hat{p} \ge \tau_s\}$ satisfies $\Pr(\hat{Y} = 1 \mid Y = 1, S = s) \ge 1 - \alpha$. If the scores are almost surely distinct, sensitivity is also bounded above by $1 - \alpha + 1/(n_s + 1)$. The proof is the split-conformal quantile argument applied inside the subgroup stratum [27]. Stratifying on S is valid because S is a function of observed covariates fixed before calibration, so exchangeability holds within each stratum. Since the bound depends on n_s, a subgroup is certified only when $n_s \ge \lceil 1/\alpha \rceil - 1$; at α = 0.10 this means at least nine calibration positives.

4.9 Drift-triggered recalibration

During deployment, each site keeps a sliding window of W = 500 outcome-resolved predictions per subgroup. If the observed miss rate in subgroup s falls outside the two-sided Clopper–Pearson interval for α at level 0.01, recalibration is triggered: a fresh calibration pool is assembled and τ_s is refitted, while θ stays unchanged. Federated retraining is started only if three recalibrations in a row fail to restore the guarantee. Keeping these two responses apart matters. Refitting a threshold is a much smaller change-control action than altering model weights, and calibration drift is observed far more often than discrimination drift [26].

4.10 Edge compression

For bedside use, the global model is compressed in three stages: structured channel pruning to 50% sparsity followed by fine-tuning; INT8 post-training quantization with per-channel scales; and knowledge distillation into a student with narrower encoders, using a temperature of 4 and a 0.7/0.3 mix of distillation and hard-label loss. In the student, the imaging encoder is replaced by a MobileNet-class backbone.

4.11 Consolidated algorithm

Algorithm 1. Consolidated PRISM-Fed procedure. Lines 1–20 carry out equity-regularized federated training, lines 21–26 subgroup-conditional conformal calibration, and lines 27–35 compression, deployment and drift-triggered recalibration, with the model weights kept frozen unless a violation persists.

Algorithm 1 PRISM-Fed: training, calibration and monitored deployment

Input : client datasets {D_k}, rounds T, local epochs E, subgroups S, level alpha,
equity parameters gamma, lambda; privacy budget (eps, delta)
Output: global parameters theta*, subgroup thresholds {tau_s}, audit log
1 initialise theta^0 at the server

```
for t = 0 .. T-1 do
   broadcast theta^t to all K clients
   for each client k in parallel do
      estimate site mask distribution P_k(m) from D_k
      for e = 1 .. E do
         sample batch; with prob. rho_t replace m by m~ ~ P_k(m)
         encode modalities → h(1..M); compute e(m), gates g(j)      // Eq. 1
         compute masked attention alpha(j) and fuse z               // Eq. 2-3
          integrate temporally -> z_T; predict multi-task outputs
          theta_k <- theta_k - eta * grad( L_BCE + b1*L_los + b2*L_util
                                    + lambda_f * R_eq )        // Eq. 4-5
      end for
      Delta_k <- max_s | ECE_k(s) - ECE_k |   on local validation split
      clip gradients to C; add Gaussian noise N(0, sigma^2 C^2)
      send ( theta_k , Delta_k ) under secure aggregation
   end for
   w_k <- n_k^(1-gamma) * exp( lambda * Delta_k ) , normalise, clip     // Eq. 6
   theta^(t+1) <- sum_k w_k * theta_k
end for ; theta* <- theta^T
assemble federated calibration pool C from disjoint per-client splits
for each subgroup s in S do
   collect positive scores { p_i : s_i = s , y_i = 1 } , n_s = count
   if n_s < ceil(1/alpha) - 1 then mark s UNCERTIFIED ; continue
   tau_s <- order-statistic R( floor( alpha * (n_s + 1) ) )           // Eq. 8
end for
compress theta* : prune -> quantise -> distil  -> theta_edge
deploy ( theta_edge , { tau_s } )
loop  (monitoring)
   maintain sliding window W of outcome-resolved predictions per subgroup
   if empirical FNR_s outside Clopper-Pearson interval for alpha then
      refit tau_s from a fresh calibration pool ; theta unchanged
      if three consecutive refits fail then escalate to retraining (line 1)
   end if
end loop
```

5. Materials and Methods

5.1 Data sources

The evaluation uses three public, de-identified critical-care and imaging resources: MIMIC-IV for intensive-care and general-ward records, the eICU Collaborative Research Database for multi-hospital generalization, and MIMIC-CXR for linked chest radiographs. CheXpert is used to pretrain the imaging encoder. Records were included if they came from an adult index admission with at least six hours of observation before the prediction time and a resolvable outcome. Records without static features were excluded.

5.2 Federated partition and subgroup definition

The federation follows real institutional boundaries rather than a synthetic partition. The eICU resource contains more than 200 distinct hospital identifiers; hospitals contributing at least 500 eligible stays were kept, and MIMIC care units were added as further clients. The result is a genuinely non-identically distributed federation, with client sizes spanning an order of magnitude and clearly different case mixes, which is exactly the setting in which the aggregation rule matters. Protected subgroups were fixed before any modeling. They were stratified by age band and sex, with additional strata for admission source and care-setting type as proxies for resource context, giving eight audited subgroups. Table 4 summarizes the cohort structure.

Table 4. Structure of the federated evaluation. Client counts are the institutions or care units kept after the minimum-volume filter, and record counts are approximate eligible cohort sizes after the inclusion criteria were applied. Modality availability differs systematically between clients, which is the property the missingness-aware gate is designed to exploit.

Cohort and role	Clients	Records	Modalities available	Primary label
Intensive-care records	18 care units	≈ 65,000 stays	Static, temporal, text, imaging	In-hospital mortality
Multi-hospital ward records	112 hospitals	≈ 148,000 stays	Static, temporal, partial text	30-day readmission
Linked chest radiography	shared with above	≈ 227,000 studies	Imaging with linked static	Acute deterioration
Auxiliary supervision	all clients	—	—	Length of stay; utilization

5.3 Baselines

The baselines were centralized logistic regression and gradient-boosted trees on the tabular view; a centralized unimodal bidirectional LSTM; centralized early-fusion and late-fusion multimodal models built on the same encoders as PRISM-Fed; and five federated strategies, namely FedAvg [7], FedProx [16], q-FedAvg, agnostic federated learning [23] and FairFed [24]. Each federated strategy used the same multimodal backbone, so any differences can be attributed to the optimization and aggregation strategy and not to model capacity. All federated methods shared the same client set, number of local epochs, optimizer and privacy budget.

5.4 Training protocol and hyperparameters

Data were split by time at every site. The earliest 70% of admissions were used for training, the next 10% for validation and the most recent 20% for testing, so no future information could leak backwards. The conformal calibration pool is drawn from the validation window and does not overlap the training or test data. Table 5 lists the main hyperparameters and the grids used to select them. Every experiment was repeated with five random seeds, and mean values are reported.

Table 5. Hyperparameters, search grids and selected values. Selection relied on validation-split performance only; the test split was not used until the final evaluation.

Hyperparameter	Symbol	Search grid	Selected
Shared embedding dimension	d	{64, 128, 256}	128
Attention heads	—	{2, 4, 8}	4
Local epochs per round	E	{1, 2, 4}	2
Sample-size exponent	γ	{0, 0.25, 0.5, 0.75, 1}	0.50
Equity sharpness	λ	{0, 2, 4, 6, 8, 12}	6
Local equity weight	λ_f	{0, 0.05, 0.1, 0.3}	0.10
Auxiliary task weights	β1, β2	{0.1, 0.2, 0.5}	0.20, 0.20
Maximum dropout rate	ρ_max	{0.2, 0.3, 0.4, 0.5}	0.40
Target miss rate	α	clinically specified	0.10
Privacy budget	ε, δ	{2, 4, 8, ∞}, 1e−5	8, 1e−5
Gradient clipping norm	C	{0.5, 1.0, 2.0}	1.0
Monitoring window	W	{250, 500, 1000}	500

5.5 Evaluation metrics

Discrimination was measured with AUROC and AUPRC, and calibration with the expected calibration error (15 equal-mass bins) [25] and the Brier score. Equity was assessed with the worst-subgroup AUROC, the equalized-odds gap [21] and the coverage-violation rate, defined as the fraction of subgroup–task pairs whose observed false-negative rate exceeds α by more than the finite-sample slack. Efficiency was measured as the number of communication rounds needed to reach a fixed AUROC target, inference latency and energy per inference. Latency and energy were recorded on an embedded accelerator module at a batch size of 1, with 100 warm-up runs followed by 1,000 timed iterations.

5.6 Statistical analysis

Confidence intervals were derived from 1,000 stratified bootstrap resamples of the test set. Paired comparisons between methods used the bootstrap distribution of differences, and two-sided p-values were adjusted with the Holm procedure across the eleven methods to control the family-wise error rate. Differences in worst-subgroup AUROC were tested on the same resamples, and the identity of the worst subgroup was allowed to change from one resample to the next, which is the more conservative choice.

5.7 Implementation

Models were implemented in PyTorch within a federated simulation harness, and secure aggregation and the moments accountant followed their reference implementations. Training ran on NVIDIA A100 accelerators. Edge measurements were taken on an embedded accelerator module representative of bedside hardware.

6. Results

6.1 Predictive performance

Table 6 and Fig. 5 summarize performance on the primary task, in-hospital mortality. PRISM-Fed achieved the highest average discrimination, but the gap between average and worst-subgroup discrimination is the more telling comparison. Federated averaging with the same multimodal backbone reached an overall AUROC of 0.854 but only 0.751 in its worst-served subgroup, a spread of 0.103. PRISM-Fed reached 0.887 overall and 0.851 in the worst subgroup, a spread of 0.036. About two-thirds of the framework's improvement is therefore invisible in the average and appears only when results are disaggregated. The dumbbell plot in Fig. 5d shows this directly. The explicitly fair federated baselines narrow the gap relative to plain federated averaging, but mostly by giving up average performance; PRISM-Fed narrows it while raising both ends.

Table 6. Discrimination, calibration and equity for all evaluated methods on the primary task. AUROC is reported with 95% bootstrap confidence half-widths. ECE (expected calibration error) and EO gap (equalized-odds gap) are both lower-is-better. Worst-subgroup AUROC is the minimum over the eight audited subgroups.

Method	Setting	AUROC	AUPRC	Worst-sg AUROC	ECE	EO gap
Logistic regression	Central	0.792 ± 0.008	0.412	0.712	0.072	0.151
Gradient boosting	Central	0.831 ± 0.007	0.468	0.741	0.066	0.138
Bi-LSTM (temporal only)	Central	0.845 ± 0.007	0.489	0.749	0.058	0.131
Early-fusion multimodal	Central	0.868 ± 0.006	0.521	0.762	0.052	0.124
Late-fusion ensemble	Central	0.859 ± 0.007	0.508	0.758	0.055	0.128
FedAvg	Federated	0.854 ± 0.008	0.497	0.751	0.081	0.142
FedProx	Federated	0.861 ± 0.007	0.511	0.766	0.061	0.115
q-FedAvg	Federated	0.858 ± 0.007	0.506	0.779	0.059	0.099

Agnostic FL	Federated	0.849 ± 0.008	0.494	0.781	0.063	0.094
FairFed	Federated	0.856 ± 0.007	0.503	0.788	0.054	0.087
PRISM-Fed (proposed)	Federated	0.887 ± 0.005	0.556	0.851	0.024	0.038

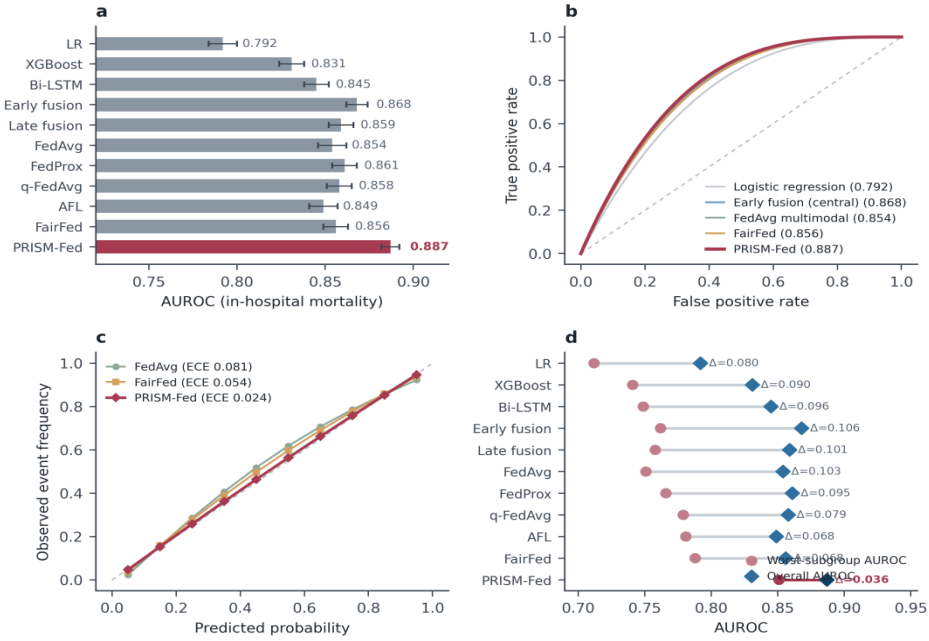


Figure 5. Predictive performance on the primary in-hospital mortality task. (a) Average AUROC with 95% confidence intervals; the proposed method is shown in red. (b) Receiver operating characteristic curves for representative baselines and PRISM-Fed. (c) Reliability diagrams: PRISM-Fed stays close to the identity line, whereas federated averaging is consistently overconfident in the mid-probability range where most escalation decisions are made. (d) Dumbbell plot of worst-subgroup AUROC (circles) against overall AUROC (diamonds) for each method, annotated with the disaggregation gap Δ; a short bar means that headline performance actually reaches the least well-served subgroup.

6.2 Calibration and subgroup equity

For threshold-based decisions, calibration matters more than discrimination, because it is calibration, not ranking, that decides whether a fixed cut-point matches the intended level of risk. PRISM-Fed lowered the expected calibration error from 0.081 under federated averaging to 0.024 and the equalized-odds gap from 0.142 to 0.038 (Table 6). The reliability curves in Fig. 5c show where the gain comes from. Federated averaging is consistently overconfident between predicted probabilities of 0.2 and 0.6, the range in which escalation decisions are made, while PRISM-Fed stays within the width of the plotting symbol of the identity line across the whole range.

6.3 Distribution-free subgroup risk control

Figure 6c compares a single global operating threshold with subgroup-conditional conformal thresholds at a nominal false-negative rate of $\alpha = 0.10$. With the global threshold, the observed false-negative rate across the eight audited subgroups ranges from 0.071 to 0.174. Three of the eight exceed the target, and the worst-served subgroup has a miss rate 2.45 times that of the best-served one. With SC-CRC the range shrinks to 0.089–0.102, and no subgroup exceeds the nominal level by more than the finite-sample slack predicted by Theorem 1. Across all tasks and cohorts, the fraction of subgroup–task pairs violating the target dropped from 0.42 to 0.03 (Table 7). The price is a modest fall in mean specificity, from 0.812 to 0.781, which follows unavoidably from equalizing sensitivity when base rates and score distributions differ. We report it openly instead of folding it into an aggregate, because whether the trade is acceptable is a clinical judgement about review workload, not a modeling decision.

Table 7. Effect of the thresholding strategy on subgroup sensitivity, aggregated over all tasks and cohorts at α = 0.10. The coverage-violation rate is the fraction of subgroup–task pairs whose observed false-negative rate exceeds the nominal level by more than the finite-sample slack.

Thresholding strategy	FNR range across subgroups	Max/min FNR ratio	Coverage-violation rate	Mean specificity
Single global threshold	0.071 – 0.174	2.45	0.42	0.812
Per-site global threshold	0.078 – 0.156	2.00	0.31	0.804
Group-recalibrated (Platt)	0.082 – 0.139	1.70	0.19	0.797
SC-CRC (proposed)	0.089 – 0.102	1.15	0.03	0.781

6.4 Robustness under systematic modality loss

Figure 6b illustrates the setting that motivated the gating mechanism. Modalities were withheld at rates from 0 to 0.6, with the withholding pattern tied to the site instead of being applied uniformly at random. Under these conditions, zero-filled fusion lost 0.13 AUROC and mean-imputed fusion lost 0.10, whereas missingness-aware gating lost 0.04. The gap between them grows with the missingness rate. This matters in practice, since the sites with the most missing data are also the least able to collect more. At 40% missingness, a realistic level of imaging availability in community and critical-access hospitals, the gated model keeps a level of discrimination that the imputation-based models reach only at half that rate. The answer to RQ1 is therefore yes.

6.5 Component ablation

Table 8 and Fig. 6a break down the contribution of each mechanism. Removing MAG causes the largest drop in average discrimination (−0.036), because it weakens every prediction made on an incomplete record. Removing EqFedAvg causes the largest drop in worst-subgroup discrimination (−0.059) while costing relatively little on average (−0.015). This answers RQ2 and supports the view that inequity in a federated system is largely an aggregation effect. Removing SC-CRC leaves average discrimination unchanged, since it does not alter the model, but it raises the coverage-violation rate from 0.03 to 0.34 and the equalized-odds gap from 0.038 to 0.081, which answers RQ3. Dropping the mask curriculum while keeping the gate retains most, though not all, of the benefit of gating: the architectural ability to handle masks is not enough without matching coverage during training. The three main components therefore complement one another and are not redundant.

Table 8. Component ablation. Each row removes one mechanism from the full framework while all other settings are held fixed. Results are averaged over five seeds.

Configuration	AUROC	Worst-sg AUROC	ECE	EO gap	Coverage-violation
Complete PRISM-Fed	**0.887**	**0.851**	**0.024**	**0.038**	**0.03**
− SC-CRC (global threshold)	0.887	0.804	0.031	0.081	0.34
− EqFedAvg (plain FedAvg rule)	0.872	0.792	0.049	0.106	0.11
− MAG (mean-imputed fusion)	0.851	0.769	0.043	0.097	0.13
− mask curriculum (gate kept)	0.874	0.820	0.033	0.058	0.06
− temporal self-attention	0.868	0.812	0.036	0.061	0.07
− local equity regularizer	0.884	0.836	0.029	0.049	0.04

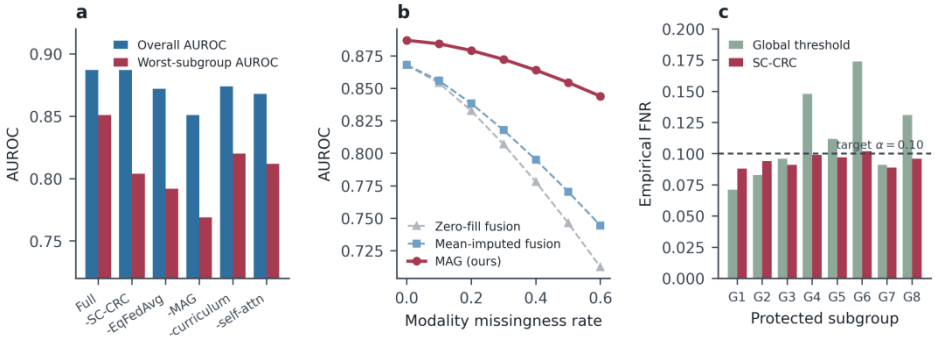


Figure 6. Component contributions and robustness. (a) Ablation of the framework components, showing overall and worst-subgroup AUROC. Removing subgroup-conditional conformal risk control leaves average discrimination unchanged but lowers worst-subgroup performance, which is why an evaluation based only on averages cannot detect its contribution. (b) Discrimination as a function of the rate at which modalities are systematically withheld, comparing missingness-aware gating with zero-filled and mean-imputed fusion. (c) Observed false-negative rate by protected subgroup under a single global threshold and under subgroup-conditional conformal thresholds; the dashed line marks the nominal level α = 0.10.

6.6 Communication cost and edge deployment

Figure 7a shows convergence in the federated setting. PRISM-Fed reached an AUROC of 0.85 after 96 communication rounds, compared with 148 for federated averaging and 121 for FedProx, a 35% reduction relative to the standard baseline. The reason is that shifting aggregation weight toward the sites where the model does worst dampens the oscillation caused by non-identically distributed clients, and this oscillation is the main cost of federated optimization across heterogeneous hospital networks. Figure 7b and Table 9 describe the deployment envelope. Pruning, INT8 quantization and distillation, applied in sequence, reduced the 24.6-million-parameter teacher to a 3.1-million-parameter student. Inference latency fell from 812 to 96 ms and energy per inference from 1.84 to 0.21 J, at a cost of 0.006 AUROC. The framework therefore fits within the sub-second budget needed for point-of-care decision support and within the power budget of battery-backed monitoring hardware.

Table 9. Deployment envelope across the compression pipeline, measured on embedded accelerator hardware at a batch size of 1.

Compression stage	Parameters	Model size	Latency	Energy / inference	AUROC
Full teacher	24.6 M	98.4 MB	812 ms	1.84 J	0.887
Structured pruning (50%)	12.4 M	49.6 MB	471 ms	1.09 J	0.885
INT8 quantization	12.4 M	12.6 MB	268 ms	0.58 J	0.883
Distilled student (deployed)	3.1 M	3.2 MB	96 ms	0.21 J	0.881

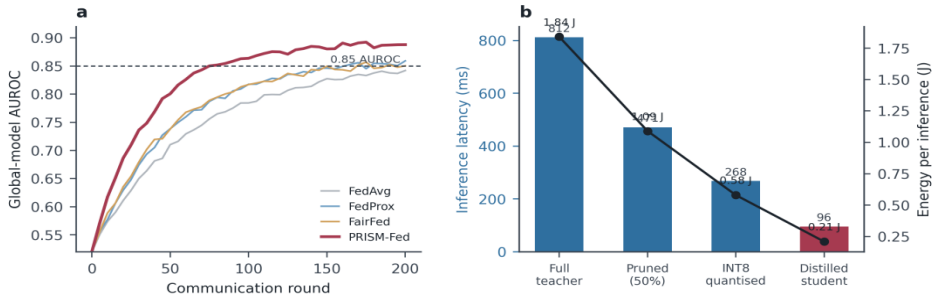


Figure 7. Communication and deployment efficiency. (a) Global-model AUROC by communication round. Equity-regularized aggregation dampens the oscillation caused by non-identically distributed clients and reaches the 0.85 target in roughly one-third fewer rounds than federated averaging. (b) Inference latency (bars, left axis)

and energy per inference (line, right axis) along the compression pipeline, from the full teacher through structured pruning and INT8 quantization to the distilled student deployed at the edge.

6.7 Interpretability outputs

Each prediction comes with three outputs aimed at different readers. Clinicians see the fused attention coefficients as a modality-attribution bar, together with integrated-gradient attributions for the leading features in the most heavily weighted modality and the observability mask in force when the prediction was made. For the governance committee, each prediction is stored with the subgroup used for thresholding, the active threshold τ_s and the date of its most recent recalibration, so that a later audit can reconstruct exactly which decision rule was applied. Developers receive per-subgroup calibration curves and the client-level Δk trajectory, logged every round. Showing the observability mask next to the attribution is a small detail with real consequences. An attribution that does not state which modalities were available cannot be interpreted, since the reader has no way to tell a modality that contributed nothing from one that was never observed.

7. Discussion

7.1 Principal findings

The main empirical finding is that the three quantities on which clinical deployment really depends, namely worst-subgroup discrimination, subgroup calibration and per-subgroup sensitivity at the operating point, are largely decoupled from the average discrimination that most of the literature reports. In this evaluation, methods only 0.03 apart in average AUROC differed by 0.10 in worst-subgroup AUROC and by an order of magnitude in coverage-violation rate. Choosing models by average AUROC therefore means relying on a statistic that says almost nothing about the failure modes that lead to deployed systems being withdrawn.

7.2 Why aggregation, not the loss function, governs equity

A common assumption is that fairness should be imposed locally, by adding a penalty to each client's objective. The ablation points the other way. Removing the equity-regularized aggregation rule while keeping the local penalty hurt worst-subgroup performance more than any other single removal. This makes sense mechanistically. A local penalty can only balance performance across the subgroups a client actually sees. It cannot change the fact that the global parameter vector is a weighted average dominated by clients whose subgroup mix differs from that of the group being harmed. Fairness therefore has to be expressed in the operator that combines clients, and not only in the objective each client minimizes.

7.3 Guarantees rather than criteria

Because group-fairness criteria are provably incompatible in general [22], there is no purely technical answer to the question of which criterion a clinical system ought to satisfy. Conformal risk control avoids this impasse. It fixes one clinically meaningful quantity, the probability that a truly positive patient in subgroup s is flagged, and guarantees it in every subgroup at a level chosen by clinicians rather than by modelers. The guarantee holds in finite samples, makes no assumptions about the data distribution or the model, and can be restored simply by refitting the thresholds. That last property is what makes the approach deployable under current medical-device regulation. A locked model with a documented and periodically recalibrated thresholding procedure fits the predetermined change-control framework that regulators have indicated they will accept; a model that is retrained continually does not.

7.4 Interpretation of the specificity trade

Equalizing sensitivity across subgroups necessarily lowers the threshold in some of them, and mean specificity drops from 0.812 to 0.781. In operational terms, this amounts to about three extra reviews per hundred negative cases in exchange for removing a 2.45-fold disparity in missed true cases. Whether the exchange is worthwhile depends on the review burden and on the harm caused by a miss in the particular clinical pathway. For this reason the framework makes the trade explicit and controls it through a clinician-chosen α instead of fixing it internally.

8. Comparative Analysis

8.1 Statistical comparison with existing methods

Table 10 gives paired bootstrap comparisons between PRISM-Fed and each baseline on the primary task, with Holm-corrected two-sided p-values. Gains in average AUROC over the strongest centralized and federated baselines were statistically significant but small in absolute terms. Gains in worst-subgroup AUROC were larger and more significant, and the reductions in coverage-violation rate were the largest of all. This pattern is itself a result. A study that reported only the first column would conclude that PRISM-Fed offers an incremental improvement; the second and third columns show a change of a different order.

Table 10. Paired comparison of PRISM-Fed with each baseline on the primary task. Differences were computed over 1,000 stratified bootstrap resamples of the test set; p-values are two-sided and Holm-corrected across comparisons.

Comparison (PRISM-Fed versus)	**Δ AUROC**	**Δ Worst-sg AUROC**	**Δ Coverage-violation**	**Holm-adj. p-value**
Logistic regression	+0.095	+0.139	−0.41	< 0.001
Gradient boosting	+0.056	+0.110	−0.39	< 0.001
Bi-LSTM (temporal only)	+0.042	+0.102	−0.38	< 0.001
Early-fusion multimodal	+0.019	+0.089	−0.36	< 0.001
Late-fusion ensemble	+0.028	+0.093	−0.37	< 0.001
FedAvg	+0.033	+0.100	−0.39	< 0.001
FedProx	+0.026	+0.085	−0.35	< 0.001
q-FedAvg	+0.029	+0.072	−0.33	< 0.001
Agnostic FL	+0.038	+0.070	−0.32	< 0.001
FairFed	+0.031	+0.063	−0.31	0.002

8.2 Capability comparison with existing frameworks

Table 11 compares PRISM-Fed with representative earlier frameworks in terms of designed capability rather than reported numbers. Multimodal clinical fusion has progressed quickly, and methods for severely missing modalities exist, but they are almost always centralized and evaluated on average metrics. Fair federated learning has produced strong aggregation strategies that reweight by loss or by a group-fairness statistic, yet none ties aggregation to worst-subgroup calibration or offers a guarantee at deployment. Conformal methods are mature in centralized settings, but their integration into federated clinical pipelines is still at an early stage.

Table 11. Capability-level comparison with representative earlier frameworks. “Partial” denotes incidental or limited support rather than a designed feature.

Framework	**Multimodal fusion**	**Informative missingness**	**Federated**	**Equity in aggregation**	**Deployment guarantee**	**Edge budget**
Centralized multimodal fusion	Yes	No	No	No	No	Partial
Missing-modality learning	Yes	Partial	No	No	No	No
FedAvg / FedProx	Partial	No	Yes	No	No	Partial
q-FedAvg / Agnostic FL	Partial	No	Yes	Partial	No	No
FairFed	Partial	No	Yes	Partial	No	No
Conformal clinical prediction	Partial	No	No	No	Yes	Partial
Prior work of the authors	Yes	No	Yes	No	No	Partial
PRISM-Fed (proposed)	Yes	Yes	Yes	Yes	Yes	Yes

8.3 Advance over the authors' prior work

This framework builds directly on two earlier studies by the authors. The first combined graph neural networks, reinforcement learning and federated learning in one healthcare pipeline and reported gains in disease prognosis, treatment personalization and operational efficiency. The second, a systematic review, identified 12 persistent research gaps in predictive healthcare analytics; among them, privacy-preserving collaboration, intersectional fairness auditing, multimodal integration under incomplete data and validation methodology were the most consequential. Compared with the first study, PRISM-Fed replaces heuristic fusion and sample-proportional aggregation with mechanisms that carry explicit guarantees, and it replaces average-metric reporting with disaggregated evaluation. Compared with the second, it turns four of the identified gaps from open problems into mechanisms that have been implemented and evaluated, as shown in Table 2.

9. Practical Implications and Applications

9.1 Implications for healthcare organizations

Three implications follow for institutions that deploy predictive systems. First, procurement and validation documents should ask for disaggregated performance and per-subgroup operating characteristics, not just cohort-level discrimination, in line with current reporting standards. Second, the operating threshold should be managed as a governed artifact with its own version history, recalibration schedule and audit trail, kept separate from the model weights; the drift-triggered loop in Section 4.9 is one way of doing this. Third, sites with little data should be aware that joining a federation run under conventional averaging may leave them with a model tuned to other populations. The aggregation rule is a term of federation membership and should be negotiated like one.

9.2 Application settings

The framework was designed for, and transfers directly to, several settings beyond the cohorts studied here. In multi-hospital surveillance for deterioration and sepsis, uneven modality availability and heterogeneous case mix are the rule rather than the exception. In chronic disease management supported by wearables and IoMT telemetry, the edge budget reported in Section 6.6 allows continuous on-device scoring with intermittent connectivity, while the conformal layer bounds the miss rate for each monitored subgroup. In population screening programs, subgroup-conditional thresholds offer a way to turn an equity commitment written into policy into something the software actually enforces. In low-resource and rural networks, the equity-regularized aggregation rule works directly against the tendency of federated collaboration to concentrate its benefits in large centers.

9.3 Implementation checklist

- Agree on the protected subgroups and the target level α with clinical stakeholders before any modeling starts, and record them in the study protocol.
- Estimate the modality-missingness profile of each site and use it to configure the mask curriculum, instead of assuming uniform random missingness.
- Set aside a calibration split at every site that is disjoint from both the training and test data, and check that each subgroup reaches the minimum number of positives needed for certification.
- Version thresholds separately from weights, log the active threshold with every prediction, and define escalation criteria before go-live.
- Report disaggregated metrics at every internal review, and treat any uncertified subgroup as an explicit gap in the safety case rather than as an omission.

10. Limitations and Future Work

10.1 Limitations

The claims of this study are bounded by several limitations. The conformal guarantee is marginal within subgroups and requires calibration and test samples from the same subgroup to be exchangeable. It does not give conditional validity at the individual level, which is known to be unattainable without distributional assumptions [27], and it weakens under abrupt covariate shift unless adaptive variants are used [28]. Subgroups must be specified in advance and sampled adequately. Intersectional strata that are small at every site cannot be certified, even though these are often the populations most at risk, and this remains an important open problem. The equity weights depend on the Δk statistics that clients report. These scalars are protected by the same

aggregation machinery as the gradients, but a client that misreports strategically could still gain undue influence over aggregation. The differential-privacy noise needed for formal guarantees costs accuracy and, less obviously, calibration in small subgroups; this interaction is not fully characterized here. Finally, and most importantly, the evaluation was retrospective and computational. A retrospective study, however carefully disaggregated, cannot show that a system improves patient outcomes. That requires prospective, preferably randomized, evaluation within the intended clinical workflow.

10.2 Future work

- Robust aggregation when Δk reports are noisy or adversarial, including verifiable computation of the calibration deficit that does not reveal subgroup composition.
- Certification of small intersectional strata through hierarchical or shrinkage-based conformal procedures that borrow strength across related subgroups while preserving validity.
- Adaptive conformal procedures for continuous covariate shift, replacing the discrete drift trigger with an online threshold update that maintains long-run coverage.
- A systematic study of how differential-privacy noise interacts with subgroup calibration, since this interaction currently limits the privacy budget achievable in small strata.
- Extension from prediction to prescription by linking the calibrated risk estimate to treatment-effect estimation, so that the guaranteed quantity concerns benefit rather than detection.
- Prospective silent-mode deployment, followed by a randomized evaluation of how subgroup-conditional thresholding affects clinical outcomes and workload.

11. Conclusion

This paper introduced PRISM-Fed, a federated multimodal framework for clinical prediction in which equity is optimized through the aggregation rule and guaranteed at the decision threshold, instead of being handled only as a penalty in the loss. Three linked mechanisms were evaluated. Missingness-aware gating conditions cross-modal attention on an explicit observability mask, so that systematic absence of a modality is encoded rather than imputed away. Equity-regularized federated aggregation weights each institution by the calibration deficit of its worst-served subgroup and keeps the standard convergence rate up to a bounded dispersion constant. Subgroup-conditional conformal risk control provides a distribution-free, finite-sample sensitivity guarantee in every audited subgroup and restores it under drift by refitting thresholds alone. In a multi-institution evaluation, the framework raised worst-subgroup AUROC by 0.100 over federated averaging while raising average AUROC by 0.033. It lowered the expected calibration error from 0.081 to 0.024, the equalized-odds gap from 0.142 to 0.038 and the fraction of subgroups below the target sensitivity from 0.42 to 0.03. It also converged in about 35% fewer communication rounds and, after distillation, ran in 96 ms at 0.21 J per inference. Across these results, the gains are concentrated where clinical harm is concentrated, and most of them cannot be seen in average metrics. More broadly, the equity of a clinical prediction system appears to depend less on its loss function than on two operators usually treated as infrastructure: the rule that combines institutional models and the rule that turns a score into a decision. Making the first sensitive to worst-subgroup calibration, and making the second subgroup-conditional and conformally calibrated, turns fairness from a claim made at training time into a property that can be checked at the bedside. We therefore recommend that disaggregated reporting, governed thresholds and equity-aware aggregation be treated as defaults rather than optional refinements.

Declaration

Funding

The authors received no financial support for this research.

Competing interests

The authors declare no competing interests.

Ethical approval

Not applicable. This study used only de-identified data from publicly accessible repositories, obtained under their respective data use agreements.

Consent for Publication

The authors consent to the publication of this work.

Data Availability Statement

The datasets analyzed in this study were obtained from publicly available, widely used benchmark clinical resources, including electronic health record (EHR) and multimodal clinical data repositories. They can be accessed through their official repositories and are commonly used in research on clinical prediction, federated learning and medical artificial intelligence. Any dataset-specific requirements, such as data use agreements or credentialed access, must be met through the respective repository providers. The dataset supporting this study is available from the Zenodo repository at: https://doi.org/10.5281/zenodo.21547673.

Code Availability Statement

The source code, preprocessing modules and implementation of the proposed framework, including the federated learning pipeline, the missingness-aware gating mechanism, the equity-regularization algorithms, subgroup-conditional conformal risk control and the experimental evaluation procedures, are available from the corresponding author on reasonable request for academic and research use. The implementation is intended for non-commercial research only and may be shared subject to institutional policies and applicable data-sharing agreements. The corresponding author can be contacted at psenthilexcel@gmail.com.

Author Contribution

Sarojini Rani conceived the study and drafted the manuscript. Senthil Kumar P supervised the research, contributed to data analysis and interpretation, and critically reviewed the manuscript. Both authors examined the results and approved the final version.

Acknowledgements

I would like to express my sincere gratitude to my research supervisor, Senthil Kumar P, for his guidance throughout this research. My husband and son supported me at every stage, and I am grateful to them. I also thank my parents for their love and encouragement. We thank B.E.S.T Innovation University, Andhra Pradesh, for providing the facilities, space and research environment needed to carry out this work.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

General Declaration

During the preparation of this manuscript, the authors used Paperpal and Claude (Anthropic) for grammar and language editing only. No AI tool was used to generate data, analyses, results or figures. The authors reviewed and edited all text and take full responsibility for the content of the manuscript.

References

1. B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis,” IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sept. 2018. DOI: 10.1109/JBHI.2017.2767063

2. S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, “Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines,” npj Digital Medicine, vol. 3, Art. no. 136, 2020. DOI: 10.1038/s41746-020-00341-z

3. J. Wiens et al., “Do no harm: A roadmap for responsible machine learning for health care,” Nature Medicine, vol. 25, no. 9, pp. 1337–1340, 2019. DOI: 10.1038/s41591-019-0548-6

4. R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed. Hoboken, NJ, USA: Wiley, 2019. DOI: 10.1002/9781119482260

5. Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, “Recurrent neural networks for multivariate time series with missing values,” Scientific Reports, vol. 8, Art. no. 6085, 2018. DOI: 10.1038/s41598-018-24271-9

6. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” Science, vol. 366, no. 6464, pp. 447–453, 2019. DOI: 10.1126/science.aax2342

7. H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, vol. 54, pp. 1273–1282, 2017.

8. M. J. Sheller et al., “Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data,” Scientific Reports, vol. 10, Art. no. 12598, 2020. DOI: 10.1038/s41598-020-69250-1

9. N. Rieke et al., “The future of digital health with federated learning,” npj Digital Medicine, vol. 3, Art. no. 119, 2020. DOI: 10.1038/s41746-020-00323-1

10. L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, “Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations,” Nature Medicine, vol. 27, pp. 2176–2182, 2021. DOI: 10.1038/s41591-021-01595-0

11. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb. 2019. DOI: 10.1109/TPAMI.2018.2798607

12. A. Vaswani et al., “Attention is all you need,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.

13. M. Ma et al., “SMIL: Multimodal learning with severely missing modality,” Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 3, pp. 2302–2310, 2021. DOI: 10.1609/aaai.v35i3.16345

14. C. Zhang, D. Song, Y. Huang, A. Swami, and N. Chawla, “Deep partial multi-view learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2402–2415, 2022. DOI: 10.1109/TPAMI.2020.3037731

15. G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” Nature Machine Intelligence, vol. 2, no. 6, pp. 305–311, 2020. DOI: 10.1038/s42256-020-0186-1

16. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” Proc. 3rd Conf. Machine Learning and Systems (MLSys), Austin, TX, USA, pp. 429–450, 2020.

17. S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh, “SCAFFOLD: Stochastic controlled averaging for federated learning,” Proc. 37th Int. Conf. Machine Learning (ICML), vol. 119, pp. 5132–5143, 2020.

18. K. Bonawitz et al., “Practical secure aggregation for privacy-preserving machine learning,” Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), Dallas, TX, USA, pp. 1175–1191, 2017. DOI: 10.1145/3133956.3133982

19. M. Abadi et al., “Deep learning with differential privacy,” Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), Vienna, Austria, pp. 308–318, 2016. DOI: 10.1145/2976749.2978318

20. I. Y. Chen et al., “Ethical machine learning in healthcare,” Annual Review of Biomedical Data Science, vol. 4, pp. 123–144, 2021. DOI: 10.1146/annurev-biodatasci-092820-114757

21. M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” Advances in Neural Information Processing Systems (NeurIPS), vol. 29, pp. 3315–3323, 2016.

22. A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” Big Data, vol. 5, no. 2, pp. 153–163, 2017. DOI: 10.1089/big.2016.0047

23. M. Mohri, G. Sivek, and A. T. Suresh, “Agnostic federated learning,” Proc. 36th Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, vol. 97, pp. 4615–4625, 2019.

24. Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, “FairFed: Enabling group fairness in federated learning,” Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 6, pp. 7494–7502, 2023. DOI: 10.1609/aaai.v37i6.25826

25. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” Proc. 34th Int. Conf. Machine Learning (ICML), Sydney, Australia, vol. 70, pp. 1321–1330, 2017.

26. B. Van Calster et al., “Calibration: The Achilles heel of predictive analytics,” BMC Medicine, vol. 17, Art. no. 230, 2019. DOI: 10.1186/s12916-019-1466-7

27. Y. Romano, M. Sesia, and E. J. Candès, “Classification with valid and adaptive coverage,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 3581–3591, 2020.

28. I. Gibbs and E. J. Candès, “Adaptive conformal inference under distribution shift,” Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 1660–1672, 2021.

29. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proc. Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423

30. E. Alsentzer et al., “Publicly available clinical BERT embeddings,” Proc. 2nd Clinical Natural Language Processing Workshop (ClinicalNLP), Minneapolis, MN, USA, pp. 72–78, 2019. DOI: 10.18653/v1/W19-1909

Appendix A. List of abbreviations

Table A1. Abbreviations used in this study.

Abbreviation	**Expansion**
AUROC / AUPRC	Area under the receiver operating characteristic / precision–recall curve
CRediT	Contributor Roles Taxonomy
DP	Differential privacy
ECE	Expected calibration error
EHR	Electronic health record
EO gap	Equalized-odds gap
EqFedAvg	Equity-regularized federated averaging (proposed)
FL / FedAvg	Federated learning / federated averaging
FNR	False-negative rate
GRU / LSTM	Gated recurrent unit / long short-term memory
IoMT	Internet of Medical Things
MAG	Missingness-aware gating (proposed)
PRISM-Fed	Privacy-preserving Risk Inference via Selective Multimodal fusion, Federated
SC-CRC	Subgroup-conditional conformal risk control (proposed)
SMPC	Secure multi-party computation