

A Blockchain-Governed Clinical Decision Support Framework with Federated Learning and Explainability

Yukta Vishnoi
NMIMS University, India

Abstract—Clinical Decision Support Systems (CDSS) are widely used to support diagnosis, risk prediction, and treatment planning. However, many existing systems are limited by institution-centric data handling, weak auditability, and a lack of clear explanations at the point of clinical decision-making. Although blockchain-based federated learning approaches improve the security of model exchange, they often treat consent management, autonomous reasoning, model drift handling, and clinician-focused explainability as separate components. To address this gap, this study proposes CDSS-AgentXAI, a blockchain-enhanced agentic AI framework for multi-institutional clinical decision support. The framework combines smart contract-based consent control, privacy-preserving federated learning, decision-time explainable inference, and drift-aware model adaptation within one auditable pipeline. It specifies a permissioned Hyperledger Fabric layer for consent and audit tracking, FedProx aggregation with differentially private updates, and an agentic decision module that combines reinforcement learning with a neural network. Clinician-facing explanations are specified using SHAP, LIME, and counterfactual reasoning, and ADWIN-based drift monitoring is specified to trigger feedback-driven retraining. The federated learning and privacy layer was evaluated on 67,989 ICU stays from MIMIC-IV v3.1 across three simulated hospital nodes, with seven configurations compared over five random seeds. Test AUC-ROC ranged from 0.883 to 0.896, and FedProx with differential privacy (0.893 ± 0.003) did not differ significantly from the non-private and centralized models.

Keywords—Clinical Decision Support System (CDSS); federated learning; Explainable AI (XAI); Agentic AI

I. INTRODUCTION

Clinical decision support systems are built to reduce diagnostic uncertainty [1], support timely interventions, and standardize clinical reasoning across diverse healthcare settings. In recent years, machine learning, deep learning, and natural language processing have helped CDSS move away from fixed, static rule-based alerts, and toward predictive and adaptive decision support. Still, when these systems shift from passive recommendation to agentic or semi-autonomous clinical AI, the research question changes: the system must do more than predict well. It also needs to operate under valid consent, explain its recommendations, adapt when data change, and maintain a verifiable audit record [2].

A conventional CDSS typically operates within a single hospital information system. Patient records are stored centrally. Fixed rules or black-box models produce alerts, and clinicians either accept or ignore the output. This structure is inadequate for cross-institutional learning and weak for

patient-centered consent. When patient consent is withdrawn, the AI workflow may not automatically reflect the updated status. When models are retrained, the provenance of the model update may not be recorded in a manner that can be independently verified. When hospital-specific data distributions shift, the model may continue to operate without recalibration.

CDSS-AgentXAI addresses this need by designing the CDSS as a governed workflow instead of an isolated predictor. In the proposed design, a ConsentGate smart contract checks permission before the agent queries patient features. SHAP, LIME, and counterfactual evidence are produced inside the decision loop before a recommendation is shown. Federated aggregation records model provenance on-chain before a model update is accepted. Drift monitors trigger retraining when performance degrades. The architecture links each recommendation to four verifiable elements: consent status, model version, explanation evidence, and clinician feedback. This study presents the framework as a conceptual architecture and reports the experimental validation of its federated learning and differential privacy layer on MIMIC-IV.

A. Novelty and Contribution

The novelty of this study lies in the integration and operational ordering of the components. Prior work includes blockchain, federated learning, XAI, and reinforcement learning separately, but the proposed framework makes these elements depend on each other in the clinical decision path. In the proposed design, the AI agent cannot act without valid consent. The clinician does not receive a recommendation without an explanation. The model cannot be updated without a provenance record. Clinician feedback is used for drift-aware retraining instead of being stored passively. Fig. 1 shows the conceptual difference between the traditional institution-centric CDSS and the proposed blockchain-augmented agentic CDSS, in which consent enforcement, explainability, auditability, scalability, and adaptive decision making are combined.

The contributions of this study are as follows. Items 1), 2), and 4) are architecture-level specifications. Items 3) and 5) are supported by the experiments reported here.

- A patient-centric smart-contract consent layer that verifies consent before every agentic CDSS action and stores tamper-evident audit records is specified.
- A safety-masked agentic module that combines reinforcement learning, a clinical neural network, and

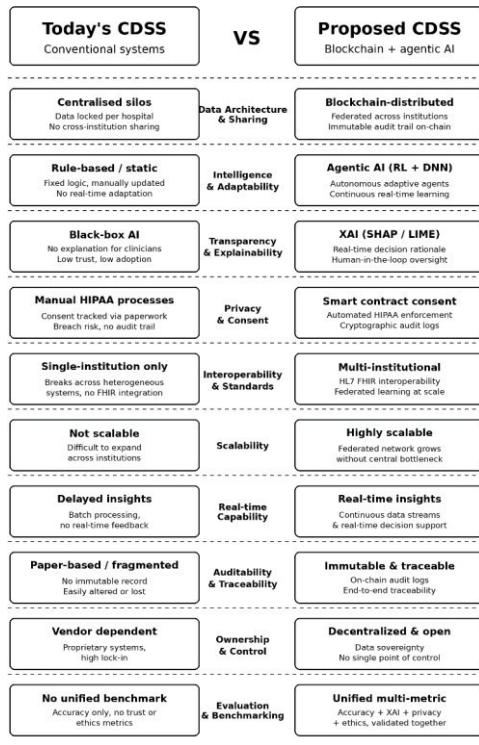


Fig. 1. Comparative positioning of conventional CDSS and proposed blockchain-augmented agentic CDSS.

rule-based constraints is specified, with its state space, action space, reward, and training procedure given in the Agentic AI layer section.

- A federated learning pipeline with FedProx and DP-SGD is implemented and evaluated on MIMIC-IV with five seeds, exact statistical tests, and measured privacy accounting.
- Decision-time SHAP, LIME, and counterfactual explanation is specified. In the experiments, only the agreement between SHAP attributions of the neural network and of XGBoost is measured.
- A reproducible experimental protocol is provided, with the cohort definition, temporal split, node partition, leakage controls, and negative findings reported in full.

II. RELATED WORK

A. Blockchain and Consent Governance in Healthcare

Blockchain-based healthcare systems have been studied for electronic health record sharing [3], access control, auditability, consent management, and model provenance. Permissioned blockchain platforms, such as Hyperledger Fabric [4], are more appropriate for clinical environments than public proof-of-work networks because they support identifiable organizations, private channels, access policies, and configurable consensus. However, many blockchain healthcare frameworks use ledgers as passive records of access or as stores for model hashes [5]. They do not necessarily force the AI decision module to verify consent at the time of clinical action. This distinction

is important because an audit trail can record misuse after it occurs, whereas an operational consent gate prevents the agent from acting when the consent is invalid.

B. Federated Learning in Healthcare

Federated learning allows hospitals to collaboratively train models without exchanging raw patient information. This is important because patient records are protected by legal, ethical, and institutional constraints. Conventional FedAvg [6] performs well when local data are similar, but hospital data are naturally non-IID because institutions differ in patient demographics, coding practices, ICU protocols, and disease prevalence. FedProx [7] adds a proximal term that reduces local divergence from the global model. In clinical CDSS, this stability is necessary because one hospital-specific distribution should not dominate the global model behavior [8-11].

C. Agentic AI for Clinical Decision Support

Agentic AI extends conventional prediction by selecting actions, maintaining the state, responding to feedback, and updating behavior over time. Reinforcement learning, especially PPO [12], is suitable for sequential decision-making environments because it can optimize a policy under controlled updates. However, in healthcare, an RL agent cannot be treated as a general autonomous agent [13]. Its actions may influence treatment and, therefore, must be constrained by consent, safety rules, clinician review, and audit logging. This study treats agentic decision-making as a human-in-the-loop recommendation process, not an unrestricted autonomous clinical actor.

D. Explainable AI in High-Stakes Healthcare

XAI methods such as SHAP [14], LIME [15], and counterfactual explanations are widely used to interpret model predictions. A limitation is that many systems generate explanations post-hoc: prediction occurs first, and the explanation is produced later [16-18]. This is weak for clinical review because an explanation is needed when the clinician evaluates the recommendation. CDSS-AgentXAI treats explanations as part of the decision object. The explanation output was displayed before the clinician's action, and the explanation hash was written to the audit record [19-21].

E. Drift Detection and Continual Adaptation

Clinical data distributions change over time owing to new protocols, seasonal disease patterns, changing demographics, coding changes, and local workflow differences. A static model may degrade silently in the presence of these shifts. Drift detection algorithms, such as ADWIN [22], use adaptive windows over streaming error values to detect statistically meaningful changes. When integrated with federated learning, drift detection triggers local retraining only when statistically significant changes are detected, thereby reducing unnecessary computation while improving responsiveness to evolving clinical patterns.

F. Summary of Related Work

Table I presents a literature-based positioning and gap mapping of major research streams. It describes how blockchain healthcare, federated learning, clinical XAI, agentic AI, and CDSS evaluation shape the design rationale of CDSS-AgentXAI.

TABLE I. LITERATURE-BASED POSITIONING AND GAP MAPPING

Research stream	Typical contribution	Main limitation	Gap addressed by CDSS-AgentXAI
Blockchain healthcare	Audit logs, consent records, model hashes	Often separate from real-time AI action	ConsentGate blocks the agent before data access
Federated learning	Multi-site training without raw data sharing	FedAvg unstable for non-IID hospital data	FedProx with drift-triggered local retraining
Clinical XAI	SHAP, LIME and counterfactual explanations	Usually post-hoc and optional	Decision-time explanation stored as evidence
Agentic AI/RL	Adaptive action selection	Autonomy can be unsafe without governance	Safety-masked PPO constrained by consent
CDSS evaluation	Accuracy, AUC and usability metrics	No unified governance score	Multi-dimensional evaluation protocol (proposed)

III. RESEARCH GAP AND MOTIVATION

The central research gap is not the absence of individual technologies. Blockchain, federated learning, explainable AI, reinforcement learning, and drift detection are all mature research directions. The gap is instead the absence of a coherent CDSS architecture in which these technologies operate together in the same governed decision loop. Most existing frameworks either focus on privacy-preserving model training, explainability, or auditability, but do not connect consent verification, action selection, explanation, and model provenance into one enforceable workflow. This gap produces five practical weaknesses in the literature. First, a blockchain-FL system may provide model provenance but still allow a decision module to operate without obtaining fresh patient consent. Second, an XAI-enabled model may explain decisions only after the recommendation has been shown or acted upon. Third, a federated model may aggregate data across hospitals but remain static under clinical drift. Fourth, many CDSS papers report accuracy and AUC but omit trust, fairness, and compliance, which are essential for real-world, high-stakes deployment. Fifth, audit records seldom store the explanation and clinician feedback behind a decision, so the decision cannot be fully reconstructed.

The motivation for CDSS-AgentXAI is to transition from a model-centered CDSS toward an evidence-centered CDSS. A recommendation should not be treated as just a probability score. Instead, it should be a checkable decision object that includes the patient consent state, the active model version, the clinical features that influenced the decision, the explanation evidence, the clinician response, and the next learning signal. Table II presents the research gap and motivation matrix, and it highlights what is missing or limited in current CDSS work. Each limitation is then connected to the relevant design response, inside the proposed framework, so the mapping is easier to follow.

TABLE II. RESEARCH GAPS AND MOTIVATION MATRIX

Gap ID	Observed limitation	Research implication	Proposed response
G1	Consent is stored but not enforced in the AI loop	The agent may act despite missing or expired consent	ConsentGate smart contract checks every action
G2	XAI is mostly post-hoc	Clinician receives prediction without timely explanation	In-loop SHAP, LIME and counterfactuals
G3	Federated models are often static	Clinical drift reduces reliability over time	ADWIN-triggered retraining and provenance logging
G4	Evaluation focuses on accuracy only	Trust, privacy, fairness and latency are under-reported	Multi-dimensional evaluation protocol (proposed)
G5	Audit records lack explanation and feedback evidence	Decisions cannot be reconstructed fully	AuditLog stores model, consent, explanation hash and feedback

IV. PROPOSED METHODOLOGY

The proposed methodology is an end-to-end clinical AI pipeline in which each stage depends on the preceding governance check. The pipeline starts with collecting multisource clinical data and performing privacy-preserving preprocessing. Consent and audit checks are performed before feature engineering and model training. Local hospital nodes train models without sharing raw records, so the data stay in place. Encrypted and differentially private model updates are aggregated using FedProx. The global model is then loaded into an agentic decision module that generates recommendations under safety constraints. Drift detection, alert calibration, explanation, clinician feedback, and blockchain-secured deployment complete the cycle. Fig. 2 shows the high-level flow. The experiments in this study cover the federated training stage.

A. Data Acquisition and Preprocessing

The implemented pipeline uses structured MIMIC-IV tables only: ICU stays, admissions, patients, laboratory events, and chart events. MIMIC-IV is de-identified at source. Missing values were imputed with training-set medians, and features were standardised with training-set means and variances. The fuller preprocessing described in the conceptual design (isolation-forest outlier detection, KNN and MICE imputation, dynamic resampling, diagnoses, procedures, medications, imaging summaries, and clinical notes) was not used in the experiments. Class imbalance was not corrected by resampling. The decision threshold was selected on the validation set.

B. Consent-Aware Clinical Feature Construction

In the proposed design, the patient-state vector is built only after consent verification, so that unauthorised feature access is prevented by the pipeline and is not left to administrative procedure. The pipeline takes structured and unstructured variables only when the smart contract returns a valid consent state, and every feature transformation is logged to support reproducibility and later audit.

Eq. (1) defines the patient state vector, which maps structured and unstructured clinical variables into one feature representation for the downstream models. In the experiments, only the demographic, vital-sign, and laboratory terms were used.

$$s_i = [x_{\text{demographic}}, x_{\text{vitals}}, x_{\text{labs}}, x_{\text{diag}}, x_{\text{proc}}, x_{\text{note}}] \quad (1)$$

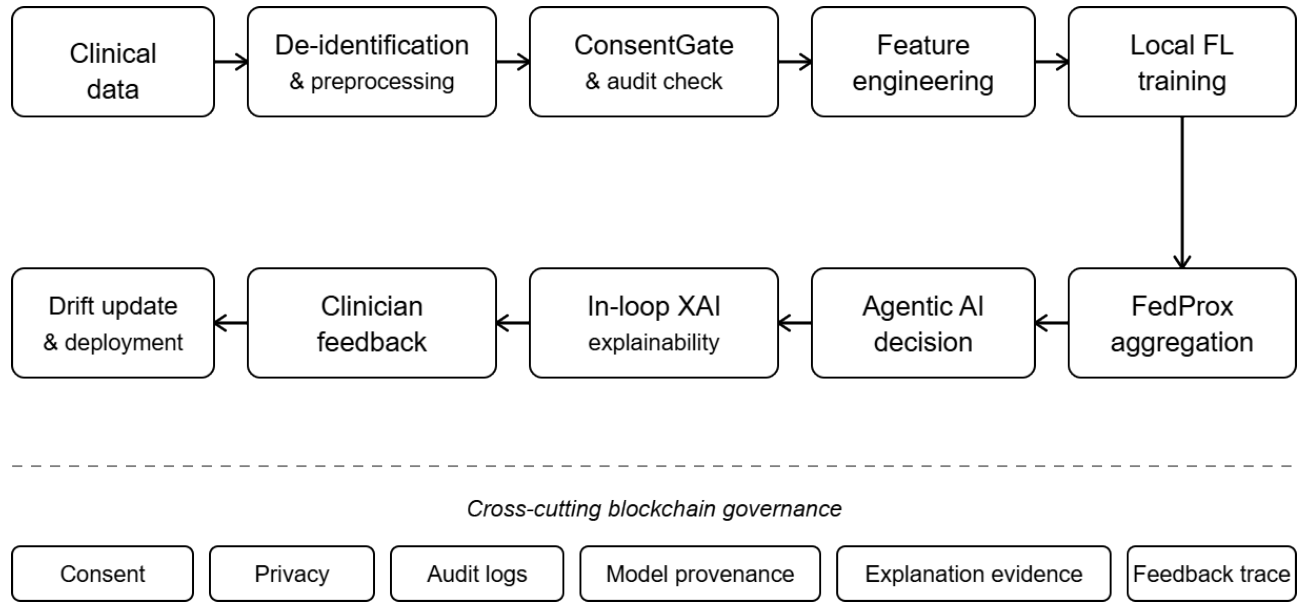


Fig. 2. High-level methodology flow of the blockchain-augmented agentic CDSS framework.

TABLE III. METHODOLOGICAL WORKFLOW OF CDSS-AGENTXAI

Stage	Input	Processing	Output
Preprocessing	Raw EHR tables	De-identification, imputation, outlier handling	Patient-state vectors
Consent gate	Patient ID, action and time	Smart-contract verification	Allow/deny signal
Local training	Node-local data	Supervised and RL-based training	Local model update
Aggregation	Encrypted DP updates	FedProx aggregation	Global model version
Decision	Patient-state vector	PPO, DNN, safety mask	Recommendation candidate
Explanation	Prediction context	SHAP, LIME, counterfactuals	Explanation evidence
Audit and feedback	Clinician action	AuditLog and drift update	Retraining signal

C. Federated and Agentic Decision Workflow

A model is trained locally at each institution, and only privacy-preserving model updates are shared. The agentic decision module uses the global model to generate candidate actions. Each action passes through three gates: consent, safety, and explanation. The recommendation is shown to the clinician only after it is consent-compliant, clinically constrained, and accompanied by explanatory evidence. This workflow is an architecture-level specification. The experiments in this study cover the local training and aggregation stages.

The clinician's response is written to the audit layer and reused as feedback for later updates. Table III summarises the workflow of CDSS-AgentXAI: the input, the processing, and the output of each stage, from preprocessing to audit logging and feedback-based retraining.

D. Algorithmic Workflow

Algorithm 1 summarises the operating logic of the proposed framework. Governance is placed inside the computational workflow: consent verification is a blocking condition and not an administrative note. An explanation is not a report generated after an event. It is created before the clinician's response.

Model provenance is not a manual version number. It is a cryptographic ledger entry associated with recommendations.

V. SYSTEM ARCHITECTURE

The system architecture is divided into seven layers: data layer, consent and audit layer, feature layer, federated learning layer, agentic decision layer, explainability layer, and feedback-driven deployment layer.

Algorithm 1 Governance-Embedded Federated Agentic CDSS

Require:

Hospital nodes $H = \{h_1, \dots, h_n\}$; patient records R ; consent registry C ; global model M ; audit ledger L

Ensure:

Consent-gated recommendations; cryptographic provenance; clinician-facing XAI; drift-aware retraining

```
1: for each hospital node  $h_i \in H$  do
2:   De-identify raw clinical records; encode as patient-state vectors  $x_p$ 
3:   for each query (patient  $p$ , action  $a$ , timestamp  $t$ ) do
4:     Execute ConsentGate( $p, a, t$ )  $\triangleright$  blocking condition
5:     if consent invalid then
6:       Log denial metadata to  $L$ ; terminate query  $\triangleright$  no features exposed
7:     end if
8:     Train local model on node-local data  $D_i$  with supervised + policy objectives
9:     Clip gradients; add differential privacy noise  $\epsilon$ ; encrypt weight updates  $\Delta w_i$ 
10:    Transmit  $\Delta w_i$  to aggregation server
11:  end for
12: end for
13: Aggregate  $\{\Delta w_i\}$  via FedProx; produce global model  $M^*$ 
14: Write ( $\text{hash}(M^*)$ , timestamp, participating nodes) to ModelProvenance  $\triangleright$  immutable ledger
15: Load  $M^*$  into agentic decision module; generate candidate clinical actions  $A$ 
16: Apply consent, safety, and contraindication masks to  $A$   $\triangleright$  pre-display filtering
17: Generate SHAP, LIME, and counterfactual explanations  $E$  inside decision loop  $\triangleright$  not post-hoc
18: Display (recommendation, confidence,  $E$ , alternatives) to clinician
19: Record (clinician response,  $\text{hash}(E)$ , decision metadata) in AuditLog  $L$ 
20: Update drift monitor; if error distribution shifts then trigger local retraining end if
```

This design separates clinical data governance from model training and decision delivery while keeping them connected through auditable events. Raw patient records stay at the local site. The blockchain stores only metadata, consent states, model hashes, decision hashes, and explanation hashes, so no clinical data are written to it. Fig. 3 shows the system architecture: clinical data acquisition, blockchain-based consent and audit control, feature engineering, federated aggregation, agentic decision making, drift detection, explainability, evaluation, and the secure deployment layers.

A. Blockchain Consent and Audit Layer

The `ConsentGate` contract checks whether the patient has given permission for the requested action, at the requested

time, and for the requesting institution. If consent is not valid, the query is stopped and a denial metadata entry is logged. If consent is valid, the agent receives the patient-state vector. The `AuditLog` contract records the model version, the timestamp, a pseudonymised patient identifier, a clinician identifier, the action type, the consent state, and an explanation hash. The `ModelProvenance` contract stores cryptographic hashes of the aggregated global model versions. These contracts are specified at the architecture level. Their deployment and measurement are planned.

B. Agentic AI Layer

The agentic layer uses a PPO-based policy network that selects a recommendation from a constrained action set. The



Fig. 3. Detailed system architecture showing preprocessing, blockchain audit, federated aggregation, agentic decision, drift detection, XAI and deployment.

specification below defines the agentic module at the architecture level. Its training and evaluation are planned as future work. State: the patient-state vector s_t (21 features in the present cohort), the risk probability and confidence of the global neural network, and the consent flag. Action space: three discrete alert-priority actions, namely no alert, routine review, and urgent review. Transitions: each patient episode is a single decision step in the retrospective setting, so the problem reduces to a contextual bandit unless a multi-step monitoring task is defined. This should be stated if the policy is later trained. Safety mask: a rule-based mask removes the no-alert action when the network risk probability exceeds the validation-selected threshold, and removes every action when consent is invalid. Clinician feedback: accept or override is coded as +1 or -1 in the clinician term of the reward. Reward: Eq. (8), whose weights $(\alpha, \beta, \gamma, \lambda)$ are not fixed here and would be selected on validation episodes before any training. Interaction with the neural network: the network is trained first and frozen. The policy takes its output as part of the state, so the policy cannot alter the predictive model. Training: PPO with clip coefficient 0.2 on retrospective episodes with simulated clinician feedback. Because the policy chooses among a masked set of three alert levels on top of a fixed risk model, its added value over a thresholded risk score

is to be tested directly by removing the policy and the mask in an ablation, which is planned.

C. Human-in-the-Loop and Feedback Layer

The clinician remains the final decision-maker. The system shows the recommendation together with the risk score, the confidence, the main SHAP signals, LIME local evidence, and counterfactual alternatives. The clinician can accept the output, reject it, or request a different explanation path. Overrides are treated as learning signals and are tracked through prediction error so that drift can be monitored. This closes the loop between clinical judgment and model adaptation in a measurable and auditable way. Fig. 4 shows the data flow: local clinical data stay within hospital nodes, while federated learning, blockchain audit logging, XAI generation, clinician review, and feedback-driven retraining form a governed decision cycle.

D. Module-Level Architecture Specification

Table IV gives the module-level architecture specification. It lists the main duty of each system layer and the failure mode it is meant to prevent.

Each layer is mapped to a failure mode that is commonly discussed for high-stakes healthcare AI: privacy failure, expla-

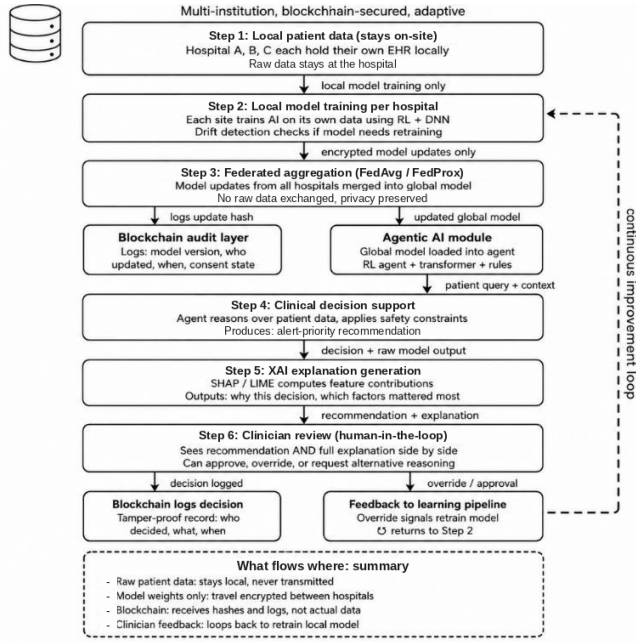


Fig. 4. Data flow of the proposed CDSS showing local data, federated learning, blockchain audit, XAI and feedback retraining.

TABLE IV. MODULE-LEVEL ARCHITECTURE SPECIFICATION AND FAILURE PREVENTION ROLE

Layer	Primary responsibility	Failure prevented
FHIR/data adapter	Maps clinical tables, notes and observations into a common representation	Schema mismatch and feature leakage
ConsentGate contract	Authorises or blocks agent access using current consent state	Unauthorised patient data use
Local trainer	Learns from institution-local data without raw sharing	Cross-hospital privacy exposure
FedProx aggregator	Combines updates while limiting non-IID drift	Dominance by one hospital distribution
Agentic policy	Generates candidate recommendation actions	Static and non-adaptive alerts
Safety mask	Blocks clinically unsafe or out-of-scope actions	Uncontrolled autonomous behaviour
XAI engine	Generates decision-time SHAP/LIME/counterfactual evidence	Black-box recommendation
Audit logger	Stores consent, model, decision and explanation hashes	Non-reconstructable clinical decision
Feedback monitor	Routes overrides and errors to drift detection	Silent model degradation

nation failure, drift failure, audit failure, and autonomy failure. The experiments in this study cover the federated learning layer. The remaining layers are specified at the architecture level.

VI. MATHEMATICAL FORMULATION

This section formalises the main components of the framework and shows how governance, prediction, explanation, and adaptation are connected. Eq. (1) to (3), (6) to (8), (10) to (12) describe the conceptual design. Eq. (4), (5), (9), (13), and (14) correspond to components used in the experiments.

A. Consent-Constrained Action Rule

Eq. (2) gives the consent constrained access mechanism, and it specifies if a requested clinical action is acceptable, depending on patient identity, the action category, and the consent status tied to the specific time window.

$$C(p, a, t) = \begin{cases} 1 & \text{if } \text{consent}(p, a, t) \text{ is valid} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Eq. (3) captures the agent action authorization rule, which restricts the AI agent to proceed only after the consent function yields a valid authorization condition.

$$\text{AgentAction}(p, a, t) = \text{allowed only when } C(p, a, t) = 1 \quad (3)$$

Here, p denotes the patient, a denotes the requested action and t denotes the time of action. The agent receives the patient-state vector only when the consent function returns one.

B. FedProx Aggregation

$$\min_w \left\{ \frac{1}{K} \sum_{k=1}^K F_k(w) + \frac{\mu}{2} \|w - w_{\text{global}}\|^2 \right\} \quad (4)$$

Eq. (4) is the FedProx objective, which stabilises federated learning under non-IID data by penalising deviation of the local model from the global model.

FedProx [7] limits the drift of local models from the global model. In the experiments, $\mu = 0.01$ was used and FedAvg corresponds to $\mu = 0$. The original FedProx study evaluated non-clinical benchmarks and does not give a clinical recommendation for μ . The value was therefore fixed in advance as a small coefficient and was not tuned on the test set. A sensitivity run with DP-SGD on (seeds 42, 7, and 13, with 10 rounds) gave mean AUC values of 0.8938, 0.8937, 0.8937, 0.8937, and 0.8928 for $\mu = 0, 0.001, 0.01, 0.1$, and 1. The proximal coefficient therefore had no measurable effect on AUC between 0 and 0.1 under DP, and $\mu = 1$ did not improve it.

C. Differential Privacy Update

$$\tilde{g} = \frac{1}{B} \left[\sum_i \text{clip}(g_i, C) + \mathcal{N}(0, \sigma^2 C^2 I) \right] \quad (5)$$

Eq. (5) is the DP-SGD gradient used for local training: per-sample gradients g_i of a batch of expected size B are clipped to norm C , Gaussian noise is added to their sum, and the result is passed to the optimiser. The experiments used $C = 1.0$ and $\sigma = 1.1$ with Poisson batch sampling and a Rényi-DP accountant per node composed over all communication rounds. The privacy budget was measured and not targeted: after 20 rounds of 5 local epochs, the worst-case node reached $\epsilon = 3.84$ at $\delta = 10^{-5}$.

D. PPO Clipped Objective and Safety-Masked Action

$$L_{PPO}(\theta) = \mathbb{E}_t [\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) A_t)] \quad (6)$$

Eq. (6) is the PPO clipped surrogate objective, which supports controlled policy updates in the agentic decision module.

$$a_t = \pi_\theta(s_t) \odot M_{safety} \odot C(p, a, t) \quad (7)$$

Eq. (7) represents the safety and consent masked action formulation, which makes sure the agent's recommendation stays within clinical safety constraints and only uses actions consistent with valid patient consent.

The candidate action is multiplied by a safety mask and a consent mask before it can be displayed. This expresses governance as part of the model output rather than as a policy statement outside the model.

E. Reward Function

$$R_t = \alpha R_{accuracy} + \beta R_{clinician} - \gamma R_{risk} - \lambda R_{unexplained} \quad (8)$$

Eq. (8) is the reward of the agentic module: a combination of predictive accuracy, clinician feedback, a risk penalty, and an explanation penalty. The weights will be selected when the policy is trained.

The reward is not based on accuracy alone. It also includes clinician acceptance, a risk penalty, and an explanation penalty, which discourages high-confidence recommendations that are poorly explained.

F. SHAP Feature Attribution

$$\varphi_i(f, x) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)] \quad (9)$$

Eq. (9) represents the SHAP feature-attribution setup, estimating the contribution of every clinical feature, to a specific recommendation for the individual.

SHAP values estimate how much each feature contributes to a specific recommendation. In this framework, explanation values are generated before the clinician response and are hashed into the audit trail.

G. ADWIN Drift Detection

$$\text{drift} = \text{TRUE} \text{ if } |\mu(W_0) - \mu(W_1)| \geq \sqrt{\frac{1}{2m} \ln\left(\frac{4n}{\delta}\right)} \quad (10)$$

Eq. (10) represents the ADWIN drift-detection condition, it is used to spot statistically meaningful variations in the prediction-error stream, and to decide when to trigger adaptive retraining.

H. Composite Evaluation Score

$$S_{CDSS} = 0.30ACC + 0.20XAI + 0.20TRUST + 0.15SEC + 0.10ETHICS + 0.05LAT \quad (11)$$

Eq. (11) defines an exploratory composite score. Its weights are author-chosen, its components (accuracy, explanation fidelity, trust, security, ethics, latency) have different scales and meanings, and the consent and audit components are defined only for the proposed system, so a comparison with baselines is not meaningful. It should be used only as an exploratory summary after a weight-sensitivity analysis, and individual metrics should always be reported alongside it.

I. Combined Optimisation View

For the complete CDSS-AgentXAI framework, the learning objective can be written as a constrained optimisation problem in which predictive loss is minimised while privacy, consent, and explanation constraints remain active. This formulation is conceptual: the terms were not optimised jointly in the experiments, which trained the prediction model with the binary cross-entropy loss only.

$$L_{total} = L_{pred} + \lambda_1 L_{PPO} + \lambda_2 L_{drift} + \lambda_3 L_{fairness} + \lambda_4 L_{explanation} \quad (12)$$

Eq. (12) represents the overall optimization target of CDSS-AgentXAI, integrating supervised prediction loss, policy optimization, drift adaptation, fairness regularization, and explanation quality in one objective. Here, L_{pred} represents supervised clinical prediction loss, L_{PPO} represents the policy optimisation term, L_{drift} penalises degradation under distribution shift, $L_{fairness}$ penalises subgroup imbalance and $L_{explanation}$ penalises low-fidelity or unstable explanations. Consent is treated as a hard constraint rather than a soft penalty.

J. Computational Complexity

The computational cost of the framework is the sum of local training, federated aggregation, inference, explanation, and blockchain logging. If H is the number of hospital nodes, M the model size, T the number of trees or explanation samples, and d the patient-state dimension, aggregation costs $O(HM)$ and inference for a two-layer network of width h costs approximately $O(dh + hZ)$. The cost of explanation depends on the explainer. Blockchain logging adds a constant computation per decision on the model side, but network finality adds waiting time in practice.

VII. DATASET DESCRIPTION

Experiments use MIMIC-IV v3.1 [23], a de-identified database of admissions to a single tertiary hospital in Boston. Only structured tables were used: ICU stays, admissions, patients, laboratory events, and chart events. Diagnosis codes, procedures, medications, clinical notes, and ClinicalBERT embeddings [24], which build on BERT [25], belong to the conceptual design and were not used in the experiments. Table V lists the input variables. The pipeline produced 21 features in total, each computed as the mean of all measurements recorded inside the first 24 hours of the ICU stay.

TABLE V. INPUT VARIABLES USED IN THE EXPERIMENTS

Dataset component	Variables	Processing role
Demographics and admission	Age, sex, admission-type indicators	Baseline risk
Laboratory values	Sodium, potassium, creatinine, haemoglobin, white cell count, lactate (24-hour mean)	Physiological state
Vital signs and neurological status	Heart rate, systolic and diastolic blood pressure, respiratory rate, temperature, SpO ₂ , Glasgow Coma Scale components (24-hour mean)	Deterioration signal

TABLE VI. COHORT DEFINITION AND COUNTS

Design item	Value
Source ICU stays	94,458 stays, 66,444 patients, 85,242 admissions
After age filter (≥ 18 years)	94,458 stays
After minimum stay of 24 hours	73,900 stays
After first ICU stay per admission (final cohort)	67,989 stays, 55,209 patients, 67,989 admissions
Outcome	Death 24–72 hours after ICU admission: 2,703 (3.98%)
Split (stays / deaths)	Train 47,663 / 1,939, validation 11,608 / 394, test 8,718 / 370
Observation window and prediction horizon	24 hours and 48 hours
Missing-value handling	Training-set median imputation and standardisation
Leakage prevention	Features from the first 24 hours only, temporal split, patient-level separation verified, and imputation, scaling, and threshold fitted without test data

A. Cohort Construction

The cohort unit is the first ICU stay of each hospital admission. Adults (age ≥ 18 years) whose ICU stay lasted at least 24 hours, which is the length of the observation window, were included. Stays shorter than 24 hours, which include most early deaths, are therefore excluded. The prediction task is binary: death during the hospital admission between hour 24 and hour 72 after ICU admission, that is, within a 48-hour horizon that follows the 24-hour observation window. Input features use only measurements recorded inside the observation window. The split is temporal: patients with anchor year groups 2008–2016 form the training set, 2017–2019 the validation set, and 2020–2022 the test set. MIMIC-IV assigns one anchor year group to each patient, so no patient appears in more than one split. This was verified programmatically. Imputation medians and standardisation statistics were computed on the training set only, and the decision threshold was selected on the validation set only. The test period overlaps the COVID-19 pandemic, so the test set is expected to contain a distribution shift relative to the training period. Table VI summarises the cohort.

B. Multi-Institution Partition Design

Three virtual hospital nodes (N1, N2, N3) were simulated. First care units were mapped to three care groups (cardiac, medical, surgical), and each group was assigned a home node. Within each care group and outcome class, patients were allocated to the nodes with proportions drawn from a Dirichlet distribution with concentration $\alpha = 0.5$ plus a home-node bias of 2.0. The partition seed was 42. It was the first seed that met the minimum-size constraints (at least 200 stays per node and at least 100, 20, and 30 deaths per node in the training, validation, and test sets). N1 holds 17,667 stays (mortality 2.99%,

cardiac 8,793, surgical 8,317, medical 557), N2 holds 28,189 stays (3.26%, medical 23,686, cardiac 2,638, surgical 1,865), and N3 holds 22,133 stays (5.67%, surgical 15,843, cardiac 5,985, medical 305). The resulting heterogeneity is moderate: the Jensen–Shannon divergence between nodes computed in the data-preparation step has a mean of 0.031 and a maximum of 0.047.

VIII. EXPERIMENTAL SETUP

Seven configurations are compared (Table VII). The evaluation addresses three questions: what differential privacy costs in discrimination, whether the proximal term of FedProx changes the result, and how federated training compares with local-only and centralized training.

A. Node Partitioning and Training Protocol

Each node trains locally for 5 epochs per round (Adam, learning rate 10^{-3} , batch size 64), and the server aggregates the parameters with weights proportional to local sample size for 20 communication rounds. In the DP configurations, local training uses DP-SGD. All federated runs of one seed start from the same initial global model.

B. Baseline Models and Evaluation Dimensions

Table VII lists the configurations. M1 is the proposed federated configuration. The others isolate individual factors.

TABLE VII. BASELINE MODELS USED IN THE COMPARATIVE EVALUATION

Code	Configuration	Purpose of comparison
M1	FedProx + DP-SGD	Proposed federated learning configuration ($\sigma = 1.1$, $C = 1.0$)
M2	FedProx without DP	Isolates the cost of differential privacy
M3	FedAvg + DP-SGD	Isolates the proximal term under DP
M4	FedAvg without DP	Non-private federated baseline
M5	Local-only (one model per node)	No federation
M6	Centralized DNN (same network)	Pooled-data reference for the same network
M7	Centralized XGBoost	Strong tabular reference

The primary metric is AUC-ROC. AUPRC, F1, precision, recall, and specificity are reported at a threshold chosen on the validation set by maximising F1. Explanation fidelity and timing, clinician trust, override rate, latency, blockchain finality, throughput, security, and fairness metrics belong to the planned clinician study and deployment stages.

C. Hyperparameter and Infrastructure Configuration

Table VIII lists the settings of the executed experiments.

D. Statistical Validation Protocol

All metrics are reported over five independent runs with seeds 42, 7, 13, 99, and 2024. The seed affects initialisation, mini-batch order, and DP noise. The cohort partition and the test set are fixed. Paired t-tests across seeds are the primary test. Wilcoxon signed-rank results are given for completeness. With $n = 5$ its smallest attainable two-sided p-value is 0.0625. Holm correction is applied within each metric over the six

TABLE VIII. SETTINGS OF THE EXECUTED EXPERIMENTS

Component	Parameter	Value
Task	Observation window and horizon	24 hours and 48 hours
Network	ClinicalDNN	21–128–64–1, LayerNorm, dropout 0.3 and 0.2
FedProx	Proximal coefficient μ	0.01 ($\mu = 0$ for FedAvg)
Aggregation	Rule	Sample-size-weighted parameter average
Federated training	Rounds and local epochs	20 and 5
Optimiser	Adam and batch size	Learning rate 10^{-3} and 64
Differential privacy	Clipping norm C , noise multiplier σ , and δ	1.0, 1.1, 10^{-5}
Differential privacy	Accountant	Rényi DP (Opacus), per node, composed over rounds
Centralized and local	Epochs	Up to 30, validation-based model selection
XGBoost	Trees, depth, learning rate	≤ 600 (early stopping 30), 5, 0.05
Seeds	Random seeds	42, 7, 13, 99, 2024
Threshold	Selection	Maximum F1 on the validation set

comparisons of the proposed configuration. Confidence intervals use the t-distribution. A stratified bootstrap with 1,000 resamples of the test set, applied to the first seed, quantifies test-set sampling uncertainty.

$$t = \frac{\text{mean}(d)}{s_d/\sqrt{n}} \quad (13)$$

Eq. (13) is the paired t-test statistic, where d is the per-seed difference between two configurations and $n = 5$.

$$CI_{95} = \text{mean}(x) \pm t_{0.975, n-1} \cdot \left(\frac{s}{\sqrt{n}} \right) \quad (14)$$

Eq. (14) is the t-based 95% confidence interval. With $n = 5$ seeds, $t_{0.975,4} = 2.776$. The normal value 1.96 is not used.

IX. RESULTS AND DISCUSSION

This section reports the executed experiments. All values come from the notebook released with the study. Every configuration was run with five seeds on the same temporal test set (8,718 stays, 370 deaths). Because the cohort partition and the test set are fixed, the standard deviations over seeds reflect training randomness and do not include test-set sampling variation, which is quantified separately with a bootstrap.

A. Clinical Prediction Performance

Table IX reports test performance. F1, recall, and specificity use the threshold selected on the validation set. Accuracy is not reported because the test mortality is 4.2% and accuracy is dominated by the majority class.

All configurations lie between 0.883 and 0.896 AUC. After Holm correction, federated FedProx with DP-SGD (0.8932) did not differ significantly from FedProx without DP (0.8955), the centralized DNN (0.8950), FedAvg with DP (0.8932), local-only training (0.8888), or centralized XGBoost (0.8899) in AUC (Table XII). The centralized DNN is a reference and not a strict upper bound. Without DP, FedProx exceeded FedAvg by 0.0122 AUC ($p = 0.0003$), and FedProx with DP exceeded FedAvg without DP by 0.0098 (Holm-adjusted $p =$

0.033). Under DP, FedProx and FedAvg gave the same AUC to four decimals, so the proximal term had no measurable effect once DP noise was added. FedAvg with DP had a higher AUC than FedAvg without DP (0.8932 versus 0.8833). Clipping and noise may act as regularisation. Local-only training had a lower AUC than the federated DP model by 0.0044 ($p = 0.057$, not significant), but a higher AUPRC (0.2751 versus 0.2503, Holm-adjusted $p = 0.027$). The nodes are large and moderately heterogeneous (mean Jensen–Shannon divergence 0.031), which limits what federation can add. AUPRC values (0.23 to 0.28) are 5 to 6 times the test mortality of 0.042.

B. Discrimination and Test-Set Uncertainty

AUC-ROC measures ranking quality across all thresholds, whereas F1, recall, and specificity depend on one operating threshold. Because the validation-selected threshold changes between seeds, recall varies more across seeds (SD up to 0.067) than AUC (SD at most 0.0040). Test-set sampling uncertainty is larger than both. The stratified bootstrap 95% confidence interval of AUC for the first seed is 0.8777–0.9075 for FedProx with DP, 0.8803–0.9093 for FedProx without DP, 0.8774–0.9085 for FedAvg with DP, 0.8643–0.8967 for FedAvg without DP, 0.8690–0.9029 for local-only, 0.8823–0.9094 for the centralized DNN, and 0.8739–0.9029 for centralized XGBoost (widths 0.027 to 0.034). The largest difference between configurations (0.012) is smaller than these intervals. Paired tests across seeds therefore describe the effect of training randomness on one fixed test set and do not establish differences at population level.

C. Privacy Accounting

Table X reports the DP accounting. DP-SGD was implemented with Opacus using per-sample clipping ($C = 1.0$), Gaussian noise ($\sigma = 1.1$), Poisson sampling at batch size 64, and a Rényi-DP accountant per node composed over all rounds. The privacy budget is a measured outcome: 20 rounds of 5 local epochs gave $\epsilon = 3.84$ at $\delta = 10^{-5}$ for the worst-case node. The value is identical for FedProx and FedAvg because the proximal term depends only on the public global model and adds no privacy cost.

The measured $\epsilon = 3.84$ is a moderate budget. DP lowered the mean AUC of FedProx by 0.0024 (0.8955 to 0.8932, paired t-test $p = 0.136$, Holm-adjusted $p = 0.38$), which is not statistically significant with five seeds. For FedAvg, the DP runs had a higher AUC than the non-DP runs (+0.0099). Clipping and noise act on a model that is small relative to the training set, which may explain the small effect. In the sensitivity run (seeds 42, 7, and 13, with 10 rounds), raising the noise multiplier σ from 0.5 to 1.1 to 2.0 lowered the worst-node ϵ from 23.53 to 2.63 to 1.12 and the mean AUC from 0.8963 to 0.8937 to 0.8884, a decrease of 0.0079 over this range. These ϵ values are lower than the 3.84 of the main experiment because the sensitivity run used 10 rounds instead of 20.

D. Provenance Ledger Simulation

Table XI reports a software simulation of model provenance. After each communication round, the SHA-256 hash of the global model parameters was appended to an in-process

TABLE IX. TEST PERFORMANCE ON THE TEMPORAL TEST SET (MEAN $\pm$ SD OVER FIVE SEEDS, WITH THE 95% CI T-BASED, 4 DEGREES OF FREEDOM)

Configuration	AUC-ROC (mean $\pm$ SD)	95% CI	AUPRC	F1	Recall	Specificity
M1: FedProx + DP-SGD	0.8932 $\pm$ 0.0033	[0.8891, 0.8972]	0.2503 $\pm$ 0.0109	0.3578 $\pm$ 0.0079	0.5168 $\pm$ 0.0668	0.9393 $\pm$ 0.0123
M2: FedProx, no DP	0.8955 $\pm$ 0.0007	[0.8946, 0.8964]	0.2498 $\pm$ 0.0041	0.3704 $\pm$ 0.0076	0.5751 $\pm$ 0.0323	0.9321 $\pm$ 0.0072
M3: FedAvg + DP-SGD	0.8932 $\pm$ 0.0032	[0.8892, 0.8972]	0.2503 $\pm$ 0.0109	0.3589 $\pm$ 0.0072	0.5611 $\pm$ 0.0534	0.9306 $\pm$ 0.0111
M4: FedAvg, no DP	0.8833 $\pm$ 0.0027	[0.8800, 0.8867]	0.2347 $\pm$ 0.0072	0.3544 $\pm$ 0.0057	0.5724 $\pm$ 0.0221	0.9265 $\pm$ 0.0049
M5: Local-only	0.8888 $\pm$ 0.0040	[0.8838, 0.8937]	0.2751 $\pm$ 0.0109	0.3601 $\pm$ 0.0073	0.4962 $\pm$ 0.0393	0.9440 $\pm$ 0.0100
M6: Centralized DNN	0.8950 $\pm$ 0.0009	[0.8939, 0.8961]	0.2542 $\pm$ 0.0043	0.3702 $\pm$ 0.0064	0.5638 $\pm$ 0.0486	0.9343 $\pm$ 0.0100
M7: Centralized XGBoost	0.8899 $\pm$ 0.0021	[0.8872, 0.8925]	0.2548 $\pm$ 0.0059	0.3470 $\pm$ 0.0073	0.5011 $\pm$ 0.0651	0.9385 $\pm$ 0.0142

TABLE X. PRIVACY ACCOUNTING (RÉNYI DP, COMPOSED OVER ALL ROUNDS, WORST-CASE NODE)

Configuration	Clipping norm C	Noise multiplier σ	Rounds $\times$ local epochs	ϵ (worst node, $\delta = 10^{-5}$)
M1: FedProx + DP-SGD	1.0	1.1	20 $\times$ 5	3.84
M3: FedAvg + DP-SGD	1.0	1.1	20 $\times$ 5	3.84

TABLE XI. PROVENANCE LEDGER SIMULATION

Metric	Value	Status	Interpretation
Blocks appended	20 (one per round)	Measured (simulation)	One model hash per communication round
Chain verification	Valid	Measured (simulation)	Recomputed hashes match the stored chain
Mean time per block	0.57 ms	Measured (simulation)	In-process hashing only, excluding network and consensus
Fabric finality, throughput, consent denial, audit completeness	—	Planned	Hyperledger Fabric deployment stage (future work)

hash chain. No consensus, network, finality, or throughput measurement was made.

The result shows only that hash-chaining model versions costs little computation in a single process. It does not support claims about latency, finality, throughput, consent enforcement, or audit completeness on Hyperledger Fabric. These properties remain design objectives and are listed as future work.

E. Federated Training Progression

Fig. 5 shows the mean test AUC after each communication round (shaded band: ± 1 SD over five seeds). Without DP, FedAvg is highest after the first round (0.8966) and declines to 0.8833 at round 20, whereas FedProx stays between 0.8935 and 0.8966 throughout (0.8955 at round 20). The decline of FedAvg is consistent with drift of the local models under heterogeneous nodes. With DP, both methods start lower (0.831 after round 1 and 0.874 after round 2), exceed 0.889 at round 6, and reach 0.893 at round 20. The FedProx and FedAvg curves under DP differ by at most 0.0003 in every round, so they overlap in the figure. The near-maximal AUC of the non-DP models after one round suggests that the prediction task is learned quickly from 21 aggregated features.

The curves in Fig. 5 show only the effect of the number of communication rounds. ADWIN-triggered retraining is part of the architecture specification, and its evaluation is planned.

F. Statistical Comparison

Table XII reports paired tests across the five seeds. With $n = 5$ the smallest attainable two-sided Wilcoxon p-value is

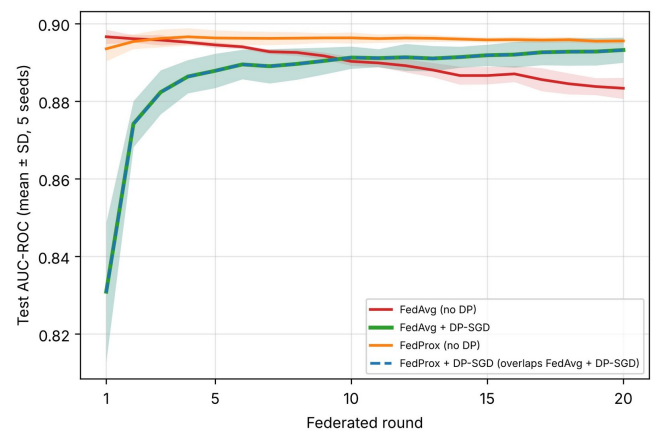


Fig. 5. Test AUC-ROC after each communication round (mean over five seeds, shaded band, ± 1 SD). The FedProx + DP-SGD curve (dashed) coincides with the FedAvg + DP-SGD curve.

0.0625, so the paired t-test with Holm correction is the primary test. The Holm-adjusted p-value is in brackets.

Significance across seeds does not establish a difference at population level, because all seeds share one test set (see the bootstrap intervals above). After Holm correction, two AUC or AUPRC comparisons of M1 remain significant: M1 versus FedAvg without DP on AUC (+0.0098), and M1 versus local-only on AUPRC (local-only higher by 0.0248). Effects of 0.01 AUC or less should be read as small.

X. COMPARATIVE ANALYSIS

This section compares CDSS-AgentXAI with existing CDSS design patterns. The comparison is qualitative and describes design intent, not measured performance. Centralised systems are easier to implement but do not support cross-institution learning or patient-controlled consent at scale. Federated CDSS designs protect raw data but typically lack in-loop explanation and smart-contract consent enforcement. Blockchain-only provenance models are strong on auditability but do not by themselves improve model adaptation or clinical trust. CDSS-AgentXAI combines these aspects in one governed workflow. Fig. 6 shows the traditional centralised CDSS data flow and its failure points: data silos, limited

TABLE XII. PAIRED COMPARISONS OF TEST AUC-ROC ACROSS FIVE SEEDS

Comparison	Metric	Mean difference	Test	Exact p (Holm-adjusted)	Interpretation
M2 versus M4 (FedProx versus FedAvg, no DP)	AUC-ROC	+0.0122	Paired t, n = 5	0.0003 (not adjusted)	FedProx higher without DP
M1 versus M2 (cost of DP)	AUC-ROC	−0.0024	Paired t, n = 5	0.136 (0.384)	No significant cost of DP
M1 versus M3 (proximal term under DP)	AUC-ROC	−0.0000	Paired t, n = 5	0.128 (0.384)	No effect of the proximal term under DP
M1 versus M4 (FedAvg, no DP)	AUC-ROC	+0.0098	Paired t, n = 5	0.0055 (0.033)	M1 higher than FedAvg without DP
M1 versus M5 (local-only)	AUC-ROC	+0.0044	Paired t, n = 5	0.057 (0.285)	Not significant
M1 versus M6 (federation gap)	AUC-ROC	−0.0019	Paired t, n = 5	0.279 (0.384)	No significant gap to the centralized DNN
M1 versus M7 (XGBoost)	AUC-ROC	+0.0033	Paired t, n = 5	0.079 (0.317)	Not significant
M1 versus M5 (local-only)	AUPRC	−0.0248	Paired t, n = 5	0.0045 (0.027)	Local-only higher in AUPRC

control over consent, weak auditability, static rules, and limited explainability.

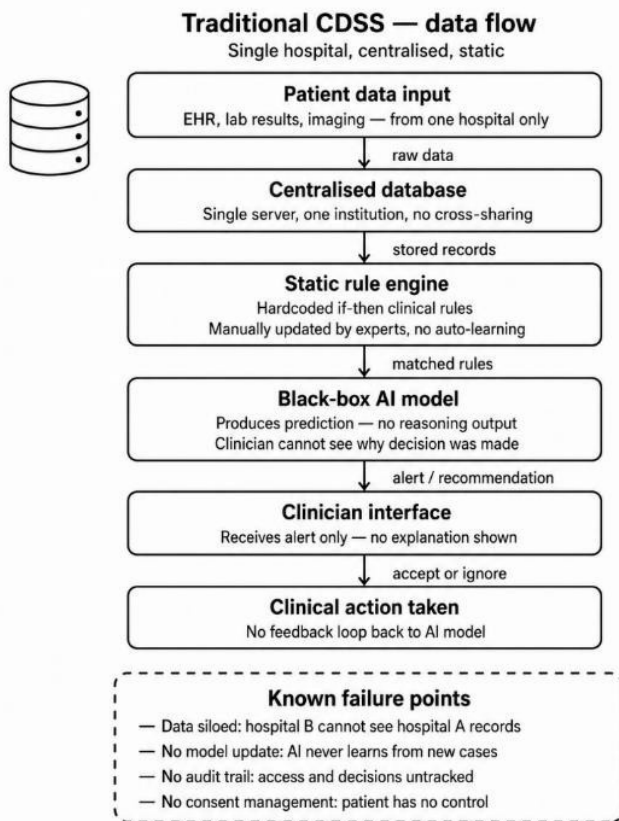


Fig. 6. Traditional centralised CDSS data flow and failure points.

Table XIII compares CDSS-AgentXAI with conventional CDSS and blockchain-FL-only designs with respect to data sharing, consent enforcement, explainability, model updating, auditability, and evaluation scope. The entries describe the design of each approach.

XI. ABLATION STUDY

Table XIV reports the component comparisons of the federated learning and privacy layer: the effect of DP, of the proximal term, and of federation. Ablations of the other components (in-loop XAI, ADWIN, consent contract, policy, and safety mask) belong to the planned validation stages.

TABLE XIII. COMPARATIVE ANALYSIS OF CDSS-AGENTXAI AGAINST CONVENTIONAL AND BLOCKCHAIN-FL DESIGNS

Feature	Conventional CDSS	Blockchain-FL only	CDSS-AgentXAI
Data sharing	Single institution or centralised	Model updates shared	Raw data local DP updates only
Consent enforcement	Manual/documentary	Often separate from model	Smart contract before every action
AI behaviour	Rule-based or static prediction	Federated prediction	Agentic decision with safety mask
Explainability	Absent or post-hoc	Usually limited	In-loop SHAP/LIME/counterfactuals
Model update	Manual retraining	Fixed aggregation	Drift-triggered retraining
Audit trail	Partial logs	Model hashes	Consent, model, decision and explanation hashes
Evaluation focus	Accuracy and alerts	Privacy and aggregation	Accuracy, XAI, trust, privacy, ethics and latency

Without DP, FedProx exceeded FedAvg by 0.0122 AUC (M2 versus M4, $p = 0.0003$). With DP, FedProx and FedAvg were indistinguishable (difference below 0.0001, $p = 0.128$), so the benefit of the proximal term did not persist under DP noise. Removing DP changed the AUC of FedProx by +0.0024 ($p = 0.136$), and removing federation (centralized DNN) by +0.0019 ($p = 0.279$). Neither change is significant with five seeds. These comparisons concern only the federated learning and privacy layer. Ablations for the policy, the safety mask, the consent contract, in-loop XAI, and drift detection are planned.

XII. EXPLAINABILITY ANALYSIS

Explanation timing (in-loop versus post-hoc), fidelity, stability, and clinical usability are planned as future work. In-loop generation of SHAP, LIME, and counterfactual evidence is specified at the architecture level. The executed analysis computed SHAP attributions for the FedProx + DP-SGD network (KernelExplainer, 100 background and 200 explained test samples) and for centralized XGBoost (TreeExplainer), and compared the feature rankings with the Spearman rank correlation. The correlation of mean absolute SHAP values between the two models across the 21 features was 0.829. For the network, the five features with the largest mean absolute SHAP value were SpO2, white cell count, the Glasgow Coma Scale verbal and eye components, and systolic blood pressure. Age ranked eighth and lactate eleventh, and the admission-type indicators ranked low. SHAP values are on the predicted-probability scale (0.0006 to 0.0008 for the top ten features) and

TABLE XIV. EXECUTED COMPONENT COMPARISONS ON TEST AUC-ROC (FIVE SEEDS, PAIRED T-TEST). DIFFERENCES ARE VARIANT MINUS M1

Variant	Configuration	AUC-ROC (mean $\pm$ SD)	Difference from M1	Exact p	Interpretation
M1	FedProx + DP-SGD (reference)	0.8932 $\pm$ 0.0033	—	—	—
M2	FedProx, DP removed	0.8955 $\pm$ 0.0007	+0.0024	0.136	No significant cost of DP
M3	FedAvg + DP-SGD (proximal term removed)	0.8932 $\pm$ 0.0032	−0.0000	0.128	No effect of the proximal term under DP
M4	FedAvg, no DP	0.8833 $\pm$ 0.0027	−0.0098	0.0055	Lower than M1 (Holm-adjusted $p = 0.033$)
M6	Centralized DNN (federation removed)	0.8950 $\pm$ 0.0009	+0.0019	0.279	No significant federation gap

are small because mortality is 4%. A high rank correlation indicates agreement between two model families on feature importance. It does not show that the explanations are faithful or clinically useful.

A. Representative Explanation Case Study

The example below is illustrative and was not generated from the trained model. For a high-risk patient, the system would show the risk tier, the top contributing features, local sensitivity, and counterfactual alternatives. A high-risk deterioration recommendation could be explained by raised lactate, reduced oxygen saturation, and abnormal respiratory rate. A counterfactual explanation could show that the risk tier would decrease if oxygen saturation stabilised and lactate fell. Table XV lists the explanation outputs of the proposed system and their audit treatment.

TABLE XV. EXPLAINABILITY OUTPUTS AND THEIR ROLE IN AUDITABILITY

Explanation output	Purpose	Audit treatment
SHAP feature ranking	Shows global/local feature contribution	Hash stored with decision
LIME local approximation	Shows local sensitivity around patient state	Hash stored with decision
Counterfactual alternatives	Shows what change would alter the recommendation	Summary hash stored
Clinician explanation rating	Captures usability and trust	Stored as feedback meta-data
Override reason	Explains disagreement between clinician and agent	Used for active learning signal

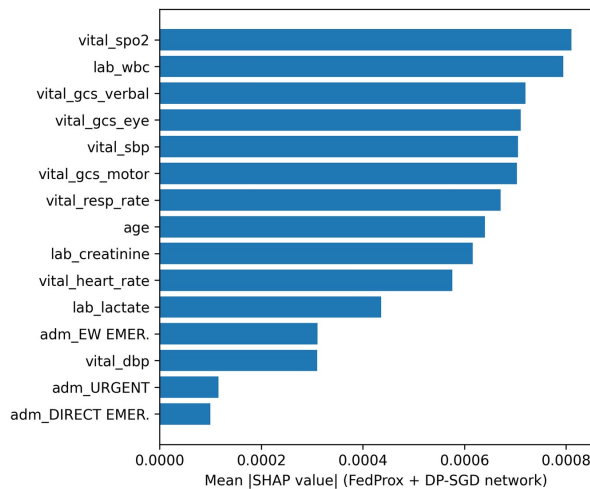


Fig. 7. Mean absolute SHAP values of the 15 features with the largest values for the FedProx + DP-SGD network (200 test samples, 100 background samples).

Fig. 7 shows the 15 features with the largest mean absolute SHAP value for the FedProx + DP-SGD network.

B. Explanation Stability and Audit Trail

Explanation stability across model versions and the consistency of hashed explanations are design goals. In the proposed design, explanation hashes and clinician feedback are stored so that a later audit can check whether explanations remain consistent across model versions.

XIII. SECURITY / PRIVACY / ETHICAL DISCUSSION

A. Security Design

Security is enforced through layered controls. Raw patient records are never written on-chain. Only consent metadata, model hashes, decision hashes and explanation hashes are stored. Data at rest should be encrypted using AES-256 or an equivalent institutional standard, while node-to-node communication should use TLS. Hyperledger Fabric [4] channels and access control lists restrict which organisations can read or write specific ledger states. The agent must re-check consent for every new action rather than relying on a cached approval.

B. Privacy-Preserving Learning

In the proposed design, privacy is protected by local training, encrypted update transmission, gradient clipping, and differential privacy noise. In the experiments, only local training, clipping, and noise were implemented. The privacy budget is the measured $\epsilon = 3.84$ ($\delta = 10^{-5}$, $\sigma = 1.1$, $C = 1.0$, see Table X). This value is the one that applies to the experiments. Differential privacy can reduce model performance. In the experiments, it lowered the mean AUC of FedProx by 0.0024, which was not significant (Table XII). AES-256 and TLS are deployment recommendations and were not part of the experiments.

C. Threat Model

The threat model covers unauthorised patient access, model update tampering that is malicious or accidental, inference leakage from those updates, explanation manipulation, and silent drift. ConsentGate covers unauthorised access, Model-Provenance handles model tampering, differential privacy reduces update leakage, explanation hashing limits post-decision alteration, and ADWIN reduces quiet degradation. That explicit mapping of threats boosts the overall research quality of the security part [26-28].

D. Ethical and Clinical Governance

The clinician remains responsible for the final clinical action. The system provides recommendation support with explanation and traceability [29-33]. Ethical compliance would be reviewed through transparency, human oversight, fairness, accountability, and audit completeness. Subgroup fairness should be measured across clinically meaningful groups. If subgroup performance differs beyond a predefined threshold, the model should be flagged for recalibration before deployment [34][35]. Table XVI maps the main deployment risks to the technical controls and governance mechanisms of CDSS-AgentXAI.

TABLE XVI. SECURITY, PRIVACY AND ETHICAL RISK MITIGATION MATRIX

Risk	Potential harm	Mitigation in CDSS-AgentXAI
Invalid consent	Unauthorised use of patient data	ConsentGate blocks the query
Model drift	Outdated recommendations	ADWIN and local retraining
Black-box output	Low clinician trust	In-loop SHAP/LIME/counterfactuals
Privacy leakage	Reconstruction from updates	DP clipping, noise and encrypted sharing
Unfair performance	Subgroup harm	Fairness metrics and recalibration trigger
Audit failure	Decision cannot be reconstructed	Consent, model, decision and explanation hashes

XIV. LIMITATIONS

This study has the following limitations. First, the experiments cover the federated learning and privacy layer. Validation of the consent contract, the PPO policy and safety mask, ADWIN drift monitoring, Hyperledger Fabric deployment, in-loop explanation, and clinician trust is planned. The value of the policy over a thresholded risk score remains to be tested. Second, the data come from one hospital, and the three nodes are simulated partitions of that data with moderate heterogeneity. This does not represent independent institutions. Third, the models use 21 aggregated features from a 24-hour window and no free text, codes, or temporal sequences, so absolute performance is moderate (AUC about 0.89, AUPRC about 0.25). Fourth, the differences between configurations are small relative to test-set sampling uncertainty, and the seeds share one test set. Fifth, the test period (2020–2022) overlaps the COVID-19 pandemic and may include a distribution shift that was not analysed. Sixth, the privacy budget ($\epsilon = 3.84$) is moderate. Seventh, fairness across subgroups was not evaluated. Eighth, SHAP and LIME can add latency for complex models, and blockchain brings infrastructure, governance, and maintenance requirements.

The conclusions apply to retrospective prediction on MIMIC-IV and do not support claims of clinical readiness, clinician trust, or patient safety. Such claims would need prospective monitoring, clinician studies with ethics approval, fail-safe controls, model rollback, and continuous governance review.

XV. CONCLUSION AND FUTURE WORK

This study presents CDSS-AgentXAI, a conceptual framework for blockchain-governed, agentic, and explainable clinical decision support, and evaluates its federated learning

and privacy layer on MIMIC-IV. The framework links each recommendation to consent verification, model provenance, explanation evidence, and clinician feedback by design.

In the executed experiments, federated training with DP-SGD (measured $\epsilon = 3.84$) reached a test AUC of 0.893, which did not differ significantly from federated training without DP (0.896), the centralized DNN (0.895), or centralized XGBoost (0.890). It exceeded FedAvg without DP (0.883) and did not exceed local-only training in AUPRC. Without DP, FedProx outperformed FedAvg. Under DP the two were indistinguishable. The value of the framework therefore rests on governance, privacy, and traceability by design, which are the subject of the planned validation, and not on gains in predictive accuracy over centralized or local training.

A. Future Work Roadmap

Future work will implement and train the PPO agent with the safety mask and test it against a thresholded risk score, deploy the consent and provenance contracts on Hyperledger Fabric and measure latency and throughput, evaluate ADWIN on the temporal shift of the test period, run a clinician usability study with ethics approval, evaluate fairness across subgroups, and validate externally on other institutions. Table XVII lists these stages.

TABLE XVII. FUTURE WORK ROADMAP FROM THE PRESENT RETROSPECTIVE EVALUATION TOWARD CLINICAL VALIDATION

Phase	Future validation task	Expected evidence
Phase 1	Implement and train the PPO agent and safety mask, and add policy and mask ablations	Policy results over seeds with exact statistics
Phase 2	Deploy ConsentGate and provenance contracts on Hyperledger Fabric	Latency, finality, throughput, consent-denial tests
Phase 3	Run ADWIN drift monitoring on the temporal shift of the test period	Detection delay and retraining effect
Phase 4	Conduct a clinician usability study with ethics approval	Trust scores, acceptance, inter-rater agreement
Phase 5	Validate externally and, after ethics approval, prospectively	External validity and clinical safety evidence

Each phase produces a different type of evidence: policy evidence, deployment evidence, drift evidence, usability evidence, and external or prospective clinical evidence.

REFERENCES

- [1] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, and B. Middleton, "Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality," *J. Am. Med. Inform. Assoc.*, vol. 10, no. 6, pp. 523–530, Nov.–Dec. 2003, doi: 10.1197/jamia.M1370.
- [2] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success," *BMJ*, vol. 330, no. 7494, p. 765, Apr. 2005, doi: 10.1136/bmj.38398.500764.8F.
- [3] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "MedRec: Using blockchain for medical data access and permission management," in *Proc. 2nd Int. Conf. Open Big Data (OBD)*, Vienna, Austria, Aug. 2016, pp. 25–30, doi: 10.1109/OBD.2016.11.
- [4] Hyperledger Foundation, "Hyperledger Fabric Documentation," Linux Foundation, 2023. [Online]. Available: <https://hyperledger-fabric.readthedocs.io>

- [5] S. Vijayalakshmi, Savita, S. P. Gayathri, and S. Janarthanan, "Blockchain security for artificial intelligence-based clinical decision support tool," in *Internet of Things, Artificial Intelligence and Blockchain Technology*, R. L. Kumar, Y. Wang, T. Poongodi, and A. L. Imoize, Eds. Cham, Switzerland: Springer, 2021, pp. 209–240, doi: 10.1007/978-3-030-74150-1_10.
- [6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, vol. 54, pp. 1273–1282. [Online]. Available: <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [7] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst. (MLSys)*, Austin, TX, USA, 2020, vol. 2. Online. Available: https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html
- [8] N. Rieke et al., "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, no. 1, p. 119, Sep. 2020, doi: 10.1038/s41746-020-00323-1.
- [9] M. J. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, p. 12598, Jul. 2020, doi: 10.1038/s41598-020-69250-1.
- [10] U.S. Department of Health and Human Services (HHS), "Health Insurance Portability and Accountability Act (HIPAA) Security Rule," 45 C.F.R. Parts 160 and 164, 2003. [Online]. Available: <https://www.hhs.gov/hipaa/for-professionals/security/index.html>
- [11] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021, doi: 10.1561/22000000083.
- [12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, Jul. 2017, doi: 10.48550/arXiv.1707.06347.
- [13] J. Wiens et al., "Do no harm: A roadmap for responsible machine learning for health care," *Nat. Med.*, vol. 25, no. 9, pp. 1337–1340, Sep. 2019, doi: 10.1038/s41591-019-0548-6.
- [14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 4768–4777. [Online]. Available: <https://dl.acm.org/doi/10.5555/3295222.3295230>
- [15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [16] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
- [17] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. Madai, and the Precise4Q consortium, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 310, Nov. 2020, doi: 10.1186/s12911-020-01332-6.
- [18] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
- [19] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, Feb. 2017, doi: 10.48550/arXiv.1702.08608.
- [20] M. Prosseri, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nat. Mach. Intell.*, vol. 2, no. 7, pp. 369–375, Jul. 2020, doi: 10.1038/s42256-020-0197-y.
- [21] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018. [Online]. Available: <https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>
- [22] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proc. 7th SIAM Int. Conf. Data Mining (SDM)*, Minneapolis, MN, USA, Apr. 2007, pp. 443–448, doi: 10.1137/1.9781611972771.42.
- [23] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Sci. Data*, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.
- [24] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clinical NLP Workshop (at NAACL)*, Minneapolis, MN, USA, Jun. 2019, pp. 72–78, doi: 10.18653/v1/W19-1909.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, vol. 1, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [26] C. Dwork, "Differential privacy," in *Automata, Languages and Programming (ICALP 2006)*, Lect. Notes Comput. Sci., vol. 4052, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Germany: Springer, 2006, pp. 1–12, doi: 10.1007/11787006_1.
- [27] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, Vienna, Austria, Oct. 2016, pp. 308–318, doi: 10.1145/2976749.2978318.
- [28] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nat. Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020, doi: 10.1038/s42256-020-0186-1.
- [29] HL7 International, "HL7 FHIR Release 4 Specification," Health Level Seven International, 2019. [Online]. Available: <https://hl7.org/fhir/R4/>
- [30] National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023. [Online]. Available: <https://doi.org/10.6028/NIST.AI.100-1>
- [31] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-57959-7.
- [32] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council — Artificial Intelligence Act (AI Act)," *Official Journal of the European Union*, L Series, Jul. 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL_202401689
- [33] World Health Organization (WHO), "Ethics and governance of artificial intelligence for health: WHO guidance," Geneva, Switzerland: WHO, Jun. 2021. [Online]. Available: <https://www.who.int/publications/i/item/9789240029200>
- [34] L. A. Celi et al., "Sources of bias in artificial intelligence that perpetuate healthcare disparities — a global review," *PLOS Digit. Health*, vol. 1, no. 3, p. e0000022, Mar. 2022, doi: 10.1371/journal.pdig.0000022.
- [35] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2021, doi: 10.1145/3457607.