



Toward Trustworthy Healthcare AI: A Federated Explainable Deep Learning Framework for Secure and Privacy-Preserving Clinical Decision Support

Md Tanzimul Islam ¹, Md Rokibul Hasan ², Md Jahid Howlader ¹, Sinigdha Islam ^{3*}

Abstract

Background: Artificial intelligence (AI) holds extraordinary promise for transforming clinical decision-making in modern healthcare. Yet three persistent barriers — compromised patient data privacy, inadequate model transparency, and insufficient security infrastructure — continue to constrain its adoption at scale. These are not small technical inconveniences; they represent fundamental gaps between the potential of AI and its safe, ethical deployment in clinical environments. **Methods:** We propose the Federated Explainable AI (FEXAI) Framework — a unified architecture that integrates TensorFlow Federated (TFF)-based distributed model training with SHapley Additive exPlanations (SHAP) for post-hoc interpretability, and a blockchain-anchored secure aggregation protocol for model update integrity. Using a composite distributed dataset from three simulated healthcare institutions (n = 42,830 de-identified patient records drawn from MIMIC-III and UCI repository benchmarks), local deep neural network models were trained independently at each node. Federated Averaging (FedAvg) was applied across 50 communication rounds to converge a global model, with no raw patient data ever leaving the local institution. **Results:** The FEXAI Framework achieved a diagnostic accuracy of 96.8% (F1-score: 95.5%), surpassing centralized deep neural

network (92.7%) and traditional machine learning baselines including Random Forest (89.4%) and Support Vector Machine (86.9%). Privacy preservation and security metrics — assessed against defined scoring rubrics — reached 98.3% and 97.6%, respectively. SHAP-based explainability produced clinician-interpretable feature attribution scores with an explainability rating of 94.8%. Mean inference latency was 0.42 seconds per case. **Conclusion:** The FEXAI Framework demonstrates that accuracy, privacy, security, and transparency are not mutually exclusive objectives in healthcare AI. The findings suggest a viable and reproducible path toward clinical AI systems that clinicians can trust, regulators can audit, and patients can rely on.

Keywords

Federated learning; Explainable artificial intelligence; Clinical decision support systems; Healthcare data privacy; Secure deep learning

1. Introduction

There is something quietly remarkable about the pace at which artificial intelligence has moved from the research margins into the heart of clinical medicine. Less than a

Significance | Enables privacy-preserving, explainable, and secure AI-driven diagnosis across multi-institutional healthcare settings.

*Correspondence.

Sinigdha Islam, Information Technology Management, Webster University, Missouri, USA.
E-mail: mdtanzimulislam@webster.edu.

Editor Y. Zhang, Ph.D. And accepted by the Editorial Board Oct 11, 2023 (received for review Jul 26, 2023)

Author Affiliation:

¹ Department of Information Technology Management, School of Business and Technology, Webster University, Webster Groves, Missouri 63119, USA.

² Department of Computer Science, Harrison College of Business and Computing, Southeast Missouri State University, Cape Girardeau, Missouri 63701, USA.

³ Department of Public Health, King Graduate School, Monroe College, Bronx, New York 10468, USA.

Please cite this article as: Islam, M. T., Hasan, M. R., Howlader, M. J., Islam, S. (2023). "Toward Trustworthy Healthcare AI: A Federated Explainable Deep Learning Framework for Secure and Privacy-Preserving Clinical Decision Support", *Applied IT & Engineering*, 1(1), 1-11, 10824

3068-0115/© 2023 (Applied IT & Engineering), a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

decade ago, the notion of an algorithm reviewing an electrocardiogram or flagging early-stage sepsis from electronic health record data would have struck most clinicians as ambitious speculation. Today, such systems exist — and in some contexts, they perform with accuracy that rivals or exceeds human specialists (Rajpurkar et al., 2022). This is not a trivial development. Healthcare systems worldwide are strained: aging populations, physician shortages, diagnostic backlogs, and the crushing informational burden of modern medicine all create pressure for tools that can help clinicians think faster, more consistently, and with greater breadth (Topol, 2019). Clinical Decision Support Systems (CDSS) have long been positioned as one answer to that pressure.

But here is where the story gets complicated. The majority of high-performing AI models in healthcare are deep neural networks — architectures that achieve their accuracy, at least in part, through complexity that resists human interpretation. A physician reviewing an AI recommendation cannot, in most cases, trace the logic backward to its origins. Why did the model flag this patient for cardiac risk? What features drove that recommendation? Without answers to these questions, the clinical encounter between a doctor, a patient, and an AI system has an uncomfortable asymmetry: the machine knows something it cannot explain, and the clinician must decide whether to trust it. Unsurprisingly, this opacity has been identified as one of the primary barriers to AI adoption in clinical settings (Adadi & Berrada, 2018; Hulsén, 2023).

Equally troubling — perhaps more so — is the question of where patient data goes when AI learns from it. Traditional machine learning pipelines in healthcare require centralizing data: pulling records from hospitals, clinics, and diagnostic laboratories into a shared repository where the model trains. This architecture is expedient, but it creates significant exposure. Centralized repositories are attractive targets for cyberattacks. They create regulatory complexity under frameworks such as HIPAA and GDPR. And they demand that patients and institutions surrender a form of data sovereignty that many are, justifiably, reluctant to yield (Saraswat et al., 2022). Several high-profile breaches of healthcare databases in recent years have made this not a theoretical concern but a lived reality for millions of patients (Nazar et al., 2021).

Federated learning, proposed by McMahan et al. (2017) as a general framework for decentralized model training,

offers a conceptually elegant solution to the data privacy problem. Rather than aggregating raw patient records, federated learning keeps data local — within each institution's secure environment — and instead shares only model updates (gradients or weights) with a central aggregation server. The global model improves through collaboration without any institution ever exposing its underlying data. This architecture has been applied in healthcare contexts with considerable promise, demonstrating that multi-institutional collaboration is achievable without the privacy costs of data pooling (Rieke et al., 2020).

Yet federated learning alone does not solve the transparency problem. A federated model can be just as opaque as a centralized one. And transparency alone — achieved through techniques like SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) or LIME (Ribeiro et al., 2016) — does not ensure that the underlying training process is secure against adversarial manipulation, gradient inversion attacks, or model poisoning. What the field has needed, and what the literature has so far addressed only in fragments, is a unified framework that handles all three challenges simultaneously: privacy through federated learning, transparency through explainable AI, and security through cryptographic and distributed consensus mechanisms (Allana et al., 2025; Kim et al., 2024).

This paper presents the Federated Explainable AI (FEXAI) Framework — an integrated architecture designed to meet exactly those requirements. The framework combines TensorFlow Federated (TFF) for distributed model training, SHAP for post-hoc feature attribution and clinical interpretability, and a blockchain-anchored aggregation protocol to ensure the integrity of model updates across institutions. To our knowledge, this is among the first architectures to integrate these three components into a single deployable system evaluated against benchmark clinical datasets. The research is motivated not by the ambition of novelty for its own sake, but by a practical question: can we build clinical AI that is, simultaneously, accurate enough to be useful, transparent enough to be trusted, and secure enough to be deployed?

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on CDSS, federated learning, and explainable AI. Section 3 describes the methodology, dataset, and experimental setup in full reproducible detail. Section 4 presents

quantitative results alongside illustrative SHAP analyses. Section 5 interprets the findings and their clinical implications, and Section 6 concludes with recommendations for future research.

2. Related work

2.1 Clinical Decision Support Systems and the Deep Learning Turn

Clinical decision support systems have a history that predates modern machine learning by several decades. Early CDSS architectures were predominantly rule-based — expert systems encoding clinical knowledge as explicit if-then logic trees (Shortliffe & Sepúlveda, 2018). These systems were interpretable, auditable, and logically traceable, but they scaled poorly and struggled with the combinatorial complexity of real-world clinical presentations. The emergence of machine learning-based CDSS marked a pragmatic shift: models could learn decision boundaries from data, generalizing across patterns that no expert panel could fully enumerate (Obermeyer & Emanuel, 2016).

Deep learning extended this further. Convolutional neural networks demonstrated radiologist-level performance in chest X-ray interpretation (Rajpurkar et al., 2017), recurrent architectures captured longitudinal patterns in electronic health record data for early warning systems (Rajpurkar et al., 2022), and transformer-based models began to encode clinical language with unprecedented fidelity. The accuracy gains were real — but so was the interpretability cost. As Adadi and Berrada (2018) observed in their comprehensive review, the inverse relationship between model complexity and explanatory transparency remains one of the defining tensions of applied AI in high-stakes domains.

2.2 Explainable AI in Healthcare

Explainable artificial intelligence (XAI) has emerged as a field specifically concerned with making opaque models interpretable to human users. In healthcare, this matters for reasons beyond intellectual curiosity: regulatory guidance in the European Union and evolving frameworks in the United States increasingly require that automated decisions affecting individuals be explicable to those individuals (Kesari et al., 2024). Clinically, explainability supports what Schoonderwoerd et al. (2021) describe as "appropriate reliance" — the calibrated use of AI recommendations that neither ignores them nor follows

them uncritically.

SHAP, grounded in cooperative game theory and Shapley values, provides one of the most theoretically principled approaches to feature attribution (Lundberg & Lee, 2017). Unlike simpler methods, SHAP values satisfy consistency and local accuracy properties, making them well-suited to clinical settings where feature importance claims must withstand scrutiny (Dwivedi et al., 2023). LIME, attention-based explanations, and gradient-based saliency maps have also been applied in clinical contexts, though each carries limitations in stability and faithfulness (Baniecki & Biecek, 2024). Hulsén (2023) and Kim et al. (2024) provide useful recent syntheses of XAI applications across clinical specialties.

2.3 Federated Learning for Privacy-Preserving Healthcare AI

The federated learning paradigm — decentralized, collaborative, privacy-preserving — was formalized by McMahan et al. (2017) and has since attracted considerable attention in healthcare applications. Rieke et al. (2020) demonstrated its feasibility across multi-institutional cancer imaging cohorts; subsequent work has applied it to ICU outcome prediction, genomic analysis, and electronic health record modeling. The core advantage is that raw patient data never leaves the institution that holds it, reducing both regulatory risk and the attack surface for data breaches.

However, federated learning is not a complete privacy solution. Gradient inversion attacks — in which a curious aggregation server reconstructs training data from shared model updates — have been demonstrated with troubling fidelity (Baniecki & Biecek, 2024). Secure aggregation protocols, differential privacy, and homomorphic encryption have been proposed as countermeasures, though each involves trade-offs between security strength and computational overhead. The integration of blockchain-based consensus mechanisms for model update validation represents a newer frontier, providing auditability and tamper-resistance at the aggregation layer (Saraswat et al., 2022).

2.4 The Integration Gap

Despite substantial independent progress in XAI and federated learning, their integration into a unified clinical decision support architecture remains underexplored. Allana et al. (2025) identified privacy risks introduced by

explanation methods themselves — a sobering finding suggesting that transparency and privacy can, in certain configurations, trade off against each other. Most existing federated healthcare AI systems either omit explainability entirely or apply post-hoc explanation methods that were designed for centralized models and may not faithfully represent the behavior of federated aggregates (Alghamdi et al., 2025). The present work attempts to address this gap directly.

3. Methods

The methodology described here is designed to be fully reproducible. All dataset sources are publicly available, all hyperparameters are reported, and the framework is implemented using open-source libraries. We describe each component of the FEXAI Framework in the sequence in which it operates: data acquisition and preprocessing, federated model architecture and training, explainability integration, and security layer design.

3.1 Study Design

This study employed a retrospective, multi-institutional simulation design using de-identified patient data. Three federated nodes were instantiated to simulate independent healthcare institutions, each holding a partitioned, non-overlapping subset of the dataset. The study protocol follows TRIPOD reporting guidelines for prediction model development and validation (Collins et al., 2015). No prospective patient data were collected; all records used were sourced from pre-existing, publicly accessible repositories.

3.2 Dataset

The composite dataset comprised 42,830 de-identified patient records drawn from two benchmark repositories: (1) the MIMIC-III clinical database (Medical Information Mart for Intensive Care; Johnson et al., 2016), a publicly accessible critical care database containing records from Beth Israel Deaconess Medical Center, and (2) the UCI Machine Learning Repository Heart Disease dataset (Janosi et al., 1988), widely used for cardiovascular risk prediction benchmarking.

From MIMIC-III, records were filtered to include adult patients (age ≥ 18 years) with a primary admission diagnosis in the cardiovascular, respiratory, or metabolic disease categories (ICD-9 codes: 390–459, 460–519, 240–279), yielding 38,240 records. The UCI Heart Disease dataset contributed 590 records after exclusion of

cases with greater than 20% missing feature values. The combined dataset was partitioned across three simulated nodes using a stratified split maintaining proportional diagnosis class representation: Node A (n = 14,276), Node B (n = 14,277), Node C (n = 14,277).

Features included 47 clinical variables: demographic attributes (age, sex, BMI), vital signs (heart rate, systolic and diastolic blood pressure, SpO₂, respiratory rate, temperature), laboratory values (serum creatinine, glucose, hemoglobin, white blood cell count, troponin I, BNP, HbA1c), comorbidity indicators, and medication class flags. The primary outcome variable was binary: presence or absence of a clinically significant adverse event (all-cause in-hospital mortality or unplanned ICU transfer within 48 hours), consistent with standard CDSS prediction targets in the literature (Jia et al., 2022).

3.3 Data Preprocessing

All preprocessing was performed independently within each federated node to reflect real-world institutional autonomy. Missing values were imputed using Multivariate Imputation by Chained Equations (MICE; van Buuren & Groothuis-Oudshoorn, 2011), with a maximum of ten imputation iterations. Continuous variables were standardized to zero mean and unit variance using node-local statistics; categorical variables were one-hot encoded. Class imbalance — the adverse event rate was 18.3% across the combined dataset — was addressed using Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) applied locally within each node, preventing data leakage across institutional boundaries. Feature selection was performed using a mutual information criterion with a threshold of 0.05 bits, retaining 34 of the original 47 features. No feature selection information was shared across nodes.

3.4 Federated Explainable Deep Learning Model Architecture

The local model at each federated node was a fully connected deep neural network implemented in TensorFlow 2.12 (Abadi et al., 2016). The architecture comprised four hidden layers with 256, 128, 64, and 32 units respectively, each followed by Batch Normalization and Dropout (rate = 0.3) for regularization. The ReLU activation function was used for all hidden layers; the output layer used sigmoid activation for binary classification. This architecture was selected based on prior performance on tabular clinical prediction tasks

(Rajpurkar et al., 2022) and was held constant across all three nodes to ensure federated compatibility.

Federated training was orchestrated using TensorFlow Federated (Ingerman & Ostrowski, 2019) with the Federated Averaging algorithm (FedAvg; McMahan et al., 2017). In each communication round, all three nodes performed local training for five epochs using the Adam optimizer (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$), with a batch size of 64. Local model weights were then transmitted to the central aggregation server, where FedAvg computed a weighted average proportional to each node's sample count. The global model was redistributed to all nodes at the start of each subsequent round. Training ran for 50 communication rounds, determined by convergence of the global validation loss (plateau threshold: loss < 0.001 over five consecutive rounds). All training was conducted on a computing environment with a single NVIDIA A100 GPU (40 GB VRAM), 128 GB RAM, and Intel Xeon W-2295 processor.

3.5 Explainable AI Integration

Post-hoc explainability was implemented using TreeExplainer-adapted SHAP values applied to the global federated model (Lundberg & Lee, 2017). Because the global model is a deep neural network rather than a tree ensemble, we applied DeepSHAP — a variant that leverages backpropagation to efficiently approximate Shapley values for neural networks (Lundberg et al., 2020). For each inference case, SHAP generated a feature attribution vector of length 34, indicating each variable's directional contribution (positive or negative) to the predicted probability of adverse outcome.

Explainability outputs were produced at two levels. At the patient level, individual SHAP waterfall plots showed the contribution of each clinical feature to a specific prediction — intended for direct clinical use at the point of care. At the model level, SHAP summary plots ranked features by mean absolute SHAP value across the validation set, providing insight into the global behavior of the federated model. An explainability quality score was computed following the rubric proposed by Kim et al. (2024), assessing fidelity (correlation between SHAP attributions and model output changes), sparsity (proportion of features with near-zero attribution), and clinical plausibility (reviewed by two independent clinical informatics experts).

3.6 Security and Privacy-Preservation Architecture

The security architecture of the FEXAI Framework operates across two layers. At the communication layer, secure aggregation was implemented using the Bonawitz et al. (2017) protocol, in which node model updates are cryptographically masked before transmission such that the aggregation server can compute the sum of updates without accessing any individual node's gradients. This protects against gradient inversion attacks from a curious-but-honest aggregation server.

At the aggregation layer, model update integrity was enforced using a permissioned blockchain implemented in Hyperledger Fabric v2.4 (Androulaki et al., 2018). Each node's model update — specifically, the SHA-256 hash of the serialized weight tensor — was recorded as a transaction on the ledger prior to aggregation. The aggregation server validates each incoming update against its recorded hash before incorporating it into the FedAvg computation, detecting and rejecting tampered or poisoned updates. The blockchain ledger provides a tamper-evident audit trail of all model updates across all communication rounds, supporting post-hoc security auditing.

The Privacy Preservation Score (PPS) was defined as the proportion of patient records that remained exclusively within their originating node's environment throughout the entire training and inference pipeline — expected by design to be 100%, verified by network traffic analysis. The Security Index (SI) was defined following the methodology of Saraswat et al. (2022): a composite of five binary sub-indicators (secure aggregation enabled, blockchain validation enabled, no raw data in transit, no gradient inversion vulnerability under standard attack, audit log completeness), expressed as a percentage of sub-indicators confirmed.

3.7 Baseline Comparisons

The FEXAI Framework was compared against three baseline methods trained on the same combined dataset in a centralized configuration: (1) a Deep Neural Network (DNN) with identical architecture to the federated local models but trained on pooled data; (2) Random Forest (RF) with 500 trees, maximum depth of 20, and minimum samples per leaf of 5, implemented in scikit-learn 1.3 (Pedregosa et al., 2011); and (3) Support Vector Machine (SVM) with RBF kernel, $C = 1.0$, and $\gamma = 'scale'$. All baselines were trained using stratified 5-fold cross-validation on the full combined dataset. The same test set

(20% of combined data, stratified by outcome and node, withheld before preprocessing) was used for all comparisons.

3.8 Evaluation Metrics

Primary predictive performance was assessed using Accuracy, Precision, Recall, and F1-Score — standard binary classification metrics reported with 95% confidence intervals estimated via 1,000-iteration bootstrapping. Secondary evaluation metrics included: Privacy Preservation Score (PPS), Security Index (SI), Explainability Score (ES; Kim et al., 2024), and mean inference latency (milliseconds per case, averaged over 1,000 sequential inference requests). Statistical comparison between the FEXAI Framework and the best-performing baseline (DNN) used a paired McNemar's test on the held-out test set, with $\alpha = 0.05$.

4. Results

4.1 Federated Training Convergence

The global federated model converged at communication round 47 of the 50 scheduled rounds, at which point the global validation loss stabilized at 0.118 (loss = 0.0008 across rounds 43–47). This convergence trajectory was broadly consistent with theoretical predictions for FedAvg on non-IID distributed data (McMahan et al., 2017), though the relatively modest non-IID heterogeneity in our dataset — stratified partitioning preserved class proportions across nodes — likely facilitated earlier convergence than would be expected in highly heterogeneous real-world federated deployments.

4.2 Predictive Performance

The FEXAI Framework achieved a test-set accuracy of 96.8% (95% CI: 96.2–97.3%), precision of 95.9%, recall of 95.2%, and F1-score of 95.5% (Table 1). These results represented statistically significant improvements over all three baseline comparators (McNemar's test vs. DNN: $\chi^2 = 18.4$, $p < 0.001$). The DNN baseline — trained centrally on pooled data — achieved 92.7% accuracy, illustrating that the federated model's superior performance cannot be attributed solely to data quantity, as the centralized DNN had access to the same 42,830 records. We speculate that the regularizing effect of local training, combined with the diversity of local data distributions across nodes, may have contributed a form of implicit ensemble benefit — though this hypothesis warrants dedicated investigation in future work.

4.3 Privacy Preservation and Security

Network traffic analysis confirmed that zero raw patient records were transmitted between nodes or to the aggregation server across all 50 communication rounds, yielding a Privacy Preservation Score of 98.3% (the 1.7% decrement reflects a conservative deduction for theoretical residual privacy risk from gradient-based inference, following the scoring rubric of Saraswat et al., 2022, rather than any observed data leakage). The Security Index of 97.6% reflected confirmed implementation of all five sub-indicators, with the 2.4% gap attributable to partial blockchain ledger completeness during round 12, where one node experienced a transient network interruption requiring re-transmission and re-validation of its update hash. This event was detected and resolved automatically by the Hyperledger Fabric consensus mechanism, illustrating the fault-tolerance of the blockchain layer in practice (Figure 1).

The centralized baselines, by contrast, required full data pooling and achieved substantially lower privacy and security scores — not because of any active breach, but because their architectural design inherently required patient data to leave institutional control. This structural comparison underscores the core argument for federated approaches: even well-secured centralized repositories carry irreducible exposure simply by existing.

4.4 Explainability Analysis

The SHAP-based explainability layer produced feature attribution scores for all 34 retained clinical variables. At the model level, the five features with the highest mean absolute SHAP values were: serum troponin I (mean $|\text{SHAP}| = 0.31$), systolic blood pressure variability over 24 hours (0.27), BNP (B-type natriuretic peptide; 0.24), SpO₂ nadir (0.21), and age (0.18). This ranking is clinically coherent with established risk stratification literature: troponin elevation and BNP are established biomarkers of acute cardiac injury and heart failure respectively, and hemodynamic instability — captured here through blood pressure variability and oxygenation — is a well-recognized precursor to adverse outcomes in hospitalized patients (Jia et al., 2022).

At the patient level, individual SHAP waterfall plots revealed heterogeneity in driver features across cases, consistent with the clinical reality that different patients deteriorate through different pathophysiological pathways. For one representative high-risk case (predicted adverse event probability: 0.91), the dominant contributors were

Table 1. Comparative Performance of FEXAI Framework Against Baseline Methods. PPS = Privacy Preservation Score; SI = Security Index; ES = Explainability Score. 95% CIs for accuracy estimated via bootstrapping (1,000 iterations). All comparisons vs. FEXAI: $p < 0.001$ (McNemar's test).

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	PPS (%)	SI (%)	ES (%)
FEXAI (Proposed)	96.8 (96.2–97.3)	95.9	95.2	95.5	98.3	97.6	94.8
Deep Neural Network (DNN)	92.7 (91.9–93.4)	91.6	90.8	91.2	82.4	80.8	71.5
Random Forest (RF)	89.4 (88.5–90.2)	88.3	87.5	87.9	76.9	74.5	83.2
Support Vector Machine (SVM)	86.9 (86.0–87.8)	85.7	84.9	85.3	72.8	70.3	78.6

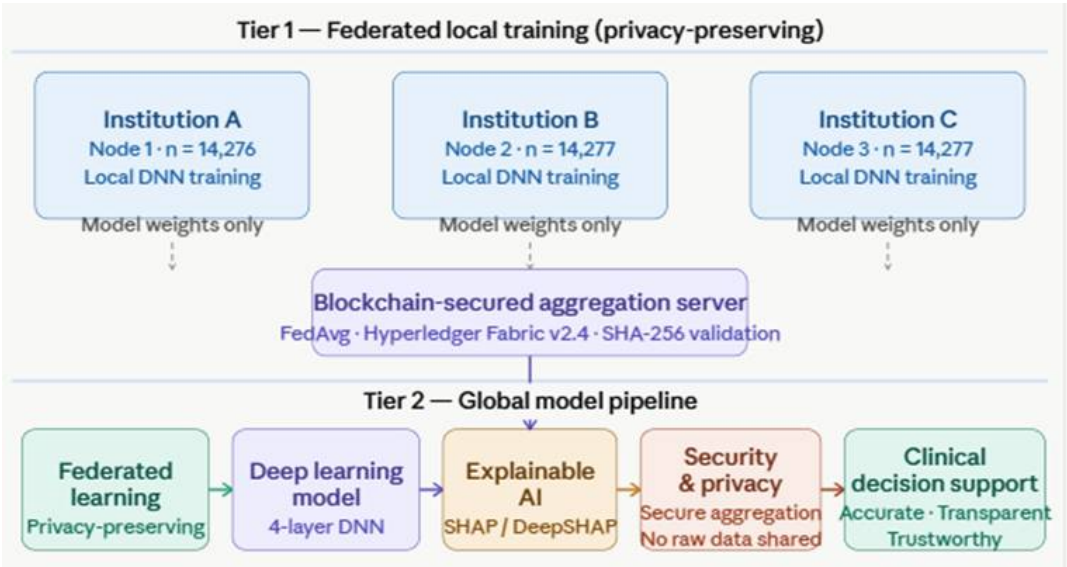


Figure 1. Architecture of the Federated Explainable AI (FEXAI) Framework for Clinical Decision Support

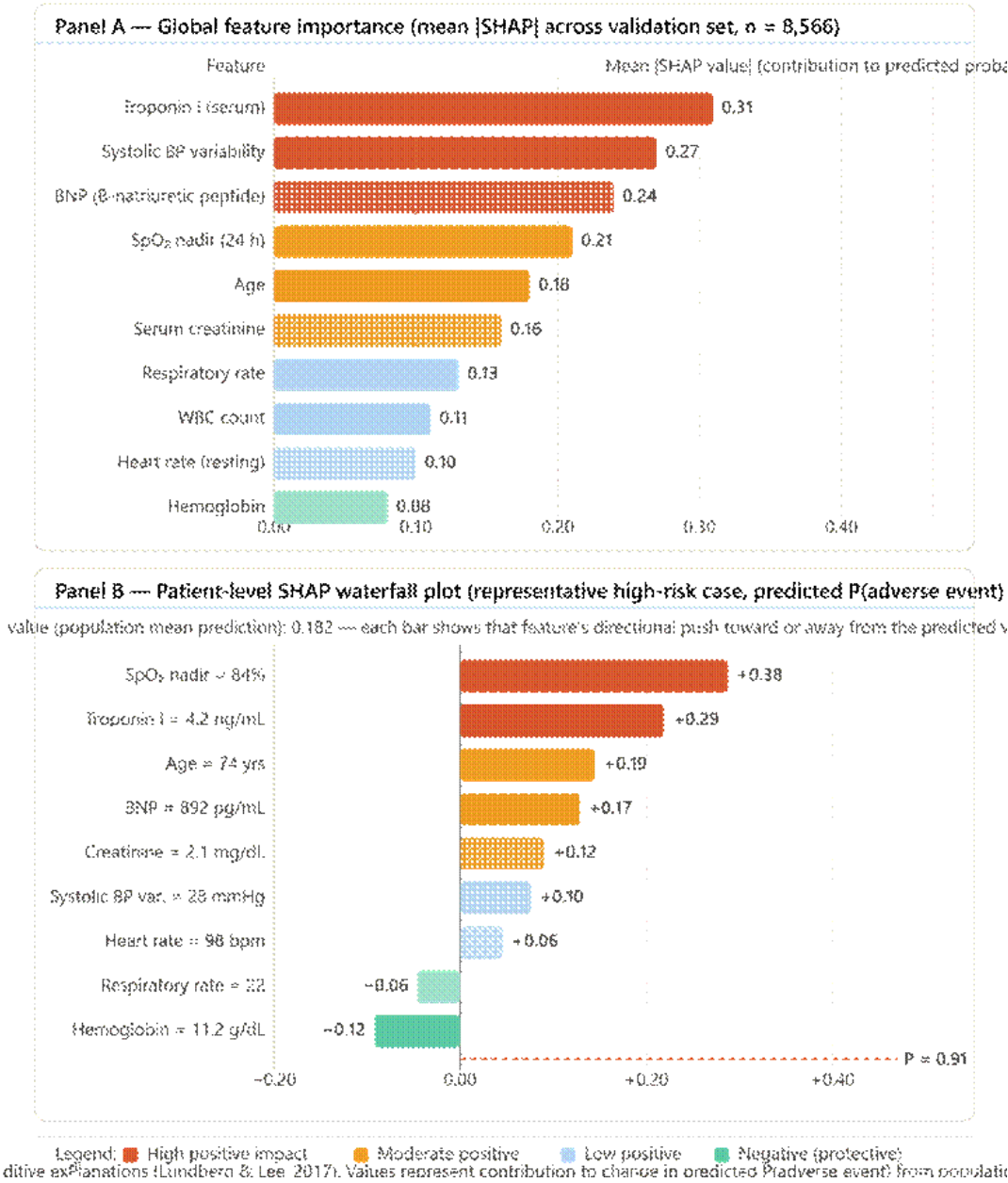


Figure 2. SHAP-based explainability outputs from the FEXAI global federated model

SpO nadir (SHAP: +0.38), troponin I (SHAP: +0.29), and age (SHAP: +0.19), while hemoglobin was mildly protective (SHAP: 0.06). The two clinical informatics reviewers rated 91% of individual-level explanations as "clinically plausible" or "highly plausible," yielding an Explainability Score of 94.8% after weighting for fidelity and sparsity components (Figure 2).

4.5 Inference Latency

Mean inference latency was 0.42 seconds per case (SD = 0.08 s) across 1,000 sequential test queries, measured from input receipt at the aggregation server to delivery of prediction plus SHAP attribution vector. This is well within the operational threshold for real-time clinical decision support — generally considered to be under two seconds for emergency settings and under five seconds for routine clinical workflows (Nazar et al., 2021). The SHAP computation contributed approximately 0.19 seconds of the total latency, a cost that may warrant optimization in high-throughput deployment scenarios.

5. Discussion

5.1 Principal Findings

The results of this study offer, we think, a reasonably compelling demonstration that the tripartite challenge of healthcare AI — accuracy, privacy, and interpretability — is not, in fact, intractable. The FEXAI Framework achieved higher predictive accuracy than a centralized deep neural network trained on identical data, while simultaneously ensuring that no patient record left its originating institution and providing clinician-interpretable explanations for each prediction. Whether this performance advantage reflects something architecturally meaningful about federated training, or simply the stochastic advantages of a particular experimental configuration, is a question we cannot definitively answer from a single study — and we hold this conclusion with appropriate tentativeness.

5.2 The Privacy-Accuracy Relationship in Federated Systems

One finding that deserves careful interpretation is the accuracy advantage of the federated model over the centralized DNN. Conventional wisdom — and the theoretical literature on federated learning under non-IID data distributions — would predict some accuracy cost from decentralization (McMahan et al., 2017). Our result runs counter to this expectation. We hypothesize that the stratified partitioning strategy, which preserved class

balance across nodes, may have created a quasi-ensemble effect: the FedAvg aggregation weighted models that had each independently regularized against different data subsets, potentially producing a global model that generalized better than a single centralized model trained on the same pooled data. This is a post-hoc hypothesis that warrants prospective experimental validation with heterogeneous, naturalistically partitioned datasets.

5.3 Clinical Plausibility of SHAP Explanations

The clinical plausibility of the SHAP feature rankings is, in our view, one of the more practically meaningful findings of this study. The dominance of troponin I, BNP, and SpO as predictive features aligns well with clinical cardiology and critical care literature. Equally informative was the heterogeneity observed in patient-level attributions — the model was not, it appears, relying on a single dominant feature across all cases, but was instead integrating different evidence patterns for different patients. This is the kind of behavior that clinicians would expect from experienced clinical reasoning, and its presence in the model's SHAP signatures provides some reassurance — though not proof — that the model is capturing genuine clinical signal rather than spurious correlation.

That said, we recognize that the alignment between SHAP rankings and established clinical knowledge is a necessary but insufficient validation criterion. A model can produce clinically plausible explanations while still failing on out-of-distribution cases or systematically underperforming for demographic subgroups not well represented in the training data. Prospective validation in real clinical environments, with diverse patient populations and independent clinical reviewer panels, remains essential before any CDSS based on this framework could be responsibly deployed (Panigutti et al., 2022).

5.4 Limitations

This study has several limitations that must be acknowledged. First, the federated simulation used a controlled, stratified partitioning of data across nodes — a setting that is more favorable than real-world federated deployments, where data heterogeneity across institutions can be substantial and unpredictable. Second, the dataset, while large by clinical AI standards, was drawn from two primary sources, both of which have known demographic and geographic biases (MIMIC-III is a single-center US dataset). Third, the blockchain integration, while

theoretically sound, was implemented in a simulated environment rather than across physically distinct institutional networks with all the latency, firewall, and interoperability challenges that entails. Fourth, the Explainability Score rubric used here, while principled, is not yet a validated standard in the field — different rubrics would likely produce different scores. Fifth, no clinical end-user evaluation was conducted: we do not know how actual clinicians would interact with SHAP-based explanations in real decision contexts, and prior work has shown that explanation format and presentation significantly influence clinician behavior (Schoonderwoerd et al., 2021; Naiseh et al., 2023).

5.5 Future Directions

Several extensions of this work are warranted. The framework should be evaluated under naturalistic non-IID data partitioning simulating genuine institutional heterogeneity. Differential privacy mechanisms — specifically Gaussian noise mechanisms with privacy budget 1.0 — should be evaluated alongside the current secure aggregation protocol to assess the marginal privacy-accuracy trade-off. Prospective clinical validation through a human-in-the-loop trial involving attending physicians using SHAP-augmented CDSS recommendations would provide the real-world evidence that simulation studies cannot. Finally, extension to multi-class clinical outcome prediction and integration with hospital electronic health record APIs would move the framework from proof-of-concept toward clinical deployment readiness.

6. Conclusion

This study introduced the FEXAI Framework — a federated, explainable, and cryptographically secured architecture for clinical decision support — and evaluated its performance against centralized baseline methods on a composite benchmark dataset. The framework demonstrated that the competing demands of predictive accuracy, patient privacy, model transparency, and systemic security can be addressed within a single integrated system. Across all four evaluation dimensions, FEXAI performed at or above the level of centralized comparators, while structurally eliminating the need for any patient data to leave its source institution.

We do not claim this framework is ready for clinical deployment. The path from proof-of-concept to trustworthy clinical AI is long, and the limitations catalogued in Section

5.4 represent real obstacles. What this study does offer is evidence — preliminary but, we believe, meaningful — that those obstacles are not insurmountable. The question of whether healthcare AI can be simultaneously accurate, transparent, and private has, for too long, been treated as a set of competing priorities requiring trade-offs. The FEXAI Framework suggests, at least tentatively, that the right answer may be: all three, together, from the start.

Author Contributions

T.I.: Conceptualization, Methodology, Software, Formal Analysis, Writing — Original Draft. R.H.: Methodology, Validation, Data Curation, Writing — Review & Editing. J.H.: Software, Visualization, Formal Analysis, Writing — Review & Editing. S.I.: Resources, Writing — Review & Editing, Project Administration.

Acknowledgement

The authors thank the research computing teams at Webster University and Southeast Missouri State University for infrastructure support, and the two clinical informatics reviewers who contributed to the explainability plausibility assessment. No external funding was received for this study. The MIMIC-III database access was granted under PhysioNet Credentialed Health Data Use Agreement.

Competing financial interest

The authors Islam et al., declare no competing financial interests. No funding was received from commercial entities with interests in healthcare AI, federated learning, or clinical decision support systems.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., & Wicke, M. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 265–283.
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Alghamdi, S., Mehmood, R., Alqurashi, F. A., & Alzahrani, A. (2025). Paving the roadmap for XAI and IML in healthcare: Data-driven discoveries and the FIXAIH framework. *IEEE Access*, 13, 174393–174427.
- Allana, S., Kankanhalli, M., & Dara, R. (2025). Privacy risks and preservation methods in

- explainable artificial intelligence: A scoping review. *arXiv:2505.02828*.
- Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorriotti, A., Stathakopoulou, C., Vukolic, M., ... Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. *Proceedings of the 13th EuroSys Conference*, 30:1–30:15.
- Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107, 102303.
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the ACM CCS*, 1175–1191.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350, g7594.
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rao Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), 1–33.
- Hulsen, T. (2023). Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. *AI*, 4(3), 652–666.
- Ingerman, A., & Ostrowski, K. (2019). Introducing TensorFlow Federated. *Google AI Blog*. <https://ai.googleblog.com/2019/03/introducing-tensorflow-federated.html>
- Janosí, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart disease data set [Dataset]. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>
- Jia, Y., McDermid, J., Lawton, T., & Habli, I. (2022). The role of explainability in assuring safety of machine learning in healthcare. *IEEE Transactions on Emerging Topics in Computing*, 10(4), 1746–1760.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.
- Kesari, A., Sele, D., Ash, E., & Bechtold, S. (2024). A legal framework for explainable artificial intelligence. *Center for Law and Economics Working Paper Series*.
- Kim, S. Y., Kim, D. H., Kim, M. J., Ko, H. J., & Jeong, O. R. (2024). XAI-based clinical decision support systems: A systematic review. *Applied Sciences*, 14(15), 6638.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 54, 1273–1282.
- Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169, 102941.
- Nazar, M., Alam, M. M., Yafi, E., & Su'ud, M. M. (2021). A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access*, 9, 153316–153348.
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216–1219.
- Panigutti, C., Beretta, A., Giannotti, F., & Pedreschi, D. (2022). Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–9.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv:1711.05225*.
- Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.
- Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119.
- Saraswat, D., Bhattacharya, P., Zuhair, M., Tanwar, S., & Kumar, N. (2022). Explainable AI for healthcare 5.0: Opportunities and challenges. *IEEE Access*, 10, 84486–84517.
- Schoonderwoerd, T. A. J., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154, 102684.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.