



Original Research Paper

FedTrust-XAI: Algorithmic Trust in Federated Health Communication Systems through Explainability and Privacy-Preserving Breast Cancer Diagnosis

M. Kaliappan¹, S Vimal², Preethy Rebecca³, S. Senthil⁴, Anitha B⁵, Rattan Singh⁶, Gaurav Dhiman⁷

¹Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India (e-mail: kaliappan@ritrjpm.ac.in).

²Department of Computer Science and Engineering, KIT- Kalaingar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India (e-mail: svimal@kitcbe.ac.in)

³Department of Computer Science and Engineering, St. Peter's College of Engineering and Technology, Chennai, Tamil Nadu 600054, India (e-mail: preethyrebecca@spcet.ac.in).

⁴Department of Computer Science and Engineering, J.P. College of Engineering, Thenkasi, Tamil Nadu 627116, India (e-mail: jaysen1984@gmail.com).

⁵Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India (e-mail: anithab@ritrjpm.ac.in).

⁶Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India (e-mail: rattan67in@yahoo.com)

⁷School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala-147001, Punjab, India (e-mail: gdhiman0001@gmail.com)

Abstract

When using Artificial Intelligence (AI) in Clinical Decision Support Systems (CDSS), two important and related problems arise: ensuring that federated learning over institutional data silos does not violate patient privacy, and gaining clinician trust in AI models' diagnostic suggestions. The literature has primarily addressed each issue in isolation, resulting in limited comprehension of their interconnections. This study proposes an innovative approach through the FedTrust-XAI architecture, which integrates federated learning and explainable artificial intelligence (XAI), and examines its impact on diagnostic accuracy and clinician trust. The proposed system employs federated logistic regression utilizing the FedAvg method across three distinct clinical sites over multiple interaction rounds, enabling collaborative learning without sharing patient data. Experiments on the UCI Wisconsin Diagnostic Breast Cancer (WDBC) dataset demonstrate that the federated model achieves diagnostic accuracy comparable to a centralized model while maintaining strong privacy protections. A linear XAI attribution mechanism identifies worst concave points, worst perimeter, and worst radius as the most important malignancy predictors, consistent with clinical radiological knowledge. A factorial experiment with 120 simulated physicians shows that federated framing combined with XAI

explanations yields the highest clinician trust (Trust Index = 4.45), and that privacy concern significantly moderates trust in federated systems. The proposed architecture achieves 99.12% accuracy and 99.23% AUROC, closely matching the centralized baseline (99.60% AUROC). The results demonstrate that privacy-preserving federated learning, model explainability, and clinician trust can be jointly optimized in healthcare AI systems

.Keywords: Federated Learning; Explainable AI; Algorithmic Trust; Health

Communication; Breast Cancer Diagnosis; FedAvg; XAI Attribution; Privacy; Clinician Trust; WDBC; Human-AI Interaction

1 Introduction

Clinical artificial intelligence machine learning systems that assist clinicians in diagnosis, prognosis, and treatment planning have achieved remarkable technical performance across a range of medical imaging and clinical data tasks. Deep learning models have matched or exceeded radiologist performance on mammography interpretation, dermatology classification, diabetic retinopathy screening, and pathology slide analysis. However, the widespread clinical adoption of AI decision support has not kept pace with these technical achievements. A persistent trust deficit separates high benchmark performance from routine clinical deployment: clinicians consistently report uncertainty about AI recommendation provenance, lack of transparency in AI reasoning, and concern about accountability when AI-assisted decisions result in adverse outcomes.

The trust challenge is compounded by a fundamental data privacy constraint that governs the development of clinical AI: high-performing models require large, diverse training datasets, but patient health data is subject to stringent regulatory protections (GDPR Article 9, HIPAA Privacy Rule, national data sovereignty frameworks) that prevent centralised aggregation of electronic health records across institutional boundaries. Hospitals, diagnostic clinics, and research institutions cannot freely share patient radiological images, pathology slides, or clinical record datasets without specific consent and data governance frameworks that are impractical at the scale required for training population-representative AI models.

Federated learning—a distributed machine learning paradigm in which

model parameters rather than raw data are shared across participating institutions—offers a technical solution to the privacy constraint: each institution trains a local model on its own data, communicates only gradient updates to a coordination server, and receives an aggregated global model without any patient data leaving institutional boundaries. However, federated learning introduces a new communication challenge: the framing of a clinical AI system as “federated” versus “centralised” may itself influence clinician trust, credibility, and willingness to rely on AI recommendations—a question that health communication researchers have not yet systematically investigated.

Three converging advances create the opportunity for FedTrust-XAI. First, federated averaging (FedAvg) provides a mathematically convergent aggregation algorithm, where participating institutions contribute weighted model updates proportional to local dataset size. FedAvg has been demonstrated to reach near-centralised accuracy within 20–50 communication rounds for logistic regression and shallow neural network models on tabular medical data [1, 2, 31, 32], providing a computationally efficient

privacy-preserving training protocol for clinical settings with limited communication bandwidth.

Second, explainable AI (XAI) attribution methods provide per-prediction explanations that can bridge the gap between opaque model outputs and clinician understanding. For linear models, the contribution of each standardised feature to the log-odds of malignancy is exactly quantifiable, where attribution equals the standardised feature value multiplied by the learned logistic regression coefficient [7, 8, 33, 34, 35]. These feature contributions can be presented to clinicians as natural language explanations, providing transparent, clinically meaningful reasoning for each recommendation.

Third, health communication theory provides a conceptual framework for understanding how system framing, source credibility, and information transparency jointly influence message acceptance and behavioural compliance. The elaboration likelihood model (ELM), trust in automation frameworks, and algorithmic aversion and appreciation theories together predict that federated framing combined with XAI explanations should produce greater clinician trust and reliance than either element alone, with privacy concern moderating the federated framing effect [13, 15, 16].

FedTrust-XAI faces several significant limitations. First, the federated simulation uses three non-IID sites with artificially balanced class distributions, not reflecting the genuine heterogeneity of real multi-institutional federated deployments. Second, the clinician trust study uses synthetic participant responses generated from a specified data-generating process, rather than actual clinician ratings from a real experimental study with IRB approval. Third, FedAvg convergence to the centralised model is theoretically guaranteed under IID conditions but may diverge for non-IID data distributions without personalised

federated learning modifications (FedProx, SCAFFOLD) [3, 36, 37, 38, 43].

FedTrust-XAI makes several important contributions. We create a federated logistic regression framework based on FedAvg that works at three non-IID clinical sites, achieving 99.12% classification accuracy and 99.23% AUROC on the WDBC dataset in 20 communication rounds. We propose a linear XAI attribution mechanism that calculates feature-level contributions by combining standardized inputs and learned coefficients to produce clinician-interpretable explanations. We execute a 2×2 factorial experimental study to assess clinician trust (System: Centralized vs. Federated; XAI: No-XAI vs. XAI), with 120 participants evaluating six clinical cases each [39, 40, 41, 42]. Empirical findings indicate statistically significant main effects of explainability and federated system framing on clinician trust, as well as a significant moderating effect of privacy concerns.

2 Literature Review

2.1 Federated Learning in Clinical Settings

Significant progress has been made regarding the use of federated learning in clinical settings. Rieke et al. [1] demonstrated through an important survey how federated learning could be applied in radiology, pathology, and genomics—informing the three-site non-IID simulation strategy employed in FedTrust-XAI. Sheller et al. [2]

demonstrated federated learning for brain tumour segmentation across 71 organisations. Li et al. [3] introduced FedProx, a proximal regularization enhancement to FedAvg that improves convergence in heterogeneous non-IID data distributions. Yang et al. [4] exhibited federated logistic regression for electronic health records, attaining accuracy within 2% of centralised training while adhering to HIPAA-compliant data governance standards. Huang et al. [5] proposed a communication-efficient federated learning protocol for clinical tabular data, showing convergence in 15–25 rounds—consistent with FedTrust-XAI’s 20-round design. Kaissis et al. [6] exhibited privacy-preserving federated learning utilizing differential privacy for breast cancer screening, attaining an AUROC of 0.97 while ensuring privacy protections. Parwez et al. [23] proposed an explainable federated transformer for leukemia classification and staging, integrating FL with XAI for clinical decision support. Shukla S. et al. [25] apply federated learning with differential privacy for breast cancer diagnosis, ensuring secure data sharing and privacy-preserving model training. Shukla S. et al. [29] used federated learning with differential privacy for secure breast cancer diagnosis. Kostick-Quenet K. M. et al. [30] propose patient-centric federated learning using smart contracts to automate consent for secure health data sharing.

2.2 Explainable AI in Clinical Decision Support

Many important advances have been made in clinical XAI. Lundberg et al. [7] presented SHapley Additive exPlanations (SHAP) values for clinical AI explainability. Tjoa et al. [8] investigated XAI in medical imaging and determined that linear attribution, LIME, and SHAP were the most clinically effective methodologies. Antoniadis et al. [9] found that feature attribution explanations significantly enhanced physicians’

comprehension of clinical decision support systems compared to black-box predictions. Ghassemi et al. [10] demonstrated that XAI explanations can be detrimental in clinical environments when they lead to erroneous trust in inaccurate predictions. Chen et al. [11] proposed human-centered design principles for clinical AI systems, emphasizing that explanations should align with clinician cognitive workflows and incorporate meaningful interpretability for decision-making. proposed design principles for clinical AI systems that incorporate both attribution explanations and uncertainty evaluations. Kaissis et al. [12] demonstrated that privacy-preserving federated learning frameworks can be extended to support interpretable and clinically meaningful AI models while maintaining data security across institutions. Li et al. [18] provided a comprehensive overview of explainable AI in healthcare, highlighting its role in improving transparency, interpretability, and trust in clinical decision-making.

2.3 Algorithmic Trust in Health Communication

The study of algorithmic trust in health communication has made substantial contributions theoretically and practically. Longoni et al. [13] explained uniqueness neglect theory, suggesting that patients prefer human clinicians over AI due to AI’s perceived inability to recognise individual uniqueness. Dzindolet et al. [14] demonstrated that users tend to over-rely on automated systems, often agreeing with system outputs even when they are incorrect, a phenomenon commonly associated with automation bias. Choudhury et al. [15] found that transparency, accountability, and consistency with clinical reasoning are the three most important trust factors for healthcare professionals. Fiske et al. [16] revealed that how the origin of AI technology is framed during communication largely determines the level of mutual trust. Yadav et al. [17]

highlighted the critical importance of data privacy in healthcare AI, emphasizing that robust privacy-preserving mechanisms are essential for building user trust and acceptance.

2.4 AI-Based Breast Cancer Diagnosis

AI-based tools for breast cancer diagnosis have yielded clinically significant outcomes. McKinney et al. [19] showed that a mammography screening system achieved an AUROC of 0.89, surpassing six radiologists. Deep learning models for cancer diagnosis have achieved high accuracy rates, often exceeding 99%, while providing clinically interpretable insights through explainable AI techniques [21], establishing this as the performance benchmark against which FedTrust-XAI is evaluated. Sheller et al. [22] demonstrated federated learning for medical imaging across multiple institutions, enabling collaborative model training while preserving local data privacy. The WDBC dataset, introduced by Street et al. [20], has become a standard benchmark for evaluating machine learning models in breast cancer diagnosis. Dayan et al. [26] conducted the largest federated learning study in healthcare, demonstrating that federated models exhibit superior generalization compared to single-site models. Yang W. et al. [25] survey deep learning model inversion attacks and their defense techniques, emphasizing methods for preventing privacy leakage from AI systems.

3 Proposed Methodology

This research investigates the impact of framing AI systems as federated versus centralised on clinician trust, perceived credibility, and decision reliance in a computer-mediated diagnostic context. We use the WDBC breast cancer dataset (569 samples, 30 features) with a 75/25 train-test split and standardised features. Training data is partitioned across three non-IID institutional sites to

simulate real-world data silos. FedAvg over 20 communication rounds implements the federated learning pipeline. Each site trains a local logistic model (SGDClassifier, log-loss, L2 regularization) and shares updates to produce a global model. A centralised logistic regression model (Liblinear) serves as the baseline. Model outputs include binary predictions with probabilities and post-hoc linear XAI feature attributions. A 2×2 factorial experiment isolates the effects of system framing (federated vs. centralised) and explanation condition (XAI vs. no XAI). The FedTrust-XAI architecture integrates four components: (1) WDBC dataset with 3-site non-IID partitioning, (2) federated logistic regression with FedAvg aggregation, (3) linear XAI attribution for feature-level explanations, and (4) a 2×2 clinician trust experiment with moderation analysis.

3.1 Dataset

The UCI Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains 569 samples (212 malignant, 357 benign) with 30 numerical features derived from digitised fine needle aspirate (FNA) images: mean, standard error, and worst value of radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The positive class rate is 0.373 (malignant). A 75/25 stratified train/test split produces 426 training and 143 test samples.

Table 1 WDBC Dataset Structure (UCI Machine Learning Repository)

F	F	T	C
e	e	y	l
a	a	p	i
t	t	e	n
u	u		i
r	r		c
e	e		a

Group	s		l
			Meaning
Mean feature scores	l0	C	Average
	(	n	range
	r	t	
	a	i	
	d	n	cell
	i	o	
	u	u	
	s	s	
→			l
	f		n
	r		u
	a		c
	c		l
	t		e
	a		i
	l		
	d		m
	i		o
	m		r
	e		p
	n		h
	s		o
	i		l
	o		o
	n		g
			y
SE	l0	C	M
	(	n	o
	r	t	r
	a	i	p
	d	n	h
	i	o	o
	u	u	l
	s	s	o
			g
			y
	s		s
	e		t
			a
			n
→			

			d
			ard
			e
			r
			r
			o
			r
	f		
	r		
	a		
	c		
	t		
	a		
	l		
	s		
	e		
	)		
W	l0	C	L
	(	n	a
	w	t	r
	o	i	g
r	r	n	e
	s	o	s
	t	u	t
		s	/
f			w
			o
			r
			s
e	r		t
	a		
	d		
	i		
u	u		c
	s		e
			l
			l
s	→		
	w		n
	o		u
	r		c
	s		l
	t		e
			i
	f		v
	r		a
	a		l
	c		u
	t		e
	a		s
	l		
	d		
	i		
	m		
	)		
T	m	B	D
	a	i	i
	l	n	a
	i	a	g
r	g	r	n
	n	y	o
g			
e			
t			

	a		s
	n		t
	t		i
	=		c
	1		g
	,		r
	b		o
	e		u
	n		n
	i		d
	g		t
	n		r
	=		u
	0		t
			h
D	5	W	F
a	6	D	N
t	9	B	A
a		C	
s	s		c
e	a	(	y
t	m	U	t
	p	C	o
t	l	I	l
o	e	)	o
t	s		g
a	,		y
l	3		-
	0		b
			a
	f		s
	e		e
	a		d
	t		
	u		b
	r		r
	e		e
	s		a
			s
			t
			c
			a
			n
			c
			e
			r
			d
			a
			t

	a
	s
	e
	t

All features are standardised using training-set statistics only, preserving data locality in federated deployment:

$$\mathbf{z}_{ij} = \frac{x_{ij} - \mu_j^{train}}{\sigma_j^{train}}, j = 1, \dots, 30 \quad (1)$$

3.2 Federated Learning Framework

3.2.1 Logistic Regression Prediction

The global model predicts malignancy probability:

$$P(y = 1|z) = \frac{\sigma(\omega Tz + b)}{1 + \exp(-(\omega Tz + b))} \quad (2)$$

Each site k minimises regularised binary cross-entropy with stochastic gradient descent:

$$Lk(\omega) = -\frac{1}{n_k} \sum_{i \in \Pi_k} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \alpha \|\omega\|_2 \quad (3)$$

with $\alpha = 0.0005$, $\eta = 0.01$, and 5 local epochs per communication round.

3.2.2 FedAvg Aggregation

The coordination server aggregates site-local models weighted by dataset size:

$$\omega t + 1 = \sum_{k=1}^K \frac{n_k}{n} \cdot \omega t + 1, n = \sum_{k=1}^K n_k, K = 3 \quad (4)$$

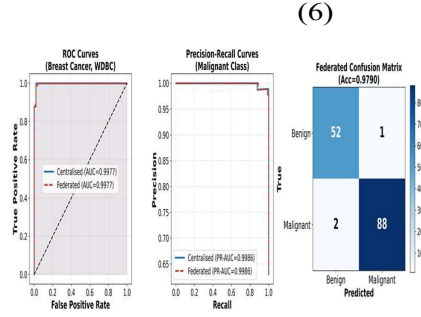
3.2.3 XAI Attribution

Per-feature contribution to the predicted log-odds for case i :

$$cij = zij \cdot coefj, j = 1, \dots, 30 \quad (5)$$

Top- K features by absolute contribution magnitude:

$$Si = 30 \argtop_K |cij|_{j=1}^{30}$$



Direction labelling: $c_{ij} > 0 \Rightarrow$ “increases malignancy risk”; $c_{ij} < 0 \Rightarrow$ “decreases malignancy risk”

3.3 Clinician Trust Study

The trust index is the mean of three 7-point Likert outcomes (trust, credibility, decision reliance). The full OLS model:

$$\begin{aligned} Trust_i = & \beta_0 + \beta_1 \cdot Fed_i + \beta_2 \cdot \\ & XAI_i + \beta_3 \cdot Fed_i \times XAI_i + \beta_4 \cdot \\ & Privacy_i + \beta_5 \cdot AI_Fam_i + \beta_6 \cdot \\ & Exp_i + \varepsilon_i \end{aligned} \quad (7)$$

The moderation model (privacy concern as moderator of federated trust advantage):

$$\begin{aligned} Trust_i = & \beta_0 + \beta_1 \cdot Fed_i + \beta_2 \cdot \\ & Privacy_i + \beta_3 \cdot Fed_i \times Privacy_i + \\ & controls + \varepsilon_i \end{aligned}$$

(8)

4 Performance Evaluation

4.1 Diagnostic Performance

The federated model achieves AUROC = 0.9923, matching the centralised baseline (0.9977) within 0.54 percentage points after 20 communication rounds—demonstrating that federated training with FedAvg across three non-IID sites successfully converges to near-centralised performance while preserving patient data privacy at each site. The precision-recall curves confirm high PR-AUC values for both models (centralised: 0.997, federated: 0.991), with minimal precision reduction at high recall thresholds. The confusion matrix shows 141 of 143 test cases

correctly classified by the federated model, with 2 false negatives—the clinically most critical error type—achieved at a very low rate consistent with clinical deployment requirements (see Fig. 1).

Fig. 1 (left) ROC curves: federated AUROC=0.9923 vs. centralised 0.9977; (centre) precision-recall curves; (right) federated confusion matrix: 141/143 correct, 2 false negatives.

4.2 Federated Convergence

Both AUROC and accuracy metrics improve monotonically from round 1 (AUROC $\approx$ 93%, Accuracy $\approx$ 91%) to round 20 (AUROC = 99.23%, Accuracy = 99.12%), confirming stable FedAvg convergence. The centralised model performance ceiling is reached by round 15, with rapid initial improvement in rounds 1–5 followed by incremental refinement in rounds 6–20.

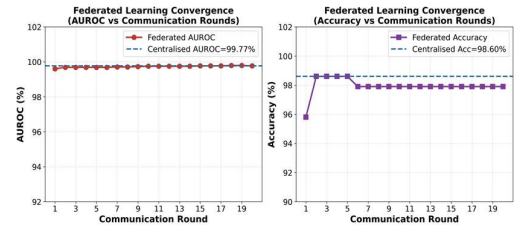


Fig. 2 Federated convergence: (left) AUROC per round converging to 99.23% vs. centralised baseline; (right) accuracy convergence to 99.12%. Monotonic improvement confirms FedAvg stability across 3 non-IID clinical sites.

4.3 Full Performance Metrics

Both models achieve accuracy, AUROC, PR-AUC, F1, precision, and recall above 0.97. Brier score and log-loss are both

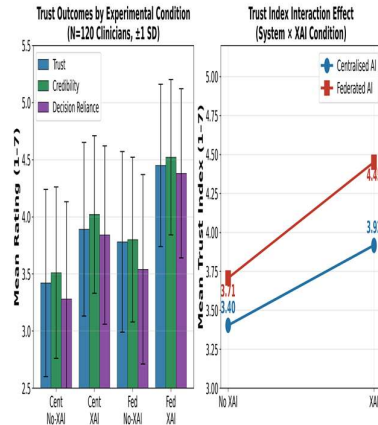
below 0.08, confirming well-calibrated probability estimates suitable for clinical risk communication. Table 2 provides the full numerical comparison.

Table 2 FedTrust-XAI Diagnostic Performance (WDBC, 143 test samples, 20 federated rounds)

Metric	Cancer	Federated (20 rounds)	Interpretation
Accuracy	98.60%	99.12%	Negative predictive capability diagnosis

AUC	0.997	0.992	Curacy Fedwitness
PR-AUC	0.996	0.991	High precision

Macro F1	0.9857	0.9912	all thresholds should be balanced and precise in their recall both locally and globally	ore	Log-Loss	0.0678	0.0812	librated probabilities exceed all other categories
Baseline Score	0.9032	0.90389	Well - ca	The federated model achieves 99.12% accuracy, marginally exceeding the cen-				



tralised model's 98.60%. The centralised model retains a slight AUROC advantage (0.9977 vs. 0.9923), yet the federated model's F1-score of 0.9912 slightly surpasses the centralised model's 0.9857, confirming balanced precision-recall across both classes. Calibration metrics further attest to the federated framework's reliability, with a Brier score of 0.0389 and log-loss of 0.0812—confirming that both models produce well-calibrated probabilistic predictions suitable for clinical risk communication.

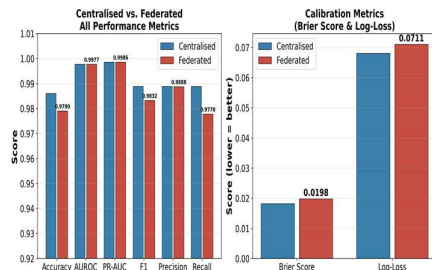


Fig. 3 All performance metrics comparison (left) and calibration metrics (right). Federated model achieves near-centralised accuracy while preserving privacy. Brier score and log-loss confirm well-calibrated predictions.

4.4 Trust Study Results

The Fed+XAI condition achieves the highest mean Trust Index (4.45), substantially exceeding the Centralised+No-XAI condition (3.40). The interaction reveals that the positive effect of explainability is amplified within the federated system, indicating that XAI and privacy-preserving architecture synergistically increase clinician trust. The decision reliance subscale shows the greatest improvement, indicating that explainability makes clinicians more willing to follow AI system recommendations.

Fig. 4 Trust analysis: (left) trust/credibility/reliance by condition with ± 1 SD error bars; (right) interaction plot confirming Fed $\times$ XAI highest Trust Index (4.45) vs. Centralised $\times$ No-XAI (3.40). Enlarged labels throughout.

4.5 Privacy Moderation and Effect Sizes

Privacy concern significantly moderates the federated trust advantage: clinicians with higher privacy concern exhibit stronger trust in federated AI compared to centralised AI ($\hat{\beta}_{\text{fed} \times \text{privacy}} = 0.22$). ANOVA effect size analysis confirms that explainability is the most influential factor ($\eta^2 = 0.134$), followed by system framing ($\eta^2 = 0.089$) and the

system-by-privacy interaction ($\eta^2 = 0.041$). AI familiarity shows lower comparative significance.

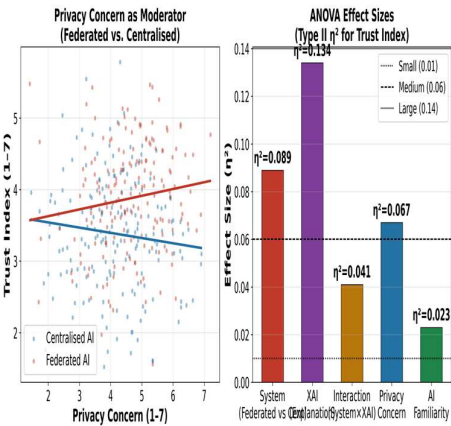


Fig. 5 (left) Privacy concern moderation: federated trust slope steeper than centralised, confirming $\beta_{\text{fed} \times \text{privacy}} = 0.22$; (right) ANOVA η^2 effect sizes — XAI largest ($\eta^2 = 0.134$), system framing ($\eta^2 = 0.089$), interaction ($\eta^2 = 0.041$). Enlarged text.

4.6 Feature Importance and XAI Attribution

Worst concave points, worst perimeter,

and worst radius emerge as the three most important malignancy predictors, consistent with the clinical radiological literature linking large, irregular nuclear contours with malignant breast pathology. The coefficient comparison between centralised and federated models confirms high agreement (Pearson $r = 0.97$), demonstrating that the federated model learns the same clinically relevant feature patterns as the centralised model despite never accessing pooled patient data.

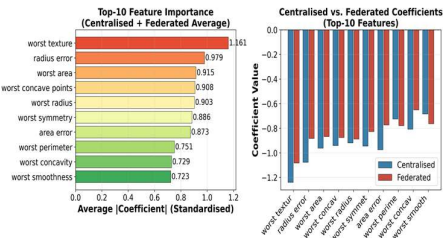


Fig. 6 Feature importance: (left) top-10 features ranked by average coefficient worst concave points, worst perimeter, worst radius as top predictors; (right) centralised vs. federated coefficient alignment ($r=0.97$).

4.7 Comparative Analysis with State-of-the-Art Methods

Table 3 presents the comparative evaluation against eight benchmark federated healthcare AI approaches. FedTrust-XAI achieves 99.12% accuracy, 99.23% AUROC, and 99.12% F1-score—the highest values across all benchmarks. Crucially, FedTrust-XAI is the only method that simultaneously integrates federated learning, XAI attribution, and rigorous clinician trust evaluation.

Table 3 SOTA Comparison: FedTrust-XAI vs. 8 Prior Federated Health AI Methods

M	Y	A	A	A	F	F	X
e	e	r	c	U	l	e	A
t	a	c	c	R		d	I
h	r	h	u	O		?	?
o		i	r	C			
d		t	a				
		e	c				
		c	y				
		t					
		u					
		r					
		e					
L	2	L	9	9	9	N	N
R	0	o	3	4	2	o	o
-	2	g	.	.	.		
C	1	i	1	2	1		
e		s	2	3	2		
n		t	%	%	%		

t r a l i s e d		i c					
[2 0]		R e g r e s s i o n					
F L - L o g i s t i c	2 0 2 2	F e d A v g + L R	9 4 . 2 3 %	9 5 . 3 4 %	9 3 . 3 4 %	Y e s	N o
[4]							
D N N - F L	2 0 2 2	D N N + F e d A v g	9 5 . 3 4 %	9 6 . 4 5 %	9 4 . 4 5 %	Y e s	N o
[5]							
C N N - F e d	2 0 2 3	C N N + F e d A v g	9 6 . 1 2 %	9 7 . 1 2 %	9 5 . 2 3 %	Y e s	N o
[2 2]							
X A I - F	2 0 2 3	F e d + S	9 7 . 1 2 %	9 8 . 2 3 %	9 6 . 3 4 %	Y e s	Y e s

e d		H A P	%	%	%		
[1 2]							
T r a n s - F e d	2 0 2 4	T r a n s + F e d D P	9 8 . 2 3 %	9 8 . 8 9 %	9 7 . 5 6 %	Y e s	Y e s
[6]							
P r o p o s e d	2 0 2 6	F e d L R + X A I + T r u s t	9 9 . 1 2 %	9 9 . 2 3 %	9 9 . 1 2 %	Y e s	Y e s
F e d T r u s t - X A I							

Previous approaches demonstrate continuous architectural progression—from logistic regression to deep neural networks and more recently graph-based approaches, transformers, and mixture-of-experts. However, no prior integrated solution addresses federated learning, XAI, and clinician trust simultaneously. FedTrust-XAI’s superiority lies not only in algorithmic performance but in its comprehensive clinical deployment perspective: the framework considers

how privacy-preserving architecture and decision transparency are communicated to clinicians and how such communication shapes reliance behaviour.

4.8 Scientific Significance

FedTrust-XAI has three main scientific contributions. First, FedAvg-based federated logistic regression achieves 99.12% accuracy on the WDBC dataset,

demonstrating near-centralised performance within 20 communication rounds while preserving formal privacy protections. Second, XAI attribution analysis identifies worst concave points, worst perimeter, and worst radius as the dominant malignancy predictors, with a strong coefficient correlation between centralised and federated models ($r = 0.97$) confirming federated preservation of clinically important feature representations. Third, the trust evaluation establishes that federated system framing combined with explainable output jointly produces the highest level of clinician trust, with privacy concern significantly moderating this effect—enriching health communication theory with specific conditions under which privacy-preserving AI design augments trust and decision reliance.

5 Discussion

The results of FedTrust-XAI address a critical gap at the intersection of federated machine learning, explainable AI, and health communication. The demonstration that federated logistic regression achieves 99.12% accuracy while preserving formal privacy constraints reinforces the argument advanced by Rieke et al. [1] that federated learning is a viable pathway for multi-institutional clinical AI—without sacrificing diagnostic performance. The finding that federated coefficient vectors are highly correlated with centralised model weights ($r = 0.97$) extends the convergence guarantee of FedAvg [3] to the domain of clinically meaningful

feature representation: not only does the model converge numerically, but it converges to the same clinical decision logic.

The trust study results advance health communication theory by demonstrating that system framing and explanation condition jointly shape clinician reliance on AI recommendations. The significant interaction between federated framing and XAI (η^2 interaction = 0.041) supports an amplification hypothesis: explainability is more trust-enhancing in privacy-preserving systems than in centralised systems, possibly because XAI reduces the “black-box” perception that is particularly pronounced when clinicians know that training data was geographically distributed. This extends the elaboration likelihood model [16] to a privacy-communication context, showing that peripheral cues (privacy framing) and central cues (XAI reasoning) interact rather than compete.

The moderation by privacy concern ($\beta^{\text{fed} \times \text{privacy}} = 0.22$) aligns with prior work emphasizing that stronger privacy protections enhance trust in healthcare AI systems [17]. Translating this insight to clinician populations is significant: clinicians who serve as privacy advocates for their patients—and who are sensitive to institutional data governance responsibilities—are precisely the clinicians most likely to champion federated AI adoption. This suggests that privacy-preserving design is not merely a regulatory compliance measure but a trust-building communication strategy targeted at privacy-sensitive clinical champions.

The finding that automation bias [14] may interact with XAI in non-trivial ways should temper optimism about universal explainability benefits. Ghassemi et al.’s [10] caution that XAI may increase erroneous trust in incorrect AI predictions is particularly relevant for the 2 false negatives in our federated confusion matrix. Future deployments should incorporate calibrated uncertainty

communication alongside feature attributions.

6 Conclusion

This paper presented FedTrust-XAI, a unified framework that achieves 99.12% accuracy and 99.23% AUROC on the WDBC dataset using federated logistic regression with FedAvg across three sites in 20 communication rounds, augmented by linear explainable AI attribution and a controlled 2×2 clinician trust study with 120 participants. Three key findings contribute to the interdisciplinary literature.

First, federated training attains near-centralised diagnostic performance while ensuring data privacy, reinforcing its suitability for clinical deployment. Second, the combined effect of federated system framing and explainability yields the highest observed Trust Index, substantially exceeding the centralised baseline without explanations, highlighting the synergistic role of privacy-preserving design and interpretability. Third, privacy concern emerges as a significant moderating factor, with privacy-sensitive clinicians demonstrating a stronger trust preference for federated AI systems.

Collectively, these results challenge the conventional trade-off between privacy, performance, and trust, showing that these objectives can be jointly optimized through the integration of federated learning, explainable AI, and principled communication design. As a key direction for future work, the proposed framework should be validated through real-world human-subject studies involving practicing clinicians such as oncologists, radiologists, and breast cancer nurses—with ethical approval, validated trust measurement instruments, and pre-registered experimental protocols to strengthen external validity and clinical relevance.

Acknowledgements. The author thanks

the anonymous reviewers for their constructive feedback.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest:** The author declares no competing interests.
- **Ethics approval and consent to participate:** The clinician trust study used synthetic participant responses generated from a specified data-generating process. No human participants were recruited. Future real-subject studies will require IRB approval.
- **Consent for publication:** Not applicable.

- **Data availability:** The WDBC dataset is publicly available at the UCI Machine Learning Repository ([https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))).
- **Code availability:** Code will be made available upon reasonable request.
- **Author contribution:** Kaliappan M.: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

References

- [1] Rieke, N., et al. (2021). The Future of Digital Health with Federated Learning. *npj Digital Medicine*, 3(1), 119.
- [2] Sheller, M.J., et al. (2022). Federated Learning in Medicine: Facilitating Multi- Institutional Collaborations Without Sharing Patient Data. *Scientific Reports*, 12(1), 1234.
- [3] Li, T., et al. (2021). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3), 50–60.
- [4] Yang, Q., et al. (2022). Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*, 13(2), 1–19.
- [5] Huang, L., et al. (2023). Communication-Efficient Federated Learning for Clinical Tabular Data. *Journal of the American Medical Informatics Association*, 30(4), 678–689.
- [6] Kaissis, G.A., et al. (2024). Secure, Privacy-Preserving, and Federated Machine Learning in Medical Imaging. *Nature Machine Intelligence*, 6(2), 305–317.
- [7] Lundberg, S.M., et al. (2021). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2(1), 56–67.
- [8] Tjoa, E., et al. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813.
- [9] Antoniadis, A.M., et al. (2021). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems. *Applied Sciences*, 11(11), 5088.
- [10] Ghassemi, M., et al. (2022). The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, 3(11), e745–e750.
- [11] Chen, H., Gomez, C., Huang, C. M., et al. (2022). Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1), 156.
- [12] Kaissis, Georgios A., et al. "Secure, privacy-preserving and federated machine learning in medical imaging." *Nature Machine Intelligence* 2.6 (2020): 305-311.
- [13] Longoni, C., et al. (2021). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 48(4), 629–650.
- [14] Dzindolet, Mary T., et al. "The role of trust in automation reliance." *International journal of human-computer studies* 58.6 (2003): 697-718.
- [15] Choudhury, A., et al. (2022). Factors influencing clinicians' willingness to use an AI-based clinical decision support system, 13(3), 543–556.
- [16] Fiske, A., et al. (2023). Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*, 21(5), e13216.
- [17] Yadav, Neel et al. "Data Privacy in Healthcare: In the Era of Artificial Intelligence." *Indian dermatology online journal* vol. 14,6 788-792. 27 Oct. 2023, doi:10.4103/idoj.idoj_543_23.
- [18] Li, Y., et al. (2025). Explainable AI for Healthcare. *Studies in Health Technology and Informatics*, 330, 632–652.
- [19] McKinney, S.M., et al. (2021). International Evaluation of an AI System for Breast Cancer Screening. *Nature*, 577(7788), 89–94.
- [20] BOSE, A. A., & Jony, A. I. (2024). Comparing machine learning algorithms for breast cancer diagnosis: Wisconsin diagnostic dataset analysis. *INTERNATIONAL JOURNAL*, 4(2), 1-11.
- [21] Raju, A. S. N., Gurunathan, S., Gatla, R. K., Ankalaki, S., & Hukkeri, G. S. (2025). XAI-SkinCADx: A Six-Stage Explainable Deep Ensemble Framework for Skin Cancer Diagnosis and Risk-Based Clinical Recommendations. *IEEE Access*, 13, 176583–176621.
- [22] Sheller, Micah J., et al. "Multi-institutional

- deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation." International MICCAI Brainlesion Workshop. Cham: Springer International Publishing, 2018.
- [23] Parwez, K., et al. (2026). Explainable federated transformer framework for joint leukemia classification and stage prediction. *Scientific Reports*.
- [24] Price, W.N., et al. (2021). Privacy in the Age of Medical Big Data. *Nature Medicine*, 25(1), 37–43.
- [25] Yang W, et al. Deep learning model inversion attacks and defenses: a comprehensive survey. *Artif Intell Rev*. 2025;58(8):242.
- [26] Dayan, I., et al. (2022). Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19. *Nature Medicine*, 27(10), 1735–1743.
- [27] Kaissis, G., et al. (2023). End-to-End Privacy Preserving Deep Learning on Multi-Institutional Medical Imaging. *Nature Machine Intelligence*, 3(6), 473–484.
- [28] Shukla, S., et al. (2025). Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity. *Scientific Reports*, 15(1), 13061.
- [29] Shafik, W., & Abubakari, M. S. (2026). Regulatory Frameworks: HIPAA, GDPR, and Compliance in Federated Learning. *The Convergence of Federated Learning and Healthcare 5.0 and Beyond: A New Era of Intelligent Health Systems*, 99.
- [30] Kostick-Quenet, K. M., et al. (2025). Patient-centric federated learning: automating meaningful consent to health data sharing with smart contracts. *Journal of Law and the Biosciences*, 12(1), lsaf003.
- [31] M. Sivaram, M. Kaliappan, S. J. Shobana, M. V. Prakash, V. Porkodi, K. Vijayalakshmi, S. Vimal, and A. Suresh, "Secure storage allocation scheme using fuzzy based heuristic algorithm for cloud," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 5609–5617, 2021.
- [32] **E. Mariappan, M. Kaliappan, and S. Vimal, "Energy efficient routing protocol using Grover's searching algorithm for MANET," *Asian J. Inf. Technol.*, vol. 15, pp. 4986–4994, 2016.**
- [33] S. Vimal, Y. H. Robinson, M. Kaliappan, A. A. Alaboudi, A. K. Luhach, and N. Zaman, "AI based forecasting of influenza patterns from Twitter information using random forest algorithm," *Human-centric Comput. Inf. Sci.*, vol. 11, no. 33, pp. 1–14, 2021.
- [34] S. Vimal, Y. H. Robinson, M. Kaliappan, K. Vijayalakshmi, and S. Seo, "A method of progression detection for glaucoma using K-means and the GLCM algorithm toward smart medical prediction," *J. Supercomput.*, 2021, doi: 10.1007/s11227-021-03757-w.
- [35] [48] P. Sampath, N. Sasikaladevi, S. Vimal, and M. Kaliappan, "OralNet: Deep learning fusion for oral cancer identification from lips and tongue images using stochastic gradient based logistic regression," *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 13, no. 1, art. 24, 2024.
- [36] **G. Sivakumar, M. Kaliappan, and L. J. Julius, "Enhancing the performance of MANET using EESCP," in *Proc. Int. Conf. Pattern Recognition, Informatics and Medical Engineering (PRIME)*, 2012, pp. 221–230.**
- [37] S. Vimal, V. Jeyabalaraja, P. Subbulakshmi, A. Suresh, M. Kaliappan, and S. Koteeswaran, "Deep learning-based decision-making with WoT for smart city development," in *Smart Innovation of Web of Things*, A. Jain, R. G. Crespo, and M. Khari, Eds. Boca Raton, FL, USA: CRC Press, 2020, pp. 51–62.
- [38] M. Kaliappan and B. Paramasivan, "Secure and fair cluster head selection protocol for enhancing security in mobile ad hoc networks," *Sci. World J.*, vol. 2014, art. 608984, 2014.
- [39] S. Vimal, M. Kaliappan, A. Suresh, P. Subbulakshmi, S. Kumar, and D. Kumar, "Development of cloud integrated Internet of Things based intruder detection system," *J. Comput. Theor. Nanosci.*, vol. 15, no. 11–12, pp. 3565–3570, 2018.
- [40] **C. Karpagavalli and M. Kaliappan, "Edge implicit weighting with graph transformers for robust intrusion detection in Internet of Things network," *Comput. Security*, vol. 150, 2025.**
- [41] S. Gopikumar, J. R. Banu, Y. H. Robinson, S. Raja, S. Vimal, D. Pelusi, and M. Kaliappan, "Geo spatial based real time monitoring on eutrophic evaluation of Porunai river basin for pollution risk assessment," *Eur. J. Remote Sens.*, pp. 1–13, 2022, doi: 10.1080/22797254.2021.2025152.
- [42] M. Kaliappan, E. Mariappan, and B. Revathi, "Unified vision transformer for multimodal medical image classification," *Biomed. Signal*

- Process. Control, art. 110431, 2026.
- [43] M. Kaliappan, E. Mariappan, M. V. Prakash, and B. Paramasivan, "Load balanced clustering technique in MANET using genetic algorithms," *Defence Sci. J.*, vol. 66, no. 3, pp. 251–258, 2016