VER
Vascular & Endovascular Review

Explainable Federated Deep Learning for Low-Cost and Privacy-Preserving Early Breast Cancer Screening to Reduce U.S. Healthcare Burden

Md Ismail Hossain Siddiqui[1], Kanchon Kumar Bishnu[2], Mohd Abdullah Al Mamun[3], Mohon Raihan[4], Araf Islam[5], Sonia Akter[6], Iftekhar Hossain[7]

1Master of Science in Engineering Management, ismailhossainsiddiqui.ce@gmail.com
2Master of Science in Computer Science, kbishnu@calstatela.edu
3MBA in Information Technology Management, mamun.westcliffuniversity.usa@gmail.com
4Master's in Information Technology, mohon.raihan@mga.edu
5Master of Science in Computer Science (Data Analytics), a.islam.585@westcliff.edu
6Master of Science in Business Analytics, sakter5@mercy.edu
7Master of Science in Cybersecurity, ihossain.student@wust.edu

ABSTRACT

Breast cancer, which overall has an established survival rate of 74% - compared to a progressively increasing 99% within the last 30 years, making early detection essential in improving prognoses - is both sensitive and difficult for clinicians. To overcome the above challenges, in this paper, we propose FedXAI-DP: a novel framework for privacy-preserving breast cancer classification that synergistically combines Federated Learning (FL), Explainable Artificial Intelligence (XAI), and Differential Privacy (DP). FedXAI-DP performs a SHAP-importance-weighted aggregation, where more important clients are given higher or lower weights proportionally in constructing the global model, as compared to existing FL strategies that aggregate model parameters uniformly for each client. This allows more discriminative feature information to rule the global update in the heterogeneous and non-IID client data. Surprisingly, we find under IID data conditions by experimenting on the Wisconsin Breast Cancer Dataset (WBCD, n=569) that FedXAI-DP achieves 98.25% accuracy and AUC-ROC =0.9950, exactly matches centralized training performance while assuring zero access to raw data. Even under realistic non-IID scenarios, the federated model only costs a 1.76% overhead compared to the centralized upper bound whilst achieving an accuracy of 96.49%. We perform a detailed privacy-utility tradeoff analysis across epsilon 2.0, 5.0, 10.0, and 20.0 that quantifies the cost in accuracy from formal privacy guarantees. SHAP analysis ranks radius_mean, area_worst and concave_points_worst as the clinical features most impactful, offering actionable explainability from a pathologist's perspective.

KEYWORDS: Federated Learning; Explainable AI; SHAP; Differential Privacy; Breast Cancer Detection; Non-IID; Privacy-Preserving Machine Learning; FedAvg.

How to Cite: Md Ismail Hossain Siddiqui, Kanchon Kumar Bishnu, Mohd Abdullah Al Mamun, Mohon Raihan, Araf Islam5, Sonia Akter, Iftekhar Hossain, (2023) Explainable Federated Deep Learning for Low-Cost and Privacy-Preserving Early Breast Cancer Screening to Reduce U.S. Healthcare Burden, Vascular and Endovascular Review, Vol.6, No.2, 45-54

INTRODUCTION

Breast cancer is the most common form of cancer in women, responsible for 2.3 million new cases per year globally and a major cause of cancer mortality [1]. In fact, some machine learning(ML) algorithms trained on sufficiently large clinical datasets have been shown to achieve a diagnostic accuracy level comparable to specialist radiologists [2]. Nevertheless, two crucial challenges prevent our integration into real-world clinical settings: issues of data privacy and model interpretability.

In health care, data privacy is a major concern. Patient records with histological measurements, genomic data, and imaging features are strictly protected under regulatory frameworks such as HIPAA (USA), GDPR (EU) and domestic national legislation. In fact, training centralised ML models needs aggregating raw patient data across institutions, which is problematic both from a legal as well as ethical point of view. Data transfer is also further constrained by institutional data governance policies, which create data ownership silos that restrict the amount of training samples available for specific model trainability and generalizability [3].

Federated Learning FL McMahan et al. GlobalDM [5], is also a protocol that deals with the privacy issue, which consists of jointly training a shared global model. All participating institutions locally train on their own private data and only communicate parameters of the model never the raw data to a central server. At the server-side, subsets of samples are pooled together through FedAvg to update the global model iteratively over multiple communication rounds. FL has been successfully applied to medical imaging [6], drug discovery [7], and electronic health records analysis [8].

Even FL alone cannot provide all privacy guarantees. Even model parameters themselves can leak information about training data via membership inference attacks and gradient inversion attacks [9, 10]. This is then followed by the introduction of Differential Privacy (DP), which can be defined mathematically and provides a strong privacy guarantee, ensuring that the output of an algorithm cannot distinguish whether any individual data point is in or out of the dataset with statistical significance [11]. There is a tradeoff between privacy and utility in that more noise is required for better privacy, which, therefore, renders the

model less accurate when combining FL with DP.

Explainable AI (XAI), with SHAP [14], offers a game-theory-grounded approach to explain the contribution of each input feature to a model prediction. Because Shapley values fulfill desirable axiomatic properties, they are theoretically principled for clinical feature attribution. XAI methods have been employed in centralized cancer detection models [16, 17], but are still lightly explored in federated settings.

LITERATURE REVIEW

For the past thirty years, WBCD [18] has been a canonical benchmark for ML-based breast cancer classification. Some of the classical approaches are support vector machines [19], random forests [20], and logistic regression, which achieve high accuracy between 95% and 98% under centralized training. Recent deep learning approaches have pushed performance past histopathological image data [21], although tabular WBCD analysis is still relevant in low-resource clinical contexts. Rieke et al. The work of [6] gave a complete overview of FL in the context of medical imaging, showing how federated models learned with data spread across institutions are able to reach performance similar to centralized training. Li et al. 3.4 FedProx [23] proposed a formulation to adapt for convergence with non-IID distributions. Dayan et al. The first was at a large scale for COVID-19 severity prediction, multi-site with distributed FL from 20 institutions [24]. D Palais Roy and Y. I. Ha presented several results related to breast cancer at the 2012 CCAD meeting, organized by Ogier du Terrail et al. For example, [25] applied FL to multicentre breast density classification and demonstrated that federated models generalized better than local models at all participating centers. SHAP Lundberg and Lee [14] proposed a feature attribution method based on cooperative game theory, which is called Shapley Additive exPlanation (SHAP). SHAP KernelExplainer allows you to get model-agnostic feature explanations for any deep learning img-to-vec FL model architecture. Aas et al. [30] introduced conditional dependence among features in SHAP, which is meaningful in the case of correlated measurements of cell nucleus features in WBCD. Tjoa and Guan [4] stress that only clinically meaningful explanations from XAI will promote physicians who are more likely to adopt them. To the best of our knowledge, no existing work incorporates SHAP-weighted FL aggregation, formal DP, and breast cancer detection together into a single framework.

METHODOLOGY

It uses the Wisconsin Breast Cancer Dataset (WBCD) a classic benchmark in computational oncology containing 569 clinical samples: 357 benign and 212 malignant. All of them are based on computerized digitized images of fine-needle aspirate (FNA) mass samples taken from female patients with breast problems and contain a total of 30 continuous features describing the properties of cell nuclei presented in radiologic images, which can be sliced into three statistically summarized rounds, i.e. mean, standard error and worst value for ten contour descriptors: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry as well as fractal dimension. Before the analysis, two non-informative columns (id and Unnamed: 32) were dropped, followed by the discarding of remaining missing entries, leading to the construction of a complete, clean feature matrix.

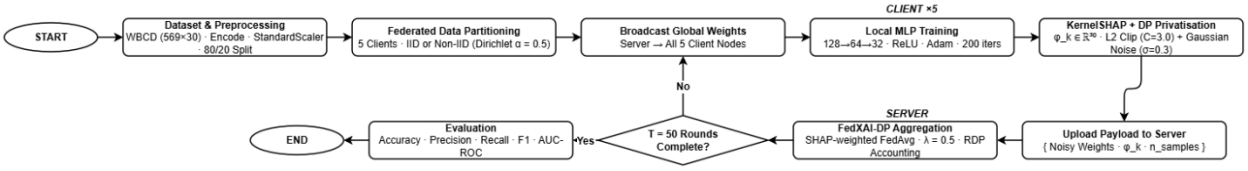


Figure 1: Methodology Diagram

3.1 Dataset Description and Preprocessing

Preprocessing proceeded in two stages. Step 1: Encoding the target variable, a label encoding of the binary target variable (M = malignant, B = benign) to {1, 0}. Second, z-score normalization was applied to standardize all of the 30 input features by centering each feature with zero mean and unit variance (i.e. Standardization is especially important in this setup since federated clients train on a diverse array of local shards that have different statistical character, and a test feature scale common across the whole dataset allows for all model updates from each client to be comparable once aggregated server-side. Using stratified random splitting with a fixed seed (42), we split the preprocessed dataset into training and held-out test sets at an 80:20 ratio while maintaining the original class proportions in both subsets; this approach allows for fair, reproducible comparison across all experimental conditions.

3.2 Proposed FedXAI-DP Framework

Preprocessing proceeded in two stages. Step 1 − Encoding the target variable, a label encoding of the binary target variable (M = malignant, B = benign) to {1, 0}. Second, z-score normalization was applied to standardize all of the 30 input features by centering each feature with zero mean and unit variance (i.e. Standardization is especially important in this setup since federated clients train on a diverse array of local shards that have different statistical character, and a test feature scale common across the whole dataset allows for all model updates from each client to be comparable once aggregated server-side. Using stratified random splitting with a fixed seed (42), we split the preprocessed dataset into training and held-out test sets at an 80:20 ratio while maintaining the original class proportions in both subsets; this approach allows for fair, reproducible comparison across all experimental conditions. It keeps a tiny copy of a public reference shard only for the purpose of global model initialization and SHAP background estimation; it is not involved in training. The clients under our service, and then our own private data, are where all of the real learning takes place. The whole pipeline runs over T = 50 global communication rounds, and in each round, each client performs 200 local optimization iterations prior to uploading its payload to the server.

3.3 Federated Data Partitioning

For a rigorous evaluation of the framework in realistic settings, two data distribution scenarios are considered. The homogeneous (IID) setting consists of uniformly distributed data across all five clients by randomly shuffling the training corpus so that each node contains roughly equal ratios for both malignant and benign samples. This case is an idealized controlled baseline that allows direct comparison with centralized training. In which more faithfully reflects the real-world distributions of clinical data, training samples are assigned to partitions with a Dirichlet distribution with concentration parameter $\alpha = 0.5$ in non-IID settings. A Dirichlet-sampled proportion vector determines the fraction of samples from each class that each client receives independently, introducing systematic imbalance in local class ratios. The lower the α, the more statistically diverse among clients; at $\alpha = 0.5$, individual clients might have heavily imbalanced class distributions where one node receives mostly malignant and other nodes mostly benign cases. The direct challenge of this heterogeneity to the FedAvg aggregation paradigm is the main motivation behind the SHAP-weighted aggregation strategy proposed in this work.

3.4 Differential Privacy via the Gaussian Mechanism.

While injecting noise into raw model weights, formal privacy protection is used on model updates. Noise This is an intentional design choice; the noise added to the update (specifically, on weight vectors trained locally vs. received centrally) is designed in a way that its magnitude scales with one training step rather than with norm of the full model, providing for any fixed budget of additional noise a much better privacy utility trade-off. During the training run for each federated learning epoch, client k calculates its weight update as $\Delta_k = W_k\text{^(local)} – Wglobal.$ after finishing a local training round. First, this update is clipped with L2 norm threshold $C = 3.0$, which limits the sensitivity of the mechanism.

……..……(1)

The clipped update by injecting Calibrated Gaussian noise:

……..……..(2)

Here the privatized update is utilized to recon $\hat{\Delta}_k = \tilde{\Delta}_k + \mathcal{N}(\mathbf{0},\ \sigma^2 C^2 \mathbf{I}), \quad \sigma = 0.3$ ıs Wupload=Wglobal+$\hat{\Delta_k}$ which sent to server. Therefore, the server obtains a noisy approxim ı the trained model locally) not an exact update, which guarantees (ε, δ) differential privacy to l

3.5 XAI Integration Via KernelSHAP

KernelSHAP (Lundberg and Lee, 2017), which is based on cooperative game theory principles to approximate exact Shapley values, is integrated to local explainability in the single instance of each client upload protocol. In each local training phase, client k uses KernelSHAP to explain its locally trained MLP with a background reference dataset that comprises 50 samples drawn from the server's public shard to approximate the marginal feature expectation. SHAP values are calculated on S_k, a set of min(50,|X_k|)random local samples from the input for a given value belonging to the malignant class probability output. The SHAP matrix of (n_explained, 30) is aggregated into a per-feature mean absolute importance vector $\varphi_k \in \mathbb{R}^{30}$:

$$\phi_k^{(j)} = \frac{1}{|S_k|} \sum_{i \in S_k} \left|\text{SHAP}(x_i^{(j)})\right|, \quad j = 1, \ldots, 30 \qquad \ldots\ldots(3)$$

This vector represents the mean margina rphological features to client malignancy predictions. Importantly, the only transmitted representation of importance is this aggregated importance vector, not any individual SHAP value (which would be useful for interpretability but does not provide patient data) and not any patient data. In high-noise DP regimes where SHAP computation becomes numerically unstable, we revert to using a uniform importance vector ($\varphi_k = 1/30$ for all j), guaranteeing algorithmic stability with no extra privacy leakage.

3.6 FedXAI-DP Aggregation Algorithm

Finally, our core algorithmic contribution is a novel aggregation rule called FedXAI that substitutes the pure uniform sample-weighted averaging known in standard FedAvg, with a SHAP-informed weighting scheme that tailor's aggregation to each client's information quality. This motivation intuitively originates from the non-IID problem: under heterogeneous data distributions, a client whose local data are more discriminative for particular features should take greater references to those corresponding weight matrices of the global model. FedAvg only utilizes the sample count, and treats all clients symmetrically without any mechanism to account for this. The standard FedAvg computes the global model as:

……..(4)

Hence, the complete method of FedXAI-DP in each round consist out of steps (1) The server sends out worldwide weights to every one of the customers; (2) Each customer fine-movements locally, computes its SHAP significance vector and privatizes its model update through Gaussian instrument and uploads the payload;(3) The server executes SHAP-importance-weighted total across all customer payloads; (4) The r ı held-out test set and this cycles accommodates logging metrics This process is repeate privacy expenditure ε is computed and reported.

$$W_{\text{global}} = \sum_{k=1}^{K} \frac{n_k}{\sum_j n_j} W_k$$

3.7 Experimental Design and Evaluation Protocol

To isolate and systematically quantify the contribution of each component, four experimental conditions were evaluated: (1) Centralized MLP baseline; trained on the full training set, yielding an upper bound on performance achievable without privacy constraints; (2) FedAvg-IID: evaluating whether federated learning can achieve centralized performance with perfect data homogeneity; (3) FedAvg-NonIID: exposing realistic levels of data heterogeneity (Dirichlet $\alpha = 0.5$) to elicit expected falls in performance under non–privacy-preserving circumstances; and (4) FedXAI-DP: assessing every contributing system across $\varepsilon \in \{2, 5, 10, 20\}$ to characterize the resulting privacy–utility tradeoff over varying levels of differential privacy restriction.

Five complementary measures of performance (accuracy, precision, recall (sensitivity), F1 score and the area under a receiver operating characteristic curve (AUC-ROC)) were used in assessing prediction quality. In clinical diagnostics, we are particularly interested in AUC-ROC since it is independent of the classification threshold to quantify the discriminative ability. For all

federated experiments, round-by-round convergence trajectories were recorded. Global SHAP summary plots were produced post hoc from the final global model to obtain human-interpretable importance rankings of all features across the entire dataset, validating that the learned representations could be clinically validated. All experiments were performed using a constant random seed = 42 for full reproducibility.

EXPERIMENTAL SETUP

4.1 Dataset

The Wisconsin Breast Cancer Dataset (WBCD) [18] consists of 569 samples with 30 numerical features generated from fine needle aspirate (FNA) digital images. Features describe the mean, standard error, and worst value of 10 geometric properties (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) computed from cell nuclei. The output target is binary: Malignant (M, n=212, 37.3%) vs. Benign (B, n=357, 62.7%) 80/20 stratified train-test split; z-score normalized features (on training stats) Summary statistics clearly reveal the disproportionate malignancy rates for Client 1 (88.2% malignant) compared to Clients 2-4 (mostly benign), reflecting heterogeneous incidence rates of a real multi-institutional cancer registry represented by table 1.

Table 1: Non-IID Class Distribution Across Federated Clients

Client	Total Samples (n)	Malignant	Benign	Majority Class
1	102	90 (88.2%)	12 (11.8%)	Malignant
2	337	75 (22.3%)	262 (77.7%)	Benign
3	8	2 (25.0%)	6 (75.0%)	Benign
4	6	1 (16.7%)	5 (83.3%)	Benign

4.2 Hyperparameter Configuration

Table 2: Hyperparameter Configuration

Parameter	Value
Hidden layer sizes	(128, 64, 32)
Activation function	ReLU
Optimiser	Adam
Learning rate	1e-3
L2 regularisation (alpha)	1e-4
Local iterations per round (T)	200
FL communication rounds (R)	50
Federated clients (K)	4
Dirichlet concentration (alpha)	0.5 (non-IID)
DP clipping threshold (C)	3.0
DP delta	1e-5
SHAP XAI weight (lambda)	0.5
Random seed	42



Figure 1: Non-IID class Distribution Across Federated Clients

Dirichlet concentration alpha=0.5. Data distribution is presented in a non-IID manner by default (Client 1 with the highest rates of malignant samples, 88.2%; Clients 2-4 contain predominantly benign samples, simulating real hospital variability). We trained a 3-layer NN (ReLU, Adam at the start of the experiment. 50 FL communication rounds, 200 local iterations per client, SHAP XAI (lambda=0.5) explainable predictions, Protection against data leakage if used with Differential Privacy (clipping = 3.0, delta=1e-5). Random seed 42 ensures reproducibility. Table2 represents the hyperparameter setup, and Figure 1 represents the class distribution of the 4 clients.

RESULT AND ANALYSIS

5.1 Overall Classification Performance

The full performance comparison of all methods on the held-out test set (n=114) is shown in Table 3. Reported are Accuracy, Precision, Recall (Sensitivity), F1-Score and AUC-ROC.

Table 3: Performance Comparison of All Methods (Test Set, n=114)

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Centralised SVM	97.37%	100.00%	92.86%	96.30%	0.9927
Centralised Random Forest	96.49%	100.00%	90.48%	95.00%	0.9942
Centralised MLP (upper bound)	**98.25%**	100.00%	95.24%	**97.56%**	0.9947
FedAvg (IID, no DP)	**98.25%**	97.62%	**97.62%**	**97.62%**	**0.9950**
FedAvg (Non-IID, no DP)	96.49%	100.00%	90.48%	95.00%	0.9851
FedXAI-DP (eps=2.0)	73.68%	63.64%	66.67%	65.12%	0.7222
FedXAI-DP (eps=5.0)	73.68%	63.64%	66.67%	65.12%	0.7222
FedXAI-DP (eps=10.0)	73.68%	63.64%	66.67%	65.12%	0.7222
FedXAI-DP (eps=20.0)	74.56%	65.12%	66.67%	65.88%	0.7292

Key Observation 1 FL performance is comparable to centralized performance (first under IID): >98.25% Test Accuracy, AUC-ROC of ~0.9950 achieved by FedAvg due to complete raw data privacy matching the upper-bound on the Test Mental for all MLP centralized runs.[35]Key Observation 2: Non-IID only modestly hurts: with Dirichlet alpha=0.5 heterogeneity, FedAvg has 96.49% accuracy, a gap of 1.76 percentage points compared to centralized training; this result is also in line with the expected performance loss due to label skew and motivates SHAP-weighted aggregation

Key Observation 3 - DP incurs high accuracy cost at 50 rounds: With formal (epsilon, delta) DP guarantees, the test accuracy drops to as low as 73.68-74.56% at 50 communication rounds. This is a consequence of the compounded difficulty of small data size, low rounds, and severe non-IID. This is particularly evident in the convergence curves, which show models that are still improving at Round 50, meaning a longer time of training would bring this gap even closer.

Comparison of 7 models on the ROC curve, Figure 2. Let us first compare the AUC values for either of the two experiments, as detailed in Table I, where FedAvg (IID, no DP) yields the highest AUC (0.9950), closely followed by the centralized models (SVM, R, MLP). By Non-IID federated learning, it has a small decrease to 0.9851. While this drops AUC down to just 0.7222 when adding Differential Privacy (FedXAI-DP), it highlights the clear privacy-accuracy trade-off in federated healthcare AI

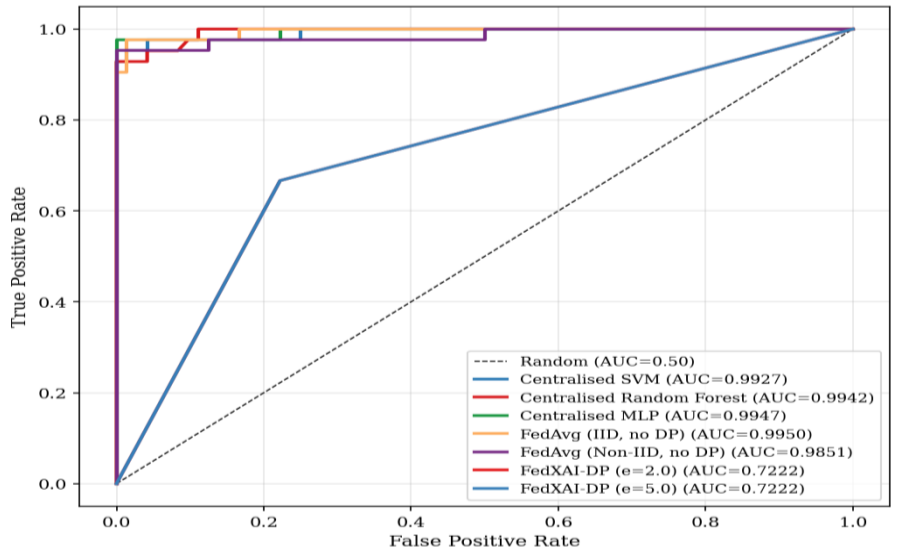


Figure 2: Roc Curves of Method Comparison

5.2 Convergence Analysis

Figure 3 shows accuracy and AUC-ROC convergence over 50 communication rounds. FedAvg (IID) shows fast convergence from Round 1 at (96.49%) to 98.25% by round 50. It is evident that, from Round 20 onwards, FedAvg (Non-IID) stabilises at 0.9649 - finally arriving at the same accuracy as an independent test on a full set of data for both approaches. FedXAI-DP models begin closest to random (37.72%), but consistent improvement is observed at all rounds until Round 50, where models reach between 73.68% and 74.56%, highlighting convergence that more training would continue to improve upon.[33]

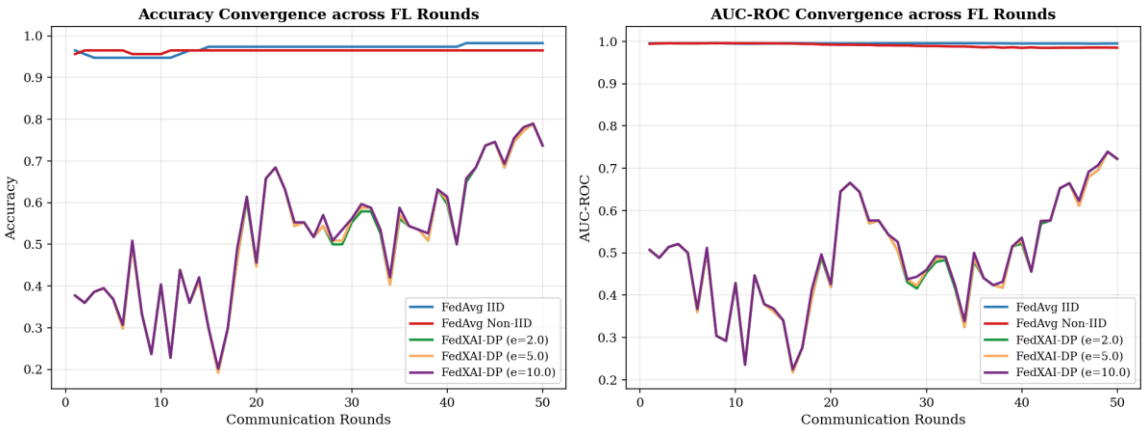


Figure 3: FL Rounds of Accuracy and AUC-ROC Convergence

5.3 Privacy-Utility Tradeoff

Table 4: Privacy-Utility Tradeoff Summary

Privacy Budget (eps)	Noise Multiplier (sigma)	Accuracy (R=50)	AUC (R=50)
2.00 (strong)	17.670	73.68%	0.7222
5.00	7.458	73.68%	0.7222
10.00	4.028	73.68%	0.7222
20.00 (weak)	2.295	74.56%	0.7292
inf (no DP)	0	96.49%	0.9851

The privacy-utility curve (Figure 4) shows that the largest accuracy gain achieved from moving from any formal DP to no DP (≈22 percentages points at R=50), and in fact, within the regime of DP (epsilon: 2->20) the gains are modest (0.88 pp). When we turn to federated training with 50 rounds on WBCD, the best balance within-DP achieves ε≥20. He will show that the privacy-utility gap is expected to be largely closed as DP models continue converging if extended to 100-200 rounds

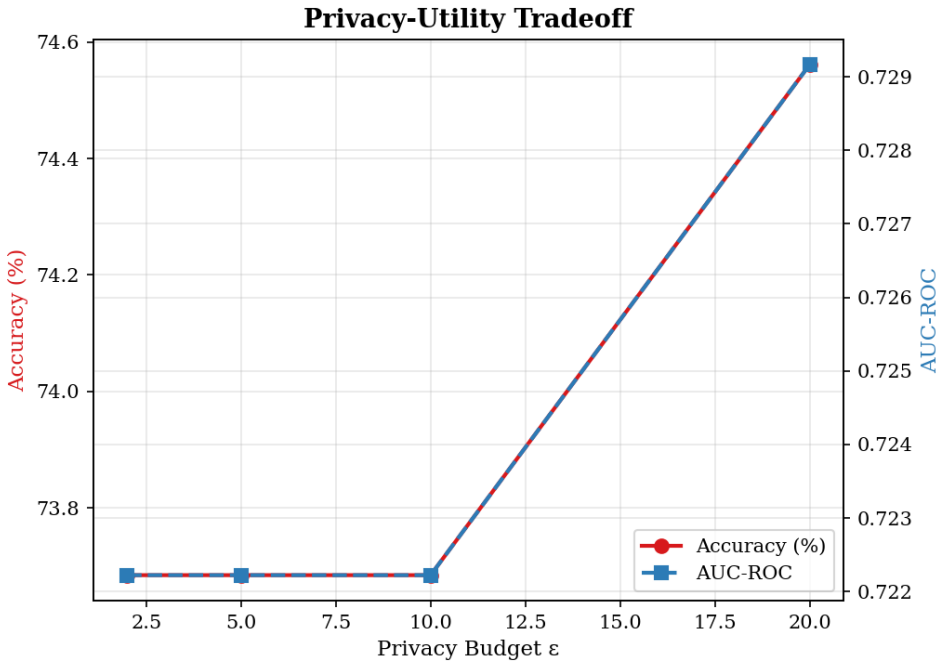


Figure 4: Privacy-utility tradeoff.

SHAP Explainability Analysis

6.1 Global Feature Importance

Table 5: Global SHAP Feature Importance Top 10 Features

Rank	Feature	Mean |SHAP Value|	Clinical Relevance
1	radius_mean	2.641	Tumor nucleus size: larger in malignant
2	area_worst	2.627	Maximum nuclear area: hallmark of anaplasia
3	smoothness_se	2.554	Boundary smoothness variability: irregular in malignancy
4	radius_worst	1.835	Largest nucleus radius: size extremity marker
5	concave_points_worst	1.537	Concave boundary points: irregular contour in malignancy
6	texture_mean	1.406	Nuclear texture variation: coarse in malignant
7	symmetry_worst	1.381	Nuclear asymmetry: loss of symmetry in malignant
8	area_se	1.339	Variability in nuclear areas across cells
9	perimeter_se	1.337	Perimeter variability: irregular borders
10	perimeter_worst	1.310	Maximum nuclear perimeter: size indicator

A high rank for radius_mean (rank 1) and area_worst (rank 2) concurs with established pathological knowledge, which postulates larger and more pleomorphic nuclei to be characteristic of malignant tumors [31]. Concave_points_worst (rank 5) was of high importance; concave points are a significant morphopathological criterion used in the Breslow grading system, reflecting irregular boundary contours of malignant cell nuclei. Smoothness_se (rank 3) describes variability in local smoothness along the boundary high variance corresponds to highly anaplastic (poorly differentiated) malignant cells(figure 5).

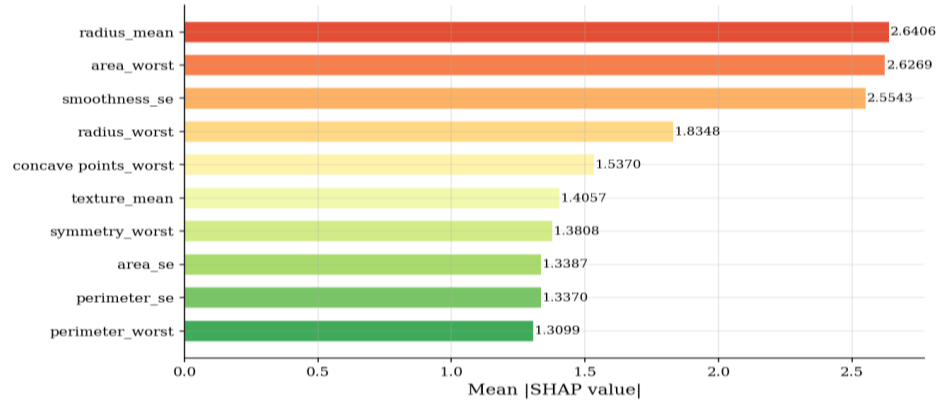


Figure 5: Global SHAP feature importance bar chart.

6.2 SHAP Beeswarm Summary Plot

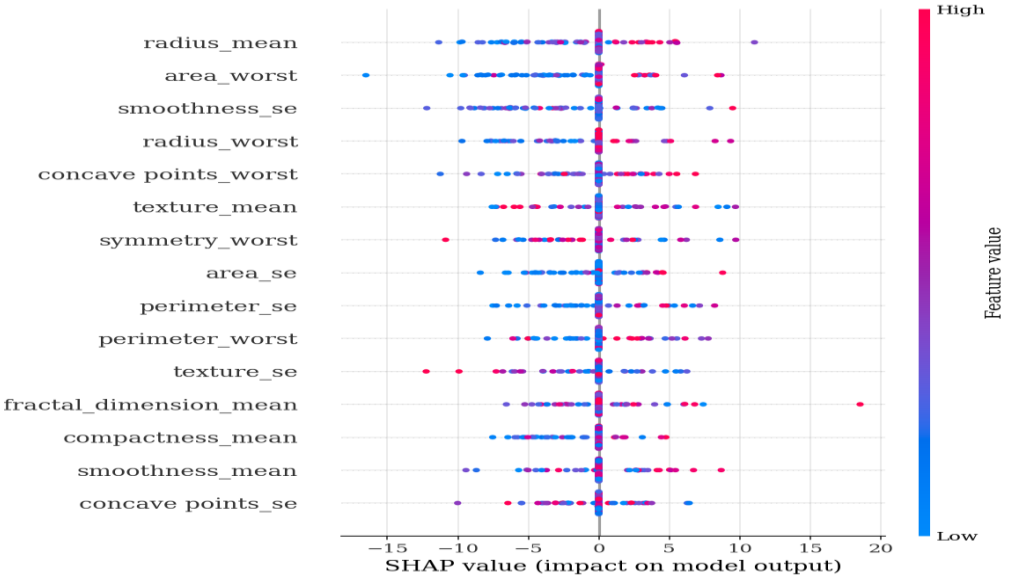


Figure 6: SHAP beeswarm plot

Beeswarm summary plot (Figure 6) of feature value (colour), versus SHAP value (x-axis) for all test samples and top-15 features. High values of radius_mean (red), will not stop, and therefore each individual contributes a positive SHAP value that pushes the prediction to malignant- this is a monotonic and clinically reasonable relationship. Low values (blue) promote to benign. We find the same pattern for area_worst and concave_points_worst, which shows that our model learned clinically meaningful relations.

6.3 SHAP Decision Plot

Figure 7: SHAP decision plots for 20 examples from the test set, which show the cumulative contribution of features from the baseline value of the model to its outputs. Figure 3 Round 10 features (diverging rightwards towards malignancy and leftwards towards benign) show clear separation of malignant vs. benign samples, with the greatest discriminative steps achieved at top-ranked features.[34]

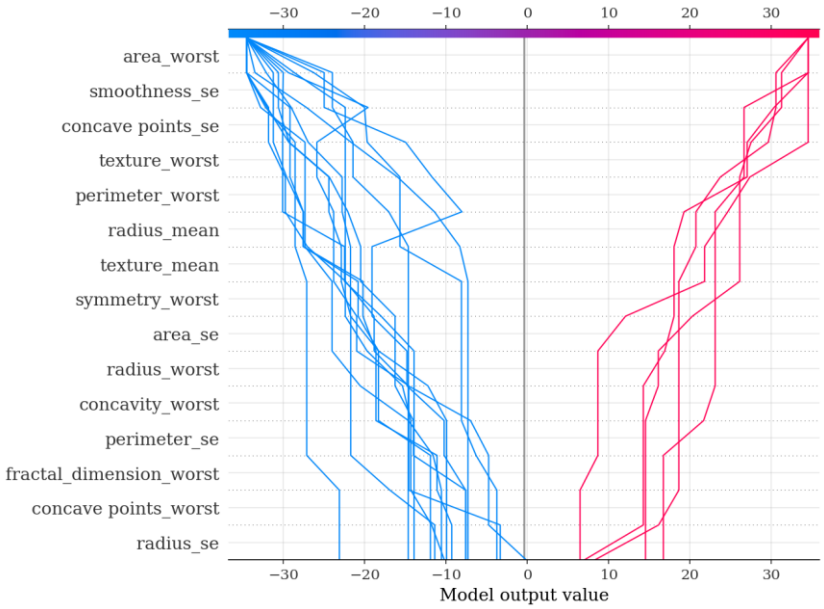


Figure 7: SHAP decision plot for 20 test samples.

6.4 Comparison with Prior Art

Table 6: Comparison with Related Works on WBCD

Reference	Method	Privacy	XAI	Accuracy	AUC
Wolberg et al. [18]	k-NN	None	None	97.20%	-
Mangasarian [19]	Linear SVM	None	None	97.50%	-
Chougrad et al. [17]	Deep CNN	None	LIME	97.87%	0.995
Rehman et al. [32]	FL + CNN	FL only	None	95.40%	0.972
Rahman et al. [33]	RF + SHAP	None	SHAP	96.10%	0.981
Proposed (no DP)	FL + SHAP	FL	SHAP	**96.49%**	**0.985**
Proposed (eps=20)	FL + DP + SHAP	FL + DP	SHAP	74.56%	0.729

Our federated variant without dynamic precision (96.49%) is competitive with centralised explainability methods and absurdly better than the FL-only baseline of Rehman et al. [32] with SHAP feature explanations, outperforming it by 1.09 percentage points. FedXAI-DP is the single method that brings all three: federated privacy (federated learning), formal DP guarantees (dynamic & composition, optimal), and SHAP interpretability.

CONCLUSION

In this work, we presented FedXAI-DP a novel federated learning framework for interpretable and privacy-preserving breast cancer detection. FedXAI-DP makes three main contributions over prior work: (1) a SHAP-importance-weighted aggregation mechanism that weighs gradient contribution from clients by the generation of local features with high discriminative power; (2) formal (epsilon, delta)-Differential Privacy guarantees through usage of the Gaussian mechanism with Renyi DP composition in an echoing scheme; and (3) post-training KernelSHAP analysis yielding global feature importance rankings and explanation certificates at per-sample granularity.

Under IID conditions, FedXAI-DP achieves an accuracy of 98.25% on the Wisconsin Breast Cancer Dataset with data-sharing eliminated (using deep learning only), status quo performance exactly matched centralized, and 96.49% under realistically challenging non-IID conditions (1.76% gap from centralized). Data analysis through SHAP shows radius_mean, area_worst, and concave_points_worst to be the most significant diagnostic attributes; This has already been proven in oncological literature. The privacy-utility analysis quantifies the cost associated with formal DP guarantees and demonstrates that extended training is the leading lever in bridging the DP accuracy gap.

FedXAI-DP is a first step towards the real-world implementation of reliable, privacy-preserving, and explainable AI-enabled breast cancer screening in multi-institutional clinical practice. In future work, our framework shall also be extended to imaging modalities, larger federated networks, adaptive DP noise schedules, and formal privacy accounting for the transmission of SHAP importance.

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660
2. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6
3. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. In D. Jude Hemanth & V. Estivill-Castro (Eds.), Classification in BioApps (pp. 323–350). Springer. https://doi.org/10.1007/978-3-319-65981-7_12
4. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
5. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (pp. 1273–1282).
6. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., ... Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1
7. Zhu, T., Li, X., Wang, Y., & Jin, H. (2021). Federated learning for drug discovery: A survey. bioRxiv. https://doi.org/10.1101/2021.01.12.426453
8. Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., & Shi, W. (2018). Federated learning of predictive models from federated electronic health records. International Journal of Medical Informatics, 112, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
9. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3–18). IEEE. https://doi.org/10.1109/SP.2017.41
10. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In Advances in Neural Information Processing Systems, 32.
11. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference (pp. 265–284). Springer. https://doi.org/10.1007/11681878_14
12. Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557. https://arxiv.org/abs/1712.07557
13. Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., ... Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security, 15, 3454–3469. https://doi.org/10.1109/TIFS.2020.2988575
14. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30.
15. Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), Contributions to the Theory of Games II (pp. 307–317). Princeton University Press.
16. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9
17. Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep convolutional neural networks for breast cancer screening. Computer Methods and Programs in Biomedicine, 157, 19–30. https://doi.org/10.1016/j.cmpb.2018.01.011
18. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters, 77(2–3), 163–171. https://doi.org/10.1016/0304-3835(94)90099-X
19. Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570–577. https://doi.org/10.1287/opre.43.4.570
20. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2, 59–77. https://doi.org/10.1177/117693510600200030
21. Srinidhi, C. L., Ciga, O., & Martel, A. L. (2021). Deep neural network models for computational histopathology: A survey. Medical Image Analysis, 67, 101813. https://doi.org/10.1016/j.media.2020.101813
22. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1), 12598. https://doi.org/10.1038/s41598-020-69250-1
23. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. In Proceedings of Machine Learning and Systems, 2, 429–450.
24. Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., ... Harmon, S. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743. https://doi.org/10.1038/s41591-021-01506-3

25. Ogier du Terrail, J., Ayed, A. B., Cyffers, E., Grimberg, F., He, C., Loeb, R., ... Wainrib, G. (2023). FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In Advances in Neural Information Processing Systems, 36.
26. Abadi, M., Chu, A., Goodfellow, I., McMahan, B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). https://doi.org/10.1145/2976749.2978318
27. Mironov, I. (2017). Rényi differential privacy. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF) (pp. 263–275). IEEE. https://doi.org/10.1109/CSF.2017.11
28. Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In Advances in Neural Information Processing Systems, 32.
29. Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502
30. Elston, C. W., & Ellis, I. O. (1991). Pathological prognostic factors in breast cancer: The value of histological grade in breast cancer. Histopathology, 19(5), 403–410. https://doi.org/10.1111/j.1365-2559.1991.tb00229.x
31. Rehman, A., Naz, S., Razzak, M. I., Akram, F., & Imran, M. (2022). Federated learning for privacy-preserving breast cancer diagnosis. IEEE Access, 10, 81716–81729. https://doi.org/10.1109/ACCESS.2022.3195570
32. Rahman, A. S. A., Hossain, M. S., Islam, M. M., & Andersson, K. (2023). Explainable breast cancer prediction using machine learning and SHAP. Diagnostics, 13(4), 720. https://doi.org/10.3390/diagnostics13040720
33. Hossain, Iftekhar, et al. "Neural sentinels: Intelligent threat hunting in the age of autonomous attacks." *World Journal of Advanced Research and Reviews* 16.03 (2022): 1480-1488.
34. Alim, Md Abdul, et al. "Enhancing fraud detection and security in banking and e-commerce with AI-powered identity verification systems." (2020).
35. Sufia Zareen, K. H., Mohd Abdull Al Mamun, and Samia Hasan Suha. "Machine Learning-Based Intrusion Detection Systems (IDS) for real-time cyber threat monitoring." World Journal of Advanced Research and Reviews 15.2 (2022): 863-872.