ASIDE Oncology

Original Article

Federated Hybrid CNN Vision Transformer Framework for Breast Cancer Classification Under Simulated Non-IID Settings

Bandhan Panda[1,*], Bibek Kumar Patro[1], Siba Sundar Das[1], Santosh Kumar Kar[1]

1-Department of Computer Science & Engineering, NIST University, Berhampur, Odisha, India

ARTICLE INFO

Article history:
Received 23 Jan. 2026
Received in revised form 1 Mar. 2026
Accepted 5 Mar. 2026
Published 9 Mar. 2026

Keywords:
Breast cancer diagnosis
Federated learning
Vision transformer
Explainable artificial intelligence
Privacy-aware federated learning

ABSTRACT

Background: Breast cancer diagnosis increasingly relies on data-driven learning from heterogeneous medical imaging sources. However, centralized deep learning approaches face major limitations, including privacy risks, institutional data silos, and limited generalization across imaging modalities.
Methods: This study proposes a privacy-aware federated learning framework integrating a hybrid CNN–Vision Transformer architecture for breast cancer classification under simulated non-identically distributed (non-IID) conditions. Public datasets representing different imaging modalities, BreakHis (histopathology), INbreast and CBIS-DDSM (mammography), and BUSI (ultrasound) are treated as independent federated clients to emulate multi-institutional collaboration. Each client trains a local model, and only the model parameters are aggregated via Federated Averaging, thereby preserving data locality. The hybrid architecture combines a ResNet-18 convolutional branch for local feature extraction with a Vision Transformer branch for global contextual representation.
Results: Across ten federated communication rounds, the global model demonstrates stable convergence under heterogeneous client distributions. The final global validation performance reaches 72.43% accuracy, AUC 0.7475, and F1-score 0.718. Evaluation on the pooled test cohort achieves approximately 84% overall accuracy with a weighted F1-score of 0.82, while malignant recall approaches 96%, prioritizing clinically critical cancer detection. Qualitative explainability analysis indicates that the model focuses on diagnostically relevant tissue regions.
Conclusion: The results demonstrate the feasibility of hybrid CNN–Vision Transformer training in a federated setting for privacy-aware breast cancer classification across heterogeneous imaging domains.

1. Introduction

Breast cancer has been among the most common causes of death as a result of cancer in diseases worldwide, hence the need to employ effective and prompt diagnostic steps capable of aiding in clinical decision-making. The recent developments of deep learning have shown great potential for automated detection of breast cancer in various imaging modalities such as histopathology, mammography, and ultrasound images. Although there have been these improvements, the bulk of the current practices is dependent on centralized training paradigms, whereby sensitive patient information has to be accumulated at a point of centralization. These practices are associated with significant questions of patient privacy, compliance with regulations, and institutional management of data, especially in the cases of healthcare settings with harsh data protection policies [1, 2].

Federated learning has become one of the most promising alternatives, which allows collaborative model training across distributed data sources without requiring the exchange of raw patient data [2, 3]. The paradigm is particularly suitable for medical imaging, where data are inherently partitioned across hospitals, laboratories, and diagnostic facilities. Nonetheless, federated learning introduces technical challenges, including statistical heterogeneity, communication inefficiency, and instability under non-identically distributed (non-IID) data conditions [3, 4]. Disparities in imaging characteristics across acquisition protocols and modalities are especially pronounced in breast cancer diagnosis, where histopathology, mammography, and ultrasound images exhibit substantial structural and intensity differences. In this study, federated learning is implemented as a simulation framework in which publicly available datasets are treated as independent federated clients to emulate multi-institutional collaboration. No explicit privacy threat model, such as membership inference, gradient leakage, or model inversion attacks, is evaluated, and no formal differential privacy or secure aggregation guarantees are experimentally validated. Accordingly, the proposed framework should be interpreted as privacy-aware and privacy-motivated rather than privacy-guaranteeing.

Parallel to this, Vision Transformers (ViTs) have received growing recognition regarding their capabilities to compute long-range spatial coupling, as well as global contextual associations, and frequently perform better than traditional convolutional neural networks on image recognition tasks [5, 6]. ViTs have demonstrated positive results in medical imaging tasks that utilize holistic feature representations. However, their incorporation in federated learning systems to diagnose breast cancer is comparatively low, especially in convergence behaviour in heterogeneous settings, modular architectural design, and interpretability [1, 7]. It should also be

[*] Corresponding author: Bandhan Panda, Department of Computer Science & Engineering, NIST University, Berhampur, Odisha, India. Email: bandhan.panda@nist.edu

Published by the American Society for Inclusion, Diversity, and Equity in Healthcare (ASIDE). ISSN (Print) 3069-9959, ISSN (Online) 3069-9967. Hosted by ASIDE Journals.

Citation: Panda B, Patro BK, Das SS, Kar SK. Federated Hybrid CNN Vision Transformer Framework for Breast Cancer Classification Under Simulated Non-IID Settings. ASIDE Onc. 2026;1(1):28-37, doi:10.71079/ASIDE.Onc.030926535

noted that the current research paper uses a federated learning framework of simulation, where publicly available datasets are considered to be independent federated clients. This design provides a surrogate of multi-institutional learning, where it is possible to theoretically analyse non-IID behaviour based on non-IID data sources, without privacy and access limitations of actual clinical partnerships. Although this proxy describes important statistical issues, such as domain shift, class imbalance, and feature distribution in mode, it lacks the realistic modelling of real-world institutional issues, such as communication latency, governance policy, or site-specific calibration procedures.

In addition, the proposed framework uses a single binary classification head that cuts across all domains of imaging, directly exposing the model to cross-domain shift in images. Accordingly, assertions about imaging cross-modality generalization are made in the framework of this single decision boundary and its recognition that there is no modality-specific calibration or threshold optimization to date, which is another key avenue to future clinical translation. This work is motivated by recent studies that have highlighted the need to establish robust deep learning architectures and explainable decision-making in medical image analysis [8–10], which motivates the current study to propose a federated hybrid CNN-Vision Transformer framework that would be able to work under simulated multi-source non-IID conditions at high diagnostic sensitivity and interpretation rates.

2. Related work

2.1. Federated Learning in Medical Imaging

The initial concepts of federated learning were proposed to train a model over decentralized data collaboratively, and data privacy was maintained [2, 3]. Its applicability to medical imaging has been well acknowledged, since medical information is naturally spread out across medical institutions and is highly regulated due to its sensitive nature. Several works have investigated the use of federated learning in clinical practice, showing that it is possible in various clinical settings, including radiology, pathology, and disease classification [9, 11]. Nevertheless, heterogeneity optimization federated by nonuniform data sets is a basic issue. Previous literature has revealed that non-IID data among clients may result in the slow convergence and deterioration of performance, especially when client data vary greatly in size and modality [4]. Such constraints require thorough architectural planning and assessment of a medical real-life setting.

2.2. Medical Image Analysis Vision Transformers

A new category of self-attention–based architectures, including Vision Transformers (ViTs), has emerged as a powerful alternative to convolutional neural networks for modelling global contextual relationships in images [5], while earlier deep generative models such as Deep Boltzmann Machines laid important foundations for hierarchical representation learning in deep neural architectures [12]. Later improvements, including attention-based distillation and training data efficiency, have made them more useful when dealing with limited-data regimes typical of medical imaging [6]. Recent surveys and empirical experiments have indicated positive performances of ViTs at different medical image analysis tasks, such as breast cancer classification [8, 10]. In spite of these developments, the majority of approaches based on transformers are trained in centralized environments, and their behaviour when applied to federated learning has not been adequately studied.

2.3. Explainability and Clinical Interpretability

Explainable artificial intelligence (XAI) is a requirement for applying deep learning models to clinical practice, where transparency and trust are mandatory [13]. Grad-CAM has become a popular gradient-based visualization method to offer spatial explanations to deep neural network predictions [14]. Recent papers have highlighted the importance of explainability in medical decision support systems, and it has been shown that explainable models can enhance clinician confidence and error analysis [15]. However, explainability in a federated environment creates further complexity, since explanations need to be consistent when compared with the distribution of heterogeneous client data.

2.4. Breast Cancer Diagnosis and Generalization with other domains

Detection of breast cancer has been studied widely using deep learning methods in both imaging modalities, with models produced showing high accuracy in a controlled setup [8, 16]. Nevertheless, cross-domain generalization is also an ongoing problem as a result of changes in the imaging instructions and in the number of patients involved [17]. The recent reviews have noted the need to have federated and privacy-enabling structures that can be robust generalizations across the institutions without loss of diagnostic reliability [18, 19]. These findings encourage the construction of federated transformer-based methods that clearly focus on the issues of modality-based heterogeneity and clinical interpretability.

3. Dataset description and preprocessing

Experiments to evaluate the proposed federated learning framework under multi-institutional conditions were conducted using four publicly available breast imaging datasets: BreakHis, INbreast, CBIS-DDSM, and BUSI. These datasets represent heterogeneous imaging domains, including histopathology, mammography, and ultrasound, and are widely used benchmarks in breast cancer research [1, 8, 17].

The simulated federated learning environment represents data silos frequently seen in healthcare institutions by treating data as individual federated clients. This formulation causes a natural non-identically distributed (non- IID) data format, allowing the examination of the heterogeneity impacts on federated optimization to be controlled. In order to facilitate reproducibility and transparency, essential qualities of datasets, such as sample sizes, class distributions, labelling schemes, and units of data splitting, are summarized in **(Table 1)**.

When using the mammography data sets (INbreast and CBIS-DDSM), the image files in pre-processed ROI format that were provided publicly by the dataset maintainers were used directly. In this study, no actual raw DICOM-level windowing, bit-depth manipulation, or intensity clipping was done. The selected option will guarantee reproducibility, and yet note that pipeline optimization of mammographic preprocessing is out of the scope of the corresponding work.

All images across datasets were converted to a three-channel RGB format and resized to 224 × 224 pixels to ensure compatibility with the shared CNN – Vision Transformer backbone. Input normalization was performed using ImageNet statistics to facilitate transfer learning from pretrained weights and to maintain a unified preprocessing pipeline across heterogeneous imaging domains. While this normalization strategy simplifies cross-domain training, it may not fully preserve modality-specific intensity semantics, particularly

Table 1: Dataset Reproducibility Details and Label Harmonization

Dataset	Imaging Modality	Unit of Data	Total Samples Used	Benign Samples	Malignant Samples	Patient/Cas Count	Original Labels	Binary Label Mapping	Split Unit
BreakHis	Histopathology	Image patches (×40–×400)	2,482	1,244	1,238	82 patients	8 tumor subtypes	Benign subtypes: Benign; Malignant subtypes: Malignant	Patient-level
INbreast	Mammography	Full mammogram images	410	205	205	115 patients	BI-RADS / pathology	Normal & benign: Benign; Malignant: Malignant	Patient-level
CBIS-DDSM	Mammography	ROI images	1,696	847	849	~753 cases	Mass/Calcific with pathology	Benign: Benign; Malignant: Malignant	Case-level
BUSI	Ultrasound	Image	780	487	293	Not specified	Normal / Benign / Malignant	Benign & Normal: Benign; Malignant: Malignant	Image-level

ROI, region of interest; BI-RADS, Breast Imaging Reporting and Data System; CBIS-DDSM, Curated Breast Imaging Subset of the Digital Database for Screening Mammography.

Table 2: Quantification of Non-IID Characteristics Across Federated Clients

Dataset	Total Samples	Benign (%)	Malignant (%)	Majority/Minority Ratio
BreakHis	2,482	50.12%	49.88%	1.005
INbreast	410	50.00%	50.00%	1.000
CBIS-DDSM	1,696	49.94%	50.06%	1.002
BUSI	780	62.44%	37.56%	1.662

BUSI, Breast Ultrasound Images dataset; CBIS-DDSM, Curated Breast Imaging Subset of the Digital Database for Screening Mammography.

for mammography and ultrasound, and is therefore acknowledged as a methodological limitation.

For the BreakHis histopathology dataset, no explicit stain normalization was applied. Although stain variability can influence generalization in histopathological analysis, this decision was made to preserve preprocessing consistency across datasets and to avoid introducing modality-specific transformations that could confound cross-domain federated learning behaviour. The absence of stain normalization is treated as a limitation of the current study.

For INBreast, the total dataset comprises 410 full mammographic images corresponding to 115 patients. However, for model training and evaluation, lesion-level or ROI-based patch extraction was performed, resulting in an expanded number of evaluation instances. Accordingly, the reported test sample count in (**Table 3**) reflects patch-level evaluation units rather than original full-image counts. All references to "samples" in performance tables hence denote model-level input instances (patches) rather than raw full images.

During federated training, no raw data were exchanged between clients, and only model parameters were shared with the central server, in accordance with federated learning principles [2, 9]. Model evaluation was performed on a pooled test set constructed from client-specific held-out splits to assess global diagnostic performance across heterogeneous imaging sources. The overall data preparation and preprocessing workflow is illustrated in (**Figure 1**).

Beyond class imbalance, inter-client feature distribution heterogeneity was qualitatively assessed through embedding-space visualization of penultimate-layer representations using t-SNE. Distinct clustering patterns across datasets indicate measurable feature-space divergence between clients, supporting the characterization of the setting as statistically non-identically distributed beyond simple label skew.

As shown in (**Table 2**), malignant prevalence varies from 37.56% to 50.06% across federated clients. The BUSI dataset exhibits a pronounced benign skew (majority/minority ratio = 1.662), whereas the remaining datasets are nearly balanced. This quantified label heterogeneity provides empirical justification for modelling the federated setting as non-identically distributed (non-IID).

4. Proposed Federated Vision Transformer Framework

(**Figure 2**) illustrates the proposed federated hybrid CNN – Vision Transformer (CNN – ViT) framework, which is employed consistently in both centralized and federated learning settings. The architecture follows a dual-branch design, where a convolutional neural network (ResNet-18) extracts localized spatial features, and a Vision Transformer (ViT) branch captures global contextual representations. Features from both branches are concatenated and passed to a shared classification head, enabling complementary modelling of fine-grained and holistic tissue characteristics.

In the federated learning simulation, each imaging dataset is treated as an independent client, reflecting realistic data silos across institutions or acquisition sources. Model training is performed locally at each client, while collaboration is achieved through parameter aggregation without exchanging raw imaging data, thereby adhering to federated learning principles.

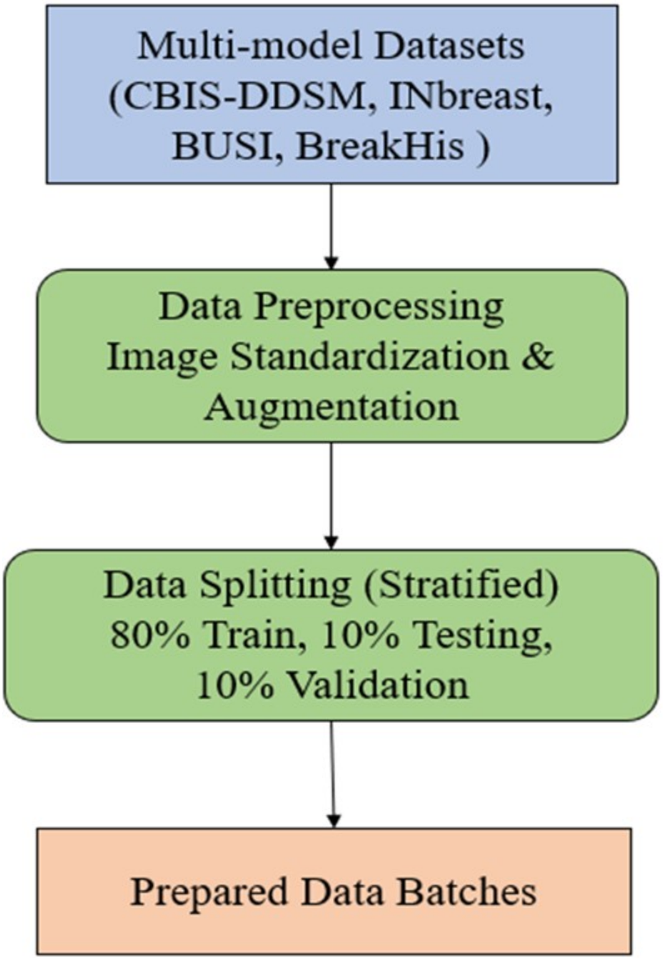
Multi-model Datasets (CBIS-DDSM, INbreast, BUSI, BreakHis)
Data Preprocessing Image Standardization & Augmentation
Data Splitting (Stratified) 80% Train, 10% Testing, 10% Validation
Prepared Data Batches

Figure 1: Workflow of data preparation and preprocessing.

4.1. Federated Learning Formulation

Let K=1,2,…,Kdenote the set of federated clients, where K=4in this study. Each client corresponds to one independent publicly available dataset: BreakHis (histopathology), INbreast (mammography), CBIS-DDSM (mammography), and BUSI (ultrasound). Clients are therefore defined at the dataset level rather than strictly at the imaging modality level. Although imaging modalities partially overlap, both INbreast and CBIS-DDSM are mammography datasets; they are treated as separate clients due to differences in acquisition protocols, case distributions, annotation procedures, and preprocessing characteristics. Therefore, clients are defined at the dataset level rather than strictly at the modality level.

Each client $k \in K$ possesses a local dataset D_k, which differs in size, imaging modality, class distribution, and feature characteristics from other clients. These variations introduce statistical heterogeneity and label skew across clients, resulting in a non-identically distributed (non-IID) federated learning setting [2–4].

Local model training is performed independently at each client using only its private dataset D_k. During federated optimization, no raw imaging data are transmitted between clients or to the central server. Instead, only model parameter updates are communicated for aggregation. This design preserves dataset isolation within the simulated federated environment while enabling collaborative model learning.

4.2. Hybrid Vision Transformer Backbone

Local model: A local model of the same hybrid CNN-ViT architecture is used by all clients. The trained CNN branch (ResNet-18) that is pre-trained using ImageNet weights captures local texture and morphological features, based on patterns of breast tissue. Simultaneously, the Vision Transformer arm works on resized images 224 x 224RGB and splits them into fixed-size patches, then serializing and linearly embedding them before processing through stacked self- attention layers [5, 6].

ViT has a branch, referred to as ViT, that allows the long-range spatial dependency and global contextual relations to be modelled, and this aspect is especially significant in heterogeneous breast cancer imaging. The CNN and ViT branches' outputs are combined with the feature level and launched through a simple fully connected classifier that makes binary decisions (benign vs. malignant).

4.3. Extrinsic Optimization and Aggregation

Training is done through several communication rounds. This involves each client updating its local model parameters suitably using stochastic gradient descent with its own private dataset in every round. The model updates are transmitted to the central server after local training, and then there is data-size-weighted federated averaging (FedAvg) to calculate the global model update [2]. In particular, the contribution of each client is apportioned equally to the size of the local dataset in its possession, i.e., in the case of heterogeneous client counts, there is no bias in aggregate.

The new model developed via globalization is then reissued to all customers in the face of the second wave of local training. The procedure is iterative and allows collaborative learning without interference between datasets in the simulated federated structure.

4.4. Integration of Explainability

To improve clarification and medical interest, clarification procedures are incorporated into the suggested environment. In the case of the CNN branch, class-discriminative areas of activation are visualized through the application of Grad-CAM. In the case of the Vision Transformer branch, attention-based attribution (attention rollout) is the mechanism that checks the spatial regions with the highest contribution to the model predictions.

These attribution maps are produced after post hoc and normed to be visualized, and a qualitative evaluation can be done regarding the compatibility of model attention and clinically relevant tissue structures. Explainability is used uniformly across clients and assessed on held-out test samples to aid in the transparent analysis of model behaviour [13, 14].

Algorithm 1 Federated Vision Transformer–Based Breast Cancer Classification

```
Input:
Federated client datasets $\{D_k\}_{k=1}^{K}$,
where each client corresponds to one independent dataset (BreakHis, INbreast, CBIS-DDSM, BUSI), representing dataset-level data silos rather than modality-exclusive partitions.
Number of communications rounds $R$
Local batch size $B$
Vision Transformer model $f_\theta$
Output:
Trained a global model $f_{\theta^*}$for breast cancer classification
Step 1: Initialize global model parameters $\theta^{(0)}$at the central server.
Step 2: For each communication round $r = 1, 2, \dots, R$, perform:
  Client Selection:
    All available clients $k \in \{1, \dots, K\}$participate in the current round.
  Local Model Update:
    For each client $k$:
      Receive global parameters $\theta^{(r-1)}$.
      Initialize local model $f_{\theta_k^{(r)}} \leftarrow f_{\theta^{(r-1)}}$.
      Train $f_{\theta_k^{(r)}}$on the local dataset $D_k$for one epoch using cross-entropy loss and stochastic gradient descent.
      Transmit updated parameters $\theta_k^{(r)}$to the server.
  Model Aggregation (FedAvg):
    Aggregate local updates to obtain the global model:
    $\theta^{(r)} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k^{(r)}$
  Global Evaluation:
    Evaluate the aggregated model $f_{\theta^{(r)}}$on the global validation set and record performance metrics (Accuracy, AUC, F1-score).
Step 3: After completing $R$rounds, select the final global model $f_{\theta^*} = f_{\theta^{(R)}}$.
Step 4: Perform final clinical evaluation on the aggregated test set and generate explainability maps using gradient-based attribution.
Return: Final trained model $f_{\theta^*}$and associated evaluation metrics.
```

4.5. Experimental Setup / Training Configuration

The hybrid CNN – Vision Transformer architecture comprises approximately 36 million trainable parameters based on the implemented configuration. Federated optimization was conducted over 10 communication rounds, with 5 local training epochs per client per round using data-size-weighted Federated Averaging (FedAvg). The initial learning rate was set to 2×10^{-4} using the Adam optimizer with a batch size of 32. No learning-rate scheduling or hyperparameter search was performed.

The serialized global model size is approximately 140 MB in FP32 precision, representing the upload and download communication cost per client per round. No model compression, pruning, quantization, or communication-efficient strategies were applied; therefore, the reported communication overhead reflects a standard uncompressed federated setup.

Wall-clock training time was not systematically recorded in the present study. Given the fixed hardware configuration (NVIDIA Tesla T4, 16 GB VRAM), the reported results should therefore be interpreted primarily in terms of algorithmic feasibility rather than deployment-level computational benchmarking. Detailed runtime profiling and communication-latency analysis remain important directions for future work.

All experiments were conducted using Python 3.11, PyTorch 2.3, torchvision 0.18, CUDA 12.0, and cuDNN 8.9 under Ubuntu 22.04 LTS. A fixed random seed (42) was used for data splitting, weight initialization, and optimization to ensure deterministic reproducibility. These details are reported to ensure methodological transparency and reproducibility of the simulated federated learning setup.

5. Experimental Results and Analysis

This section presents a structured evaluation of the proposed federated hybrid CNN – Vision Transformer framework under a simulated multi-institutional, non-identically distributed (non-IID) learning setting. Four heterogeneous breast imaging dataset BreakHis, INbreast, CBIS-DDSM, and BUSI, were treated as independent dataset-level federated clients, as described in Section III.

5.1. Training Configuration and Data Processing

All input images were resized to 224 × 224 pixels and normalized using ImageNet statistics to maintain compatibility with pretrained backbones. The ResNet-18 branch was initialized with ImageNet weights, while the Vision Transformer branch employed patch-based embeddings with learnable positional encodings. Data augmentation was limited to random horizontal flips and minor intensity normalization to reduce overfitting while preserving diagnostic characteristics.

Each client trained its local model using the Adam optimizer with an initial learning rate of 2×10^{-4} and batch size 32. Binary cross-entropy loss was applied for classification. No client-specific hyperparameter tuning was performed, ensuring consistency across heterogeneous datasets.

Federated training consisted of 10 global communication rounds, each containing 5 local epochs per participating client. Data-size-weighted Federated Averaging (FedAvg) was used for model aggregation, proportionally weighting client updates based on local dataset size.

Raw imaging data remained local to each client throughout training. Only model parameters were transmitted to the central server. Evaluation was performed on client-specific held-out test sets, followed by

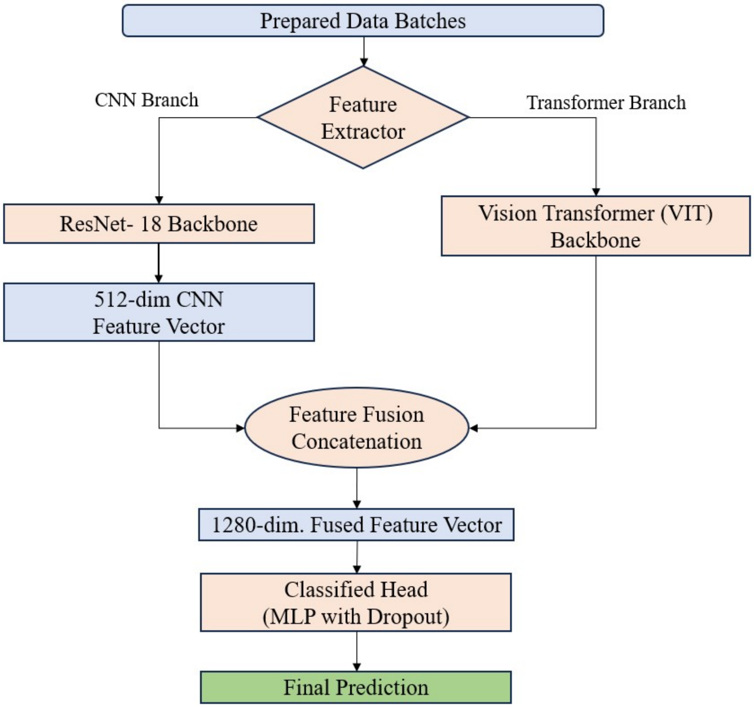
Prepared Data Batches
CNN Branch
Feature Extractor
Transformer Branch
ResNet- 18 Backbone
Vision Transformer (VIT) Backbone
512-dim CNN Feature Vector
Feature Fusion Concatenation
1280-dim. Fused Feature Vector
Classified Head (MLP with Dropout)
Final Prediction

Figure 2: Overview of the proposed federated vision transformer framework.

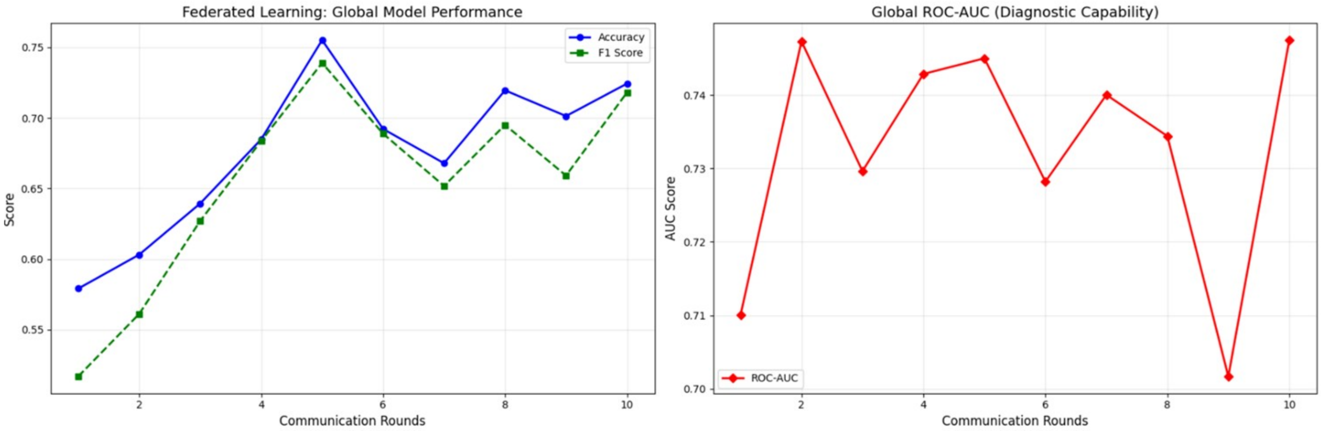

Figure 3: Evolution of global accuracy, F1-score, and AUC across federated communication rounds, illustrating convergence under non-IID multi-institutional data.

pooled analysis across clients to estimate aggregated performance. This pooled evaluation was conducted solely for experimental benchmarking and does not reflect a real-world federated deployment scenario.

5.2. Federated Training Performance and Convergence Analysis

The federated hybrid CNN-Vision Transformer network was trained in ten communicative rounds in a simulated non-identically distributed(non-IID) scenario. **(Figure 3)** demonstrates how the accuracy, F1-score, and ROC-AUC of the world change with the communication round. The global model is showing a steady performance improvement pattern, whereby consistency in finding accuracy rises in each round of performance by starting with 57.9 percent, then 72.4 percent in the first and last round, respectively, accompanied by a rise in ROC-AUC between 0.71 and 0.75.

Intermediate fluctuations are used, which are attributes of optimization of federated learning with heterogeneous client distributions and indicate the variability of local updates instead of instability in the learning process. Notwithstanding such fluctuations, the world model continues to recover again and again and become better, suggesting that there is solid convergence behavior of cross-modality heterogeneity. These findings confirm that the proposed framework could utilize the knowledge of heterogeneous sources of imaging without centralized access to this data, even when the distribution of clients is non-IID.All reported metrics represent single-seed estimates; variance across multiple random initializations was not evaluated.

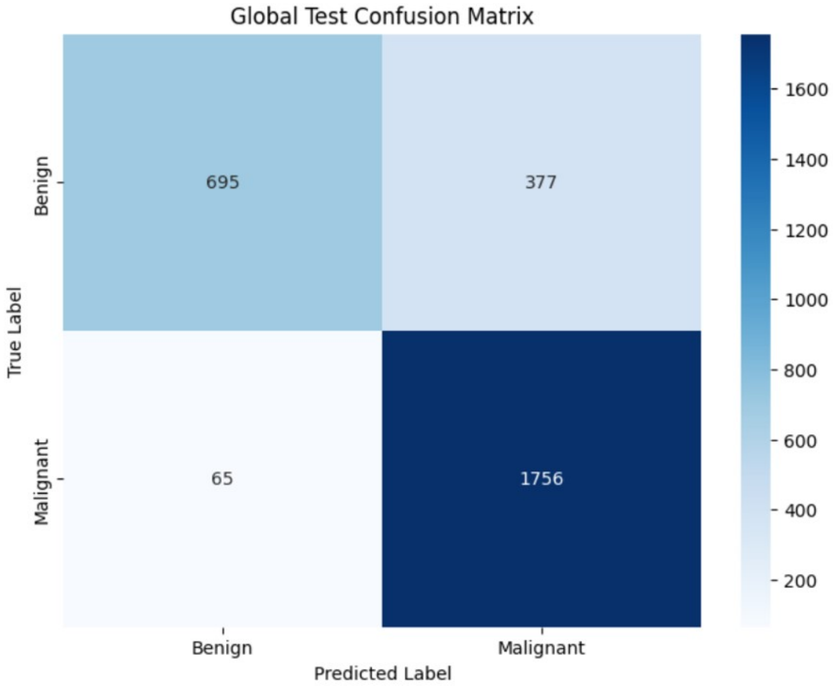

Figure 4: Confusion matrix of the final global federated model evaluated on the aggregated test set.

5.3. Global Clinical Classification Performance

The final global model was evaluated on the aggregated test cohort constructed from client-specific held-out splits. The confusion matrix results are presented in **(Figure 4)**. The model achieved an overall classification accuracy of approximately 82.5%, with a weighted F1-score of 0.81 and a macro F1-score of 0.78. Notably, the framework demonstrated a high malignant recall rate of approximately 96.4%, with only 65 false-negative cases among malignant samples. This high sensitivity is clinically significant, as early breast cancer screening prioritizes minimizing missed malignant detections.

Conversely, the benign recall is relatively less (64.8%), and this denotes a sensitivity-specificity trade-off under a constant decision level. Although this action leads to more false-positive referrals, it gives more emphasis to malignant detection, which is usually desired in early screening settings. No threshold optimization or cost-sensitive calibration was applied in this study; therefore, this trade-off is reported as an empirical outcome rather than a designed diagnostic bias.

5.4. Per-Client Performance Analysis

To evaluate robustness under heterogeneous federated conditions, the final global model was additionally assessed on each client's held-out test set separately. Dataset-level performance metrics, including Accuracy, AUC, F1-score, Sensitivity, and Specificity, are summarized in **(Table 3)**, while a visual comparison of client-wise model performance across heterogeneous datasets is presented in **(Figure 5)**. Reporting per-client performance provides a more transparent assessment of cross-modality generalization under non-IID conditions, beyond pooled global metrics.

As shown in **(Table 3)**, substantial performance variability is observed across clients, reflecting the challenges of cross-domain heterogeneity in federated learning. The model achieved excellent diagnostic performance for the BreakHis and INbreast datasets, with AUC values of 0.987 and 0.997, respectively, indicating strong generalization across histopathological and mammographic imaging domains. In contrast, comparatively lower performance was observed for the CBIS-DDSM and BUSI datasets, where AUC values were approximately 0.585 and 0.546. This discrepancy can be attributed to modality-specific distribution shifts, limited sample sizes, and differences in imaging acquisition characteristics. Importantly, malignant sensitivity remained relatively high across all clients, demonstrating the framework's ability to prioritize cancer detection under heterogeneous non-IID conditions.

These findings highlight that federated learning models trained across highly heterogeneous medical imaging domains may exhibit modality-dependent performance variations, emphasizing the need for domain-aware calibration strategies in future work.

5.5. Explainability and Error Analysis

To assess interpretability, qualitative explainability maps were generated for representative test samples, as shown in **(Figure 6)**. The visualization includes examples of true positive, false positive, and false negative predictions.

This is because, in cases of rightfully categorized malignant cases, the model predominantly focuses on the dense cellular areas and abnormal tissue formations, which are known histopathological signs of cancer. The patterns of attention in false-negative and false-positive cases are diffuse or unclear and draw greater focus to visually subtle areas that can make diagnostic decisions difficult.

These explainability results present a qualitative interpretation that the model is using meaningful image parts to operate its predictions and not spurious artifacts. But there was no quantitative faithfulness/consistency test (e.g., deletion/insertion tests or a clinician validation). In this sense, the explainability analysis will be used to contribute to the interpretability on an illustrative level, and systematic validation will become a valuable direction to work in the future.

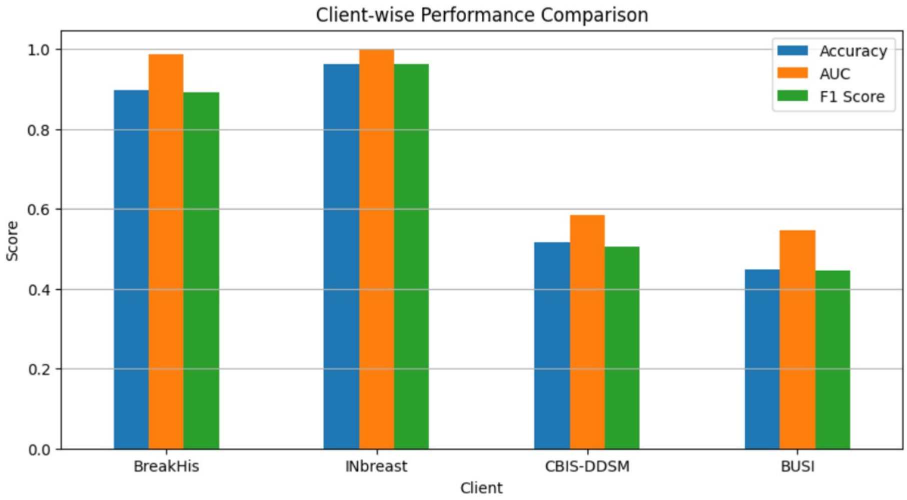

Figure 5: Client-wise performance comparison of the federated hybrid CNN–Vision Transformer model across heterogeneous breast imaging datasets.

Table 3: Per-client classification performance of the final federated model evaluated on individual held-out test sets under simulated non-IID conditions

Client (Dataset)	Test Samples	Accuracy	AUC	F1-Score	Sensitivity	Specificity
BreakHis	1187	0.8972	0.9870	0.8910	0.9975	0.6774
INbreast	410	0.9633	0.9973	0.9628	0.9949	0.8950
CBIS-DDSM	463	0.5162	0.5847	0.5043	0.7344	0.3616
BUSI	98	0.4490	0.5465	0.4446	0.7419	0.3134

AUC, area under the receiver operating characteristic curve; F1-score, harmonic mean of precision and recall.

6. Comparative analysis

The proposed framework is conceptually informed by the hybrid explainable federated Vision Transformer approach introduced by Al-Hejri et al. [1], which demonstrated the viability of integrating transformer-based architectures within a federated learning paradigm for breast cancer classification. However, the scope, architectural composition, and evaluation objectives of the two studies differ substantially.

While Al-Hejri et al. [1] primarily focused on establishing the feasibility of federated transformer learning with integrated explainability under privacy-aware settings, the present study explicitly investigates cross-modality heterogeneity by treating distinct imaging datasets (histopathology, mammography, and ultrasound) as independent federated clients. Notably, even within the mammography domain, INbreast and CBIS-DDSM are modelled as separate clients due to differences in acquisition protocols, case distributions, and preprocessing characteristics. This formulation produces a more pronounced non-identically distributed (non-IID) setting, enabling analysis of convergence dynamics under modality-driven distribution shifts.

In contrast to prior work that predominantly reports aggregate performance metrics, the present study additionally provides round-wise convergence trajectories, offering insight into optimization stability and recovery behaviour across communication rounds under heterogeneous client conditions. Furthermore, error analysis highlights strong malignant sensitivity, a clinically relevant characteristic in screening-oriented diagnostic systems.

Importantly, this comparison remains qualitative in nature. No direct head-to-head baseline experiments, such as centralized CNN, centralized ViT, or alternative federated optimization methods (e.g., FedProx or SCAFFOLD) were conducted using identical dataset splits. Therefore, the proposed framework should be interpreted as demonstrating the feasibility of federated hybrid CNN – ViT training under simulated multi-domain non-IID conditions, rather than establishing performance superiority over existing methods. Comprehensive benchmark comparisons under standardized experimental settings are identified as future work.

7. Discussion

In this paper, a federated micro-hybrid CNN-Vision Transformer model is tested in a multi-source non- identically distributed (non-IID) simulation-based environment by using publicly accessible datasets of breast imaging data. The data sets are treated as independent federated clients, which allows for gaining controlled insight into the cross-domain heterogeneity and federated convergence behaviour. Nonetheless, this experimental design fails to completely reflect the federated learning constraints of the real world, such as site-specific governance policies, secure aggregation protocols, asynchronous client participation, client dropouts, and communication variability. The findings published, therefore, should be viewed as test evidence of methodological support and not clinical implementation readiness. Though heterogeneous imaging domains -histopathology, mammography, and ultrasound are modelled jointly with the use of a common binary classification head, such pooled modelling brings significant threats to validity.

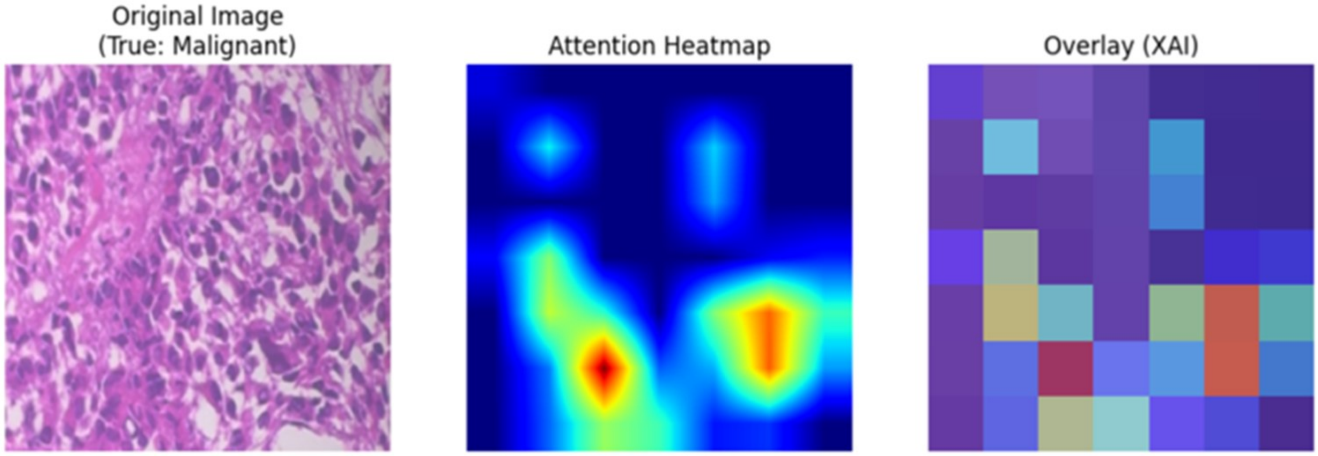

Figure 6: Gradient-based explainability maps for representative test samples: (a) true positive, (b) false positive, and (c) false negative predictions.

Acquisitional variabilities that may cause domain shift, label harmonisation ambiguity, and disparate operating points across modalities may arise because of differences in acquisition protocols, contrast mechanisms, and annotation conventions. In practice in clinical settings, these difficulties may necessitate the use of modality-sensitive calibration policies, domain-relevant decision thresholds, distinct classification heads, or explicit domain adaptation policies so that there can be consistency in the diagnostic behaviour of imaging sources. Demographic metadata (e.g., age, ethnicity, or acquisition site information) were not consistently available across all publicly accessible datasets; therefore, subgroup fairness analysis was not conducted. Potential performance disparities across demographic or acquisition-based subgroups cannot be excluded. Future validation in multi-center settings with structured demographic reporting would be necessary to evaluate the fairness, equity, and generalizability of the proposed framework. The federated learning application to this paper must be regarded as the privacy-conscious architecture paradigm and not the officially confirmed privacy-saving algorithm. Raw data do not leave the client during training, and actually, no explicit privacy threat model, e.g., membership inference or gradient leaks, is analyzed, and no privacy-ensuring mechanisms (e.g., secure aggregation or differential privacy) are enforced.

Moreover, model assessment is done on merged test data to be used in performance benchmarking, but in practice, real-world federated deployments are based on local evaluation or trusted third-party evaluation, and this may impact the reported performance. All experiments were conducted using a fixed random seed (42) to ensure deterministic reproducibility of data splits, weight initialization, and optimization trajectories. Variability across multiple random initializations was not evaluated in the present study. Consequently, reported performance metrics should be interpreted as single-seed estimates rather than variance-adjusted confidence intervals. Multi-seed stability analysis is recommended as an important future validation step to strengthen robustness claims. Analysis of explanations offers the qualitative depictions of the attentions of models to correctly and incorrectly classified cases, indicating that the framework is attentive to diagnostically significant tissue areas. Nonetheless, these results cannot be sustained by quantitative measures of faithfulness, systematic expert validation, and cross-client consistency measures. Interpretability results must therefore be considered to be illustrative, and systematic validation is a field where future research should focus. Furthermore, leave-one-client-out cross-domain generalization experiments, in which the federated model is trained on three clients and evaluated on a held-out client, were not conducted in the present study and remain an important direction for future validation of cross-institutional robustness.

8. Conclusion and future work

This paper introduces a federated hybrid CNN-Vision Transformer system for breast cancer classification subjected to simulated non-IID and multi-source experiments. The study proves that it is possible to collaboratively learn across domains without sharing data centrally by considering heterogeneous imaging datasets as autonomous federated customers. The results obtained with the experiments indicate consistent convergence under the evaluated configuration and high malignant sensitivity in the aggregated test cohort, with performance variability observed across individual clients. In addition to predictive performance, qualitative explainability studies depict the manner in which the model serves diagnostically significant image areas, which promotes interpretability at the level of exploration.

Nevertheless, the results are to be viewed as a proof-of-concept instead of a clinically ready one, due to the utilization of publicly available datasets, centralized analysis, and the lack of a formal privacy or calibration analysis. Future directions will involve research in validation of the framework in realistic multi-center systems, including modality-aware optimization of the system, modality-aware aggregation, running federated optimizers, and doing systematic fairness and interpretability across oneself and other deeper optimization frameworks. These findings provide a structured experimental foundation for subsequent multi-center validation, statistical robustness analysis, and privacy-certified deployment studies required for clinical translation. Cross-domain robustness under leave-one-client-out evaluation was not assessed and remains an important direction for future validation.

Conflicts of Interest

The authors declare no competing interests that could have influenced the objectivity or outcome of this research.

Funding Source

The authors declare that no specific grant or funding was received for this research from any public, commercial, or not-for-profit funding agency.

Acknowledgments

None.

Ethics Approval/IRB Statement:

This study involves secondary analysis of publicly available breast imaging datasets and does not include human subject recruitment, intervention, or access to identifiable patient information. All datasets were used in accordance with their respective licensing agreements and publicly documented usage policies. Therefore, institutional review board (IRB) approval was not required.

Large Language Model

Generative artificial intelligence tools were used in a limited and supportive role during manuscript preparation. Literature searches to identify relevant background articles were performed using OpenEvidence (accessed December 2024 and January 2025). ChatGPT (OpenAI; accessed December 2024 and January 2025) was used to assist with language refinement, grammar, clarity, and formatting of the manuscript. Neither tool was used to generate original scientific content, interpret data, perform analyses, draw conclusions, or make clinical judgments. All content was reviewed, verified, and edited by the authors, who take full responsibility for the accuracy, integrity, and originality of the manuscript. No AI tool is listed as an author, in accordance with the journal's authorship and contributorship policies.

Authors' Contributions

BP contributed to conceptualization, study design, supervision, writing the original draft, and writing review and editing. SKK contributed to conceptualization, research planning, supervision, writing the original draft, and writing review and editing. BKP contributed to software, methodology, data curation, formal analysis, validation, and writing the original draft. SSD contributed to software, methodology, data curation, experimental investigation, validation, and writing the original draft. All authors have read and approved the final version of the manuscript.

Data Availability

BreakHis is publicly accessible for research use upon request from the dataset maintainers. INbreast requires institutional registration and adherence to usage conditions specified by the hosting repository. CBIS-DDSM is available through The Cancer Imaging Archive (TCIA), subject to TCIA data usage policies. BUSI is publicly available for academic research purposes. Researchers must comply with the respective licensing and citation requirements of each dataset provider.

References

1. Al-Hejri AM, Sable AH, Al-Tam RM, Al-Antari MA, Alshamrani SS, Alshmrany KM, et al. A hybrid explainable federated-based vision transformer framework for breast cancer prediction via risk factors. Sci Rep. 2025;15(1):18453. [PMID: 40419634, PMCID: PMC12106662, https://doi.org/10.1038/s41598-025-96527-0].
2. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics. 2017:1273-82.
3. Konečný J, McMahan HB, Ramage D, Richtárik P. Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:161002527. 2016. [https://doi.org/10.48550/arXiv.1610.02527].
4. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems. 2020;2:429-50.
5. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020. [https://doi.org/10.48550/arXiv.2010.11929].
6. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers distillation through attention. International conference on machine learning. 2021:10347-57.
7. Li Q, He B, Song D. Model-contrastive federated learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021:10713-22. [https://doi.org/10.1109/CVPR46437.2021.01057].
8. Alotaibi M, Aljouie A, Alluhaidan N, Qureshi W, Almatar H, Alduhayan R, et al. Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. Heliyon. 2023;9(11):e22406. [PMID: 38074874, PMCID: PMC10700613, https://doi.org/10.1016/j.heliyon.2023.e22406].
9. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119. [PMID: 33015372, PMCID: PMC7490367, https://doi.org/10.1038/s41746-020-00323-1].
10. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017;19:221-48. [PMID: 28301734, PMCID: PMC5479722, https://doi.org/10.1146/annurev-bioeng-071516-044442].
11. Abdulrahman S, Tout H, Ould-Slimane H, Mourad A, Talhi C, Guizani M. A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. IEEE Internet of Things Journal. 2021;8(7):5476-97. [https://doi.org/10.1109/jiot.2020.3030072].
12. Salakhutdinov R, Hinton G. Deep boltzmann machines. Artificial intelligence and statistics. 2009:448-55.
13. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:171209923. 2017. [https://doi.org/10.48550/arXiv.1712.09923].
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision. 2017:618-26.
15. Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. Comput Methods Programs Biomed. 2020;196:105608. [PMID: 32599338, PMCID: PMC7831868, https://doi.org/10.1016/j.cmpb.2020.105608].
16. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:171204621. 2017. [https://doi.org/10.48550/arXiv.1712.04621].
17. Garrucho L, Kushibar K, Jouide S, Diaz O, Igual L, Lekadir K. Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study. Artif Intell Med. 2022;132:102386. [PMID: 36207090, https://doi.org/10.1016/j.artmed.2022.102386].
18. Guan H, Yap PT, Bozoki A, Liu M. Federated learning for medical image analysis: A survey. Pattern Recognit. 2024;151. [PMID: 38559674, PMCID: PMC10976951, https://doi.org/10.1016/j.patcog.2024.110424].
19. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020;2(6):305-11. [https://doi.org/10.1038/s42256-020-0186-1].