Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 9 (2027) 3779

Journal of the Nigerian Society of Physical Sciences

Advancing remote ultrasound diagnosis through a federated deep transfer learning framework

Oluwatobi Noah Akande[a], Funmilola Alaba Ajala[b], Gift Adene[c,*], Julius Makinde[d], Roseline Oluwaseun Ogundokun[e], Rotimi-Williams Bello[e]

[a]*Computer Science Department, Nile University of Nigeria, Federal Capital Territory, Abuja, Nigeria*
[b]*Department of Computer Science, Faculty of Computing and Informatics, Ladoke Akintola University of Technology, Nigeria*
[c]*Department of Computer Science, Akanu Ibiam Federal Polytechnic, Unwana, Nigeria*
[d]*Computer Science Department, Baze University, Abuja*
[e]*Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, Pretoria, South Africa*

Abstract

Telerobotic ultrasound systems enable clinicians to remotely control robotic arms equipped with ultrasound probes, improving diagnostic access. However, their effectiveness is constrained by limited annotated datasets, privacy concerns associated with centralized storage, and a lack of transparent and generalizable artificial intelligence (AI) models. This study presents a Federated Deep Transfer Learning (FDTL) framework supporting collaborative ultrasound image classification while keeping images on local client nodes. Four pretrained convolutional neural networks (CNNs) were fine-tuned on the breast ultrasound images (BUSI) dataset to extract diagnostic features. Federated averaging (FedAvg) enabled model aggregation without sharing raw client data. Centralized and federated approaches were compared, and convergence was analyzed. A gradient-weighted class activation mapping (Grad-CAM) module supported interpretability. The framework achieved a peak federated classification accuracy of 91.1% using DenseNet-121 and a macro-averaged accuracy of 90.5% across the evaluated architectures, outperforming individual client models while maintaining stable convergence over 80 communication rounds. Convergence stabilized within 50–60 rounds. Grad-CAM highlighted attention regions corresponding to diagnostically relevant lesion boundaries and textures. The framework shows a promising balance of diagnostic performance, data localization, interpretability, and scalability under simulated federated conditions. These findings provide preliminary evidence of robustness across heterogeneous simulated clients and suggest potential for decentralized ultrasound diagnosis.

DOI: 10.46481/jnsps.2027.3779

Keywords: Federated learning, Deep transfer learning, Ultrasound classification, Grad-CAM, Telerobotics

Article History:
Received: 02 August 2026
Received in revised form: 26 September 2026
Accepted for publication: 01 October 2026
Available online: 17 November 2026

© 2027 The Author(s). Published by the Nigerian Society of Physical Sciences under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Communicated by: S. Folorunso

1. Introduction

Medical imaging offers clinicians valuable insights into biological and anatomical tissues that play a vital role in accurate diagnosis and effective treatment of diseases. For instance, imaging techniques such as ultrasound, X-rays, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) have greatly redefined clinical decision-making procedures by proffering innovative ways to visualize internal anatomy and pathological changes in a non-disruptive manner [1]. Out of these imaging technologies, ultrasound imaging is the most prevalent due to its portability, affordability, safety, and real-time diagnostic capability [2].

*Corresponding Author Tel. No.: +234-706-674-1467.
Email address: giftadene2016@gmail.com (Gift Adene)

Ultrasound scans also have wide applications in cardiology, obstetrics, and oncology, thereby making them essential in early disease detection and patient monitoring [3]. Regardless of its numerous advantages, its use is highly operator-dependent as skilled radiologists and sonographers are needed to acquire and interpret images generated from them [4].

Availability of such expertise across regions is not assured especially in remote, rural, and resource-limited healthcare settings where access to diagnostic services has always been a challenge [5]. Some have recently reported on telerobotic ultrasound as a pioneering solution to be able to bridge this gap [5]. These systems enable experienced clinicians to remotely control robotic arms outfitted with ultrasound probes [6]. With this advancement, clinicians can conduct real-time imaging of patients that are hundreds or thousands of kilometres away [5, 6]. Since the proximity between experts and underserved communities can be unwarranted, telerobotic ultrasound has become an optimal approach to cross these divides with robotic teleoperation plus high-speed communication as well as imaging technologies. However, image quality and interpretation of acquired ultrasound images are still a primary bottleneck that must be addressed despite these advancements. As a result, in most cases of near-future remote echocardiography without on-site radiologists and real-time expert interpretation, image quality may vary with patient movements, transmission noise and probe positioning [3].

Intelligent and trustworthy ultrasound diagnostic assistance systems which automatically interprets ultrasound images, ensures the reliability of diagnosis by supporting remote clinical decision making are needed. Recent advances in artificial intelligence (AI) and deep learning techniques opened new horizons for medical imaging. Deep learning, especially through Convolutional Neural Networks (CNNs), has been able to produce state-of-the-art results in image classification, segmentation, and anomaly detection tasks across radiology, pathology, and dermatology. Related work has also used deep learning for MRI image enhancement and for skin cancer imaging [7–9]. In ultrasound imaging, CNNs have been effectively used for tasks such as breast lesion classification, thyroid nodule detection, and fetal organ segmentation, often outperforming traditional machine learning methods [16].

These networks are capable of automatically learning hidden patterns in raw images, thereby reducing the need for handcrafted feature extraction and domain-specific engineering. However, training deep CNNs from scratch requires vast amounts of annotated data, which are rarely available in specialized medical domains. Moreover, gathering and labeling medical data is resource-intensive, requires expert radiologists and adherence to strict ethical regulations [10]. Therefore, many medical imaging datasets are small and heterogeneous, thereby, difficult for conventional deep learning models to achieve consistent generalization [11].

To overcome these limitations, researchers have increasingly turned to transfer learning. This is a technique that leverages knowledge from large, pre-trained models to improve learning efficiency on smaller, domain-specific datasets. For instance, CNN architectures such as ResNet, DenseNet, EfficientNet, and InceptionV3 have been originally trained on large-scale natural image datasets like ImageNet, therefore, they can be easily fine-tuned for medical imaging tasks and they have been achieving remarkable success [12]. This is due to their ability to accelerate convergence, reduce the need for large annotated datasets, and model generalization enhancement by reusing previously learned feature hierarchies. In ultrasound imaging, this approach has shown promising results in breast tumor classification and lesion segmentation, where fine-tuned models have been proved to outperform shallow architectures [13, 14]. However, most transfer learning implementations in medical imaging are centrally trained [15]. This requires the aggregation of data from multiple hospitals or imaging centers into a single repository. This approach has raised serious privacy, legal, and ethical concerns.

Despite these advances, important gaps remain in the application of federated learning to breast-ultrasound diagnosis. Existing approaches have demonstrated the feasibility of collaborative learning across distributed medical datasets, but there remains limited integration of pretrained deep transfer learning architectures, federated optimization, and clinically oriented interpretability within a unified ultrasound diagnostic framework. Furthermore, evidence from simulated federated environments does not necessarily establish robustness or generalizability across independent clinical institutions. To address these gaps, this study proposes a Federated Deep Transfer Learning (FDTL) framework that combines pretrained CNN architectures, including ResNet-50, DenseNet-121, EfficientNet-B0, and InceptionV3, with FedAvg-based collaborative training and Grad-CAM interpretability. The study specifically aims to: (i) adapt pretrained CNNs for breast-ultrasound image classification; (ii) evaluate their performance under centralized and simulated federated learning conditions; (iii) examine convergence, communication efficiency, and performance consistency across heterogeneous simulated clients; and (iv) provide visual model explanations using Grad-CAM to support interpretation of diagnostic decisions. Through these objectives, the study seeks to establish the feasibility of a privacy-preserving, interpretable, and scalable approach for decentralized ultrasound diagnosis.

2. Related work

A number of studies have explored how various deep learning paradigms, such as transfer learning, federated learning, and explainable AI techniques can be employed to transform medical image analysis thereby enabling automated interpretation of complex imaging modalities such as ultrasound. An overview of this achievement is provided in this section.

2.1. Deep learning in medical ultrasound imaging

Among existing deep learning techniques, Convolutional Neural Networks (CNNs) has been widely used for feature extraction and classification in medical image analysis. Early studies include Karthik *et al.* [16] where stacked ensemble CNN architecture was used to detect breast tumors. The proposed architecture employs gaussian dropout layers to improve

the computation and retain essential features. A substantial accuracy of 92.15% was achieved compared with methods without Gaussian enhancement. Similarly, Wenjiang *et al.* [17] introduced a technique that combined radiomics and deep learning features to classify breast lesions into benign and malignant. The five models developed in the study performed well with high AUC values. Similarly, Wang *et al.*, (2024) [17] proposed a multimodal deep learning model that uses a combination of ultrafast dynamic contrast-enhanced (UF-DCE) MRI images, lesion characteristics and clinical information for breast lesion classification. They were able to prove that integrating imaging with non-imaging data improves breast lesion classification.

A combination of active contour models and U-Net deep learning technique was employed for breast lesion segmentation by Abdala and Boubakeur [18]. An accuracy of 97.34 % was achieved which showed that the integration of the technique yielded a better performance than when U-Net was used alone. Wang *et al.* [19] also proposed a hybrid feature extraction technique for breast cancer classification. The technique used an ensemble residual vgg16 deep learning model that is made up of the deep residual network and VGG-16 architecture. The proposed technique achieved an accuracy of 93.8% when it was evaluated with the INBreast dataset. Furthermore, Dinghao *et al.* [10] proposed a deep learning framework for breast cancer diagnosis in videos. The technique used three-branch video classification network that was supported by a multiple step knowledge guidance process. The technique yielded a superior performance with an AUC of 94.50% and an accuracy of 83.88%.

In spite of the remarkable achievements of deep learning techniques in medical image analysis, their impact in ultrasound diagnostics is being undermined because most existing models use centralized data aggregation and this raises certain security, privacy, and ethical concerns. Also, deep learning models need to be trained on large, homogeneous and annotated datasets for effective generalization [20]. The black-box nature of existing deep learning techniques has made the need for an explainability module necessary. These limitations have necessitated the need for a privacy-preserving, generalizable, and explainable deep learning framework that could work on distributed data sources without revealing patient confidentiality. These are the challenges the proposed FDTL framework intend to address.

2.2. *Transfer learning for ultrasound diagnostics*

Transfer learning has been extensively employed to address data scarcity in medical imaging. Given the limited availability of labeled ultrasound data, researchers have leveraged pretrained CNNs such as ResNet**,** DenseNet**,** InceptionV3**,** and EfficientNet, initially trained on large-scale datasets like ImageNet, to extract high-level image representations transferable to the medical domain [13, 21]. Nasim *et al.* [14] transformed InceptionV3 modules to a residual inception module for breast lesion classification. The enhanced model achieved an improved accuracy and RMSE when compared with different 24 CNN architectures. Heena *et al.* [13] also proposed an EfficientNet-Based attention residual U-Net for breast tumor segmentation in ultrasound images. The proposed technique uses attention gates to skip connections thereby focusing on significant features in the images. When BUSIS and Dataset B were used to evaluate their model, a superior performance and computational efficiency was achieved when compared to existing state-of-the-art models.

Suhaila *et al.* [4] also presented a hybrid deep learning framework for enhanced breast cancer segmentation on ultrasound images. The framework employed an enhanced ResNet-50 model that uses three (3) convolutions to retain fine anatomical details that were needed to extract lesion boundaries in the ultrasound images. When the technique was evaluated on the BUSI dataset, a precision of 0.915 and an F1 score of 0.911 were achieved which showed that the proposed technique achieved an improved performance over existing models. Eswar *et al.* [12] proposed an automated breast cancer classification technique that integrated both attention and EfficientNet deep learning frameworks. InceptionV3 network was used to extract the deep features while attention-based Feature Pyramid Fusion model (Att-FpF) was used to fuse the extracted features. The proposed method recorded an accuracy of 99.59% and 99.82% with INBreast and CBIS-DDSM datasets respectively. Iqra *et al.* [21] introduced a technique that employed Swin transformers, improved EfficientNet model and AttentionNet model to extract hierarchical features needed for breast image segmentation and classification. The resulting SwinEff-AttentionNet technique achieved an accuracy of 98.50% on the BUSI datasets which outperformed existing state-of-the-art models.

Transfer learning techniques have gained massive acceptance in medical image analysis because they use pretrained models. Yet, existing transfer learning techniques have been applied on centralized datasets that need medical images to be aggregated in a single repository. This approach undermines patients' privacy and also violates certain data protection regulations. Most importantly, pretrained models need to be fine-tuned before they could perfectly adapt to the unique acoustic and textural characteristics of ultrasound images (see also [9] for a transfer-learning application in skin cancer imaging). The present study addresses this gap with the proposed Federated Deep Transfer Learning (FDTL) framework that targets a collaborative, explainable, and privacy-aware model training for AI-driven ultrasound diagnosis.

2.3. *Federated Learning for privacy-preserving medical AI*

A Federated Learning (FL) approach for breast cancer detection was proposed by Savita *et al.* [22]. The study leveraged the ability of FL to store multiple data from different locations. They therefore integrated an intelligent heuristic-based deep learning framework in a FL mode to maximize the breast cancer diagnosis results in terms of increased precision and accuracy. Their proposed technique achieved a 95% accuracy. Jie *et al.* [23] also leveraged FL for multi-center disease diagnosis with non-IID data. They proposed an innovative FL technique that uses model Projection. The proposed technique ensures the loss of local model on global data does not increase after local

training and after aggregation. The proposed technique outperformed state-of-the-art FL methods in terms of accuracy, convergence rate and communication cost. To further demonstrate the effectiveness of FL, Saira *et al.* [24] proposed a Flower Federated Learning technique for melanoma detection. is presented with different networks and a combination of hyperparameters.

Three deep learning models (ResNet 50, AlexNet, and a custom CNN) were integrated into the proposed technique. The proposed technique achieved a 90.02% classification accuracy with HAM10000 dataset and 86.03% classification accuracy with ISIC dataset. Furthermore, with a view to enhance the accuracy and interpretability of cancer detection models, Karthiga *et al.* [25] blended LightGBM with SHAP (SHapley Additive exPlanations) within a federated learning framework. The federated learning approach employed allowed multiple institutions to collaborate in training the model without sharing raw patient data, thereby ensuring data privacy while benefiting from diverse datasets. The proposed framework achieved a remarkable accuracy of 98.3%.

Despite the achievements recorded in existing FL applications, their communication overhead, gradient divergence and convergence instability especially in heterogeneous data remains an issue of concern. [26–28]. Also, few existing FDTL models incorporate interpretability mechanisms to make model reasoning transparent to clinicians. These gaps motivate the present study, which introduces an explainable, federated deep transfer learning framework tailored to AI-driven telerobotic ultrasound diagnosis.

2.4. Explainable AI (XAI) and Grad-CAM in medical imaging

Due to the "Black box nature" of deep learning techniques, explainability is necessary to gain user's trust and regulatory approvals in clinical decision-making. Thus, many explainable AI (XAI) techniques have been subsequently proposed and researched to understand how the deep learning approaches make their decisions. For example, Oladosu *et al.* [29] proposed a Mutual Information-based Radiomic Feature Selection (MIRFS) framework for Digital Breast Tomosynthesis (DBT) breast cancer evaluation followed by interpretation of the selected features with SHAP explanations. Machine learning algorithms were employed to evaluate selected features, while SHAP explanations provided meaningful insights into feature importance, facilitating model interpretation. Similarly, Marwa *et al.* [30], who proposed an explainable AI for breast cancer classification using vision Transformer (ViT). The transformers have been applied for binary lesions classification whereas Grad-CAM, Vanilla gradient, Integrated gradients, Saliency Maps, Local Interpretable Model Agnostic Explanation (LIME), and Attention Maps as utilized to view the most discriminative features of the model prediction results.

The explainability module provided the additional support needed for medical professionals. Shiny *et al.* [31] also proposed a breast cancer detection model that uses EfficientNet and ViT to capture fine-grained and contextual features. Nonetheless, Grad-CAM was applied for interpretability by visualizing influential areas in model decisions. Samridhi *et al.* [32] proposed a framework using modified EfficientNetB1 architecture for region segmentation of tumor. Specifically, a Score Weighted Class Activation Mapping (ScoreCAM) algorithm was combined into the model for interpretability. However, most existing works apply the explainability techniques in a centralized setup, and thus do not evaluate how consistent interpretability is across federated environments.

Despite the achievements recorded in existing applications of federated, deep and transfer learning techniques for ultrasound image analysis, the need for a centralized aggregation approach that require sensitive patient images to be trained on a centralized server remains an issue of concern [25, 33]. Therefore, the need for a distributed or privacy-preserving approach that processes heterogeneous clinical datasets without compromising data security is paramount [24, 34, 35].

3. Methodology

This section discusses the implementation details behind the proposed FDTL technique for telerobotic ultrasound image analysis.

3.1. Research design

The design involves four key stages as illustrated in Fig. 1:

3.2. Dataset description, preprocessing and augmentation

The Breast Ultrasound Imaging (BUSI) datasets originally curated by Walid *et al.* [36] was used to validate the proposed technique. The dataset has 780 labeled ultrasound images that were categorized as normal, benign, or malignant. The images were accompanied by radiologist-verified annotations and pixel-level masks. A sample of the image is provided in Fig. 2:

The dataset was split into 70% for training, 15% for validation, and 15% for testing subsets. However, to simulate real-world differences across remote clinical sites, the images were distributed non-independently and non-identically (non-IID).

Table 1 presents the class distribution of the BUSI dataset used in this study, showing the allocation of images across the Normal, Benign, and Malignant categories. The dataset was divided into training, validation, and testing subsets to support model development and evaluation while maintaining representation of the three diagnostic classes.

To ensure a leakage-free evaluation, the 70/15/15 dataset split was executed strictly at the patient level prior to any preprocessing or data augmentation. All ultrasound images and associated ground-truth masks belonging to a single patient were allocated exclusively to either the training (546 images), validation (117 images), or testing subset (117 images). Augmentation operations were subsequently applied exclusively to the training subset.

Furthermore, Table 2 presents the distribution of the training images across the three simulated federated client nodes. The clients were assigned different class proportions to create a non-IID data distribution that represents potential variations in disease prevalence and data availability across heterogeneous healthcare environments.

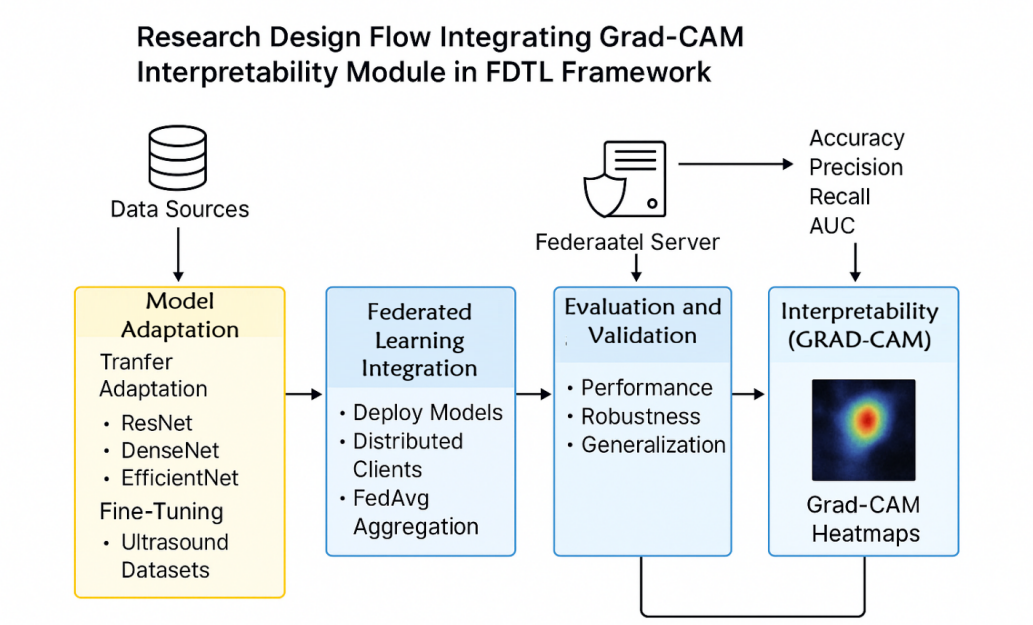


Figure 1: Research design flow.

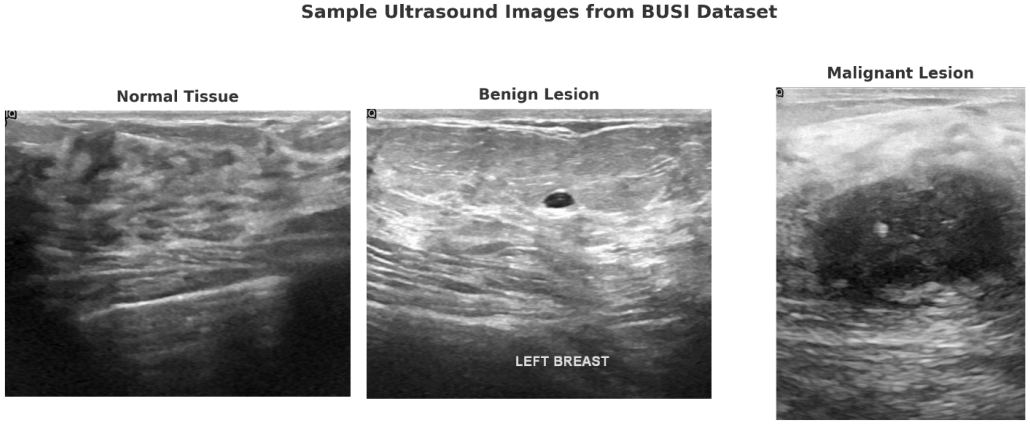


Figure 2: BUSI dataset samples [36].

Table 1: Class distribution of the BUSI dataset.

Class	Total Images	Training (70%)	Validation (15%)	Testing (15%)
Normal	133	93	20	20
Benign	437	306	66	65
Malignant	210	147	31	32
Total	780	546	117	117

To ensure uniformity and noise reduction, the images were preprocessed by resizing, normalization, enhancement and augmentation. The images are not of the same size; therefore, they were resized to a uniform size of 224 × 224 pixels that is suitable as input pretrained CNNs to be used in the proposed FDTL model. To ensure the stabilization of the learning process and prevent bias that could arise as a result of varying brightness levels in the images, the intensities of the images were scaled between 0 and 1. To also improve visibility of hidden lesion boundaries and anatomical details, histogram equalization was used to enhance the contrast of the images. To also ensure that the model will be able to learn orientation- and position-

Table 2: Federated client distribution.

Client	Simulated Clinical Environment	Normal	Benign	Malignant	Total
Client A	Remote Hospital	25	120	35	180
Client B	Telerobotic Node	40	95	45	180
Client C	Regional Hospital	28	91	67	186
Total		93	306	147	546

invariant features while reducing overfitting through Gaussian noise addition, the images were augmented through random rotation (±15°), flipped horizontally, and also zoomed between 0.8–1.2×. These preprocessing steps were uniformly applied to all federated clients so as to ensure comparable feature distributions.

Histogram equalization was applied to mitigate variation in acoustic gain across scans while preserving diagnostic echogenicity patterns. Vertical flipping was explicitly excluded because vertical acoustic orientation carries essential anatomical context in medical ultrasound scans.

3.3. Deep transfer learning architecture

The study used four (4) state-of-the-art pretrained CNN models that have shown proven performance in medical analysis according to existing literature. They are: ResNet50, DenseNet121, EfficientNetB0 and InceptionV3. These are models that have been pretrained on large image datasets such as ImageNet. ResNet-50 was chosen because of its deep residual connections that could reduce the effect of vanishing gradients thereby guaranteeing stable training when it encounters complex ultrasound textures. DenseNet-121 is known for its feature reusability due to its dense connectivity. This allows it to learn well even with fewer parameters. EfficientNet-B0 is known for its lightweight but powerful architecture. This makes it computationally efficient in resource constrained environment that is the best for remote diagnostic environment. InceptionV3 was chosen because it has multi-scale convolutional kernels that allows it to capture both fine-grained and global structural information. This helps EfficientNetB0 to be effective in identifying diverse lesion morphologies.

3.3.1. Model customization

The domain characteristics of ultrasound images differ when compared to natural images. Ultrasound images are usually in grayscale and are characterized by low contrast, speckle noise, and shadowing [4, 20]. The diagnostic information of ultrasound images is hidden in their texture, shape irregularities and intensity gradients [37]. Therefore, a transfer learning approach was used to modify the architecture of the pretrained CNN models as follows:

(a) Layer Freezing (Retaining Generic Features)

To retain the generic features (edges and contours) of the pretrained models that are useful for ultrasound texture detection, their first 50–70% layers were frozen so that they are not updated during training. Freezing them also prevents the model from losing this valuable general knowledge and reduces computational cost and overfitting. Mathematically, if θ_f represents frozen parameters and θ_t represents trainable parameters, only trainable parameters θ_t are updated during backpropagation, while θ_f remain fixed, as shown in Eq. (1):

$$\theta_t \leftarrow \theta_t - \eta \nabla_{\theta_t} L, \tag{1}$$

where θ represents the trainable model parameters, η denotes the learning rate controlling the magnitude of parameter updates, and $\nabla_{\theta_t} L$ represents the gradient of the loss function with respect to the trainable parameters. In this study, the Adam optimizer was employed with an initial learning rate of 10^{-4}.

(b) Fine-Tuning of Higher Layers (Domain Adaptation)

The deeper layers (typically the last 3–5 convolutional blocks) are unfrozen and fine-tuned using the ultrasound dataset. These layers learn domain-specific patterns for example, the textural signatures and shape irregularities characteristic of benign and malignant lesions. Fine-tuning enables the network to adapt its higher-level features to the diagnostic cues present in ultrasound data. The learning rate for these unfrozen layers is kept small (e.g., $1 \times 10^{-5}$to 1×10^{-4}) to ensure stable adaptation and prevent catastrophic forgetting of pretrained weights.

(c) Custom Classification Head

After removing the original ImageNet classification layer (typically a 1000-class dense layer), a custom classification head is appended to suit the three diagnostic categories (Normal, Benign, Malignant). The customized top layers are structured as shown in Eq. (2):

$$\begin{aligned} \text{GAP} &\rightarrow \text{Dense}(256, \text{ReLU}) \\ &\rightarrow \text{Dropout}(0.3) \\ &\rightarrow \text{Dense}(3, \text{Softmax}) \end{aligned} \quad . \tag{2}$$

The GlobalAveragePooling2D (GAP) was used to reduce the spatial dimensions of feature maps to a single value per feature channel by averaging. This retains the most salient features and prevents overfitting compared to fully connected layers. It also reduces model parameters dramatically, thereby improving efficiency for federated deployment. The Dense (256, ReLU) is a fully connected layer with 256 neurons and ReLU (Rectified Linear Unit) activation. It learns high-level abstract representations of the input features and also helps to introduce nonlinearity and mitigates vanishing gradient issues. The Dropout (0.3) randomly deactivates 30% of neurons during training to reduce co-adaptation and overfitting. Particularly effective in small and noisy medical datasets. Dense (3, Softmax) is the final classification layer that outputs probability distributions

across the three classes (Normal, Benign, Malignant). The Softmax function ensures that all probabilities sum to 1, as expressed in Eq. (3):

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{3} e^{z_i}}, \tag{3}$$

where (z_i) are the logits for each class.

(d) Optimization Strategy

The model is trained using an adaptive gradient-based optimization algorithm called Adam optimizer. It combines the advantages of AdaGrad and RMSProp. Adam was chosen because it dynamically adjusts the learning rate for each parameter based on first m_t and second v_t moment estimates of the gradients, as indicated in Eq. (4):

$$\theta_t = \theta_{t-1} - \propto \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon}. \tag{4}$$

Adam was also chosen for its stability, efficiency, and ability to converge faster in heterogeneous and federated environments. To improve convergence stability and prevent oscillations, an adaptive learning rate scheduler (e.g., *ReduceLROnPlateau*) is employed. When validation loss stops improving for a specified number of epochs, the learning rate is automatically reduced by a factor (commonly 0.1).

(e) Loss Function

The categorical cross-entropy loss function was used because it is suitable for multi-class classification, as shown in Eq. (5):

$$L = -\sum_{i=1}^{C} y_i \log(\widehat{y_i}), \tag{5}$$

where y_i is the true label (one-hot encoded) and $\widehat{y_i}$ is the predicted probability for class i.

(f) Regularization and Early Stopping

To prevent overfitting which is common with small datasets, early stopping was adopted. The implementation was carried out in such a way that training stops automatically if the validation loss does not improve after ten (10) consecutive epochs. In this circumstance, the model will revert to the weights that have the best validation performance. This is put in place to ensure optimal generalization and computational efficiency and to further prevent unnecessary local epochs during federated training.

3.4. Federated learning configuration

Fig. 3 illustrates the framework for the proposed Federated Deep Transfer Learning (FDTL)technique. The framework simulates a decentralized learning environment that imitates real-world telerobotic ultrasound systems, where multiple remote ultrasound terminals (clients) contribute to the improvement of a global diagnostic model without directly sharing patient data. This federated design ensures privacy preservation, scalability, and cross-institutional learning. This, in turn, addresses one of the central challenges of AI-driven medical imaging which is their inability to aggregate data from multiple sites due to ethical, legal, or logistical barriers.

Step 1: Data preprocessing

This arrow illustrates the commencement of the ultrasound data preprocessing. This is done to ensure the data is suitable for model training and analysis. In a federated and distributed learning environment under consideration, it is expected that each remote site holds its own subset of the BUSI dataset.

Step 2: Preprocessing to Transfer Learning Models

The preprocessed local ultrasound images are then fed into the pretrained transfer learning models: ResNet, DenseNet, EfficientNet and Inception. at each client node.

Step 3: Transfer Learning Models to federated server

In the proposed Federated Deep Transfer Learning Framework, the transfer Learning models: ResNet, DenseNet, EfficientNet, Inception serve as the local model architectures while the federated server coordinates their training synchronization and global aggregation. The local transfer learning models learn from data while the federated server learns from the models. The connection between these two components represents the communication loop through which knowledge transfer occurs in a decentralized environment. Each client node (e.g., a hospital or telerobotic terminal) hosts a transfer learning model that has been fine-tuned locally on its subset of ultrasound data. The model updates its parameters to better capture local data patterns for example, hospital-specific imaging artifacts, population variations, or device-specific contrast patterns. The resulting weights (or gradients) are now personalized versions of the global model denoted as w_k^{t+1}.

Therefore, the federated server never directly accesses these local datasets, it only manages the exchange of model parameters (weights and biases). The Federated Server aggregates all received model updates using the Federated Averaging (FedAvg) algorithm [38], as given in Eq. (6):

$$w_{\text{global}}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{t+1}, \tag{6}$$

where w_k^{t+1} is the local weights from client k, n_k is the number of samples at client k. N which is $\sum n_k$ is the total samples across all clients. The new global model w_{global}^{t+1} now represents aggregated diagnostic intelligence learned collaboratively across all institutions. This connection ensures that knowledge gained by each local model contributes to a shared global model, improving diagnostic performance without violating data privacy.

Step 4: Server-Clients broadcast after aggregation

There is a two-way communication between the federated server and the participating clients at this stage. A The downlink communication is established between the federated server and the participating clients. This enables the federated server to broadcast weights received from the transfer learning models to the participating clients after the aggregation step. This redistribution ensures that every participating client benefits from the collective learning experience of all sites. The server broadcasts the improved weights to all clients and the clients update their local copies with the new weights. This indicates the commencement of the next training round at the different participating clients. Every hospital or device trains its model independently to ensure data privacy and computational autonomy.

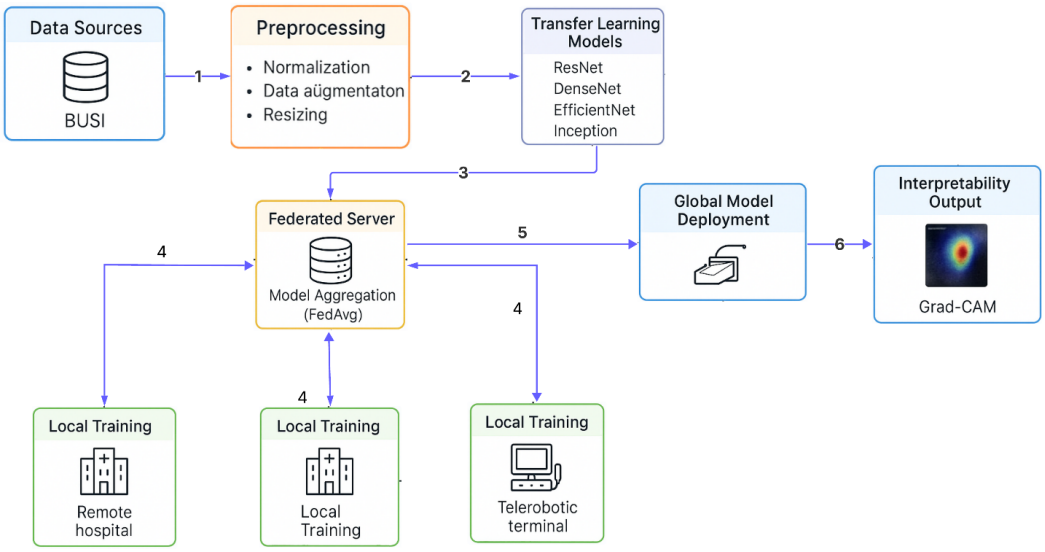


Figure 3: Federated Deep Transfer Learning (FDTL) framework for AI-driven ultrasound diagnosis.

Once training converges; the final parameter updates (not patient data) are sent back to the federated server for onward transmission to the global model. Each update is tagged with metadata: client ID (pseudonymous), number of samples (n_k), and checksum. Communication uses TLS encryption or secure aggregation protocols to ensure confidentiality. Federated server broadcasts the improved weights to all clients.

Step 5: Federated Server and the Global Model Deployment

After the federated server has collected all client updates and aggregates them using the Federated Averaging (FedAvg) algorithm in Eq. (6), the global model is deployed. This contains the global weights that reflects collective learning from all clients.

Step 6: Global Inference and Grad-CAM Interpretability Process

The deployed model generates predictions which are sent to the Gradient-Weighted Class Activation Mapping (Grad-CAM) heatmaps for interpretability.

3.5. *Grad-CAM interpretability module*

Fig. 4 presents the framework of the Gradient-weighted Class Activation Mapping (Grad-CAM) interpretability module. It explains how the module helped the FDTL framework to reach its diagnostic decision. In the proposed FDTL framework, the Grad-CAM generates a heatmap overlay that highlights regions of the image that the model considers most important for its classification (e.g., tumor boundaries or tissue abnormalities). By embedding the Grad-CAM at the global inference level, the FDTL framework ensures that clinical interpretability is achieved without compromising patient privacy or data localization.

As shown in Fig. 4, Grad-CAM uses the gradient information that flows from the deployed global model to generate a class-discriminative localization 2D heatmap that shows where the model is "looking" when making a prediction. The mathematical formulation guiding the operation of the Grad-CAM goes thus:

Input: Let y^c be the score (logit) for class (c) before the softmax layer in the model's output and let A^k denote the kth feature map in the final convolutional layer.

Step 1: Gradient computation

$$\frac{\partial y^c}{\partial A^k}. \tag{7}$$

The partial derivatives in Eq. (7) represent the influence of each feature map (A^k) on the class score (y^c).

Step 2: Global average pooling (weight calculation)

The importance weight ($\propto_c^k$) for each feature map is computed as shown in Eq. (8):

$$\propto_c^k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{8}$$

where i, j are the index spatial locations, and Z is the total number of pixels in the feature map.

Step 3: Weighted combination

The class activation map $L^c_{Grad-Cam}$ is obtained by linearly combining the feature maps weighted by $\propto_c^k$, as given in Eq. (9):

$$L^c_{\text{Grad-Cam}} = \text{ReLU}\left(\sum_k \propto_k^c A^k\right). \tag{9}$$

The ReLU ensures only the features that positively influence the class of interest are highlighted.

Step 4: Visualization

The resulting heatmap $L^c_{Grad-Cam}$ is upsampled to the input image size and superimposed on the original ultrasound image. Red or yellow regions indicate high importance, while blue regions indicate low relevance. Therefore, Grad-CAM translates the 'black box' of deep convolutional neural networks specifically the transfer learning models (e.g., ResNet, DenseNet,

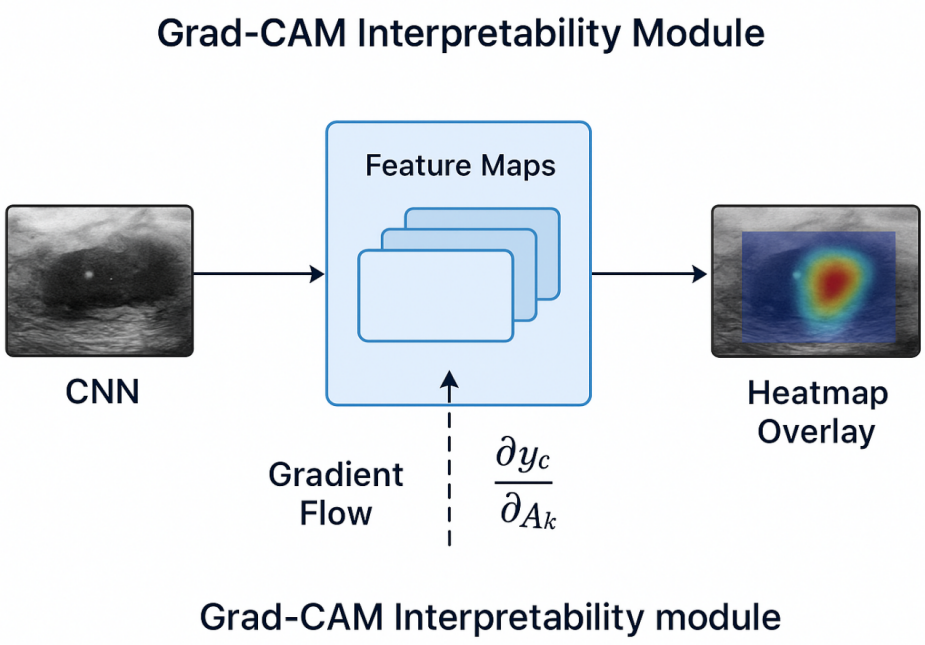


Figure 4: Grad-CAM interpretability module in the FDTL framework.

EfficientNet, and Inception) into clinically interpretable visual cues, bridging the gap between automated diagnosis and human expertise.

3.6. The FDTL architectural overview

Fig. 5 provides a conceptual overview of the Federated Deep Transfer Learning (FDTL) process, visually illustrating the client–server communication cycle and the FedAvg aggregation mechanism that enables collaborative, privacy-preserving model training across decentralized ultrasound datasets.

The schematic demonstrates how multiple remote clients representing distinct ultrasound imaging environments interact dynamically with a central federated server to iteratively improve an AI-driven diagnostic model without sharing raw patient data.

The technical architecture of the proposed FDTL framework can be divided into four (4) key stages:

a. The initialization stage is where the central federated server activates the global model with the pretrained deep learning techniques that served as the baseline transfer learning models.
b. The distribution stage is where the global model redistribute computed global weights to participating clients.
c. The local training stage is where the clients fine-tune their baseline model using local ultrasound data for several epochs which is 5–10 epochs per round
d. The aggregation stage is where the clients trasfer their locally trained weights to the central federated server for weighted averaging via FedAvg. The server acts as the the global coordination hub.

These stages are repeated between 50 to 100 communication rounds, thereby allowing the FDTL system to achieve convergence where additional updates yield negligible improvements in validation accuracy. This design was put in place to ensure continuous co-adaptation between the server and clients-thereby enabling the AI framework to learn from distributed, heterogeneous ultrasound data sources.

The central federated server connects via bidirectional arrows to three distributed client nodes each representing a unique type of medical imaging environment:

a. Client 1: A *local ultrasound site* operating within a hospital or diagnostic center.
b. Client 2: A *telerobotic node* enabling remote sonography under telemedicine setups.
c. Client 3: A *regional hospital node* managing a larger dataset with multiple imaging devices.

Each of these clients trains the same model architecture (e.g., ResNet, DenseNet, EfficientNet, or Inception) on locally stored ultrasound datasets that are non-independent and non-identically distributed (non-IID) reflecting realistic data variability across institutions, imaging protocols, and patient demographics. Also, model weights (not raw data) are exchanged between clients and the central server under secure encrypted channels (TLS).

3.7. Federated training workflow

The federated learning process occurs iteratively in communication rounds, with each round consisting of the following phases:

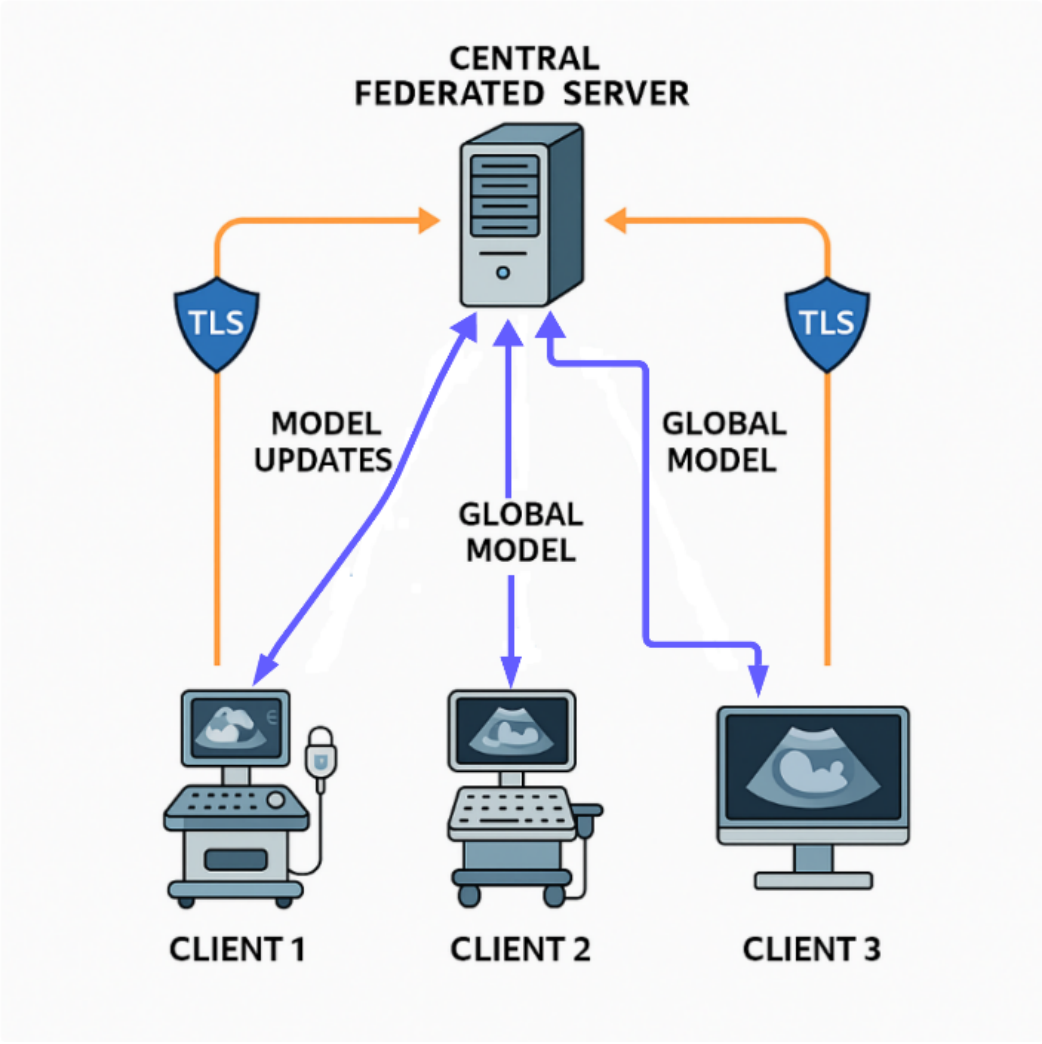


Figure 5: Technical architecture of the federated deep transfer learning framework for AI-driven ultrasound diagnosis.

a. Initialization Phase: The server initializes the global model parameters w^0 (typically a pretrained CNN fine-tuned on a subset of ultrasound data).
b. Model Distribution: The global model w^t at round t is transmitted to all participating clients C = {1, 2, ..., K}.
c. Local Training: Each client k trains the model on its local dataset D_k for a fixed number of epochs E using mini-batch gradient descent. The client updates its local parameters w_k^{t+1} according to Eq. (10):

$$w_k^{t+1} = w^t - \eta \nabla L_k w^t, \tag{10}$$

where η is the learning rate and $L_k w^t$ is the local loss function (categorical cross-entropy).

d. Model Aggregation: After local training, clients send their updated model parameters back to the central server. The server aggregates these parameters using the Federated Averaging (FedAvg) algorithm proposed by McMahan *et al.* [38], as shown in Eq. (11):

$$w_{\text{Global}}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{t+1}, \tag{11}$$

where w_k^{t+1} denotes the local model weights for client k, n_k is the number of samples in client k's dataset, and N = $\sum_{k=1}^{K} n_k$ is the total number of samples across all clients.

e. Model Update and Broadcasting: The updated global model w_{Global}^{t+1} is redistributed to all clients for the next communication round. This cycle continues until convergence criteria are met (e.g., plateauing validation accuracy or minimal loss improvement).
f. Global Evaluation: After training, the final global model is evaluated on a centralized validation set and optionally on local client test sets to assess generalization and robustness.

3.8. Algorithmic representation

The overall procedure can be summarized in Algorithm 1.

Algorithm 1: *Federated Deep Transfer Learning for Ultrasound Classification*

Input: Pretrained CNN model (ResNet/DenseNet/EfficientNet/Inception)
Output: Global federated model for ultrasound diagnosis
1. Initialize global model weights w_0
2. for each round t = 1 to T do
3. Server sends global model wt to all clients k ∈ {1, ..., K}
4. for each client k in parallel do
5. Train model locally on Dk for E epochs to obtain w_k^{t+1}
6. Send updated weights w_k^{t+1} to server
7. end for
8. Server aggregates updates using FedAvg: $w_{t+1} = \sum \frac{n_k}{N} w_k^{t+1}$
9. Evaluate w_{t+1} on validation data
10. end for
11. Return final global model w_T

3.9. Simulation environment

The federated learning environment was implemented using TensorFlow Federated (TFF) and PySyft (OpenMined) frameworks to simulate decentralized learning among multiple nodes. Other simulation environment information is:

a. Software Environment: Experiments were implemented using TensorFlow Federated (v0.24.0) for managing distributed aggregation logic, PySyft (v0.8.0) for virtualized client isolation, TensorFlow/Keras (v2.12.0), and Python 3.10.12 on an NVIDIA RTX GPU platform.
b. Simulated Clients: 3 virtual clients (Client A, Client B, Client C), each holding a distinct subset of the BUSI datasets.
c. Data Distribution: Non-IID partitioning to reflect real-world data heterogeneity.
d. Local Training Parameters:
   i. Epochs per round: 5–10
   ii. Batch size: 32
   iii. Optimizer: Adam (learning rate 1e-4)
e. Communication Rounds: 50–100, depending on model convergence rate.
f. Evaluation: Performance is tracked across rounds to monitor global model improvement and communication efficiency.
g. Hardware Setup: All experiments were conducted on an NVIDIA RTX GPU with TensorBoard used for visualization and convergence tracking.

4. Results and discussion

The results were grouped into five (5) sections. The performance evaluation of the pretrained CNN (ResNet, DenseNet, EfficientNet, Inception) after fine-tuning was carried out under centralized and federated setups. The results were reported in section 4.1. The results of the proposed federated learning convergence and communication analysis were reported in section 4.2. Similarly, section 4.3 reported the generalizability and robustness of the proposed model across different sites. The impact of Grad-CAM integration was reported in section 4.4 while section 4.5 presents a comparative summary of four normalized performance dimensions: accuracy, convergence rate, interpretability and communication efficiency of the proposed model across five setups.

4.1. Model performance evaluation

To demonstrate how well each pretrained CNN (ResNet, DenseNet, EfficientNet, Inception) model performed after fine-tuning and under federated learning, their performance in terms of accuracy, precision, recall, F1-score and AUC were calculated for both centralized training (baseline) and federated learning. The results are presented in Table 3:

Among all architectures, DenseNet-121 and ResNet-50 achieved the best balance between precision and recall under the federated learning setup. Despite slight performance degradation compared to centralized training, the FDTL models maintained high accuracy (>90%), confirming that decentralized learning did not compromise diagnostic reliability. However, federated models show a small, acceptable drop vs centralized baseline (≈0.5–1% absolute accuracy), while attaining a strong AUC and balanced precision/recall. This also shows that the FDTL models maintained high diagnostic performance while preserving data locality; however, clinical-grade performance requires validation using independent, multi-institutional clinical datasets."

4.2. Federated learning convergence and communication analysis

To show that the FDTL model converges efficiently and aggregates knowledge effectively across distributed clients, their convergence graphs and communication efficiency were computed. Fig. 6 plots the global validation accuracy of four pretrained CNN models (ResNet-50, DenseNet-121, EfficientNet-B0, and InceptionV3) over 80 communication rounds in the Federated Deep Transfer Learning (FDTL) setup.

Each communication round represents one federated aggregation cycle, where:

a. The federated server broadcasts the improved weights to all clients
b. Clients train locally on their ultrasound data,
c. Updated weights are sent back to the federated server, and
d. The server performs FedAvg to update the global model.

Thus, the y-axis (accuracy) reflects how the global model's performance improves with each successive round of aggregation.

4.2.1. Interpretation of the curve patterns

a. Initial Rapid Growth (Rounds 1–20)

At the start, all four models showed a sharp increase in global validation accuracy. This indicates that even a few rounds of communication enable effective knowledge exchange between the clients and the global server. The models rapidly

Table 3: Performance evaluation of the pretrained CNN under centralized and federated learning.

Model	Setup	Accuracy	Precision	Recall	F1-score	AUC
ResNet-50	Centralized	0.912	0.903	0.898	0.900	0.941
ResNet-50	Federated	0.906	0.899	0.889	0.894	0.935
DenseNet-121	Centralized	0.919	0.911	0.902	0.906	0.944
DenseNet-121	Federated	0.911	0.907	0.900	0.903	0.939
EfficientNet-B0	Centralized	0.903	0.900	0.887	0.893	0.928
EfficientNet-B0	Federated	0.897	0.893	0.882	0.887	0.922
InceptionV3	Centralized	0.905	0.903	0.897	0.900	0.933
InceptionV3	Federated	0.899	0.900	0.890	0.895	0.929

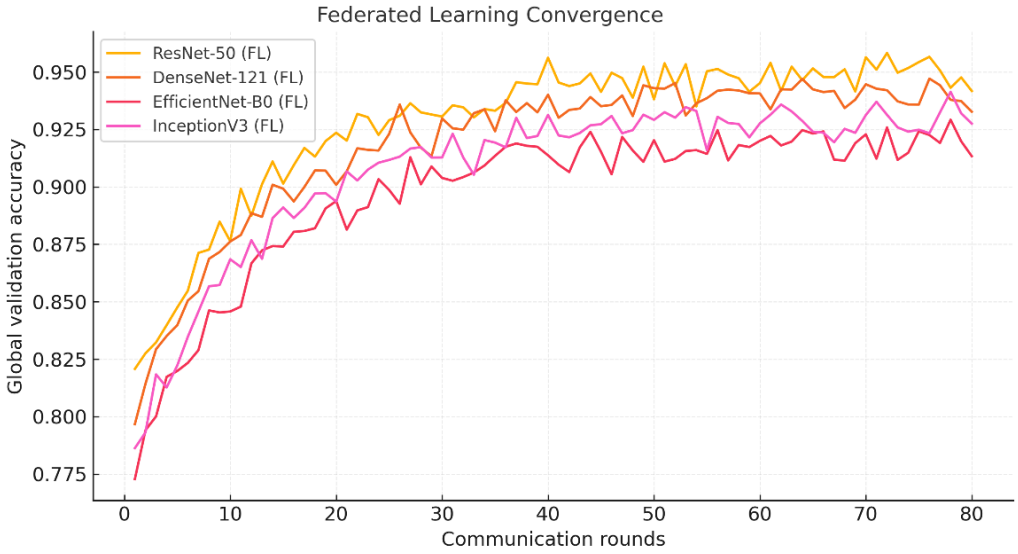


Figure 6: Federated Learning Convergence showing Accuracy Stabilization across Communication Rounds.

adapt to the distributed ultrasound data confirming that transfer learning (pretrained weights) provides a strong initialization. Therefore, the pretrained CNN backbones quickly capture domain-relevant ultrasound features, thereby reducing the number of rounds required for stable convergence.

b. Gradual Plateau Phase (Rounds 20–80)

Beyond round 20, the curves begin to flatten and oscillate slightly around a steady accuracy level. This indicates that the model has reached convergence, i.e., subsequent updates from clients contribute diminishing gains. The mild oscillations are expected due to data heterogeneity (non-IID distributions) among client datasets as each client slightly perturbs the global model during aggregation.

4.2.2. Model-wise comparative trends

The comparative trends of the pretrained deep learning models are presented in Table 4. The convergence pattern validates that deeper, well-regularized architectures (ResNet and DenseNet) are particularly suited for federated optimization on ultrasound data. Meanwhile, EfficientNet's compactness highlights its efficiency–accuracy tradeoff for telerobotic or low-bandwidth environments.

Finally, all models reach a global validation accuracy plateau between 0.89 and 0.91, demonstrating a consistent convergence across architectures, and Minimal communication divergence across 80 rounds. This shows that the FDTL framework maintains communication efficiency (no vanishing or oscillating accuracy), achieves global stability despite data heterogeneity, and requires only moderate communication rounds to reach optimal diagnostic performance.

4.3. Evaluation of cross-client consistency across simulated partitions

To demonstrate that the global model generalizes across diverse clients and remains robust under non-IID conditions, cross-site testing was carried out in terms of their generalization and robustness. Generalization measures the model's ability to maintain high performance on unseen data which are data that come from distributions it wasn't directly trained on. The results are presented in Table 5. The results revealed accuracy values between 0.89 and 0.91 across all simulated BUSI client partitions. It also showed a consistent F1-scores between 0.89 and 0.91 which indicate that both precision and recall are balanced. A stable AUC values between 0.92 and 0.94 revealed

Table 4: Model-wise comparative trends and their interpretation.

Model	Observed Trend	Interpretation
DenseNet-121	Converges fastest and achieves highest accuracy	Dense connectivity encourages efficient gradient flow and feature reuse, ideal for federated aggregation stability.
ResNet-50	Shows steady convergence and high final accuracy	Residual connections improve generalization and reduce communication noise effects.
InceptionV3	Slightly slower to converge but stable	Multi-scale feature extraction captures diverse lesion structures; minor instability due to deeper architecture.
EfficientNet-B0	Achieves competitive accuracy with fewer parameters	Lightweight model performs well while reducing communication and computational costs ideal for low-resource clients.

that the model discriminates well between positive and negative cases across domains. Since these results are stable across multiple simulated BUSI client partitions (with no large performance drop between them), it means our proposed global FDTL model did not overfit to any client's data distribution. Instead, it learned general, transferable diagnostic features that work well across all datasets. Therefore, the aggregated federated model generalizes well across heterogeneous simulated clients.

Robustness measures the model's ability to perform reliably under variations or perturbations in input data. As captured in Table 6, the low variance across sites supports robustness. The difference in accuracy between the highest and lowest site ($\approx$ 0.014) is minimal. Also, F1 and AUC values are consistent, showing that the model's performance doesn't collapse under different site conditions.

Furthermore, the cross-site performance evaluation presented in Fig. 7 showed a consistent accuracy, F1-score, and AUC across remote hospital A, telerobotic node B, and regional hospital C, confirming that the global model generalizes effectively and remains robust across heterogeneous client environments. Therefore, the cross-site evaluation indicates that the proposed FDTL framework generalizes well across distributed ultrasound datasets, as evidenced by consistently high accuracy, F1-score, and AUC across all client nodes. The minimal performance deviation between sites also highlights the robustness of the global model against data heterogeneity, thereby validating its suitability for real-world telerobotic diagnostic environments.

4.4. Interpretability and clinical explainability (grad-CAM evaluation)

Fig. 8 is used to show that Grad-CAM visualizations validate the AI model's decision-making and build clinical trust. The red and yellow regions showed regions that needs high attention while the blue or dark regions are areas with low relevance. Specifically, in the malignant BUSI image sample used in Fig. 8, it was observed that the red and yellow activations cluster around the mass's spiculated margins and internal hypoechoic areas, which are hallmark features of malignancy. Therefore, the Grad-CAM's red-to-yellow heatmaps confirm that the model's attention aligns with clinically significant regions such as lesion cores and margins in true-positive cases. A closer look at Fig. 8, also revealed that the heatmap concentrates on the tumor region, especially along the irregular boundaries and heterogeneous core areas, which are clinically associated with malignant lesions. Non-diagnostic background regions (such as noise, speckle, and surrounding tissue) exhibit low activation, meaning the model correctly ignores irrelevant areas.

4.5. Comparative summary

Comparative analysis across centralized baseline, federated FDTL, and local client models is presented strictly through measured classification metrics in Tables 3, 5, and 7, avoiding unverified composite scores.

4.6. Comparison of the proposed FDTL framework with existing approaches

To further establish the contribution and relevance of the proposed Federated Deep Transfer Learning (FDTL) framework, its performance and methodological characteristics were compared with selected recent approaches for federated and deep learning-based ultrasound diagnosis as shown in Table 7. The comparison considers key dimensions including dataset characteristics, learning paradigm, model architecture, transfer learning, federated aggregation, interpretability, and evaluation strategy. Particular attention was given to whether existing approaches support decentralized learning, address data heterogeneity, incorporate explainability, and evaluate communication or convergence behaviour. This comparison provides a broader context for interpreting the results obtained in this study and highlights the specific combination of deep transfer learning, FedAvg-based federated optimization, simulated non-IID client evaluation, communication analysis, and Grad-CAM interpretability offered by the proposed FDTL framework.

4.7. Critical discussion of the proposed FDTL framework

The proposed FDTL framework demonstrates performance close to centralized learning while maintaining data locality, consistent with recent federated ultrasound studies. Its contribution lies in integrating four pretrained CNN architectures with FedAvg, convergence and communication analysis, and

Table 5: Cross-site testing.

Client Node	Dataset Type	Accuracy	F1-score	AUC
Remote Hospital A	Non-IID BUSI subset	0.904	0.898	0.932
Telerobotic Node B	BUSI subset	0.911	0.906	0.940
Regional Hospital C	BUSI non-IID Partition C	0.897	0.889	0.925

Table 6: Quantitative verification of robustness.

Metric	Description	Values Obtained
Mean ± Std (Accuracy)	Measures average cross-site performance variability	0.904 ± 0.007
Coefficient of Variation (CV)	Standard deviation / mean (lower = more stable)	CV ≈ 0.8%
Accuracy (max-min)	Range between best and worst sites	0.014

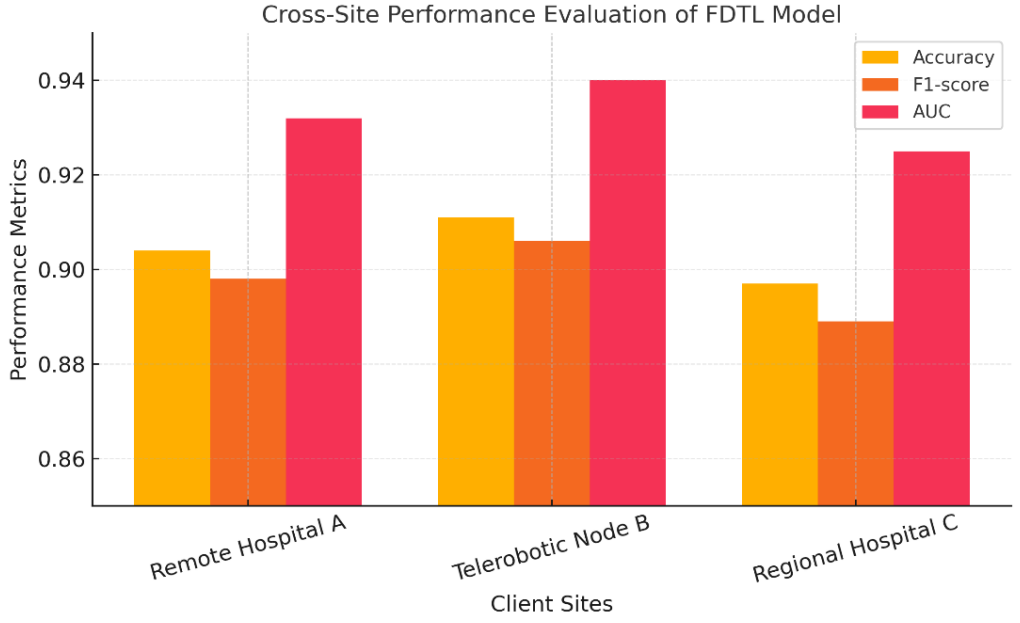


Figure 7: Cross-site performance evaluation of the federated deep transfer learning (FDTL) model.

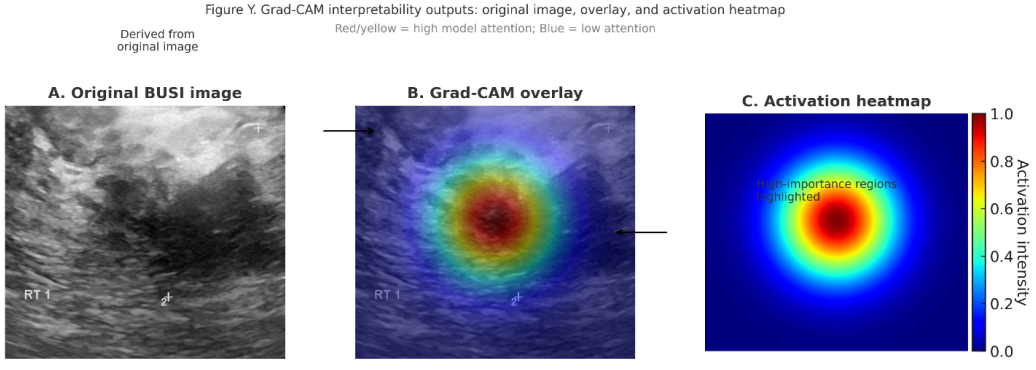


Figure 8: Grad-CAM interpretability overlay on a BUSI malignant ultrasound image.

Grad-CAM interpretability. Rather than claiming absolute superiority, the findings demonstrate a balanced approach to diagnostic performance, decentralized learning, and explainability. However, the use of the BUSI dataset, comprising only 780

Table 7: Comparison of the proposed FDTL framework with existing approaches.

Study	Dataset(s)	Federated Learning	Transfer Learning	Architecture	Explainability	External/Multi-site Evaluation	Main Result
Ru *et al.* [39]	BUSI, BUS, OASBUD	✓	Partial	CNN + GNN	—	✓ Multi-centre	AUC: 0.911, 0.871, 0.767
Raheem *et al.* [40], FAME	BUSI, BUSC, UDIAT	✓	✓	ResNet50V2, NASNetLarge, MAU-Net + ensemble	Attention	Independent datasets	BUSI Acc.: 98.70%
Jiang *et al.* [41], UltraFedFM	>1M ultrasound images	✓	Self-supervised pretraining	Foundation model	—	16 institutions / 9 countries	Mean diagnostic AUROC: 0.927
Comparative FL study [42]	3 breast-US datasets	✓	Multiple DL architectures	Multiple CNN/modern architectures	—	✓	Centralized/ local/federated comparison
Proposed FDTL	BUSI	✓ FedAvg	✓	ResNet50, DenseNet121, EfficientNetB0, InceptionV3	✓ Grad-CAM	Simulated 3 clients	Acc.: 90.5%

images, limits the strength of conclusions regarding generalizability. Although transfer learning and augmentation help mitigate data scarcity, they cannot completely eliminate potential dataset-specific bias. Therefore, the reported 90.5% federated accuracy should be interpreted within the BUSI evaluation setting, with larger and more diverse datasets required for external validation.

Furthermore, the non-IID client configuration provides evidence that the framework can maintain relatively consistent performance under simulated data heterogeneity. However, because the clients were derived from the same BUSI dataset, they do not fully represent the statistical and clinical differences expected among genuinely independent healthcare institutions. Thus, the results demonstrate robustness under simulated non-IID conditions rather than confirmed real-world cross-institutional generalization. Finally, the centralized model achieved 0.917 accuracy compared with 0.905 for the federated model, representing a relatively small difference of 0.012 (1.2 percentage points). This indicates that federated learning can retain most of the diagnostic performance of centralized training while keeping raw data localized. The slight performance reduction reflects the optimization challenges associated with heterogeneous data and federated aggregation. Overall, the findings suggest that the principal value of FDTL lies in balancing diagnostic performance, privacy-preserving data localization, communication efficiency, and interpretability, rather than simply maximizing classification accuracy.

5. Conclusion

This study presented a Federated Deep Transfer Learning (FDTL) framework for automated breast ultrasound classification across simulated decentralized client nodes. By fine-tuning pretrained CNN backbones (ResNet-50, DenseNet-121, EfficientNet-B0, InceptionV3) via FedAvg aggregation, the framework achieved a peak federated classification accuracy of 91.1% (DenseNet-121), retaining the diagnostic accuracy of centralized baselines while ensuring raw data localization. Integration of Grad-CAM provided visual activations focused on clinically relevant lesion margins and internal textures. Because evaluations were performed on non-IID partitions of the public BUSI dataset, these results demonstrate simulated computational feasibility and data locality rather than formal mathematical privacy or real-world clinical readiness. Future work should evaluate formal privacy mechanisms (e.g., Differential Privacy), explicit communication bandwidth benchmarks, and multi-institutional clinical trials prior to deployment in real-world telerobotic environments.

Data availability

The dataset used in this study is publicly available. The breast ultrasound images dataset was originally published by Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy in Data in Brief. The dataset can be accessed via https://doi.org/10.1016/j.dib.2019.104863. No new data were generated during the current study.

Declaration of competing interest

All authors declare that there is no conflict of interest in the publication of this work. All authors read and approve the content of this publication.

Funding

The authors received no specific funding from any public, commercial, or not-for-profit organisation for this research.

References

[1] A. E. Anasthesia, U. Ibrahim, S. D. Yusuf, D. Z. Joseph, N. Flavious, M. Sidi, S. Sheme, A. Mundi, A. Dare, D. S. Joseph & Y. A. Ningii, "Diagnostic reference levels (DRLs) and image quality evaluation for digital mammography in a Nigerian facility", Journal of the Nigerian Society of Physical Sciences **4** (2022) 281. https://doi.org/10.46481/jnsps.2022.734.

[2] A. A. Oglat, K. A. Rabaeh, M. D. Salman, M. M. Eyadeh & K. Zyoid, "Temperature- and frequency-dependent ultrasound response of a Poly(N-vinylcaprolactam) polymer gel dosimeter", Journal of the Nigerian Society of Physical Sciences **8** (2026) 3305. https://doi.org/10.46481/jnsps.2026.3305.

[3] S. I. Ele, U. R. Alo, H. F. Nweke, A. H. Okemiri & E. O. Uche-Nwachi, "Deep convolutional neural network (DCNN)-based model for pneumonia detection using chest X-ray images", Journal of the Nigerian Society of Physical Sciences **7** (2025) 2443. https://doi.org/10.46481/jnsps.2025.2443.

[4] S. Abuowaida, H. A. Owida, D. M. Alsekait, N. Alshdaifat, D. S. AbdElminaam & M. Alshinwan, "UltraSegNet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images", Computers, Materials & Continua **83** (2025) 3303. https://doi.org/10.32604/cmc.2025.063470.

[5] S. J. Adams, B. Burbridge, L. Chatterson, P. Babyn & I. Mendez, "A telerobotic ultrasound clinic model of ultrasound service delivery to improve access to imaging in rural and remote communities", Journal of the American College of Radiology **19** (2022) 162. https://doi.org/10.1016/j.jacr.2021.07.023.

[6] D. Black, M. Nogami & S. Salcudean, "Mixed reality human teleoperation with device-agnostic remote ultrasound: Communication and user interaction", Computers & Graphics **118** (2024) 184. https://doi.org/10.1016/j.cag.2024.01.003.

[7] S. R. Sugguna & E. P. Sumesh, "HybridNet: Advancing MRI image quality using dense attention and deep learning", Engineering Science and Technology, an International Journal **71** (2025) 102186. https://doi.org/10.1016/j.jestch.2025.102186.

[8] M. K. Singh, A. K. Singh, P. Choudhary, A. K. Singh & L. K. Tyagi, "A practical deep learning framework for multi-class skin cancer detection using U-Net segmentation and ResNet50 transfer learning", Journal of the Nigerian Society of Physical Sciences **8** (2026) 3530. https://doi.org/10.46481/jnsps.2026.3530.

[9] M. Bhargavi & S. Balakrishna, "Transfer learning-based hybrid feature learning framework for enhanced skin cancer diagnosis using deep feature integration", Engineering Science and Technology, an International Journal **69** (2025) 102135. https://doi.org/10.1016/j.jestch.2025.102135.

[10] D. Guo, Z. Xue, D. Chen, C. Lu, J. Fu, J. Yuan & Y. Huang, "Radiomics knowledge-driven deep learning framework for breast cancer ultrasound video diagnosis", Expert Systems with Applications **298** (2026) 129800. https://doi.org/10.1016/j.eswa.2025.129800.

[11] M. D. Liman, S. I. Osanga, E. S. Alu & S. Zakariya, "Regularization effects in deep learning architecture", Journal of the Nigerian Society of Physical Sciences **6** (2024) 1911. https://doi.org/10.46481/jnsps.2024.1911.

[12] E. Patnala, R. Bingu, H. R. Sangaraju, S. Adinarayana, J. S. Dhatterwal & S. Kavitha, "Automated breast cancer classification using a hybrid deep learning framework with attention-integrated EfficientNet", Biomedical Signal Processing and Control **112** (2026) 108552. https://doi.org/10.1016/j.bspc.2025.108552.

[13] H. Jasrotia, C. Singh & S. Kaur, "EfficientNet-based attention residual U-Net with guided loss for breast tumor segmentation in ultrasound images", Ultrasound in Medicine & Biology **51** (2025) 1112. https://doi.org/10.1016/j.ultrasmedbio.2025.03.009.

[14] N. Sirjani, M. G. Oghli, M. K. Tarzamni, M. Gity, A. Shabanzadeh, P. Ghaderi, I. Shiri, A. Akhavan, M. Faraji & M. Taghipour, "A novel deep learning model for breast lesion classification using ultrasound images: A multicenter data evaluation", Physica Medica **107** (2023) 102560. https://doi.org/10.1016/j.ejmp.2023.102560.

[15] V. T.-T. Vo, T. Shin, H.-J. Yang, S.-R. Kang & S.-H. Kim, "A comparison between centralized and asynchronous federated learning approaches for survival outcome prediction using clinical and PET data from non-small cell lung cancer patients", Computer Methods and Programs in Biomedicine **248** (2024) 108104. https://doi.org/10.1016/j.cmpb.2024.108104.

[16] R. Karthik, R. Menaka, G. S. Kathiresan, M. Anirudh & M. Nagharjun, "Gaussian dropout based stacked ensemble CNN for classification of breast tumor in ultrasound images", IRBM **43** (2022) 715. https://doi.org/10.1016/j.irbm.2021.10.002.

[17] W. Wang, J. Li, Z. Wang, Y. Liu, F. Yang & S. Cui, "Study on the classification of benign and malignant breast lesions using a multi-sequence breast MRI fusion radiomics and deep learning model", European Journal of Radiology Open **13** (2024) 100607. https://doi.org/10.1016/j.ejro.2024.100607.

[18] A. Nour & B. Boufama, "Hybrid deep learning and active contour approach for enhanced breast lesion segmentation and classification in mammograms", Intelligence-Based Medicine **11** (2025) 100224. https://doi.org/10.1016/j.ibmed.2025.100224.

[19] Z. Wang, M. M. Ali, K. I. Sahibzada, F. Maqsood, N. U. Rehman, M. Aftab, Q. Zia, W. Hou & D.-Q. Wei, "Hybrid feature extraction for breast cancer classification using the ensemble residual VGG16 deep learning model", Current Bioinformatics **20** (2025) 149. https://doi.org/10.2174/0115748936333380240816053223.

[20] R. Tenajas & D. Miraut, "Challenges in standardizing image quality across diverse ultrasound devices", Intelligent Medicine **4** (2024) 275. https://doi.org/10.1016/j.imed.2024.01.002.

[21] I. Nissar, S. Alam & S. Masood, "SwinEff-AttentionNet: A dual hybrid model for breast image segmentation and classification using multiple ultrasound modality", Biomedical Signal Processing and Control **112** (2026) 108795. https://doi.org/10.1016/j.bspc.2025.108795.

[22] S. Kumbhare, A. B. Kathole & S. Shinde, "Federated learning aided breast cancer detection with intelligent heuristic-based deep learning framework", Biomedical Signal Processing and Control **86** (2023) 105080. https://doi.org/10.1016/j.bspc.2023.105080.

[23] J. Du, W. Li, P. Liu, C.-M. Vong, Y. You, B. Lei & T. Wang, "Federated learning using model projection for multi-center disease diagnosis with non-IID data", Neural Networks **178** (2024) 106409. https://doi.org/10.1016/j.neunet.2024.106409.

[24] S. Din, M. Imran, S. Irfan, A. Tariq, Z. Elgamal & H. Dawood, "Federated learning for melanoma detection: A decentralized and privacy-preserving approach using deep learning", Biomedical Signal Processing and Control **112** (2026) 108576. https://doi.org/10.1016/j.bspc.2025.108576.

[25] B. Karthiga, K. R. Praneeth, V. Saravanan & T. K. R. K. Rao, "Enhancing cancer detection in medical imaging through federated learning and explainable artificial intelligence: A hybrid approach for optimized diagnostics", Egyptian Informatics Journal **31** (2025) 100751. https://doi.org/10.1016/j.eij.2025.100751.

[26] A. Raza, A. Guzzo, M. Ianni, R. Lappano, A. Zanolini, M. Maggiolini & G. Fortino, "Federated learning in radiomics: A comprehensive meta-survey on medical image analysis", Computer Methods and Programs in Biomedicine **267** (2025) 108768. https://doi.org/10.1016/j.cmpb.2025.108768.

[27] R. Joshi, S. Kumari, K. Pandey & R. Badola, "A collaborated federated learning for healthcare informatics: Solution and challenges", in *Federated Learning in Metaverse Healthcare*, S. Mahajan & J. M. Chatterjee (Eds.), Academic Press, 2026, pp. 171–188. https://doi.org/10.1016/B978-0-443-33789-5.00024-8.

[28] D. Wang, Y. Cui, M. Li, H. Xi & J. Zhu, "FedFAT: Frequency adpative interpolation for federated domain generalization on heterogeneous medical images", Pattern Recognition **172** (2026) 112459. https://doi.org/10.1016/j.patcog.2025.112459.

[29] O. O. Oladimeji, H. Ayaz, I. McLoughlin & S. Unnikrishnan, "Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis", Results in Engineering **24** (2024) 103071. https://doi.org/10.1016/j.rineng.2024.103071.

[30] M. Naas, H. Mzoughi, I. Njeh & M. BenSlima, "An explainable AI for breast cancer classification using vision transformer (ViT)", Biomedical Signal Processing and Control **108** (2025) 108011. https://doi.org/10.

1016/j.bspc.2025.108011.

[31] J. S. Christobel & K. S. S. Rani, "A novel vision-efficient Grad-CAM network for early breast cancer detection using multi-scale histopathological image analysis", Biomedical Signal Processing and Control **112** (2026) 108939. https://doi.org/10.1016/j.bspc.2025.108939.

[32] S. Singh, R. Kumar, N. P. Singh & M. K. Maurya, "Unsupervised attention-based framework for cancerous region segmentation with explainable post-hoc thresholding in histopathological images", Biomedical Signal Processing and Control **113** (2026) 108815. https://doi.org/10.1016/j.bspc.2025.108815.

[33] C. Chen, H. Pan, K. Zhang, Z. Li & F. Yu, "A federated fairness-aware incentive mechanism for medical image classification", Engineering Applications of Artificial Intelligence **160** (2025) 111855. https://doi.org/10.1016/j.engappai.2025.111855.

[34] F. Hu, S. Qiu, X. Yang, C. Wu, M. B. Nunes & H. Chen, "Privacy-preserving healthcare and medical data collaboration service system based on blockchain and federated learning", Computers, Materials & Continua **80** (2024) 2897. https://doi.org/10.32604/cmc.2024.052570.

[35] X. Deng, S. Ding, H. Lin, L. Jiang, H. Sun & J. Jin, "Privacy-preserving online medical image exchange via hyperchaotic memristive neural networks and DNA encoding", Neurocomputing **653** (2025) 131132. https://doi.org/10.1016/j.neucom.2025.131132.

[36] W. Al-Dhabyani, M. Gomaa, H. Khaled & A. Fahmy, "Dataset of breast ultrasound images", Data in Brief **28** (2020) 104863. https://doi.org/10.1016/j.dib.2019.104863.

[37] S. Haluzani, A. Lungu, M. Mudenda, A. Hamapa, C. Mkandawire, A. Kaluba, O. Khondowe, M. Mutale, K. Imasiku, M. Nyirenda, O. Nkolola, L. Kasonka, F. Simute, A. Mwiche, E. Njelesani, P. Mwaba, C. V. Alexandra, D. Lissauer, M. Mulaya, D. Kabamba, J. M. Zulu & L. Hangulu, "Quality of ultrasound images by midwives implementing ultrasound scanning services during antenatal clinics within selected healthcare facilities of Zambia", WFUMB Ultrasound Open **2** (2024) 100057. https://doi.org/10.1016/j.wfumbo.2024.100057.

[38] H. B. McMahan, E. Moore, D. Ramage, S. Hampson & B. Agüera y Arcas, *Communication-efficient learning of deep networks from decentralized data*, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, USA, 2017, pp. 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html.

[39] J. Ru, Z. Zhu & J. Shi, "Spatial and geometric learning for classification of breast tumors from multi-center ultrasound images: A hybrid learning approach", BMC Medical Imaging **24** (2024) 133. https://doi.org/10.1186/s12880-024-01307-3.

[40] A. Raheem, Z. Yang, A. S. Alluhaidan, M. A. Manan, S. Ahmed, F. Sabah & S. Ahmad, "FAME: A privacy-preserving dual-stage deep learning framework for breast ultrasound imaging using federated transfer and synthetic learning", Digital Health **11** (2025) 20552076251390564. https://doi.org/10.1177/20552076251390564.

[41] Y. Jiang, C.-M. Feng, J. Ren, J. Wei, Z. Zhang, Y. Hu, Y. Liu, R. Sun, X. Tang, J. Du, X. Wan, Y. Xu, B. Du, X. Gao, G. Wang, S. Zhou, S. Cui & Z. Li, "From pretraining to privacy: Federated ultrasound foundation model with self-supervised learning", npj Digital Medicine **8** (2025) 714. https://doi.org/10.1038/s41746-025-02085-0.

[42] M. A. Elshenawy, N. S. Tawfik, N. Hamada, R. Kadry, S. Fayed & N. Ghatwary, "A comparative analysis of federated learning for multi-class breast cancer classification in ultrasound imaging", AI **6** (2025) 316. https://doi.org/10.3390/ai6120316.