IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 5, May, 2025, pp. 1379-1388
Journal homepage: http://iieta.org/journals/isi

Federated Learning for Multi-Center Medical Image Classification Using Deep Learning Models

Noor S. Hassan[1*], Ali H. Hamad[2]

[1] Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, Baghdad 10071, Iraq
[2] Information and Communication Engineering Department, AlKhwarizmi College of Engineering, University of Baghdad, Baghdad 10071, Iraq

Corresponding Author Email: eng.noorsabah@st.nahrainuniv.edu.iq

https://doi.org/10.18280/isi.300523

Received: 17 March 2025
Revised: 16 May 2025
Accepted: 22 May 2025
Available online: 31 May 2025

Keywords:
federated learning, deep learning, healthcare, medical images, CNN, VGG16

ABSTRACT

Artificial intelligence is being applied in numerous industries, including healthcare among others. Due in great part to needs like dependable findings, data security, exact prediction, and a volume of data, among other things, research is being undertaken in the AI-enabled healthcare market. Regarding conventional deep learning models, datasets saved on a single device are used throughout the training process. Training the data calls for both highly efficient equipment and a lot of storage capacity. The work shown here suggests a federated learning approach suitable for five different customers. 9702 ultrasonic images of the gallbladder (GB) correspond with eight distinct disease types. Every client owns a part of the dataset with some unique classes from those of other clients. This is so since clients have divided the dataset. Two deep learning models applied and assessed in this work were CNN and VGG16. Clients used both models as well as the global ones. This paper proposes a possible global model solution based on the FedAvg aggregation method. The results show that VGG16 shows better outcomes in classification for both the client and the global model with a 99% accuracy rate in FL and a 94% accuracy rate for local training alone operations. CNN shows accuracy with a 99% in Florida and an 81% for local training initiatives.

1. INTRODUCTION

Physicians employ medical images, such as computed tomography (CT), magnetic resonance imaging (MRI), and X-rays, to create reliable and accurate DL models that help in the diagnosis, detection, and treatment of many diseases. Different metrics are included in these images to aid in the extraction of patient information from the medical images during analysis. However, the objective of the analysis phase is to identify the processed images using a variety of techniques, including machine learning (ML), which are frequently employed in the analysis and feature extraction of medical images used for disease classification and diagnosis [1, 2].

The classification of a multiclass medical image gallbladder disease dataset is used in this current study. The gallbladder, a vital liver organ, is crucial for digestion but can be affected by diseases like gallstones, cholecystitis, cancer, and others. Early diagnosis is essential for effective treatment, especially in severe cases like gallbladder cancer, which improves survival rates [3].

Deep Learning for medical image interpretation is crucial for improving diagnostic abilities. However, the dispersion of clinical data across healthcare facilities poses a challenge, as it cannot be combined for centralized model training due to privacy laws and data exchange difficulties. This limits the generalizability and effectiveness of diagnostic models for gallbladder issues [4].

The dataset was distributed to multiple clients using two methods: IID (Independent and Identically Distributed) and Non-IID (Independent and Identically Distributed) to replicate a real-world federated learning environment. IID was achieved by randomly dividing the dataset into equal portions for each client, ensuring identical assignments. The most optimal configuration was used to test the relative performance of the FL model [5].

In this work, we used the dataset of gallbladder diseases as the case study to estimate the proposed system model in the field of medical image classification. The dataset used is IID (Independent and Identically Distributed), which is distributed among 3 healthcare centers and is ideal for deep learning algorithms, as it allows for faster convergence and higher accuracy.

Convolutional Neural Networks (CNNs) are a prominent deep learning technique that automatically extracts features from data and improves diagnostic accuracy in medical image analysis tasks. Combining CNN architectures with transfer learning techniques improves image classification performance, recognizing important visual patterns with less preprocessing on raw data. VGG16Net, the largest CNN architecture, has a high computational load [2, 6], which is employed to classify the gallbladder diseases dataset before using FL.

Deep learning methods' predictive abilities are influenced by the amount of training data, which should come from multiple sources [7]. Obtaining sufficient data in the field of medical imaging is a significant difficulty. This difficulty might be solved by collaboration among various institutions. Sharing medical data in a single location raises a number of legal, privacy, technological, and data-ownership issues.

To solve these issues, Federated Learning (FL) enables individual hospitals to benefit from the large datasets of numerous non-affiliated institutions without centralizing the data in one location [8]. FL's privacy-preserving feature allows collaborations across various medical institutes [9] without exchanging medical data between healthcare centers. This helps to preserve privacy and gain model generality through the global model.

In an ideal FL system, medical institutions collaborate using a centralized orchestration cloud server in a trusted execution environment. FL allows hospitals and other medical institutions to keep local data by training a model across multiple medical data centers. The central server maintains a global shared model, co-owned by all participating institutions, while each institution maintains its local version. The server verifies local model quality before aggregating updates based on preset criteria. Global models iterate between aggregation servers and participating institutions, producing high-quality, converged, and performant FL models [10], which are applied to an estimated gallbladder dataset.

Aggregation algorithms are techniques for combining the outcomes of several models that have been trained on the client's end using local data. In addition to updating the global model, they manage the fusion of local client training outcomes. In a FL scenario, several aggregation techniques are employed based on objectives such as convergence rate and user privacy. Weighted aggregation, secure aggregation, momentum aggregation, clipped average aggregation, and average aggregation are a few of these strategies. Average aggregation, which averages client updates, provides benefits like simplicity and increased model accuracy, but it can't work well with non-IID data and is vulnerable to fraudulent clients and outliers. The Federated Averaging FedAvg, FedProx, FedDist, and HeteroSAg aggregation algorithms are being developed for federated learning [11].

The Federated Averaging (FedAvg) is a fundamental algorithm that uses a central server to acquire a primary model, which is then trained using local data from each client. The model parameters are then transmitted to the central server, which then aggregates global models using parameters from other clients. The aggregated global model is then distributed to the clients, marking the end of one learning round. The process continues until the model's accuracy meets the required level, with both clients and the central server maintaining the models from previous iterations [12].

Alternatively, to enable distributed and privacy-preserving FL model training, Local data is used to train the model, and only the model updates are transmitted to a central server. By reducing the amount of data sent between devices and servers and decentralizing the machine learning process, federated learning was able to reduce the risk of data breaches due to malicious attacks and malfunctions in the systems.

Privacy preservation in FL is a crucial aspect of different learning models, ensuring the privacy of training data while enabling successful model training. Robust aggregation techniques, such as FedAvg and secure aggregation, are used to address model integrity attacks, while federal transfer learning fine-tunes models on decentralized data, outlier identification and removal improve performance and model resilience. Different approaches to secure privacy-preserving methods are explored, ranging from simple averaging to more advanced methods such as secure multi-party computing and differential privacy [11, 13].

In this work, a federated deep learning models based on CNN and VGG16 has been proposed for multiple clients (multiple healthcare centers), each client has part of the dataset to train. The gallbladder dataset has been used with eight classes divided between five clients. The rest of the paper is organized as follows: in Section 2 the related work is described. Sections 3 and 4 show a description of the dataset. algorithms and methodologies used in this work are presented in Section 5, while the proposed system design is shown in Section 6. In Section 7, a discussion of the obtained results is presented. Finally, a conclusion is introduced in Section 8.

2. RELATED WORK

A number of research studies have been carried out in medical imaging analysis based on DL and FL in detection, segmentation, and classification. Federated Learning (FL) is rapidly transforming medical imaging and healthcare by enabling collaborative model training across multiple institutions while preserving patient data privacy. This section groups relevant studies based on their primary contributions to the field.

Some recent research has significantly advanced Federated Learning (FL) by focusing on the critical issues of data heterogeneity and training stability, particularly in medical imaging applications. This area of focus aims to overcome challenges posed by the non-Independent and Identically Distributed (non-IID) nature of medical data across different institutions, which can lead to training instability and reduced model performance. The overarching goal within this grouping is to develop robust FL algorithms capable of effectively learning from diverse and unbalanced datasets.

For instance, to mitigate training instability stemming from medical data heterogeneity, FedSLD (Federated Learning with Shared Label Distribution) was proposed and rigorously tested on OrganMNIST and PathMNIST datasets under various non-IID settings [14]. Similarly, HarmoFL was introduced to harmonize local and global drifts in FL on heterogeneous medical images. This approach normalizes image amplitudes in the frequency domain and utilizes client weight perturbation, with both theoretical analysis and empirical demonstrations confirming its superior convergence [15]. Furthermore, SplitAVG presented a heterogeneity-aware FL method that leverages network splitting and feature map concatenation to encourage unbiased model training, achieving results comparable to baseline centralized training even in highly heterogeneous environments [16]. The impact of data heterogeneity on FL algorithms was also empirically investigated using the COVIDx CXR-3 dataset, revealing a considerable reduction in global accuracy with non-IID data, especially for smaller datasets, while larger datasets showed improved accuracy [17].

Another area of research in FL focuses on enhancing model efficiency, overall performance, and generalization capabilities. This involves improving the computational efficiency of FL frameworks, boosting model accuracy, and ensuring that models trained via FL generalize effectively to unseen data from diverse sources. The ultimate aim is to make FL a more practical and effective solution for real-world clinical applications.

For instance, studies have investigated communication-efficient FL frameworks for multi-institutional medical image classification, demonstrating improvements in model training efficiency for methods like FedAvg and FedProx. One such study reported a 2% improvement in testing accuracy and a 28% reduction in training loss on a diabetic retinopathy dataset [18]. A comprehensive comparison between single-institution models, collaborative data sharing (CDS), and FL for brain tumor segmentation highlighted FL's superior model quality and generalization, achieving approximately 99% accuracy using the BraTS dataset from ten institutions [19].

Research has also concentrated on the application of FL with specific deep learning architectures and advanced feature engineering techniques. This explores how various deep learning models and sophisticated feature extraction methods are integrated within FL frameworks to address particular medical imaging tasks, leveraging the power of deep learning in a privacy-preserving distributed setting.

One approach utilized an ensemble of top-performing pre-trained CNN models (Inception V3, VGG19, DenseNet121) within an FL framework for brain tumor detection from MRI images, noting a trade-off where FL maintained privacy with 91.05% accuracy, compared to 96.68% for a centralized CNN [20]. The effectiveness of FL's pre-trained models has been evaluated by combining EfficientNet with CNN and incorporating traditional image processing techniques like Gray-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP) for liver CT and brain MRI image classification, demonstrating high accuracy rates (98.8% for CT and 97.4% for MRI) [21]. Additionally, a cooperative FL architecture based on the Inception-V3 model has been proposed for classifying lung and colon malignancies from histological images, achieving exceptionally high classification accuracy (99.867% for lung and 100% for colon cancer) [22].

Beyond standard diagnostic image classification, FL is demonstrating its versatility through creative applications in a range of healthcare contexts, including clinical outcome prediction and solving weakly-supervised learning problems.

For instance, an FL model trained on data from 20 global institutes was developed to predict the future oxygen requirements of symptomatic COVID-19 patients using vital signs, laboratory data, and chest X-rays. This model achieved an average AUC >0.92, showing a significant improvement in generalizability (38%) compared to single-site models [23]. For multi-site fMRI analysis, privacy-preserving FL combined with domain adaptation has been addressed, proposing a decentralized iterative optimization algorithm and a randomization mechanism for shared local model weights. This approach demonstrates promise for boosting neuroimage analysis without direct data sharing [24]. Another FL framework, incorporating differential privacy, has been proposed for improving histopathology image classification using a multiple model approach. This framework was applied to histopathology images, examining the effects of IID and non-IID distributions, healthcare provider numbers, and dataset sizes using the Cancer Genome Atlas dataset. The study concluded that FL is a reliable and efficient framework for collaborative model development, achieving performance similar to conventional training with strong privacy guarantees [25]. Finally, FedLPPA (Federated Learning with Personalized Prompt and Aggregation) offers a novel personalized FL framework to uniformly leverage heterogeneous weak supervision for medical image segmentation across different sites and annotation formats. FedLPPA maintains learnable universal and personalized prompts, integrated with sample features, and employs a dual-decoder strategy for pseudo-label generation, showing efficacy closely paralleling fully supervised centralized training [26].

The collected works demonstrate the Federated Learning potential to revolutionize medical imaging and healthcare by overcoming data silos and privacy concerns. A central theme is the development of robust FL algorithms that can effectively handle the inherent data heterogeneity across medical institutions, ensuring model stability, high performance, and strong generalization. Researchers are actively exploring various deep learning architectures and feature engineering techniques within FL frameworks to tackle specific diagnostic and predictive tasks, while also innovating new FL paradigms like personalized prompts and domain adaptation to address complex real-world challenges such as weakly-supervised learning and multi-modal data integration. Although some studies indicate potential trade-offs between strict privacy and peak accuracy compared to centralized models, the overwhelming consensus is that FL offers a powerful, ethical, and scalable solution for advancing AI in medicine by enabling collaborative intelligence without compromising patient data.

3. DATASET OVERVIEW

In the proposed system, The Gallbladder (GB) diseases dataset is a database of high-quality ultrasound pictures taken at four hospitals in Baghdad, Iraq. The data was gathered over four years from Jenin Hospital, the Al-Numan Teaching Hospital Specialized Gastroenterology Center, and the Gastroenterology Department of the City of Medicine Teaching Hospital. The information is essential for creating deep learning and machine learning algorithms that can identify and categorize gallbladder illnesses. The information supports comparison research and testing of new methods for an ultrasound image analysis to investigate the medical field and enhance patient care. The dataset was collected 10,692 high-resolution ultrasound images of the gallbladder from 1,782 individuals [27]. The images are organized into nine classes, each representing a specific gallbladder disease based on anatomical landmarks. The dataset includes images from female patients (6,246 images, average age 47 years) and male patients (4,446 images, average age 53 years) [28]. The images were acquired using cutting-edge technologies from four different ultrasound machines (Siemens Acuson X700, Philips Affiniti 70, Philips CX50 and Canon Viamo c100). The data collection took place over four years at four medical facilities in Baghdad, Iraq, with medical staff members and expert doctors contributing to the collection [29].

In this current work we are used only eight classes of gallbladder diseases dataset which are used to training and evaluation model, these selected classes or diseases (gallstones, cholecystitis, gangrenous cholecystitis, gallbladder perforation, polyps and cholesterol crystals, gallbladder adenomyomatosis, cancer, and intraabdominal and retroperitoneum problems) of 9702 images each resized to the dimension 224×224 pixels with RGB color mode. We show examples of each class in Figure 1, highlighting the differences in appearance on ultrasound at different gallbladder diseases. Table 1 indicates the whole distribution of Gallbladder disease images in diverse classes within the dataset.

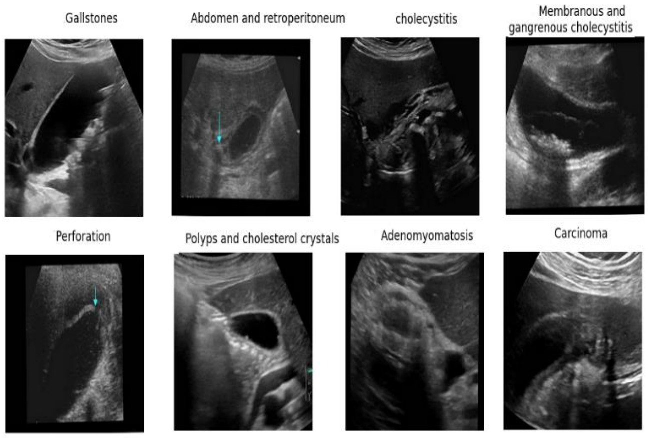

Gallstones
Abdomen and retroperitoneum
cholecystitis
Membranous and gangrenous cholecystitis
Perforation
Polyps and cholesterol crystals
Adenomyomatosis
Carcinoma

Figure 1. Sample ultrasound images from different classes

Table 1. Number of Gallbladder disease images per class

Gallbladder Diseases Classes	Number of Images	Number of Images after Balancing
Gallstones	1326	1020
Abdomen and retroperitoneum	1170	1020
Cholecystitis	1146	1020
Membranous and gangrenous cholecystitis	1224	1020
Perforation	1062	1020
Polyps and cholesterol crystals	1020	1020
Adenomyomatosis	1164	1020
Carcinoma	1020	1020

The GB, a hollow organ in the abdomen, stores bile secreted by the liver, causing numerous pathologies, with cholelithiasis being the most common. The list of frequent GB pathologies is summarized below.

Gallstones, which can be mild or severe, are tiny calcium crystals and cholesterol/bile salts that develop inside the gallbladder (GB). Cholecystitis occurs when stones in the bile duct obstruct the bile duct, trapping bile and causing inflammation. Gangrenous cholecystitis is a severe form of cholecystitis that necessitates an immediate cholecystectomy since it causes necrosis of the GB owing to problems with the blood supply. Perforation of the GB wall, a common major complication in inflamed GB, is a serious issue in people with diabetes and compromised immune systems, requiring urgent intervention for treatment. Polyps and cholesterol crystals are common causes of GB disease, with polyps occurring in 2.6%-9.9% of cases. Cholesterol polyps are benign, while endothelial polyps may develop into cancer. A major risk factor for adenomyomatosis of the GB is a persistent infection. The disease is characterized by mucosal epithelial hypertrophy, which leads to Luschka's crypts. Carcinoma is a GB cancer, a rare, female-dominated tumour, is more common in over 70-year-olds, often accompanied by gallstones. Early diagnosis and treatment are crucial for survival. Intraabdominal and retroperitoneum problems, including GB cancer spreading to retroperitoneal structures, can be detected using MRI or CT and ultrasound for further evaluation [29].

4. DATASET PREPROCESSING

The preprocessing step aimed to ensure the quality and balance of the gallbladder dataset for machine learning tasks that's includes:

1. Data balancing: This technique was applied to down sample each class to be the same size as the smallest class to mitigate this bias. The polyps and cholesterol crystals Class had the fewest images, with 1020. Random sampling was employed to select 1020 images from the remaining seven classes to balance the dataset. This resulted in a balanced dataset of 1020 images per class, leading to 8160 images (i.e., 1020 images × 8 classes). Figure 2 shows the balanced classes. This balanced dataset allowed the models to train on class-specific features effectively.

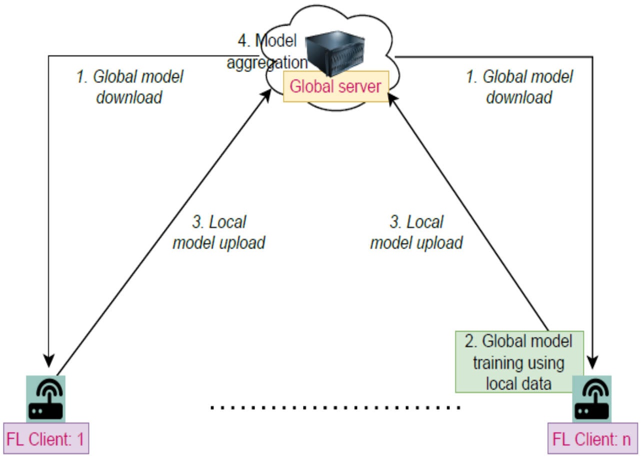


Figure 2. Traditional FL framework [30]

2. Image resizing: Resizing all images to a constant size of (224×224×3 pixels), which is an acceptable dimension that keeps the main structure of the images before feeding it into the network to reduce the number of parameters and reduce the requirement of computation power.

3. Split the dataset into training, validation, and testing: 70% of the dataset is used for training the model, 10% of the dataset is used for validation, and 20% for testing the accuracy of the model.

4. Dataset splitting: The dataset was split into three subsets, one for each client, to simulate the use of FL in medical image classification. To eliminate bias, the images in each class shuffled and distributed equally across three clients.

5. Labeling one-hot encoding: Use one-hot encoding by identifying the dataset's class labels for converting categorical information into numerical form. It is commonly used in neural networks to encode categorical data as model input, with each category represented as a vector of 0s and 1s, resulting in a clear and distinct numerical representation for each class in classification tasks.

5. METHODOLOGY

The methodology is structured to leverage advanced deep learning techniques and federated learning (FL) frameworks to enhance the classification performance of medical images for different gallblader disease detection. Two different architectures were applied, which include custom Convolutional Neural Network (CNN), VGG16 models, and combining each one of these models with an FL framework.

FL is a potential strategy for training machine learning models across several sites while protecting data privacy, especially in medical image analysis. FL enables local model training, which is essential in medical fields by allowing collaboration between institutions without sharing sensitive patient data and reducing bandwidth requirements [31].

A traditional FL includes several devices working as independent clients with a single global server. Usually, a fixed number of FL and clients (D) are selected randomly from a pool of edge devices that have shown their interest in participating in the learning process. For an iteration of training, each client (d) downloads the global learning model (θ_t) from the global server and start training the model with its own local data at time t. If md represents the local training dataset samples for client d, then $\sum_{d=1}^{D} = md = M$, where M is the total size of data samples from D number of clients. $f(\theta)$ is optimized by the FL.

$$f(\theta) = \sum_{d=1}^{M} \frac{md}{M} F_d(\theta) \tag{1}$$

$$F_d(\theta) = \frac{1}{m_d} \sum_{i \epsilon m_d} fi(\theta) \tag{2}$$

where, $fi(\theta)$ is the loss function related with sample i in the dataset of client d. During the local training, client d updates the global learning model using an optimization algorithm like Adam and stochastic gradient descent (SGD) to minimize their loss functions. Once local model training is complete, each client sends the global server the updated global model (θ_{t+1}^{d}).

$$(\theta_{t+1}^{d}) = \theta_t + \alpha_d \lambda_d \tag{3}$$

where, α_d is the learning rate and λ_d is the gradient computed at client d on its local data-set with θ_t. The global server computes the improved global model (θ_{t+1}) by combining the received local models as follows.

$$\theta_{t+1} = \sum_{d=1}^{M} \theta_{t+1}^{d} \tag{4}$$

For the next training cycle, FL clients receive the updated global model. The global model keeps going through this procedure until it converges [31].

Convolutional Neural Networks (CNNs), demonstrated excellent performance in the field of medical imaging by employing CNNs to construct an image classification model. Convolutional Neural Network (CNN) consists of 3D image input layers, convolutional layers for feature extraction, pooling layers for dimensionality reduction and compression of data, and fully connected layers to obtain distinguishable features that facilitate subsequent classification. It extracts features from images, compresses data, and learns both high-level and low-level features that are automatically extracted, eliminating the number of parameters and accelerating the system's processing [32].

Transfer learning is a process used to apply knowledge from one task to another. Feature transfer, parameter sharing transfer, and relational knowledge transfer are some of the several techniques of transfer learning. The model, like VGG16, is initialized on a pre-trained dataset and fine-tuned on the target domain. This method shortens training time, mitigates underfitting and overfitting, and enhances generalization performance. The source domain's model parameters for network initialization are taken from the ImageNet dataset. Furthermore, a new fully connected layer is built, and our dataset is used to retrain and modify every parameter in the network layers [32]. The parameter transfer approach is used in this work. This method implies pre-training a model on a different dataset (the source domain) to initialize the network and then fine-tuning the model on the GB dataset utilized for this work. This approach has the potential to reduce training time, successfully reduce underfitting and overfitting, and improve the model's generalization performance by utilizing transfer learning approaches. Accuracy (Acc), sensitivity (Sens), specificity (Spes), F1, and precision were evaluation measures. After a thorough analysis of the central server network's performance, the mean value of the best metrics. The multiclass confusion matrix evolves into an MxM matrix, where M denotes the total number of unique class labels (C0, C1, ..., CM). Their computational relationships are as follows:

$$Accuracy = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N} \sum_{j=1}^{N} C_i, j\, ()} \tag{5}$$

$$Precision = \frac{TP(C_i)}{TP(C_i) + FP(C_i)} \tag{6}$$

$$Sensitivity = \frac{TP(C_i)}{FN(C_i) + TP(C_i)} \tag{7}$$

$$F1(C_{i)} = \frac{2 * Pre(C_i) * Sense(C_i)}{Pre(C_i) + Sense(C_i)} \tag{8}$$

where,

TP(C_i) is true positive for class$_i$, FN(C_i) is false negative for class$_i$.

Among them, sensitivity and specificity measures are crucial for medical picture classification tasks. However, because they contradict each other, we decided to utilize sensitivity metrics to evaluate the model's performance in our tests.

6. PROPOSED SYSTEM

The proposed FL for multi healthcare centers has been designed for 3 clients each with a separate dataset. This dataset was sliced and simulated into three clients where each has some part of the data and distributed across multiple healthcare institutions or clients to assess the effect of federated learning (FL) on model efficiency. Each client trains their local neural networks independently using their own data, without sharing any sensitive patient information.

The technique uses two of the most widely used architectures (a custom CNN, VGG16). Every model is trained in isolation and also within an FL framework. The cloud server is distributing the initial global model to the clients, then the local model (custom CNN, VGG16) trains on data for 20 epochs and updates the local model, the clients send updates to the cloud server, then the cloud server aggregates updates from all clients and updates the global model, this process is repeated for 10 rounds. Finally, the cloud server evaluates the performance of the model's generality while preserving privacy and security.

The hyperparameters of proposed system architecture are illustrated in Table 2.

Table 2. Hyperparameter of proposed system architecture

Parameter	Value with CNN Model	Value with VGG16 Model
Convolution layer	2 with 16,32 filter size and 2*2 kernel size	Base model: Pre-trained vgg16 with image net weight
Pooling layer	Maxpooling	Global average pooling
Dropout rate	0.4	0.1 for 512 and 256 neurons, 0.3 for 128
Dense layer	1 with 128 neurons	512,256,128 neuron
Loss function	Categorical Cross entropy	Categorical Cross entropy
Activation function	Relue, SoftMax (in output layer)	Relue, SoftMax (in output layer)
Optimizer	Adam	Adam
Number of output classes	8	8
Batch size	32	32
Number of epoch	20	20
Number of round	10	10

The choice of 10 rounds in a global model convergence process ensures robust performance by integrating knowledge from all clients. It manages communication efficiency by limiting the number of rounds, which involves uploading local updates and downloading the global model. The number of rounds during the experiment satisfies the model's generality and convergence. That helps manage communication overhead, which is crucial in federated learning environments [33], especially across multiple healthcare centers.

The choice of 20 epochs for local training on each client per communication round in Federated Learning (FL) prevents overfitting and improves communication efficiency. A high number can lead to client drift and insufficient local learning, leading to intensive overfitting and massive domain shifts in the local models and, ultimately, decreasing the aggregated model's performance. Interestingly, while this scenario is a specific manifestation of the well-known non-IID, this refers to the generic situation where local data distributions are not identical and independently distributed [34]. While a low number can result in insufficient local learning and slower global convergence, the number of epochs in this work, selected by experimentation, helps to improve the model's performance on the IID GB dataset.

The Adam optimizer is chosen for both model architectures due to its robust performance and efficiency in deep learning applications. The loss function, categorical cross-entropy, is used for multi-class classification problems, minimizing the difference between the predicted probability distribution and the true one-hot encoded label.

Adam is adaptive optimizers that dynamically adjust learning rates based on gradient information. They stabilize training in noisy or non-stationary data settings. Adam adds momentum, offering precise parameter changes and quick convergence in centralized environments, especially for non-stationary objectives [35].

Global CNN Model: The model architecture consisted of two convolutional layers with filter sizes of 16 and 32, followed by max-pooling layers. A dropout layer with a rate of 0.4 was included to prevent overfitting. The model was then flattened and connected to a dense layer with 128 neurons and an output layer with 8 neurons for classification. Each client trained this model on its local dataset for 20 epochs per round. After 10 communication rounds, the global model achieved robust performance while maintaining privacy. The use of a lightweight CNN ensured that the computational was able to learn from all of its clients while never sharing any private data. The mathematical formulation of the FedAvg algorithm:

$$w_{t+1} = \frac{1}{N}\sum_{i=1}^{N} n_i w_i^t \qquad (10)$$

where, w_{t+1} is the updated global weights after round t, N is total number of clients, n_i is as number of samples at client i and w_i^t are local weights at client i after round t. It also uses the standard measures such as Acc, sens, prse, pre, recall, and loss as evaluation metrics for models to show how each model performs in centralized and FL setups. A comparative analysis outlines the workflow of the methodology describes the process of FL that uses the Federated Averaging (FedAvg) algorithm to train CNN and VGG16 global models across multiple clients. This method ensures that only model changes are sent to the central server. The algorithm starts by initializing global model weights, which are shared across all participants. Each client trains a local model on its own private dataset, customizing the global model to fit the unique data of that client. The server broadcasts these weights to all clients during each communication round, ensuring the same initial condition for local training. After training, clients upload their learned model weights to the central server, and operators aggregate to the global model using knowledge for each client while protecting privacy. The server performs a federated averaging step, updating global aggregation model weights via a weighted average of local model weights. This process is repeated for several communication rounds, resulting in a trained global model that holds knowledge from all clients without accessing their private data. Figure 3 shows the proposed federated learning system.

Algorithm 1: FL for gallbladder

```
Input: Gallbladder diseases dataset D
Output: Client and global model
Step 1: Preprocessed dataset
Step 2: Divide Dataset D for each client N {D1,D2,..DN}
Step 3: Initialize global model weights w
Step 4: Download global weight w to all clients
Step 5: For each round R:
            For each client:
                For each epoch:
                    Train local deep learning models
                End
            End
        Upload client model weight to the global model
        Aggregates weights using FedAvg
        Train global deeplearning models
    End
Step6: Download global model final weight to all clients.
Step7: End
```

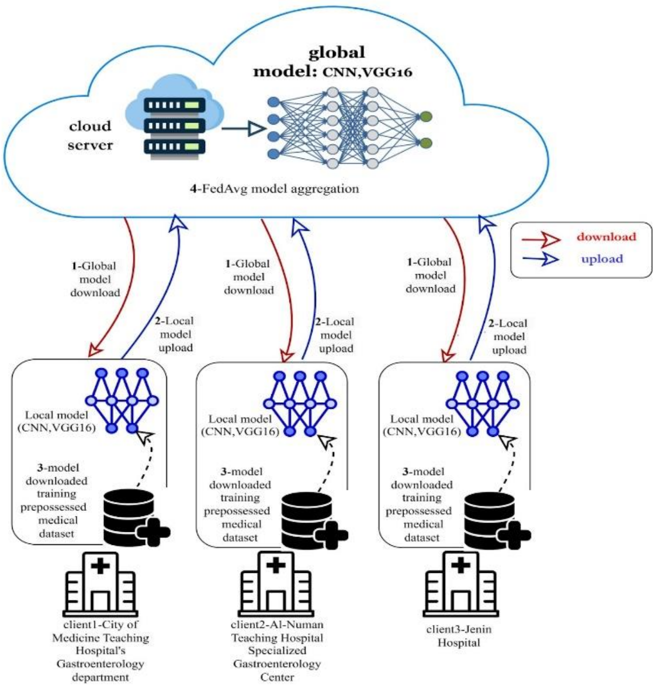


Figure 3. The proposed federated learning model

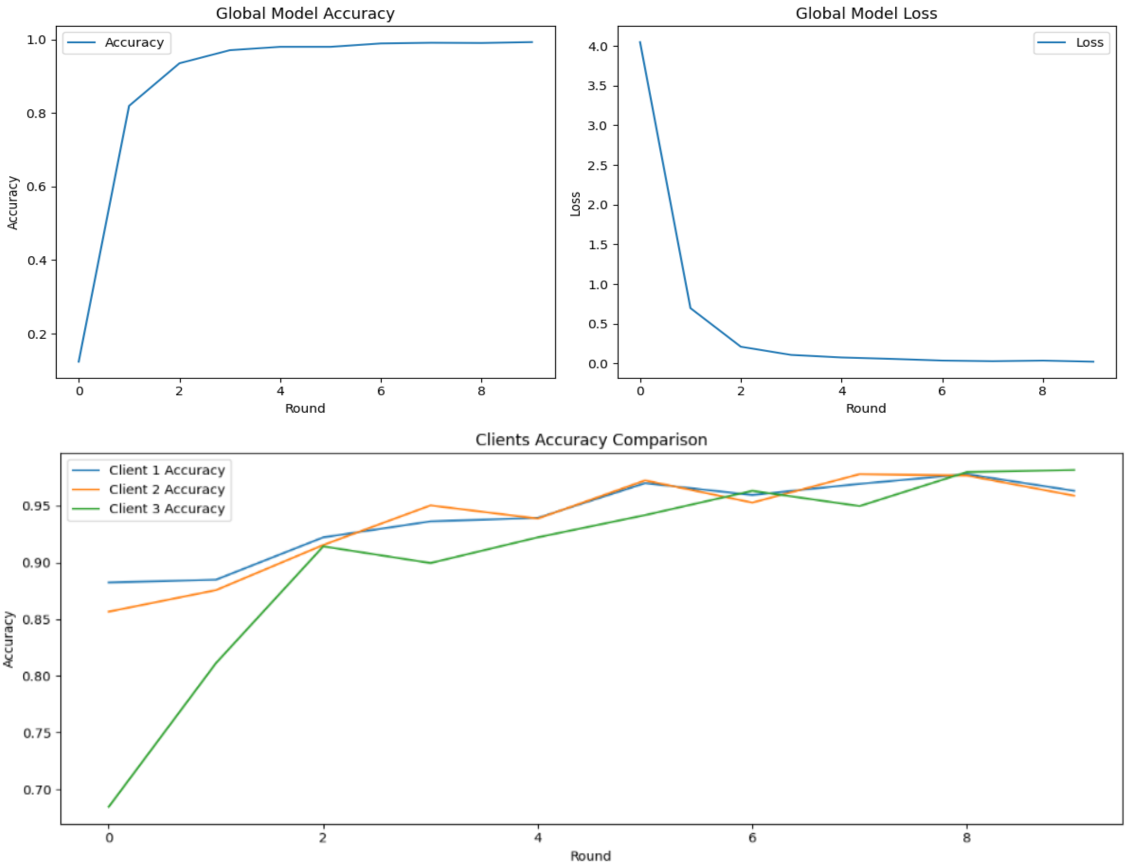
7. RESULTS AND DISCUSSION

The outcome of training and assessing model performance was covered in this section. To assess a classification model's performance, the local and global models' confusion matrix and classification report compare actual target values with predicted ones. Models without FL (local model) were categorized using sensitivity, specificity, accuracy, precision, recall, and Compared to the global model, the F1 score indicates a better prediction of more real data. Other metrics that are used to evaluate a classifier's effectiveness in machine learning include specificity and sensitivity. The proportion of positive items that our classifier correctly identified is known as sensitivity, while the percentage of negative objects that received the same classification is known as specificity. The results of training and testing are displayed in Table 3.

The local VGG16 model had a higher accuracy of 0.94% than the local CNN, with a weighted average of 0.94% over metrics. The F1 score for all classes was 0.984%, indicating a few of both false positives or false negatives. The VGG16 model was found to be a strong local model for all classes with even sensitivity and specificity, but limited weakness in identifying the 5Perforation class. The VGG16 had better performance than CNN models for most class predictions and was a good candidate for balanced classification.

Table 3. The results of training and testing of the proposed local and global model

Dataset	Models	Acc		Loss		Pre		Recall		F1-score	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
GB diseases Dataset	Local CNN	0.811	0.596	0.656	1.564	0.836	0.631	0.811	0.596	0.812	0.598
	Global FedAvg(CNN)	1.0	0.887	1.119	1.497	1.0	0.887	1.0	0.887	1.0	0.886
	Local VGG16	0.984	0.94	0.057	0.184	0.984	0.943	0.984	0.938	0.984	0.938
	Global Fedavg(VGG16)	1.0	0.99	0.0004	0.048	0.977	0.987	0.977	0.987	0.977	0.987



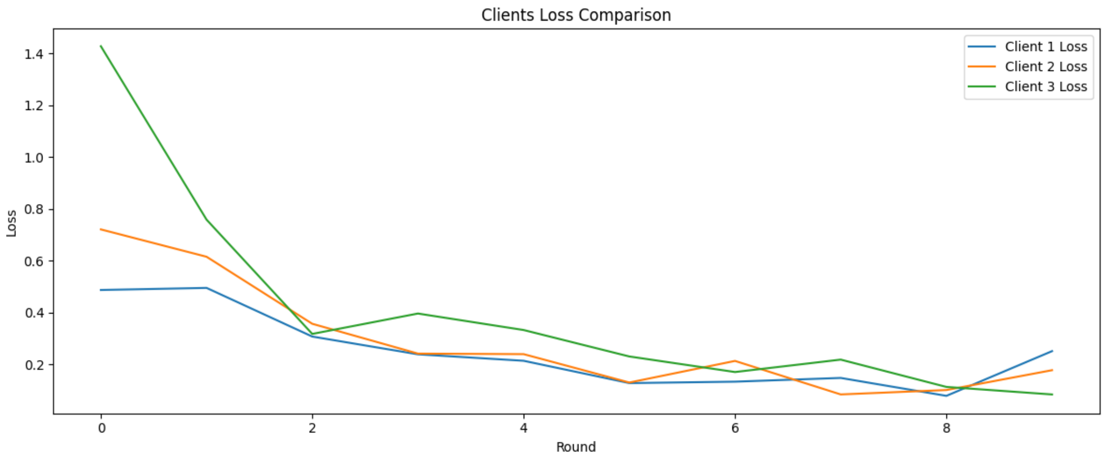


Figure 4. Accuracy and loss for the global FedAvg (VGG16) model and per each client

Table 4. Classification report

Class	Precision	Recall	F1 Score	Support
Cholecystitis	0.99	0.98	0.98	211
Carcinoma	1.00	0.97	0.98	211
Abdomen and retroperitoneum	0.99	0.99	0.99	181
Membranous and gangrenous cholecystitis	1.00	1.00	1.00	206
Perforation	0.99	1.00	0.99	190
Adenomyomatosis	1.00	1.00	1.00	202
Polyps and cholesterol crystals	0.95	1.00	0.98	217
Gallstones	0.99	0.97	0.98	214
Accuracy	0.99	0.99	0.99	1632
Macro avg	0.99	0.99	0.99	1632
Weighted avg	0.99	0.99	0.99	1632

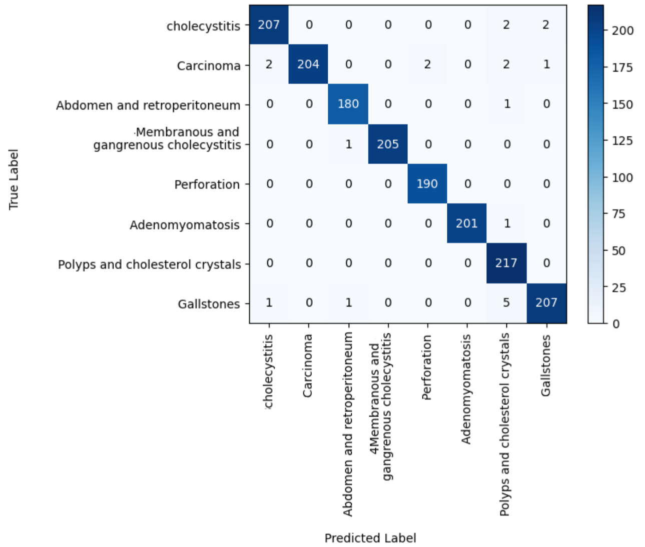


Figure 5. Multiclass confusion matrix (8 classes)

Federated Learning for the global model showed that the evaluation of the gallbladder diseases dataset revealed that most global models performed better than local models. The global FedAvg (VGG16) model achieved an accuracy of 0.997%, with a recall of 99.00%, F1-score and recall weighted average of around 98.00%, and precision weighted average of 99.00%. This model achieved a balance between avoiding false positives and identifying true positives. The global CNN model had an accuracy of 0.887%, and the global CNN model's weighted average reached 0.89%. Overall, the FedAvg aggregation method showed good performance in all global models compared to their non-federated counterparts. The Global FedAvg (VGG16) model was the most robust architecture, demonstrating perfect accuracy while maintaining privacy. Figure 4 shows the Accuracy and loss for the global FedAvg (VGG16) model and per each client after 20 rounds.

A confusion matrix is one method used to evaluate the system model's performance. When N is the number of classes, the resulting matrix is called a multiclass confusion matrix. When evaluating models that classify instances into more than two classes, this kind of confusion matrix is an extension of the confusion matrix. When dealing with problems that include more than two classes, this is quite beneficial. All of the matrix's components a class i instance's ci,j indicates how many times it was allocated to class j. The confusion matrix as a whole may have a comprehensive collection of metrics (accuracy (Acc), sensitivity (Sens), specificity (Spes), precision, and F1). The system model's performance is evaluated using a multiclass confusion matrix shown in Figure 5, with each column representing the anticipated label and each row representing the real label. Off-diagonal components indicate misclassifications, while diagonal elements indicate the number of correctly identified examples for each class, while Table 4 shows the classification report for each class.

The clinical significance of misclassifications is discussed, focusing on the impact of false positives and negatives. False positives for 'Perforation' can lead to unnecessary procedures, increased patient anxiety, and higher healthcare costs. For instance, False negatives for 'Perforation' can result in severe complications, sepsis, and even death. The model's recall for Perforation is 1.00 %, indicating no false negatives.

Figure 6 presents a comparison of model performance between the basic local learning models and global approaches of FL. Accuracy, precision, recall, and F1-score are the four assessment measures used in the comparison. Each of the models, CNN and VGG16, was tested both with and without the FL global model, allowing for a direct comparison of the effect of FL with FedAvg and pre-trained VGG16 model on system performance.

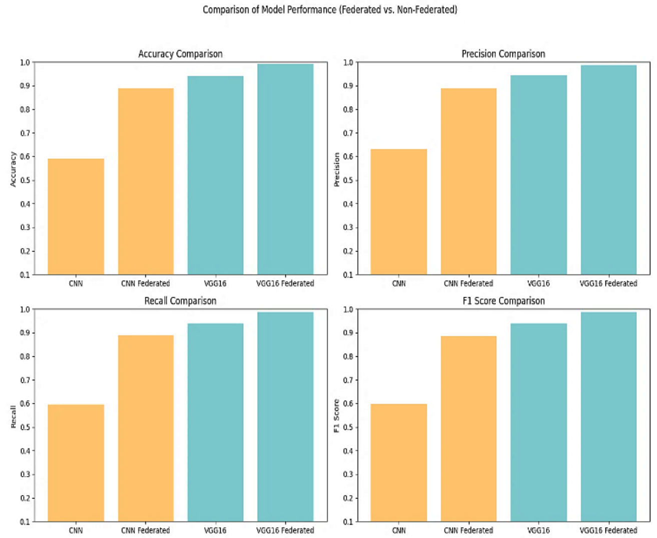


Figure 6. Comparison of model performance between the basic transfer learning and FL approaches

In the accuracy comparison (top-left), the VGG16 model with FL achieved the highest accuracy, reaching approximately 0.99%, outperforming all other models. The CNN accuracy is 0596%, and within the global FL model, it is improved to 0.887%. The precision comparison (top-right) also highlights the superiority of the VGG16 model with FL, achieving precision close to 0.987%. The recall comparison (bottom-left) illustrates the sensitivity of the models in identifying positive instances.

The VGG16 model with FL showed the highest recall, near 0.987%. Finally, the F1-score comparison (bottom-right) shows the balance between precision and recall. The VGG16 model with FL excelled with an F1 score of about 0.987%, followed by the CNN model. Figure 6 demonstrates that FL has a significant positive impact on the performance of the CNN VGG16 model, enhancing its accuracy, precision, recall, and F1 score.

8. CONCLUSION

This current study reveals that the use of a feature called FL significantly enhances the performance of deep learning models, particularly in medical image analysis. The VGG16 model, when paired with the FL global approach, achieved an impressive accuracy of 0.99%. This demonstrates FL's ability to provide collaborative model training across decentralized datasets, addressing privacy concerns and generating powerful diagnostic tools. The FL-enhanced VGG16 model is expected to be clinically adopted, enabling more reliable and widely applicable powered medical diagnosis solutions. The FL-enhanced VGG16 model's performance suggests a good probability of clinical adoption, enabling more reliable and widely applicable and powerful medical diagnosis solutions.

To make the models more practical, future research will look at how this FL approach may be used to a diversity of non-IID medical datasets from different diseases. Additionally, evaluating the efficacy of different optimizers within the FL framework may yield further performance improvements, and fusing blockchain technology with Federated Learning offers an attractive means of enhancing data provenance and ensuring even more robust privacy guarantees in future developments in medical image classification.

REFERENCES

[1] Sohan, M.F., Basalamah, A. (2023). A systematic review on federated learning in medical image analysis. IEEE Access, 11: 28628-28644. https://doi.org/10.1109/access.2023.3260027

[2] Jiang, X.Y., Hu, Z.J., Wang, S.H., Zhang, Y.D. (2023). Deep learning for medical image-based cancer diagnosis. Cancers, 15(14): 3608. https://doi.org/10.3390/cancers15143608

[3] Bozdag, A., Yildirim, M., Karaduman, M., Mutlu, H.B., Karaduman, G., Aksoy, A. (2025). Detection of gallbladder disease types using a feature engineering-based developed CBIR system. Diagnostics, 15(5): 552. https://doi.org/10.3390/diagnostics15050552

[4] Abood, R.H., Hamad, A.H. (2025). Multi-label diabetic retinopathy detection using transfer learning based convolutional neural network. Fusion: Practice and Applications, 17(2): 279-293. https://doi.org/10.54216/FPA.170221

[5] Abdalah, R.W., Abdulateef, O.F., Hamad, A.H. (2025). A predictive maintenance system based on industrial Internet of Things for multimachine multiclass using deep neural network. Journal Européen des Systèmes Automatisés, 58(2): 373-381. https://doi.org/10.18280/jesa.580218

[6] Obaid, M.H., Hamad, A.H. (2024). Internet of Things based oil pipeline spill detection system using deep learning and LAB colour algorithm. Iraqi Journal for Electrical and Electronic Engineering, 20(1): 137-148. https://doi.org/10.37917/ijeee.20.1.14

[7] Hosseini, S.M., Sikaroudi, M., Babaie, M., Tizhoosh, H.R. (2023). Proportionally fair hospital collaborations in federated learning of histopathology images. IEEE transactions on medical imaging, 42(7): 1982-1995. https://doi.org/10.1109/tmi.2023.3234450

[8] Joynab, N.S., Islam, M.N., Aliya, R.R., Hasan, A.R., Khan, N.I., Sarker, I.H. (2024). A federated learning aided system for classifying cervical cancer using pap-smear images. Informatics in Medicine Unlocked, 47: 101496. https://doi.org/10.1016/j.imu.2024.101496

[9] Wicaksana, J., Yan, Z.Q., Yang, X., Liu, Y., Fan, L.X., Cheng, K.T. (2022). Customized federated learning for multi-source decentralized medical image classification. IEEE Journal of Biomedical and Health Informatics, 26(11): 5596-5607. https://doi.org/10.1109/JBHI.2022.3198440

[10] Rehman, M.H.U., Hugo Lopez Pinaya, W., Nachev, P., Teo, J.T., Ourselin, S., Cardoso, M.J. (2023). Federated learning for medical imaging radiology. The British Journal of Radiology, 96(1150): 20220890. https://doi.org/10.1259/bjr.20220890

[11] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H.,

Raad, A. (2023). Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. Electronics, 12(10): 2287. https://doi.org/10.3390/electronics12102287

[12] Nazir, S., Kaleem, M. (2023). Federated learning for medical image analysis with deep neural networks. Diagnostics, 13(9): 1532. https://doi.org/10.3390/diagnostics13091532

[13] Shiri, I., Razeghi, B., Sadr, A.V., Amini, M., et al. (2023). Multi-institutional PET/CT image segmentation using federated deep transformer learning. Computer Methods and Programs in Biomedicine, 240: 107706. https://doi.org/10.1016/j.cmpb.2023.107706

[14] Luo, J., Wu, S.D. (2022). Fedsld: Federated learning with shared label distribution for medical image classification. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, pp. 1-5. https://doi.org/10.1109/isbi52829.2022.9761404

[15] Jiang, M.R., Wang, Z.R., Dou, Q. (2022). HarmoFL: Harmonizing local and global drifts in federated learning on heterogeneous medical images. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1): 1087-1095. https://doi.org/10.1609/aaai.v36i1.19993

[16] Zhang, M., Qu, L.Q., Singh, P., Kalpathy-Cramer, J., Rubin, D.L. (2022). SplitAVG: A heterogeneity-aware federated deep learning method for medical imaging. IEEE Journal of Biomedical and Health Informatics, 26(9): 4635-4644. https://doi.org/10.1109/JBHI.2022.3185956

[17] Babar, M., Qureshi, B., Koubaa, A. (2024). Investigating the impact of data heterogeneity on the performance of federated learning algorithm using medical imaging. Plos One, 19(5): e0302539. https://doi.org/10.1371/journal.pone.0302539

[18] Zhou, S., Landman, B. A., Huo, Y., Gokhale, A. (2022). Communication-efficient federated learning for multi-institutional medical image classification. In Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, San Diego, California, United States, Vol. 12037, pp. 6-12. https://doi.org/10.1117/12.2611654

[19] Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., et al. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1): 12598. https://doi.org/10.1038/s41598-020-69250-1

[20] Islam, M., Reza, M.T., Kaosar, M., Parvez, M.Z. (2023). Effectiveness of federated learning and CNN ensemble architectures for identifying brain tumors using MRI images. Neural Processing Letters, 55(4): 3779-3809. https://doi.org/10.1007/s11063-022-11014-1

[21] Srinivasu, P.N., Lakshmi, G.J., Narahari, S.C., Shafi, J., Choi, J., Ijaz, M.F. (2024). Enhancing medical image classification via federated learning and pre-trained model. Egyptian Informatics Journal, 27: 100530. https://doi.org/10.1016/j.eij.2024.100530

[22] Hossain, M.M., Islam, M.R., Ahamed, M.F., Ahsan, M., Haider, J. (2024). A collaborative federated learning framework for lung and colon cancer classifications. Technologies, 12(9): 151. https://doi.org/10.3390/technologies12090151

[23] Dayan, I., Roth, H.R., Zhong, A., Harouni, A., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10): 1735-1743. https://doi.org/10.1038/s41591-021-01506-3

[24] Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Medical Image Analysis, 65: 101765. https://doi.org/10.1016/j.media.2020.101765

[25] Adnan, M., Kalra, S., Cresswell, J.C., Taylor, G.W., Tizhoosh, H.R. (2022). Federated learning and differential privacy for medical image analysis. Scientific Reports, 12(1): 1953. https://doi.org/10.1038/s41598-022-05539-7

[26] Lin, L., Liu, Y.X., Wu, J.W., Cheng, P.J., Cai, Z.Y., Wong, K.K.Y., Tang, X.Y. (2024). FedLPPA: Learning personalized prompt and aggregation for federated weakly-supervised medical image segmentation. IEEE Transactions on Medical Imaging, 44(3): 1127-1139. https://doi.org/10.1109/TMI.2024.3483221

[27] Obaid, A.M., Turki, A., Bellaaj, H., Ksantini, M., AlTaee, A., Alaerjan, A. (2023). Detection of gallbladder disease types using deep learning: An informative medical method. Diagnostics, 13(10): 1744. https://doi.org/10.3390/diagnostics13101744

[28] Turki, A., Mahdi Obaid, A.M., Bellaaj, H., Ksantini, M., Altaee, A. (2024), Gallblader diseases dataset. Mendeley Data, V1. https://doi.org/10.17632/r6h24d2d3y.1

[29] Turki, A., Obaid, A.M., Bellaaj, H., Ksantini, M., AlTaee, A. (2024). UIdataGB: Multi-Class ultrasound images dataset for gallbladder disease detection. Data in Brief, 54: 110426. https://doi.org/10.1016/j.dib.2024.110426

[30] Bharti, S., Mcgibney, A. (2021). Privacy-aware resource sharing in cross-device federated model training for collaborative predictive maintenance. IEEE Access, 9: 120367-120379. https://doi.org/10.1109/ACCESS.2021.3108839

[31] Hernandez-Cruz, N., Saha, P., Sarker, M.M.K., Noble, J.A. (2024). Review of federated learning and machine learning-based methods for medical image analysis. Big Data and Cognitive Computing, 8(9): 99. https://doi.org/10.3390/bdcc8090099

[32] Liu, Z., Peng, J., Guo, X., Chen, S., Liu, L. (2024). Breast cancer classification method based on improved VGG16 using mammography images. Journal of Radiation Research and Applied Sciences, 17(2): 100885. https://doi.org/10.1016/j.jrras.2024.100885

[33] Liang, Y.P., Chen, Q.M., Zhu, G.X., Jiang, H., Eldar, Y.C., Cui, S.G. (2024). Communication-and-energy efficient over-the-air federated learning. IEEE Transactions on Wireless Communications, 24(1): 767-782. https://doi.org/10.1109/twc.2024.3501297

[34] Hung, N.N., Nguyen, T.T., Hoang, T.N., Pham, H.H., Nguyen, T.H., Le, N.P. (2025). SAFA: Handling sparse and scarce data in federated learning with accumulative learning. IEEE Transactions on Computers, 74(6): 1844-1856. https://doi.org/10.1109/TC.2025.3543682

[35] Efthymiadis, F., Karras, A., Karras, C., Sioutas, S. (2024). Advanced optimization techniques for federated learning on non-IID data. Future Internet, 16(10): 370. https://doi.org/10.3390/fi16100370