IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 February 2023, accepted 7 March 2023, date of publication 15 March 2023, date of current version 22 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257562

RESEARCH ARTICLE

A Transfer Learning Approach to Breast Cancer Classification in a Federated Learning Framework

Y. NGUYEN TAN[1], VO PHUC TINH[1], PHAM DUC LAM[2], NGUYEN HOANG NAM[3], AND TRAN ANH KHOA[3]

[1]Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam
[2]Faculty of Engineering and Technology, Nguyen Tat Thanh University, Ho Chi Minh City 700000, Vietnam
[3]Modeling Evolutionary Algorithms Simulation and Artificial Intelligence, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam

Corresponding author: Tran Anh Khoa (trananhkhoa@tdtu.edu.vn)

This work was supported by Funding no 37/qd-khcn/2023 of Nghe An Department of Science and Technology.

ABSTRACT Artificial intelligence (AI) technologies have seen strong development. Many applications now use AI to diagnose breast cancer. However, most new research has only been conducted in centralized learning (CL) environments, which entails the risk of privacy breaches. Moreover, the accurate identification and localization of lesions and tumor prediction using AI technologies is expected to increase patients' likelihood of survival. To address these difficulties, we developed a federated learning (FL) facility that extracts features from participating environments rather than a CL facility. This study's novel contributions include (i) the application of transfer learning to extract data features from the region of interest (ROI) in an image, which aims to enable careful pre-processing and data enhancement for data training purposes; (ii) the use of synthetic minority oversampling technique (SMOTE) to process data, which aims to more uniformly classify data and improve diagnostic prediction performance for diseases; (iii) the application of FeAvg-CNN + MobileNet in an FL framework to ensure customer privacy and personal security; and (iv) the presentation of experimental results from different deep learning, transfer learning and FL models with balanced and imbalanced mammography datasets, which demonstrate that our solution leads to much higher classification performance than other approaches and is viable for use in AI healthcare applications.

INDEX TERMS Artificial intelligence, synthetic minority oversampling, federated learning, transfer learning, breast cancer.

I. INTRODUCTION

According to statistics published by the International Agency for Research on Cancer in December 2020, breast cancer has overtaken lung cancer as the most diagnosed cancer worldwide [1]. Over the past two decades, the total number of people diagnosed with cancer has nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. Today, one in five people worldwide will develop cancer in their lifetime. It is estimated that the number of people diagnosed with cancer will further increase in the future: nearly 50% by 2040 compared to 2020. The number of people who die from cancer has also increased, from 6.2 million in 2000 to 10 million in 2020. Late diagnosis and a lack of access to treatment have become increasingly prevalent issues that require more attention and follow-up. Breast cancer is a malignant tumor of the breast. A tumor can be benign (non-cancerous) or malignant (cancerous). Most breast cancers begin in the milk ducts, with a small percentage of cases developing in the milk sacs or lobules. If detected and treated late, breast cancer may metastasize to the bones and other organs and the pain will multiply.

Therefore, early detection of breast cancer is critical for treating and saving patients. When the disease is in its early stages, its manifestations may not be accurate and precise; as a result, many abnormalities may be overlooked [2]. Currently, many studies apply machine learning to improve early detection, reduce the risk of death, and prolong the patient's life. However, sharing patient data is not widely considered at present due to privacy, technical, and legal issues. Security

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

and privacy techniques enable stricter protection of patient data and the use of data for research and routine clinical purposes [3], [4].

A study on breast mass classification from mammograms using convolutional neural networks (CNN) was published in 2016 [5], where the authors gave results with recall for identifying lesions estimated to be between 0.75 and 0.92, which means that up to 25% of abnormalities could remain undetected. Therefore, the ability to automatically detect lesions and predict their likelihood of malignancy would be valuable for doctors and could dramatically improve survival rates. Thus, we developed an FL base to extract features from multiple participating environments rather than a centralized learning environment. To investigate the real-world performance of FL, we conducted a study for the applied development of numerous breast cancer classification models using mammography data. An international group of hospitals and medical imaging centers joined this collaborative effort to train models in a completely decentralized fashion, without any data sharing between hospitals. This placed higher requirements on the robustness of algorithms and the selection of hyperparameters. In our study, we believed that the analysis of recall performance was more important than accuracy as false negatives can be life-threatening and false positives are likely to be viewed by humans in diagnosing breast cancer, and that is the main objective of this study.

Recently, FL has become a novel research trend in AI applications. It aims to train a machine learning (ML) algorithm across multiple decentralized nodes while holding the data samples (i.e., without locally exchanging them) [6]. Training such a decentralized model in an FL setup presented four main challenges: (i) system and data heterogeneity, (ii) pre-trained data processing, (iii) data protection and privacy, and (iv) efficiency selection of distributed ML algorithms. We addressed these challenges for breast cancer classification in the context of FL.

The first challenge was system and data heterogeneity. Different system vendors produce images with considerably different intensity profiles for the same imaging modality. To address this diversity, many recent studies have found that a data-balancing solution such as the unsupervised domain adaptation method forces the model to learn solution domain-agnostic features through adversarial learning [7] or a specific type of batch normalization [8]. However, more straightforward methods were used in the current study to address this challenge; we present a solution more efficiently balances data.

The second challenge is imperative to process the data before training because of its heterogeneity. There are many data processing methods [9], [10]. We chose transfer learning due to its many benefits, such as saving training time, better neural network performance (in most cases), and the fact that large amounts of data are not needed [11].

To address the third challenge, data protection and privacy [12], [13], many studies have incorporated more security and privacy solutions. Our solution assumes that an international group of hospitals and medical imaging centers have joined this collaborative effort to train the model in a completely decentralized manner, with no data sharing between hospitals. This places higher requirements on the robustness of algorithms and the selection of hyperparameters.

The fourth challenge concerned the distributed learning ability of the FL models [14], [15] employed. Many distributed learning models are used in FL for different applications. However, most studies focus on hypothetical data, and each model is only suitable for one dataset, which makes it difficult for researchers with practical applications as in breast cancer. To evaluate the effectiveness of these models, we tested the evaluation by other methods for comparison. The contributions of this paper are as follows:

- Design of an FL framework for breast cancer classification that includes a global server, which acts as a weight aggregator and mobile replacement edge clients in tissue training deep learning (DL). This solution is useful for AI healthcare applications and can be widely deployed in different hospitals or clinics.
- Pioneering use of a transfer learning pre-training dataset in FL for breast cancer classification. Various models in transfer learning were selected for performance evaluation, including k-nearest neighbors (kNN), AdaBoost, and eXtreme Gradient Boosting (XGB). First, the image's features are extracted using the Convolutional Neural Network (ConvNet) of the pre-trained model, and a linear classifier is used to classify the images. Next, we used data equalization techniques such as SMOTE and data augmentation in combination with ImageNet to enrich and further optimize the training data.
- With both balanced and imbalanced methods, experimental results from the Digital Database for Screening Mammography (DDSM) dataset demonstrate that our solution's FeAvg-CNN + MobileNet is much better for centralized learning, which is more than 5% recall [5] in improved performance. Moreover, the accuracy of our research results reached nearly 98%; by comparison, the maximum results were only 88.67% for the two-class cases (calcifications and masses) and 94.92% (benign mass vs. malignant mass and benign calcification vs. malignant calcification) in the study [6].

The rest of this study is organized as follows. Section II discusses related works; Section III describes the background study, and proposes together with some challenges and underlying ideas. Section IV presents an experimental evaluation of the deployed components. Finally, Section V discusses, and Section VI concludes the future work.

II. RELATED WORKS

A. DEEP LEARNING

Today, breast cancer research mainly focuses on detecting and diagnosing breast tumors using deep learning algorithms [3],

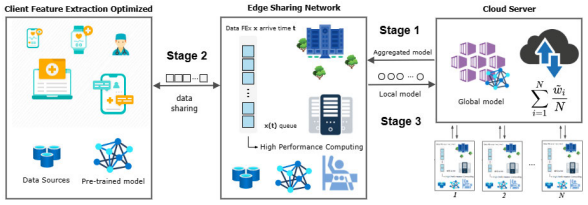

FIGURE 1. Architecture of the proposed approach federated learning settings.

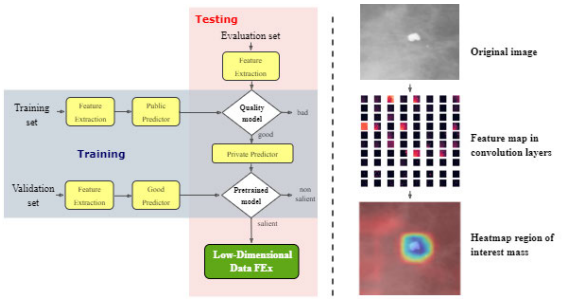

FIGURE 2. Flowchart of data processing for training and evaluation.

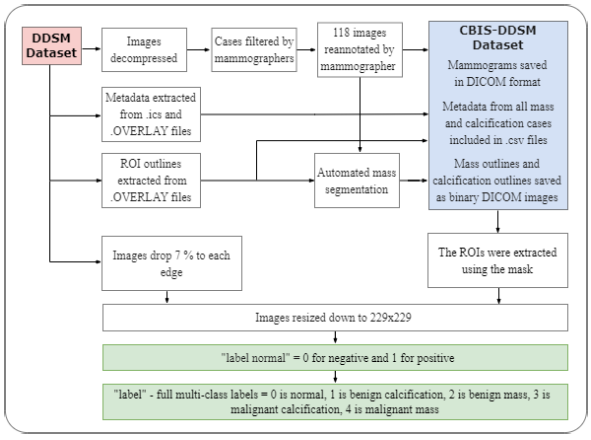

FIGURE 3. Diagram of DDSM data processing and conversion to images by ROI extraction.

[4], [16], [17], [18], [19]. However, most studies still focus on data processing, tumor prediction, and diagnosis using distributed learning models on a server. In such an environment, patient data and information must be shared for centralized processing on the server, which negatively affects privacy. Moreover, centralized processing entails more work when not all parties have access to a highly configured server.

A study [16] proposed a novel and efficient DL model based on transfer learning to automatically detect and diagnose breast cancer. The specific of this study is to use the knowledge gained while solving one problem in another relevant problem. Furthermore, in the proposed model, the features are extracted from a mammographic image analysis dataset (MIAS) using a pre-trained CNN such as InceptionV3, RestNet50, Visual Geometry Group Networks (VGG)-10, VGG-16, and Inception-V2 Restnet. To evaluate the experiments, six metrics were used and demonstrated that the transfer learning of the VGG16 model was powerful in classifying mammogram images with accuracy, sensitivity, specificity, and so on for breast cancer diagnosis. One study used a breast cancer–detection system that included principal component analysis (PCA), a multilayer perceptron (MLP), transfer learning, and a support vector machine (SVM) [17]. The authors proposed a new processing method for predicting breast cancer based on nine individual attributes and four basic machine learning methods; the final accuracy of the results was 86.97% on breast cancer coimbra dataset (BCCD). Research becomes complex when too many approaches are combined, and the results remain scarce.

Another study used CNN and US-ELM for feature extraction and clustering [3] and a mammogram segmented into several sub-regions. Then, CNN was used to extract features based on each sub-region, and unsupervised extreme learning machine (US-ELM) was employed to cluster features of sub-regions, which eventually located the region of the breast tumor. Next, the authors designed a CNN network with 20 in-depth features and other features to determine tumor density. However, the mammogram dataset only included approximately 400 women and yielded moderate accuracy.

By using DL to support the AdaBoost algorithm, paper [4] introduced an advanced technique for identifying and diagnosing breast cancer regions. Moreover, the study used the CNN network and LSTM algorithms to identify the characteristics of tumors for diagnostic tasks. The results demonstrated that the use of magnetic resonance imaging (MRI), ultrasound (US), digital breast tomosynthesis, and mammography yielded an accuracy that was too high: up to 97.2%. The previous section introduced challenges of data imbalance in breast cancer detection. A research article in [18] used a transfer learning solution to solve this issue. The primary model for breast cancer image classification is VGG-19. The results demonstrated that the accuracy was approximately 90%. The paper [19] introduced a framework for automatically evaluating areas of doubt detected in mammography screening without additional tests, especially in unnecessary biopsies, the suspected site is a benign tumor. It mainly focused on the identification of segmented regions of interest (ROIs) using a modified K-means algorithm. Next, a two-way experimental mode (BEMD) analysis algorithm was applied to extract multiple layers from the ROI. The results demonstrate that accuracy reached 98.6% when the digital mammography dataset was used.

B. FEDERATED LEARNING

Recently, FL was raised from the need to share sensitive data between service providers in various fields, such as competent healthcare and smart cities [20], [21], [22], [23], [24]. The results of some FL studies have been confirmed and applied to medical imaging, such as brain tumor segmentation, prediction of disease incidence, patients' responses to treatment and

other healthcare services, and late classification [25], [26], [27], [28], [29], [30].

Regarding breast imaging, only two papers [31], [32] have evaluated breast density classification. The authors employed a client server-based FL method with federated averaging (FedAvg) [7], which combines local stochastic gradient descent (SGD) on each site with a server that performs model averaging. However, [31] significantly downsampled the input mammograms. Although low resolutions are acceptable for density classification, the loss of detail negatively affects malignancy classification. Moreover, this study did not apply any domain adaptation techniques to compensate for the domain shift of different pixel intensity distributions. The authors [31] opted for a different approach by working on high-resolution mammograms with federated domain adversarial learning [32]. In addition, they [32] applied curriculum learning in FL to boost classification performance while improving domain alignment and explicitly handling domain shift with federated adversarial domain adaptation. The paper employs three datasets of Full Field Digital Mammography (FFDM), and the experimental results shown that the proposed memory-aware curriculum method is beneficial to further improve classification performance.

Based on the FL framework combined with CNN, the paper in [33] used CNN's federated prediction model is based on improvements in general modeling and simulation conditions on five types of cancer, the accuracy of cancer data reaches more than 90%, the accuracy is better than the tree model single model machines and linear models and neural networks. However, this study still lacked comparisons with different models rather than only MLP and did not address the issue of data imbalance and treatment.

In 2022, there is a growing trend toward using FL to predict breast cancer. Another study [34] used the Breast Cancer Histopathological Image Classification dataset (BHI) for detection. The authors first used residual neural networks for automatic feature extraction, then employed the network of Gabor kernels to extract another set of features from the dataset. They extracted two sets of features and passed the output through a custom classifier. The results showed that this method achieved more than 80% accuracy.

Unlike previous authors [31], [32], [33], [34], who proposed a FL framework for breast density classification based on deep learning models, we targeted the more complex task of breast cancer prediction based on the mammography dataset. The innovation of the current research is that we focused on accurately processing data using transfer learning. Because the data was robust, the classification results were expected to improve the treatment rate for patients. Many different models evaluated experimental results, and the most suitable model for breast cancer was selected based on the DDSM dataset. In addition to performing the results in terms of accuracy, recall, and higher F1-score compared to previous methods [5], we also analyzed diagnostic images based on histograms to enable doctors and medical staff to have a clearer view based on the displayed images. Finally, we summarize and compare our proposed method with the existing literature in Table 1.

III. THE PROPOSED METHOD

This section highlights our proposed approach in the FL framework. First, its overall structure is presented in Sub-section III-A. Next, Sub-section III-B describes he use of transfer learning for data feature export. Next, in subsection III-C, we introduce the two mammography datasets used in this study and how they were processed to improve classification quality. The FedAvg algorithm is introduced in Sub-section III-D, which we summarize in pseudocode to explain the implementation of the FL framework. Next, section IV, we evaluate the results and provide explanations.

A. AN OVERALL ARCHITECTURE OF THE PROPOSED METHOD

The current sub-section III-A presents a complete system model, including an overview diagram of how DL models work in FL and simulation designs. Fig. 1 depicts how the general behavior of the federated model was tested. The model structure includes a global server that acts as a weight aggregator and edge stations that replace mobile devices in training the deep learning model. The FL process occurs in three stages: *stage (1)* priming the initial model in the first round of FL or updating the new model after aggregating weights after the N^{th} round of learning, *stage (2)* local training with terminal data at the edge stations, and *stage (3)* aggregating the weights to the server and updating the global model. Taking advantage of transfer learning in the local environment of edge stations, here are hospitals that connect locally to machine learning-approaching technology devices using models that optimize performance, traffic conditions, etc. Homogeneous communication does not impose data labels on devices for prediction. However, it only uses personalized data features to learn, limiting transmission weights, limiting computation on edges, helping local training at the edge take place rapidly in reinforcement learning, back-distribution, cross-linking increasing data on the edge device sent to the edge increases over time.

The objective of the current study was to test the FL models' distributed learning ability. Thus, the server and edge devices were simulated by initializing a similar DL model in both the global and local phases. Therefore, the weight update between the host and the edge device was also directly be performed and ignored transmission time in the network.

At the beginning of the FL process, the starting server initiated a DL model. It sent the newly initialized set of weights to participating stations to create the first round of FL *(cloud server)*. The local model updated this set of weights and began the training process. At the input of the local model, data is the data that is optimally learned from pre-trained modeling with IMAGENET [35] on incoming edge devices (phones, computers, doctors, medical devices, etc.) *(client feature extraction)*. Due to the limitation in simulation, local learning occurred by training each edge with

TABLE 1. Summary of literature paper used for breast cancer classification.

Ref	Model	Contribution	Task performance	Dataset	Classes	Metric performance
[5]	Transfer learning and CNN	Directly classify pre-segmented breast masses	Breast mass classification	DDSM	Benign, Malignant	Accuracy, precision, and recall
[6]	CNN	Design a novel CNN model	Breast cancer diagnosis	DDSM (CBIS)	Benign, Malignant, and Calcification	Accuracy
[16]	Transfer learning (Inception-V3, Resnet 50, VGG-16/19, and Inception-V2 ResNet)	Reduce time training/ Improve the affected areas detection/ Improve the classification performance	Classification of normal and abnormal tissues in DDSM images	DDSM (MIAS)	Benign, Malignant, Normal	Accuracy, sensitivity, specificity, precision, F-score, AUC, and 10-Fold cross validation
[17]	Transfer learning techniques using SVM	The methods can effectively and powerfully examine the incidence of breast cancer/ Combined the PCA, MLP, and SVM to construct the learning algorithm	Predict breast cancer on the basis of nine individual attributes	BCCD	Benin, Malignant	Accuracy, and 10-fold cross validation
[3]	CNN, ELM	Feature fusion with convolutional neural network (CNN) deep features and unsupervised extreme learning machine (ELM) clustering	Mass detection and breast cancer classification	400 female mammograms	Benin, Malignant	Accuracy, sensitivity, specificity, TPRatio, TNRatio, AUC
[4]	Adaboost, CNN and LSTM	Combine these machine-learning approaches with the methods of selecting features and extracting them through evaluating their output using classification and segmentation techniques to find the most appropriate approach	Breast cancer detection, classification and segmentation	MRI, US, Tomosynthesis and Mammography	Benign, Malignant, Normal	Accuracy, and sensitivity, specificity
[18]	Transfer learning techniques and VGG-1	Transfer learning to improve accuracy and focus our efforts on histopathological and imbalanced image classification	Imbalanced class classification	277,524 images in RGB format	Balanced and Imbalanced	Accuracy, sensitivity, specificity, G-means, and F1-score
[19]	Deep multiple instance learning	Automate assessment of suspicious regions, detected in screening mammography	Breast cancer classification	DDSM (MIAS)	Benin, Malignant	AUC/ ROC
[31]	FL with DenseNet-121	Use of FL to build medical imaging classification models	Breast density classification	BI-RADS	Fatty, scattered, and heterogeneously, extremely denses	Kappa score
[32]	FL and curriculum learning	Curriculum Learning in FL to boost the classification/ Combined with unsupervised domain adaptation to deal with domain shift while preserving data privacy	Breast cancer classification	FFDM	Benin, Malignant	AUC/ ROC
[34]	FL, Resnet 18/50	Two input modalities obtained separately by extracting histopathology image features and pre-train/ Automate diagnosis	Breast cancer classification	BHI	IC-NST-negative and positive	Accuracy, F1-score, Precision, Recall, and specificity
Our proposed	Federated Learning (MLP, CNN), Transfer learning (MobileNet, Densenet121, Xception, Resnet50)	Propose a lightly DL model for FL framework, Reduce time training/ Improve the affected areas detection/ Improve the classification performance	Breast cancer classification	DDSM (CBIS)	Normal, Benign calcification, Benign mass, Malignant, Malignant calcification, Malignant mass	Accuracy, AUC, Precision, Recall, F1-score, Heatmap, and Accuracy for non and 4-Fold cross validation

a piece of data from the processed dataset and grouping it to mark the order of the participating stations' FL network *(edge sharing)*. Therefore, updating and training for the entire edge station were sequentially performed until the number of data groups in the divided dataset was exhausted instead of simultaneous parallel learning on all devices, as in real situations. In addition, the decentralized exchange sequential sharing system mechanism has strict privacy conditions.

During the local training phase, the pre-split data will be trained with the local model. The learning process is akin to a regular DL network training process, which includes the following steps: fitting, forward propagation, and backward propagation. Local training is completed only after all stations have concluded training with their data. The weights of all stations will be aggregated for the global model update according to the expression of the FedAvg algorithm. In addition to local weights, the number of data points in the data group used for training at each station must be collected to perform the aggregation. Therefore, in addition to storing weights, each edge station also calculates and retains the number of data points that it has trained to prepare for the synthesis process at the server. The entire local learning process at the edge stations is performed with DL models. Specifically, the network model FedAvg-ANN (MLP) and FedAvg-CNN and machine learning models kNN, AdaBoost, and XGB were used in this study.

After the local training is complete, the set of weights and data points for all stations is updated on the server. Currently, the server plays the role of aggregator and runs the algorithm with the set of weights and number of data points from the stations to find a new set of weights for the global model. Once the global model has been updated, the server checks the results of the FL round by running the classification problem on the test data, saves the test results, and moves on to the next learning round. The entire FL process is run with N given learning rounds. At each edge station, the number of epochs (DL cycles) is also performed with n pre-selected cycles. The number of FL rounds selected for the simulation is 200. The timing and predictive power of the model both depend on the number of epochs performed at each station. To check for relativity, the model was sequentially tested with one, two, and five epochs in the first 200 FL rounds. Both cases will be tested with only a DL model applied in FL.

In this study, the models selected for FL were all DL networks with shared parameters in the training process. For backpropagation, the categorical cross-entropy loss function and the SGD optimizer were used to improve the training of all models. During forward propagation, the ReLU function was used to activate hidden layer neurons, while the Softmax function was used for predictive decision-making at the output layer.

FL algorithms are fundamentally different and mainly solve the problem of data security (i.e., the connection between security hospitals). In a traditional data science workflow, data is collected on a single server and used to build and train a centralized model. FL has a centralized model that uses training under a decentralized model. Once the model is independently trained, each of the updated model weights is sent back to a central server, which combines them to create a highly efficient model. This also ensures that the data in each node complies with data security policies and protects against any data leaks or breaches. Quality data locally exists on edge devices in hospital centers around the globe and is protected by strict privacy laws. FL provides an intelligent means of connecting machine learning models to this discrete data regardless of location and, more importantly, without breaking privacy laws. Instead of taking the data to the model for general rule training, FL feeds the model to the data instead. All that is needed is the flexibility of the data storage device to commit itself to the binding process.

B. TRANSFER LEARNING FOR FEATURE EXTRACTION

In a hypothetical problem, 1,000 patients must be identified, but the data for the training consists of only approximately four images per person. Thus, there is insufficient data to train a complete machine learning model. In such cases, the models are usually pre-trained with extensive data from sources such as ImageNet, which contains 1.2 million images and 1,000 different categories. Then, in this study, the feature extractor solution was used. After extracting the features of the images using ConvNet of the pre-trained model, we used a linear classifier to classify the images. In short, the image features (e.g., calcifications, tumors, etc.) provided input for linear or logistic regressions. The overview had three feature extraction methods, which resulted in three different results for each classifier [36]:

- Histograms of oriented gradients (HOGs)
- Features extracted from the discrete cosine (DCT) domain
- Features extracted from a pre-trained CNN.

Different feature extraction methods have different advantages. HOGs are commonly used for object detection and employ gradients to provide information about edges, corners, and contents in an image. However, the decision was made to extract features from the DCT domain because the features were created to describe quality parameters in an image. The last method uses features extracted from a CNN; the network is trained on a large set of images in the object recognition task to enable it to be generalized to the tasks and another dataset for which the network has not been trained. This method was chosen because of its high performance on common tasks in organized learning. The task in Sub-section III-B requires a more detailed representation of the compact and efficient local communication of input values through an unsupervised learning scheme with transfer from the large ImageNet (transfer) by mapping the input to a selective latent through a ConvNet network of pre-trained models whose predictive output we need. The representative and symbolic learned output feature is usually low-dimensional (called FEx) and contains all of the necessary information at the calcification region, as indicated by the heatmap, and can

therefore be used as a calculation vector input function for a supervised learning model (e.g., MLP or CNN) to perform FL at the edge server. It is noteworthy that the features are learned by compressing useful information of local input data into the low-dimensional ConvNet network (MobileNet, ResNet50, Xception, etc.) from developers providing the network structure with reference open handover, always learned and maintained by local users without sending to the cloud. Moreover, the cloud server only aggregates the updated model parameters obtained by performing calculations at the edge server based on the global model and local data from users, patients, medical devices, etc., without the user's sensitive data access rights (e.g., data samples, representative low-dimensional features) to protect user privacy in federated communication and, optimally, in local sampling dispersion.

The left portion of Fig. 2 shows how the system is trained on the input set, which results in two classification models: one for quality and one for content. The reviews are sorted by rank, and images that are classified as outstanding are sent to the retrieval section. In partial retrieval, selection is made from sets of similar images, such that that only one is retrieved. The resulting images are good, outstanding, and unique; we perform further analysis by retraining the classification model to provide the best result (e.g., the right image is the result of the whole set of images' progress).

C. DATASETS

DL heavily relies on datasets to automatically extract features that uniquely characterize the various target classes. In our study, we used the digital database of mammography screenings from the University of South Florida. These datasets and step processes are described in the following sub-section III-C.

1) DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY (DDSM) DATASETS

DDSM is a database of 2,620 scanned film mammography studies [37], [38], [39], [41]. It includes normal, benign, and malignant cases with verified pathology information. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) collection includes a subset of DDSM data selected and curated by a trained mammologist. Since CBIS-DDSM only contains abnormal images, conventional scans were obtained from DDSM and combined with CBIS-DDSM scans. However, the size image is relatively small. To increase the size of the dataset, we extracted ROIs from each image, the algorithm is proposed in Algorithm 1. We aimed to classify the calcifications in which stage of the disease, so we extracted information from several available sources and introduced previous work. Previous studies in [42] did a great job extracting ROI features from a combined DDSM and CBIS-DDSM data source. However, our study is only selective regarding the number and size of images to be suitable for transfer tasks and federated for edge device environments. The ConvNets were trained to predict whether the scan was normal or abnormal and whether the abnormalities were calcifications or masses and benign or malignant. DDSM provides metadata in three files that include the patient's age, the study date, the date of digitization, the type of dense tissue, the scanner used to digitize, and the resolution of each image. In addition, cases with anomalies include an OVERLAY file that contains information about each abnormality, including the type of abnormality (mass or calcification). The entire data processing for training and evaluation is described in detail in Fig. 3 and the subsequent sub-sections III-C2.

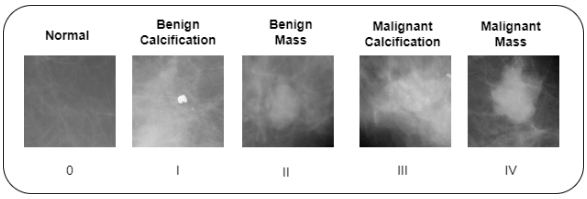

FIGURE 4. Illustrations of the calcification status of layers.

2) DATA PRE-PROCESSING

The dataset used in the study included images from the DDSM and CBIS-DDSM datasets. Images were pre-processed and converted to 299 × 299 images by extracting the ROIs. The data was stored as a `tfrecords` file for TensorFlow. The dataset contained 55,890 training examples, of which 86% were negative, and the remaining 14% were positive; they were divided into five `tfrecords` [42]. The data was also divided into training and testing in the CBIS-DDSM dataset. The test files were equally divided into test and validation data. However, the separation between the test and validation data was incorrectly performed, which resulted in the non-test files containing only volumes and the validation files containing only calcifications. The dataset consists of negative images from the DDSM dataset and positive images from the CBIS-DDSM dataset. The data was pre-processed to convert it into 299 × 299 images. The negative (DDSM) images were tiled into 598 × 598 tiles, which were then resized to 299 × 299. The positive (CBIS-DDSM) images had their ROIs extracted using the masks with a small amount of padding to provide context. Each ROI was then randomly cropped three times into 598 × 598 images, with random flips and rotations, and then the images were resized down to 299 × 299. These files should be combined for complete and balanced test data. The images were labelled in two ways: i) normal label: 0 for negative and 1 for positive; ii) full multi-class labels: 0 for normal, 1 for benign calcification, 2 for benign mass, 3 for malignant calcification, and 4 for malignant mass.

As previous work addressed the classification of predefined abnormalities, we focused on classifying images as normal or abnormal. We expected to retrain the model to classify for the whole five labels after achieving satisfactory performance illustrated through each stage. Fig. 4 and Table 2 summarize breast cancer classification stages in the DDSM

TABLE 2. Breast cancer stages in the DDSM dataset.

Category	Code	Cancer	Description
Negative	NO	NOrmal	None
Positive	BEC	BEnign Calcification	Stage I
Positive	BEM	BEnign Mass	Stage II
Positive	MAC	MAlignant Calcification	Stage III
Positive	MAM	MAlignant Mass	Stage IV

dataset. Based on the analysis results from Fig. 4 and Table 2, doctors can diagnose which stage a patient is in and can provide them with helpful advice.

3) DATA AUGMENTATION

CBIS-DDSM scans are relatively large, with an average height of 5,295 pixels and an average width of 3,131 pixels. To create usable images from full-sized scans, ROIs were extracted using a mask and sized down to 299 × 299 pixels. Each ROI was extracted in several ways:

- The ROI was extracted at the initial size of 598 × 598
- The ROI was zoomed to 598 × 598, with borders to provide context
- If the ROI was too large to fit into a 598 × 598 image, it was extracted in 598 × 598 cells with a spacing of 299

The 598 × 598 pixels were then resized to 299 × 299. To increase data samples, dataset data augmentation was used, including random positioning of the ROI in the image, random horizontal flip, random vertical flip, and random rotation. The extraction of the ROI is detailed below. Since the CBIS-DDSM dataset only contained anomalous scans, normal scans were obtained from the DDSM dataset. While the CBIS-DDSM images were reviewed and changed to remove pseudo-elements such as white borders and overlay text, the DDSM images were absent. Many variable-size contours and arrays of white color were used to obscure the patient's personal information. To remove contours, DDSM images were cropped by 7% on each side. Since the extracted CBIS-DDSM ROIs were proportional to their size instead of at a fixed zoom, the DDSM images were scaled down by a random factor of between 1.8 and 3.2, then segmented into 299 × 299 pixels with spacing between 150 and 200 pixels.

4) HISTOGRAM

It was necessary to collect different datapoints from the DDSM dataset to evaluate quality indicators because the data constantly fluctuated. Consider those randomly obtained data, it will be challenging to appreciate the whole meaning of the information they bring, and it is difficult to identify their fluctuations. To analyze and evaluate the quality situation from the collected data to draw accurate conclusions, people gather, classify, and rearrange them to represent the distribution in the form of charts. Different pixel densities (histograms) according to the characteristics of the data obtained. Based on the form of a frequency distribution by the graph, one can have accurate conclusions about the normal or abnormal situation of the quality criteria of the process.

Algorithm 1 ROI Extraction Algorithm

Input: `Slide_size` = 299; `Full_slide_size` = 598; `offset` = 60.

Get the Base File Name and the Mask Name.

Add the ROI ← **Preprocessing**

if *mask_size* ≤ *(full_slice_offset)* **then**
  image_slice = *image*[***ROI_edges***]
end if
if *mask_size* ≤ *(full_slice) / 1.5* **then**
  roi_size = *mask_size* + 20%
  image_slice = *Random_flip* and rotate(image[***ROI_edges***])
end if
if *mask_size* > *(full_slice)* **then**
  roi_size = *mask_size* + 5%
  image_slice = *Random_flip and rotate*(*image*[***ROI_edges***])
end if
return image_slice[*slice_size*]
Preprocessing(image,mask)
image ← *resize(image)* = $\frac{image}{2}$
mask ← *resize(mask)* = *image.size*
if *image* > *50,000 white pixels* **then**
  image ← *image* − *trimming edges (image)* = *20 pixels*
end if
(*center_row*, *center_col*) ← *corners(mask)*
return center_row, *center_col*, *mask*
mask_size = *int(max(mask.shape[0], mask.shape [1]))*
return center_row, center_col, mask_size
ROI_edges(*center_col, center_row, image.shape, roi_size*)
return (start_row, end_row, start_col, end_col)

From there, appropriate decisions can be made to improve the quality of the data input. Fig. 5 shows the density and pixel intensity distribution for different stages of breast cancer of DDSM dataset used in the study. The vertical axis represents the number of pixels; the higher the vertices (e.g., the label "Malignant Calcification"), the more pixels there are in that area and the greater the detail. The horizontal axis represents the average brightness of each area, which means a new color similar to gray at 18%. The image labeled "Normal" occupies the dark space because it does not contain more calcified areas (white pixels) than the layers labeled positive for breast cancer. The origin of 0 is considered the darkest (akin to black); values increase as they move to the right, and the lightest value is 255. The area between these two values represents medium brightness. Thus, the closer pixels are to a value of 0, the darker the image; conversely, the closer they are to a value of 255, the brighter the image. Pixels on the vertical column of either value will lose detail (either too dark or too light). A bright and clear image has a bell-shaped

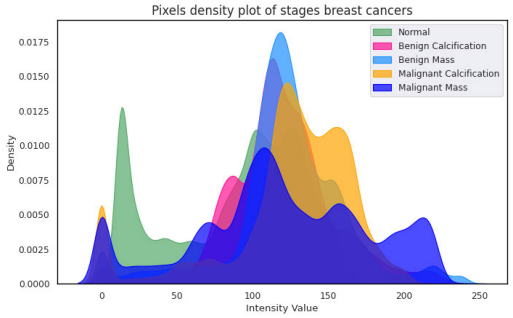

FIGURE 5. Density and pixel intensity distribution for different stages of breast cancer.

histogram, peaks in the region of medium brightness, and tapers off to the left and right regions of the graph.

5) SMOTE

From the fully labeled dataset, data balancing was performed using the SMOTE method [40]. In an unbalanced dataset, a different number of input samples represents each output layer (or target layer). The resulting classifier loss will often not be the high fact that the data is always non-IID. However, the scenario that solves this problem is quite common nowadays. Some current methods apply unbalanced data pre-processing:

- Oversampling: Oversampling involves increasing the number of samples of the smallest class up to the number of samples of the largest class and creating composite templates
- Undersampling: Undersampling involves reducing the number of samples of the largest class to the smallest class size and removing some samples from the largest class
- Class weight: This method involves specifying a weight for each class. The weight of the largest class is equal to 1, while the weight of the smallest class is equal to the largest class's sample divided by the smallest class's sample
- Decision threshold: If the predicted value is greater than the threshold, it is set to 1; otherwise, it is set to 0.

Because of the unbalanced input data, one way to address unbalanced datasets is to oversample the minority. The most straightforward approach is to duplicate examples in the minority class. However, these examples do not add any new information to the model. Instead, recent examples can be synthesized from existing measures. This data augmentation type for the minority class is called the synthetic minority oversampling technique, abbreviated as SMOTE. Since the Fig. 6 and 7 show that the class distribution predictive breast cancer before and after SMOTE use.

6) DATA SPLITING

Only around 10% of mammogram images are abnormal. To maximize the likelihood, we evaluated our training data more towards the abnormal scan direction (balance the

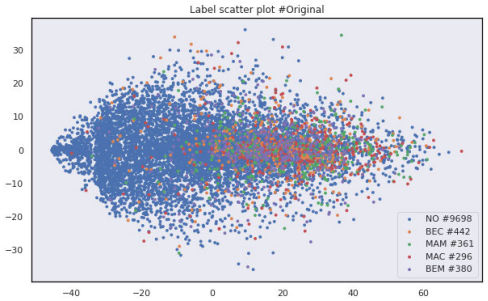

FIGURE 6. Class distribution predictive breast cancer data before SMOTE use.

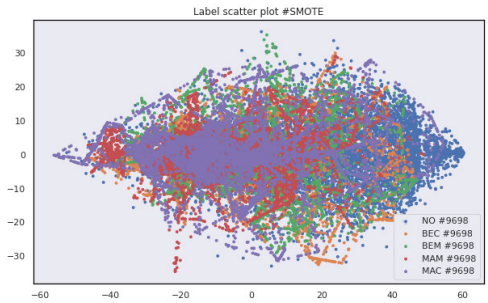

FIGURE 7. Class distribution predictive breast cancer data after SMOTE use.

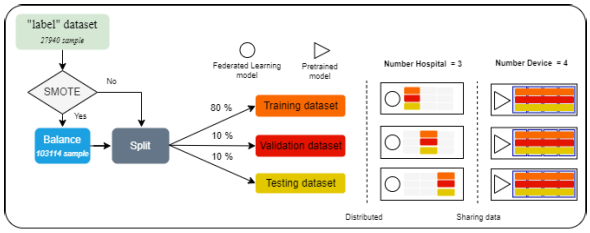

FIGURE 8. Data division process with the DDSM dataset.

abnormal minority class with normal), with a goal of 85% normal. The data was separated randomly into training and test data using existing parts of the CBIS-DDSM dataset to avoid overlap. By using the scikit-learning machine learning library, which provides the `train_test_split()` function to perform train test split. The overall data was divided randomly into training, validation, and testing data according to the following percentages: 80%, 10%, and 10%, respectively.

D. FEDAVG ALGORITHMS

The primary algorithms 2, 3, and 4, which were used to simulate the learning process in the FL model, is introduced in the sub-section III-D. The outermost for loop loops through a given number of learning loops, each being a global learning loop. In each of these loops, the program performs global parameter initialization in the first step. The feature extraction task performs the following steps for the client side: loading the optimal weighted model from ImageNet corresponding to breast cancer data, randomly partitioning the data into packets for direct prediction (i.e., taking advantage of the

model's convolutional layers to extract features), packing the data into FE_x, and sending it to the edge. There will be two processes at the edge server. The first is training for global communication between edges, while the other is local feature extraction of data from clients and linking. The model will get weight with the general model (ANN-FedAvg, CNN-FedAvg). Client updates pretrained data with ImageNet, and extracted will be reconnected. Then, edge server will train with the total data obtained to generate local weight. This is a global update algorithm that could sever, initialize weight w_0, then send to edge-sever according to the preset round-robin schedule. On the edge-server side will update the weight later in the training process at communication loops. After getting the weight list of all participating edges, the model will aggregate and average. Then update the weights again for the next round.

Algorithm 2 Local Pretrained-Model Update Feature Extracted Low-Dimensional Dataset.
M_{sub} Are Selected Model From Edge Sever;
C Client With Index k in Edge Computing Area;
B Is the Edge Batch Size From Edge Sever Request;
P_{sub} Data Point Each Node Be Divided From Edge

ClientUpdate(k, FE_x)
$M_{sub}(w_{optimized})$ ← initialize(net) from IMAGENET
/* feature extractor */
B_k ← (split P_{sub} into batches of size B_k
for *batch b in* B_k **do**
  images ← uniformly random sample b images
  X, y ← preprocess(images)
  $z \leftarrow predict(net, X) \leftarrow M_{sub}(w_{optimized})$
  $l \leftarrow loss(z, y)$
  FE_x ← *update*[*flatten*(z), *backward*(l)]
end for
return FE_x to edge server

Algorithm 3 Collaborative Local Update in Client-Edge. There Are M Edge Servers Model Are Indexed by m;
B Is the Edge Batch Size From Cloud Sever Request;
E Is the Number of Client Epochs;
P Feature Extracted Dataset From Edge Type 1-Dimensional and η Is the Learning Rate Into Each Edge Rounds

EdgeSeverUpdate(m, w, P)
received $M(W_{round})$ from global model of cloud sever and initialize to edge model
$M(W_{round}) \leftarrow initialize(net)$
for *each edge model* $m \in M$ *in parallel* **do**
  $P^m_{t+1} \leftarrow EdgeConcatenate(m, \text{FE\_x}, \sum_c k) \leftarrow ClientUpdate(k, \text{FE\_x})$ /* Pretrained & arrange global model */
  for *each edge epoch from* 1 *to* E **do**
    for *batch* $b \in B$ **do**
      $w^t_m \leftarrow \eta \nabla_l (M(w_{round}; b))$
    end for
  end for
end for
return w_m and P_m to cloud server

Algorithm 4 Global Aggregation. There Are N Edge Sever Are Indexed by i;
B Is the Edge Batch Size;
E Is the Number of Edge Epochs;
P Data Point Each Edge Held

for *each round* $t = 1, 2 \ldots$ **do**
  updated w_t send to all N edge sever each round
end for
for *each edge* $i \in N$ *in parallel* **do**
  $w^i_{t+1} \leftarrow EdgeSeverUpdate(i, w_t, P^i_t)$
  $w_{t+1} \leftarrow \sum_{i=1}^{N} \frac{\tilde{w}_l}{N}$ /* global aggregation based on parameters w_{t+1} */
end for

IV. EXPERIMENTAL VALIDATION

The current section evaluates the effectiveness of transfer learning using popular models such as MobileNet, Densenet21, Xception, and Resnet50 with two DDSM datasets on a FL framework and compares it to previous methods of breast cancer identification, including FL-CNN and FL-MLP. For fair comparison, all models are implemented in Python 3.8 with Tensorflow 2.9, and Anaconda3. The experiments were run in the machine with following specifications: GPU NVIDIA TESLA P100 2vCPU RAM 256 GB. Furthermore, in all simulations, we used the same settings: rounds = 60, edges = 3, clients = 4, frequency = 5, epochs = 3, batches = 32, length = 27,940.

A. PERFORMANCE TO COMPARE THE QUALITY CLASSIFICATION FOR FL AND CENTRALIZDED LEARNING

Since training on the dataset is prolonged, the period models choose three epochs with binary labels. The metrics evaluation as accuracy, precision, recall, and F1-score and the area under the curve (AUC) are used to evaluate the models. After a well-performing model is extracted on the shared balance dataset, three epochs with the global model will be retrained, with the previous weights reused to update the training. If the model works well, it is used to classify all five classes; it is assumed that this will allow the convolutional layers to extract the essential features that provide optimal input from the device hospital update.

We considered using transfer learning from MobileNet, Densenet121, Xception, Resnet50, or randomization models at each suitable parametric model time, but we decided that the features of the ImageNet data were adequately different from those of the ImageNet data features of X-ray scans. Therefore, it made more sense to learn the features from

scratch on this dataset. However, using transfer learning between models accelerated the training process by considering around 60 communication rounds, which saved weeks of training time to reach the global minimum loss.

In the first performance of Table 3, we compare accuracy and AUC between CL, including models like KNN, AdaBoost, and XGB, and FL models like FedAvg-MLP and FedAvg-CNN for two datasets, balanced and imbalanced data. In addition, in transfer learning, we leave the random model in the data processing so that we can compare each different model.

The AUC results with the balanced dataset show that XGB's centralized learning yielded the best results, followed by FedAvg-CNN, FedAvg-MLP, KNN, and AdaBoost, which exhibited similar accuracy. With the imbalanced dataset, XGB still demonstrated the best results, followed by FedAvg-MLP, KNN, FedAvg, and AdaBoost with AUC and similar for accuracy. Based on the results, the application of FL in breast cancer classification is a reasonable choice when it is associated with more advantages than traditional learning methods. Moreover, the results confirmed that AUC should be chosen over accuracy in breast cancer classification.

In the evaluation in Tabs. 4, 5, 6, and 7, similar to Table 3, we change the transfer learning models in data processing at the client. However, the analysis showed lower results for FL than for XGB (100% for both balanced and unbalanced data). Even when comparing the results with other methods, the FeAvg-MLP and FeAvg-CNN results combined with MobileNet transfer learning are the best compared to other classical methods.

However, in terms of other advantages, FedAvg-CNN with MobileNet transfer learning gives better results than other learning methods, including accuracy and AUC in both data types: balanced (**97,106%/99.743%**) and unbalanced (**85,741%/95,999%**). In addition to checking the quality classification performance, we must pay attention to the selection of learning models for customers. In FL, low-parameter models are necessary because models with significant parameters cannot run on low-profile clients. Based on the results in Table 3 to 7 choose MobileNet, with a capacity of 4.3M, as the lowest and most reasonable. Again, we recommend selecting the suitable learning models for research requirements; otherwise, the choice will affect the accuracy.

B. PERFORMANCE TO COMPARE THE AVERAGE ACCURACY, RECALL, AND F1-SCORE RESULTS ACROSS BREAST CANCER STAGE

Consider the results on the breast cancer stage prediction classes as in Table 8, the recall result to avoid false negative control almost reached 100% for ''IID'' when the DDSM data used about 50%, and oversampling by SMOTE the samples on three edge servers, then averaging from a total of about 600 samples per layer. Meanwhile, the scenario ''non-IID'' keeps the original data of the model entirely unchecked for the minority classes. That is, the periods do not almost correctly predict more than the threshold of 0.5.

C. PERFORMANCE TO COMPARE THE AVERAGE ACCURACY, AUC FOR K-FOLD CROSS-VALIDATION

In the sub-section III-C6, we randomly split the dataset into training, validation, and testing sub-datasets by the following percentages: 80%, 10% and 10%. This division is perfectly reasonable if we have a large amount of data. However, when there is too little data, this division will lead to the abysmal performance of the deep learning model. The reason is that some data points useful for training have been included for validation and testing, but the model still needs to learn that data point. Sometimes, the small amount of data will lead to erroneous results when validating and testing because some classes are only used in validation and testing and not in training (because the division of training and validation is entirely random). If only based on that result were to evaluate the model, it would be inaccurate. In this study, to assess whether the MobileNet model we have chosen is suitable for the DDSM dataset, we use K-fold cross-validation ($K = 4$) to evaluate. In the case of non-Fold without cross-validation, we test and use the 90% (80% training set, 10% validation set) dataset to train the process and the 10% test set, while the 10% test set is distinct. In the case of $K = 4$ with cross-validation, we use 90% (training, validation) divided every 4-fold corresponding to each K, including 75% training set (K-1) and 25% (K) for the validation set. The remaining 10% of the test set is similar to non-Fold (fair-play) that we use for testing after finishing training with cross-validation. After the evaluation process, we choose $K = 4$ because the data shared by the server is already relatively small enough for the validation process to have enough data to test. Then we train the model K times, where ''1'' part is the validation data and (K-1) the rest is the training data. The final model evaluation result will be the average of K training times evaluation results. Consider Table 9, we do non-fold normal and 4-fold cross-validation with two datasets, balanced and unbalanced, with different models.

Evaluation results based on accuracy and AUC show that it is correct to choose FeAvag-CNN and MobileNet models with advantages when using the FL framework. Then the value of accuracy and AUC (**97.91%/99.80%** in case of non-fold and **92.79%/99.04%** in case of 4-fold) balance. And accuracy and AUC (**84.93%/95.93%** in case of non-fold and **89.33%/96.99%** in case of 4-fold) balance is the highest and most stable compared to the other methods.

D. PERFORMANCE TO COMPARE THE ACCURACY ON LOCAL MODELS AT EDGE SERVERS

We performed a local comparison of training, testing, and validation data on three edge servers for 60 rounds of data without overfitting (see Fig. 9). For this test, we evenly split the data across edge servers, assuming that each server would have 33.3% of the original data. This data will be optionally distributed and have no pre-assigned label order. The data

TABLE 3. Quality classification accuracy and AUC performance assessment results for a balanced and imbalanced datasets for CL at edge servers, FL models at cloud servers, and random model in transfer learning at clients.

Methods			Accuracy and AUC		Model parameters	
Cloud server	Edge server	Client	Balanced data	Imbalanced data	Edge server	Client
-	KNN	Random model	59.647% / 88.926%	88.266% / 94.531%	-	4-26M
-	AdaBoost		56.023% / 82.006%	80.251% / 59.580%	-	
-	XGB		100% / 100%	100% / 100%	-	
FeAvg-MLP			95.107% / 99.542%	88.754% / 97.977%	2,789,349	
FeAvg-CNN			97.106% / 99.743%	85.741% / 95.999%	8,722,149	

TABLE 4. Quality classification accuracy and AUC performance assessment results for a balanced and imbalanced datasets for CL at edge servers, FL models at cloud servers, and MobileNet model in transfer learning at clients.

Methods			Accuracy and AUC		Model parameters	
Cloud server	Edge server	Client	Balanced data	Imbalanced data	Edge server	Client
-	KNN	MobileNet model	62.724% / 90.651%	88.252% / 96.638%	-	4.3M
-	AdaBoost		53.847% / 81.034%	81.554% / 59.811%	-	
-	XGB		100% / 100%	100% / 100%	-	
FeAvg-MLP			95.244% / 99.548%	88.904% / 98.153%	2,789,349	
FeAvg-CNN			**97.919% / 99.804%**	**84.933%/95.932%**	8,722,149	

TABLE 5. Quality classification accuracy and AUC performance assessment results for a balanced and imbalanced datasets for CL at edge servers, FL models at cloud servers, and Densenet121 model in transfer learning at clients.

Methods			Accuracy and AUC		Model parameters	
Cloud server	Edge server	Client	Balanced data	Imbalanced data	Edge server	Client
-	KNN	Densenet121 model	59.740% / 88.670%	87.191% / 94.110%	-	8.1M
-	AdaBoost		54.980% / 83.456%	81.640% / 60.837%	-	
-	XGB		99.999% / 100.000%	100% / 100%	-	
FeAvg-MLP			91.969% / 99.079%	87.911% / 97.931%	2,789,349	
FeAvg-CNN			96.792% / 99.652%	84.150% / 95.606%	8,722,149	

TABLE 6. Quality classification accuracy and AUC performance assessment results for a balanced and imbalanced dataset for centralized learning at edge servers, federated learning models at cloud servers, and Xception model in transfer learning at clients.

Methods			Accuracy and AUC		Model parameters	
Cloud server	Edge server	Client	Balanced data	Imbalanced data	Edge server	Client
-	KNN	Xception model	59.368% / 87.781%	88.471% / 93.230%	-	22.96M
-	AdaBoost		53.590% / 82.392%	80.865% / 59.759%	-	
-	XGB		99.999% / 100.000%	100% / 100%	-	
FeAvg-MLP			95.928% / 99.655%	87.797% / 97.425%	2,789,349	
FeAvg-CNN			93.661% / 98.080%	83.851% / 94.995%	8,722,149	

TABLE 7. Quality classification accuracy and AUC performance assessment results for a balanced and imbalanced dataset for centralized learning at edge servers, federated learning models at cloud servers, and Resnet50 model in transfer learning at clients.

Methods			Accuracy and AUC		Model parameters	
Cloud server	Edge server	Client	Balanced data	Imbalanced data	Edge server	Client
-	KNN	Resnet50 model	58.551% / 89.084%	86.160% / 95.685%	-	25.6M
-	AdaBoost		52.754% / 80.981%	79.588% / 61.160%	-	
-	XGB		99.944% / 100.000%	100% / 100%	-	
FeAvg-MLP			20.129% / 49.951%	86.674% / 92.427%	2,789,349	
FeAvg-CNN			22.562% / 53.063%	49.277% / 68.142%	8,722,149	

extraction process was performed after the client extracts. The results of the edge server training demonstrate that, with the CNN model, the distribution was only different, and the test results for edge 1 were not significantly lower than those of the other two edges. Thus, the proposed plan is suitable only when accuracy sharply increases in the first communication rounds. Federated transfer learning is the pre-trained process with the ImageNet framework that sup-ports optimizing the call training phase.

E. PERFORMANCE TO COMPARE THE GLOBAL AUC AND RECALL POINT CHART BOX PLOT WITH DIFFERENT MODELS

We compared AUC score and recall point with different models in two cases: IID and non-IID. Based on Fig. 10, we see the AUC scores of Xception-IID, DenseNet121-non-IID, and ResNet50-non-IID, the AUC scores data of Res-Net50-non-IID and Xception-IID show that the process is under performing because it tends to concentrate data (median) at a

TABLE 8. Evaluation of average precision, recall, F1-score results across classes in the case of FeAvg-CNN + MobileNet, IID and non-IID datasets.

	IID				non-IID			
Breast cancer stage	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
NO	0.996	0.896	0.946	599	0.913	0.98	0.946	2396
BEC	0.96	1.00	0.976	617	0.36	0.183	0.28	99
BEM	0.986	1.00	0.99	615	0.42	0.26	0.32	109
MAC	0.966	1.00	0.98	592	0.263	0.126	0.173	88
MAM	0.99	1.00	0.996	589	0.586	0.48	0.37	92

TABLE 9. Average accuracy and AUC of five classes using non-Fold and 4-Fold validation in the case of balanced and imbalanced datasets.

Methods			Accuracy and AUC (Balanced data)		Accuracy and AUC (Imbalanced data)	
Cloud server	Edge server	Client	non-Fold	4-Fold	non-Fold	4-Fold
FeAvg-CNN		Random model	97.10% / 99.74%	92.92% / 99.06%	85.74% / 95.99%	70.47% / 86.23%
		MobileNet model	**97.91% /99.80%**	**92.79% / 99.04%**	**84.93% / 95.93%**	**89.33% / 96.99%**
		Densenet121 model	96.79% / 99.65%	90.26% / 98.59%	84.15% / 95.60%	86.02% / 95.89%
		Xceptron model	93.66% / 98.08%	79.65% / 94.82%	83.85% / 94.99%	87.35% / 95.46%
		Resnet model	22.56% / 53.06%	22.92% / 53.97%	49.27% / 68.14%	41.39% / 65.09%

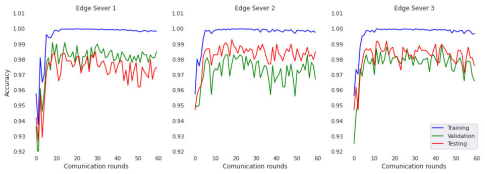

FIGURE 9. Comparison of training, testing, and validation accuracy per communication round for three edge servers with the same settings.

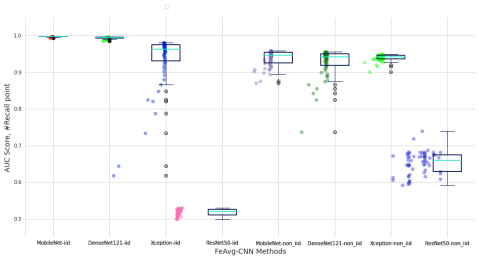

FIGURE 10. Comparing the AUC score, recall point on eight model learnings as MobileNet, DenseNet121, Xception, ResNet50 for both case IID and non-IID data distributions.

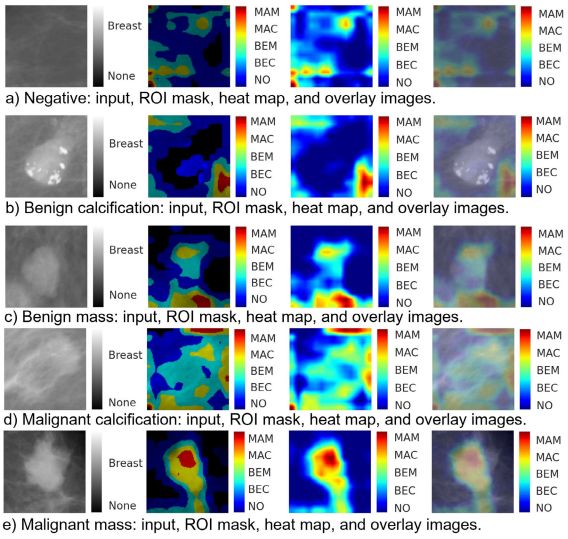

FIGURE 11. Heat map distribution for breast cancer classification stages (a to e) in the DDSM dataset after classification.

high level, large variability. Meanwhile, the quality control of MobileNet and DenseNet121 was the best because the AUC score for concentration was low and fluctuated within a narrow range. In addition, the degree of dispersion for recall points in the communication loop rapidly met the requirements in the case of MobileNet and DenseNet121. At the same time, the remaining scenarios did not converge. Once again, the effectiveness of the transfer learning method in accurately classifying the area of breast cancer can be seen.

F. PERFORMANCE A HEAT MAP BASED ON A BREAST CANCER STAGES

Consider directly at the breast cancer classification stage in DDSM data. We performed a heat map review by applying FeAvg-CNN + MobileNet (see Fig. 11). The results in order img input > thresh > ROI mask > heatmap > overlay shows an overview of the breast cancer recognition partition based on any input image, using a binary threshold to pick out the region (high, low). Next, we initialized roi_mask to define the area of interest. Finally, a heat map and overlaying are initialized after classification. The heat map provided a full heat index (MAM, MAC, BEM, BEC, NO), and live image of the central areas with prominent pixels in the image. A heatmap-based assessment can highlight the type of partition to be diagnosed and indicate that it is worthy of investigation. In addition to staging breast cancer based on the assessments described in the previous subsection, a heatmap-based assessment offers a different perspective of breast cancer and provides doctors and medical staff with another assessment method.

V. DISCUSSION

We demonstrated that FL with CNNs can be trained to determine whether part of a mammogram is abnormal, improving

multiclass classification through each stage was positive with 100% recall ability and 99.804% AUC when using pre-trained MobileNet model to extract features. Adjusting the decision threshold and communication parameters further improved the use of SGD as a weight global update function. These methods can be used in premammogram screenings to allow radiologists to focus on scans that are likely to contain multi-hospital and multinational abnormalities.

Although training on many units with each small dataset, FL yielded the same classification accuracy results as the centralized learning method. CNN is one of two topology models that achieved the best and most accessible recall performance for distributed modeling. A key advantage of combining FedAvg-CNN and MobileNet feature extraction was the ability to customize the ConvNet layer volume to quickly obtain results that were equivalent to larger networks, but long-distance connectivity was not guaranteed. Therefore, FedAvg-CNN had the advantage of adapting well to mobile devices with hardware that allows the rapid processing of medium tasks. Finally, performing simulations with a random number of stations for each round of FL demonstrated that CNN can function even in unstable network situations.

However, FL generally requires longer training time in simulations than centralized learning. The CNN model still requires many data processing steps and feature extraction to achieve high accuracy for the request and the ability to recall avoiding false positive detection. Moreover, although it works in situations with changing station numbers, the performance of the learning model associated with CNN in the report is still significantly degraded (98% down) in the real scenario with more individual data humanized and increases over time and needs time to maintain the frequency of updates.

VI. CONCLUSION AND FUTURE RESEARCH

In this study, we presented a solution for classifying breast cancer images using feature extraction from multiple participating environments instead of a centralized learning facility. The centralized environment consisted of an inter-national group of hospitals and medical imaging centers that joined collaborative efforts to train the model to be completely data-decentralized, without sharing any data between hospitals. Moreover, we focused on analyzing recall performance more than accuracy because false negatives can be life-threatening. By contrast, like studies, humans can consider false positives instead of a whole before. The results demonstrate that the accuracy was higher than that of other models. In the future, we plan to create a system that will scan the entire mammogram as input, segment it, and analyze each segment to yield results for the entire mammogram to make an end- the complete to-end for mammogram analysis. In addition to improved data processing, simulations can be extended with multiple clients or groups of clients on separate devices, and individual patient interventions share privacy security.

REFERENCES

[1] (2021). *Breast Cancer Now Most Common Form of Cancer: WHO Taking Action*. Accessed: Feb. 3, 2021. [Online]. Available: https://www.who.int

[2] (2021). *Breast Cancer Overtakes Lung Cancer in Terms of Number of New Cancer Cases Worldwide*. Accessed: Feb. 4, 2021. [Online]. Available: https://www.who.int

[3] Z. Wang, M. Li, H. Wang, H. Jiang, Y. Yao, H. Zhang, and J. Xin, "Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features," *IEEE Access*, vol. 7, pp. 105146–105158, 2019, doi: 10.1109/ACCESS.2019.2892795.

[4] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis," *IEEE Access*, vol. 8, pp. 96946–96954, 2020, doi: 10.1109/ACCESS.2020.2993536.

[5] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," 2016, *arXiv:1612.00542*.

[6] N. Mobark, S. Hamad, and S. Z. Rida, "CoroNet: Deep neural network-based end-to-end training for breast cancer diagnosis," *Appl. Sci.*, vol. 12, no. 14, p. 7080, Jul. 2022, doi: 10.3390/app12147080.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[8] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," 2019, *arXiv:1911.02054*.

[9] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," 2021, *arXiv:2102.07623*.

[10] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022, doi: 10.1109/TMC.2020.3045266.

[11] S. Bozinovski, "Reminder of the first paper on transfer learning in neural networks, 1976," *Informatica*, vol. 44, no. 3, pp. 293–295, Sep. 2020.

[12] Q. Li et al., "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023, doi: 10.1109/TKDE.2021.3124599.

[13] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, Q. S. T. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020.

[14] X. Lu, Y. Liao, P. Lio, and P. Hui, "Privacy-preserving asynchronous federated learning mechanism for edge network computing," *IEEE Access*, vol. 8, pp. 48970–48981, 2020, doi: 10.1109/ACCESS.2020.2978082.

[15] C. He, S. Li, J. So, M. Zhang, H. Wang, X. Wang, P. Vepakomma, A. Singh, H. Qiu, L. Shen, P. Zhao, Y. Kang, Y. Liu, R. Raskar, Q. Yang, M. Annavaram, and S. Avestimehr, "FedML: A research library and benchmark for federated machine learning," *arXiv*, vol. abs/2007.13518, 2020.

[16] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique," *IEEE Access*, vol. 9, pp. 71194–71209, 2021, doi: 10.1109/ACCESS.2021.3079204.

[17] H.-J. Chiu, T.-H.-S. Li, and P.-H. Kuo, "Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine," *IEEE Access*, vol. 8, pp. 204309–204324, 2020, doi: 10.1109/ACCESS.2020.3036912.

[18] R. Singh, T. Ahmed, A. Kumar, A. K. Singh, A. K. Pandey, and S. K. Singh, "Imbalanced breast cancer classification using transfer learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 83–93, Jan. 2021, doi: 10.1109/TCBB.2020.2980831.

[19] A. Elmoufidi, "Deep multiple instance learning for automatic breast cancer assessment using digital mammography," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022, doi: 10.1109/TIM.2022.3177141.

[20] T. A. Khoa, D.-V. Nguyen, M.-S. Dao, and K. Zettsu, "Fed xData: A federated learning framework for enabling contextual health monitoring in a cloud-edge network," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 4979–4988, doi: 10.1109/BigData52589.2021.9671536.

[21] T. A. Khoa, D.-V. Nguyen, P. V. Nguyen Thi, and K. Zettsu, "FedMCRNN: Federated learning using multiple convolutional recurrent neural networks for sleep quality prediction," in *Proc. 3rd ACM Workshop Intell. Cross-Data Anal. Retr.*, New York, NY, USA, Jun. 2022, pp. 63–69, doi: 10.1145/3512731.3534207.

[22] T. A. Khoa, D.-V. Nguyen, M.-S. Dao, and K. Zettsu, "SplitDyn: Federated split neural network for distributed edge AI applications," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Osaka, Japan, Dec. 2022, pp. 6066–6073, doi: 10.1109/BigData55660.2022.10020803.
[23] D.-D. Le, A.-K. Tran, M.-S. Dao, K.-C. Nguyen-Ly, H.-S. Le, X.-D. Nguyen-Thi, T.-Q. Pham, V.-L. Nguyen, and B.-Y. Nguyen-Thi, "Insights into multi-model federated learning: An advanced approach for air quality index forecasting," *Algorithms*, vol. 15, no. 11, p. 434, Nov. 2022, doi: 10.3390/a15110434.
[24] D.-V. Nguyen and K. Zettsu, "Spatially-distributed federated learning of convolutional recurrent neural networks for air pollution prediction," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 3601–3608, doi: 10.1109/BigData52589.2021.9671336.
[25] C. I. Bercea, B. Wiestler, D. Rueckert, and S. Albarqouni, "FedDis: Disentangled federated learning for unsupervised brain pathology segmentation," 2021, *arXiv:2103.03705*.
[26] W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, and M. Baust, "Privacy-preserving federated brain tumour segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 133–141.
[27] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *J. Biomed. Informat.*, vol. 99, Nov. 2019, Art. no. 103291.
[28] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101765.
[29] Y. Yeganeh, A. Farshad, N. Navab, and S. Albarqouni, "Inverse distance aggregation for federated learning with non-IID data," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Cham, Switzerland: Springer, 2020, pp. 150–159.
[30] Z. Wang, Q. Liu, and Q. Dou, "Contrastive cross-site learning with redesigned net for COVID-19 CT classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2806–2813, Oct. 2020.
[31] H. R. Roth et al., "Federated learning for breast density classification: A real-world implementation," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Cham, Switzerland: Springer, 2020, pp. 181–191.
[32] A. Jiménez-Sánchez, M. Tardy, M. A. González Ballester, D. Mateus, and G. Piella, "Memory-aware curriculum federated learning for breast cancer classification," 2021, *arXiv:2107.02504*.
[33] Z. Ma, M. Zhang, J. Liu, A. Yang, H. Li, J. Wang, D. Hua, and M. Li, "An assisted diagnosis model for cancer patients based on federated learning," *Frontiers Oncol.*, vol. 12, Mar. 2022, Art. no. 860532, doi: 10.3389/fonc.2022.860532.
[34] B. L. Y. Agbley, J. Li, M. A. Hossin, G. U. Nneji, J. Jackson, H. N. Monday, and E. C. James, "Federated learning-based detection of invasive carcinoma of no special type with histopathological images," *Diagnostics*, vol. 12, no. 7, p. 1669, Jul. 2022, doi: 10.3390/diagnostics12071669.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–252.
[36] M. Lorentzon, "Feature extraction for image selection using machine learning," M.S. thesis, Dept. Elect. Eng., Comput. Vis. Lab., Linköping Univ., Linköping, Sweden, 2017, pp. 17–18.
[37] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer Jr., R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography*. Berlin, Germany: Springer, 1998, pp. 457–460.
[38] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci Data*, vol. 4, p. 170177, Dec. 2017, doi: 10.1038/sdata.2017.177.
[39] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proc. 5th Int. Workshop Digit. Mammogr.*, M. J. Yaffe, Ed., 2001, pp. 212–218.
[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[41] R. S. Lee, F. Gimenez, and A. Hoogi, "Curated breast imaging subset of DDSM," Tech. Rep., 2016.
[42] E. A. Scuccimarra. *DDSM Mammography*. Accessed: Feb. 3, 2022. [Online]. Available: https://www.kaggle.com/datasets/skooch/ddsm-mammography

Y. NGUYEN TAN is currently pursuing the Ph.D. degree with the Faculty of Electrical and Electronic Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. He is a Lecturer with Phan Thiet University. His research interests include machine learning, image processing, and computer vision in the medical field.

VO PHUC TINH is currently pursuing the bachelor's degree with the Faculty of Electrical and Electronic Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. His research interests include the Internet of Things, embedded systems, and artificial intelligence.

PHAM DUC LAM is currently a Lecturer with Nguyen Tat Thanh University, Vietnam. His research interests include health applications that apply technologies, such as image processing, embedded systems, and the Internet of Things.

NGUYEN HOANG NAM received the Ph.D. degree from the National Chiao Tung University, Taiwan, in 2017. Currently, he is a Researcher with the MERLIN Research Group and a Lecturer with the Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. His research interests include health applications using artificial intelligence, image processing, and computer vision.

TRAN ANH KHOA received the Ph.D. degree from the University of Siena, Siena, Italy, in 2017. Currently, he is a Researcher with the MERLIN Research Group, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. His research interests include health applications using artificial intelligence and the Internet of Things.