AIMS Mathematics, 10(11): 26260–26292.
DOI: 10.3934/math.20251155
Received: 24 June 2025
Revised: 29 September 2025
Accepted: 11 October 2025
Published: 13 November 2025

http://www.aimspress.com/journal/Math

Research article

Privacy-preserving breast disease detection via federated GA-optimized ensembles learning

Karim Gasmi[1,*], Ibtihel Ben Ltaifa[2], Moez Krichen[3], Shahzad Ali[1], Omer Hamid[4], Mohamed O. Altaieb[1], Lassaad Ben Ammar[5], Manel Mrabet[5] and Mahmood Mohamed[6]

[1] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia

[2] STIH, Sorbonne Université, Paris, France

[3] ReDCAD Laboratory, University of Sfax, Sfax 3038, Tunisia

[4] Cybersecurity Department, College of Engineering and Information Technology, Buraydah Private Colleges, Buraydah 51418, Saudi Arabia

[5] Department of Computer Sciences, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia

[6] Department of Information Systems and Technology, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

* **Correspondence:** Email: kgasmi@ju.edu.sa.

Abstract: Breast cancer is still one of the leading causes of death in women, and finding it early is essential for treatment to work. This study presents a sizeable deep learning architecture for classifying and segmenting breast ultrasound images. It utilizes federated learning to maintain patients' data privacy. The proposed pipeline begins with a thorough preprocessing stage that includes scaling, normalizing, and advanced image augmentation using affine and contrast-based changes. We employ four convolutional neural network architectures for hierarchical classification: ResNet50, EfficientNet, VGG16, and Xception. First, we distinguish between typical cases and abnormal ones. We then further classify abnormal images into benign and malignant classes. We employ an ensemble technique that combines the outputs of ResNet50 and EfficientNet through a weighted average optimized by a genetic algorithm to enhance the model's resilience. This method dramatically improves the classification's effectiveness, achieving higher accuracy and reliability. We use a federated learning system with the federated averaging (FedAvg) algorithm to improve data privacy. Our federated architecture maintains high accuracy while ensuring that the raw data stays at its local source. We test it with both single-client and multi-client setups. Ultimately, we employ a hybrid architecture that combines the feature maps of ResNet50 and EfficientNet to segment images of lesions known to be malignant. This yields significant spatial agreement with expert annotations. The Dice score and intersection over union (IoU)

are two evaluation criteria that demonstrate the effectiveness of our segmentation model. This all-in-one system offers accurate and privacy-conscious breast ultrasound analysis, indicating that it could be beneficial in decentralized healthcare settings. The suggested method was tested using a standard breast magnetic resonance imaging (MRI) dataset, demonstrating its robustness and applicability in various situations. We used the accuracy, precision, recall, and F1-score to measure performance, and we found that the classification was accurate up to 96%. This paper discusses a scalable system that maintains people's privacy by utilizing ensemble learning, optimization-driven feature selection, and federated learning to aid doctors in early breast cancer detection using MRI data. This will make it easier for doctors to use this system in a broader range of medical tests.

Keywords: SDG 3; federated learning; ensemble learning; genetic algorithm; breast cancer detection; weight selection; classification
Mathematics Subject Classification: 62H30, 68T05, 68U35, 90C59, 92C50

1. Introduction

Breast cancer is still the most common type of cancer among women around the world, and it is a significant cause of cancer deaths. The World Health Organization said that in 2020 that more than 2.3 million women were diagnosed with breast cancer. This resulted in more than 685,000 deaths worldwide [1]. There is a strong link between the stage of breast cancer when it is found and the outlook and survival rate of people who have just been diagnosed with the disease. An early and accurate diagnosis can significantly lower the death rate by allowing for timely intervention and the creation of personalized treatment plans.

The diagnosis of breast cancer is challenging because of several things, such as the fact that the tumour is not the same in all cases, the breast tissues are very dense, and the quality of imaging can vary. Ultrasound is a promising imaging method because it is readily available, does not use ionizing radiation, and can distinguish between cystic and solid masses. On the other hand, the interpretation of ultrasound images depends significantly on the reader, varying from person to person and even from one observer to another. Due to these issues, there has been a growing interest in developing automated diagnostic systems that utilize artificial intelligence (AI) [2] and deep learning to enable radiologists to identify and describe breast lesions more accurately and consistently.

Medical professionals have found computer vision methods that utilize AI to be helpful. These systems can accurately locate, outline, and classify malignant tumors using medical imaging tools like mammography [3]. Traditional methods employed image processing and machine learning to extract handcrafted and basic features from images to identify likely locations. As false positive results increase and accuracy decreases, new deep learning algorithms slowly replace old tumor segmentation methods. These new algorithms are superior to the ones currently in use [4]. They combine information about the surrounding tissue and automatically extract features for tumour delineation and classification in computer-aided diagnosis systems. Convolutional neural networks (CNNs) are utilized in automated computer-aided design systems and medical imaging to detect and identify features, particularly patterns associated with breast cancer. Deep learning models are gaining popularity as computers become faster and more powerful. This is because these models can automatically get a lot of

information from medical images without needing to know anything about them or do any feature engineering first [5]. This has improved the results of the automated systems by finding a balance between how many lesions they can find in a mammogram and how accurate they are [6, 7].

Deep learning methods, particularly CNNs, have demonstrated excellent performance in visual recognition tasks such as classifying and segmenting diagnostic medical images. Several CNN designs, including ResNet, EfficientNet, VGG, and Xception, have been successfully utilized to detect breast cancer. These models can learn to distinguish between normal, benign, and malignant tissues, and they can also extract complex information from ultrasound images. On the other hand, these models only work effectively when they have access to large datasets containing a wide range of information. In the medical field, however, these types of datasets are often limited due to concerns about privacy, data silos within institutions, and regulations that restrict access.

Federated learning (FL) has become a significant solution to these problems, offering a compelling alternative to the traditional method of centralized training. FL enables many healthcare institutions or edge devices to collaborate in training machine learning models without requiring the sharing of patient data. A central server allows people to to communicate with each other and track changes to the model, while maintaining the data's safety and privacy. This decentralized approach adheres to the rules for managing medical data, providing people with access to more extensive and diverse datasets as they are distributed across various institutions. Recent studies have demonstrated that FL can be applied in other healthcare settings. Some of these applications include identifying COVID-19 [8], categorizing skin lesions [9], and segmenting brain tumors [10].

In this case, we present a complete deep learning framework for analyzing breast ultrasounds. This framework utilizes CNN-based classification and segmentation, along with the benefits of FL, to ensure the safety of individuals' privacy. Our plan includes a two-step classification process. The first step is to differentiate between normal and abnormal ultrasound images. The second step is to classify the abnormal images as benign or malignant on the basis of their characteristics. We employ a genetic algorithm to determine the optimal weights for an ensemble learning method that combines the predictions of ResNet50 and EfficientNet. This helps us create a more substantial and accurate model.

The last step of the process uses a segmentation model that combines the feature representations of the ResNet and EfficientNet backbones. After that, the malignant cases are looked at more closely. The model accurately represents lesion boundaries, making it easier to assess and make informed decisions in the clinic. Changes are made to every part of the pipeline to work in a FL setting. This is done to ensure the raw patient data remains safe at each client's site. Our contributions show a workable and effective way to use ultrasound imaging systems to decentralize the diagnosis of breast cancer.

The results of our study show that the proposed framework can classify and segment tasks with an accuracy of 0.97. This suggests that the framework is practical and can be applied in real-life clinical settings. The rest of this work is organized as follows: Section 2 provides a comprehensive overview of the existing research on breast cancer detection. In Section 3, we outline our proposed framework, which includes a complete description of the preprocessing steps, data augmentation methods, hybrid feature extraction methods, and the ensemble learning architecture that uses deep classifiers and the optimizer algorithm. In Section 4, we talk about the experimental setup, which includes the dataset, the evaluation measures, and the baseline configurations. In Section 5, we talk about the empirical results. This includes comparing individual models with our ensemble technique and examining how well the hybrid characteristics work. In conclusion, Section 5 summarizes the most important contributions and

talks about possible topics for future research.

2. Related work

Breast cancer research is a crucial area of study in medical image analysis. Digital mammography can help find breast cancer early, which makes it more likely that the person will survive. There are many ways to check for breast cancer, such as mammography [11], digital breast tomosynthesis (DBT) [12], computed tomography (CT) [13], positron emitted tomography (PET) [14], thermography [13], magnetic resonance imaging (MRI) [15], and ultrasound [16]. Radiologists often struggle to accurately judge mammography screening images without the aid of a computer-aided diagnosis (CAD) system. The goal of a computer-aided diagnosis (CAD) system is to automatically detect breast cancer tumors and distinguish between healthy and cancerous tissues [17]. Several studies have come up with automated methods for finding, classifying, and diagnosing breast cancer. These methods are now used in computer-aided design systems [18]. These studies provide an overview of the fundamental steps in image processing, including preprocessing, segmentation, feature extraction, and classification. Pre-processing images is one of the first steps in every computer-aided design system. It is essential to get the best possible results for the next steps. MRI images typically contain additional noise that must be removed before they can be sent to the computer-aided design system. Pre-processing facilitates the identification of meniscal and unusual lesions, while also enhancing the key image features for subsequent processing. Segmentation and feature extraction are two of the most critical steps in image pre-processing. In this section, we talk about the breast cancer literature and new ways to use MRI images to find, classify, and separate breast cancer.

2.1. Classification approches

Classification of breast cancer aims to find a systematic and objective prognostic for pathologists [19]. Generally, the most recurrent classification is binary: Benign tumors or cancer. Several studies have reviewed automated techniques and investigation methods used to explore breast images from different perspectives, depending on the disease status and the type of screening images [20]. Several AI and neural network methods have been proposed to enhance the effectiveness of breast cancer detection and classification. Machine learning is a popular tool for evaluating algorithms that facilitate breast cancer prediction, recognition, and classification. Breast cancer is classified with several techniques such as linear regression [21], K-nearest neighbors (KNN) [22], softmax regression [23], artificial neural networks (ANNs) [24], and support vector machine (SVM) [25]. Breast cancer is also classified using techniques such as the softmax discriminant classifier, linear discriminant analysis [26], and fuzzy C means clustering [27]. Amrane et al. [19] proposed two different classifiers: The naive bayes classifier and KNN for breast cancer classification. After comparing the two algorithms, it was found that KNN achieved a higher accuracy. In [28], the authors proposed an optimized stacking ensemble learning model called OSEL that incorporates multiple classification algorithms to determine the most effective combination for early breast cancer prediction.

Over the last few years, deep learning has become a popular tool for detecting and classifying breast cancer [29]. Several studies have demonstrated that deep learning methods can detect and identify breast cancer up to 12 months earlier than those using traditional clinical procedures [30]. Moreover,

deep learning-based techniques can be deployed to determine the most pertinent features [31,32]. Thus, various deep learning-based methods have been used in breast cancer classification, including CNN, DNNs (deep neural networks), RNNs (recurrent neural networks), DBNs (deep belief networks) [33], and AutoEncoder (AE).

Several challenges and limitations are associated with using deep learning techniques for breast detection and classification [34]. Several studies have demonstrated that the CNN model has shown greater accuracy in breast cancer detection [20, 35]. These studies applied the CNN model to extract relevant features from validated gene expression data [36]. The CNNs deployed for breast cancer diagnosis can be grouped into transfer learning-based models and de novo trained models [30]. A de novo trained model is a CNN model that was generated and trained from scratch. In contrast, transfer learning-based CNN methods utilize a trained neural network, such as AlexNet [37], the residual neural network (ResNet) [38, 39], or the visual geometry group (VGG) [40], among others.

Abunasser et al. in [20] proposed a deep learning model that adds five fine-tuned deep learning models, including Xception, InceptionV3, VGG16, MobileNet, and ResNet50, to detect and classify breast cancer into eight classes. In [41], the authors proposed a deep neural convolutional network that utilizes GoogleNet and AlexNet for breast cancer recognition, using DDSM (digital database for screening mammography) breast cancer images with three classes: Normal, benign, and malignant. Khan et al. [42] proposed a classification framework that combines three CNN architectures, including GoogLeNet, VGGNet, and ResNet, to extract different low-level features, which are then fed into a fully connected layer.

2.2. Segmentation techniques

Image segmentation methods aim to partition the initial image into multiple distinct regions, called regions of interest (ROIs) [43]. In particular, they aim to distinguish important areas from the rest of the image using specific criteria such as region or edge characteristics. In the case of MRI images, segmentation becomes crucial for delimiting pathological regions and enabling accurate diagnosis methods [44]. In the case of breast cancer, accurate segmentation of the ROI realizes precise tumor detection within the rest of the breast region [45]. As a result, a new representation with significant image characteristics is meaningful and easier to interpret.

Several segmentation methods have been employed in medical imaging, mainly focusing on breast cancer detection and diagnosis [46]. They aim to isolate the ROI within mammographic images for detecting the breast mass [47]. Methods employed in segmentation include classical, machine learning, and deep learning segmentation techniques.

Classical segmentation deploys different techniques, including edge-driven segmentation [48] and region segmentation [49]. These segmentation methods rely on two characteristics or features of images: discontinuity and similarity [43]. The edge-based methods rely on Discontinuity features that identify objects' edges. These methods are grouped into gradient operators [50, 51], Laplacian operators [52], and optimal operators (Canny) [53]. On the other side, the region based-segmentation approaches deploy different methods such as threshold methods [54], clustering methods [55, 56], graph-based methods [57], superpixels methods [58, 59], region growing methods [60], and morphological watersheds [61, 62].

Machine learning segmentation methods applied in medical image analysis require using images annotated by qualified medical experts. These methods aim to identify ROIs in MRI scans such as

tumor tissues (e.g., malignant, benign), background tissues, lesions (mass/non-mass enhancements), and fibroglandular tissue. Unsupervised and supervised machine learning methods are used for image segmentation. In unsupervised machine learning, various techniques are employed, including hierarchical k-clustering [63], k-means clustering [55], and fuzzy C-clustering. In supervised machine learning, various models are employed, including vector machines [64] and naive Bayes models [65].

Various deep learning models are deployed for mammogram image segmentation [66, 67]. For medical imaging tasks, the U-Net model [68] has gained significant attention due to its effectiveness, especially in scenarios with limited annotated data. Shen et al. [69] used the U-Net model (ResCUNet) to segment and classify mammography images. In [70], the authors proposed a novel segmentation model, the full-resolution convolutional network (FrCN), for mammogram image segmentation. In addition, they used three traditional deep learning models: InceptionResNet-V2, ResNet-50, and a feedforward CNN to classify the segmented breast lesions as benign or malignant. Hossain [71] proposed a modified U-Net segmentation network for segmenting microcalcifications in mammogram images. In [68], the authors developed a novel model called AUNet for segmenting breast masses.

2.3. Neural architecture search and modern hybrid design

A neural architecture search (NAS) automates the design of neural network architectures by concurrently defining a search space (encompassing cells, operators, or macro-topologies), a search strategy (which may involve evolutionary, reinforcement, differentiable, or surrogate-assisted methods), and an evaluation protocol (generally multi-objective, accounting for metrics such as accuracy, latency, and memory). Recent improvements that have improved efficiency and versatility include the following. Progressive evaluation with sub-population preservation retains medium and large candidates with developmental potential through an extended training time, thereby diminishing the search costs while preserving diversity [72]; gradient-guided evolutionary NAS (ENAS) blends first-order sensitivity with evolutionary exploration to accelerate convergence while averting supernet coupling [73]; and score-predictor–assisted ENAS presents a learning surrogate that maintains the candidates' ranking, enabling cost-effective selection of extensive populations [74]. In the context of breast cancer histopathology utilizing the BACH/BreakHis datasets, a bio-inspired neural architecture search employing block-based stochastic categorical-to-binary (BSCB) encoding and the Ebola optimization search algorithm (EOSA) effectively identified CNNs that rival state-of-the-art techniques, according to a multi-criteria assessment [75].

In this scenario, we emphasize privacy (no unprocessed images may leave the hospital) and inter-institutional diversity (several institutions utilize distinct populations, methodologies, and technologies). This federated architecture complicates end-to-end NAS operations due to substantial communication demands, intricate secure aggregation processes, and the global model's susceptibility to site shifts that may occur during supernet training or frequent candidate alterations. We opted for another technique. We commence with a compact assembly of robust backbones, such as ResNet50 or EfficientNet, and subsequently employ a lightweight genetic algorithm (GA) to optimize an ensemble in weight space. Each client uses its own local validation partition to ascertain the mixing weights. The server uses FedAvg to integrate the models and amalgamate the mixture weight proposals from the clients into a global mixture, all while preserving the confidentiality of the raw data. This GA-weighted ensemble under FL mitigates the impact of complementary inductive biases among models, enhances stability across locations, and adheres to regulations governing healthcare data management.

Our pipeline remains operational with NAS. If a NAS technique subsequently develops a more robust backbone through pretraining on institutional or public data, it may be incorporated as an additional expert in the ensemble. The genetic algorithm only readjusts the mixture weights throughout the federation process. NAS focuses on the types of backbones to be incorporated, whereas our GA-weighted FL architecture emphasizes integrating these backbones while considering privacy and heterogeneity. This division of effort ensures that the advantages of network-attached storage are utilized effectively while maintaining reliable and secure performance across multiple sites.

3. Methodology

Our proposed architecture features a multi-stage deep learning pipeline that utilizes FL to classify breast ultrasound images while maintaining users' privacy. The first step in the system involves thoroughly preprocessing and augmenting the ultrasound dataset. This step aims to standardize all images to a uniform size, enhance the visual features, and increase the applicability of existing models across a broader range of situations. We utilize several cutting-edge CNN architectures, including ResNet50, EfficientNet, VGG16, and Xception, to construct individual and ensemble classification models for bots. The classification work is done at two levels. The first level distinguishes between normal and abnormal ultrasound scans. The second level separates abnormal scans into benign and malignant lesions. We use an ensemble method that combines the outputs of ResNet50 and EfficientNet. A genetic algorithm selects the optimal fusion weights on the basis of a customized fitness function. This helps us make our system more substantial and better at making predictions.

We use the federated averaging (FedAvg) protocol for FL. This approach keeps patients' information private while allowing multiple people to collaborate on developing a model. Because of this, several decentralized clients can train their models in private, and the central server only gets information about weight changes. We simulate single-client and multi-client FL scenarios to evaluate the performance of our ensemble models in this distributed setup. Before the updates are sent to everyone, each client goes through several local training epochs. This ensures that the models will converge and that the gearbox will operate as efficiently as possible.

In the final step of the process, which involves segmenting malignant lesions, a hybrid model combining feature maps from ResNet50 and EfficientNet is utilized. To create accurate lesion masks, the images are combined using a weighted average, then processed through convolutional and sigmoid layers, and finally combined again. The Dice score and the intersection over union (IoU) are two evaluation measures that demonstrate the spatial alignment of manual annotations. Our pipeline is a complete way to use ultrasound imaging to diagnose breast cancer in general. It provides very accurate lesion localization, respects privacy in various settings, and offers many other features. The Figure 1 shows all of the processes that are described below.

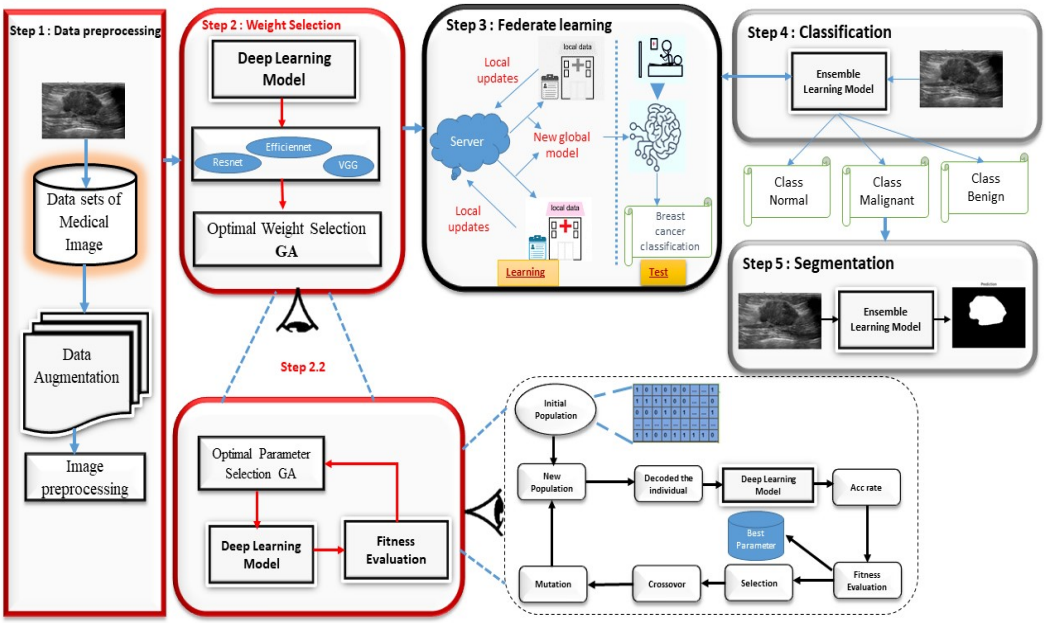

Figure 1. Proposed model for breast cancer classification and detection.

3.1. Preprocessing

Preprocessing is an essential step in deep learning pipelines that lays the groundwork for good training. We prepared the breast ultrasound image dataset for this work by combining normalization, reshaping, and encoding techniques. This was done to ensure compatibility with the latest neural networks and maintain data accuracy across classification and segmentation tasks. This was done to achieve the desired results.

- **Image resizing:** The sizes of the images were changed to fit the requirements of the chosen CNNs. For cumbersome models like ResNet50, VGG16, and Xception, the images' resolution was 224 by 224 pixels. We used a small size of 128 * 128 to reduce the extra work that lighter architectures like EfficientNet needed to perform. This was done while keeping the basic structural parts in place.
- **Normalization:** Normalization plays a pivotal role in stabilizing learning. We employed model-specific normalization routines.
  - ResNet and VGG: Standard score normalization using $x' = \frac{x-\mu}{\sigma}$.
  - EfficientNet: Pixel rescaling with $x' = \frac{x}{255}$.
- **Label encoding:** The dataset comprises multiple classification levels, first into "normal" and "abnormal", and then further classification of "abnormal" images into "benign" and "malignant". Labels were encoded in binary or one-shot vector formats.
- **Data splitting:** Stratified sampling divided the dataset into 80% training, 10% validation, and 10% testing sets.
- **Directory structuring and file integrity:** Datasets were structured by the client for FL experiments and checked for format consistency.

- **Channel adaptation:** Grayscale images were converted to red, green, and blue (RGB) via $I_{rgb} = [I, I, I]$.
- **Data pipeline optimization:** TensorFlow batching, shuffling, caching, and prefetching were used to ensure reproducibility.

3.2. Data augmentation

Data augmentation is a process used to increase the size of the dataset and provide the model with a broader range of image types to work with. This is especially useful in medical imaging, where datasets are often tiny.

Two primary libraries were used:

- ImageDataGenerator from Keras for simple affine transformations;
- Albumentations for advanced augmentation strategies with greater control.

3.3. Keras-based augmentations

- **Rotation:** Rotates the image by an angle of θ, implemented as: $R(\theta) = x\cos\theta - y\sin\theta$.
- **Translation:** Shifts the image by $\Delta x, \Delta y$, T(x, y), T(x, y) = (x + Δx, y + Δy).
- **Shear:** Applies a distortion defined as: S(x, y) = (x + k y, y).

3.4. Albumentation-based augmentations

- **Contrast stretching:** $I'(x, y) = \alpha(I(x, y) - \bar{I}) + \bar{I}$.
- **Brightness shift:** $I'(x, y) = I(x, y) + \beta$.
- **Contrast limited adaptive histogram equalization**: Enhances local contrast and preserves edge detail, which isespecially useful for lesion regions.

All augmentation functions were applied using medical imaging-specific probability thresholds and parameter ranges to preserve the meaning of the images.

3.5. Classification models

The classification pipeline in this study is divided into two parts. The first step is to distinguish between normal and abnormal breast ultrasound images. After that, the unusual cases are categorized as either benign or malignant. This two-part method makes diagnoses more specific and facilitates easier clinical decision-making hierarchies.

ResNet50, EfficientNet, VGG16, and Xception are four of the most advanced CNN architectures. They were all trained on ImageNet first. We compared all of these architectures. Each model features a dense classification head that can handle both binary and multi-class outputs. They also all received a global average pooling layer.

3.5.1. ResNet50

ResNet50 is a deep residual network with 50 layers. It leverages shortcut connections based on identity to address the issue of gradients disappearing in deep networks. The residual block is the most critical part of ResNet, and it is defined as follows:

$$y = F(x, \{W_i\}) + x.$$

The expression $F(x, W_i)$ shows the residual mapping that needs to be learnt, and the symbol x shows the same shortcut connection. This structure facilitates the training of more complex models by allowing gradients to flow directly through the network. Our tests demonstrated that ResNet50 can reliably extract spatial features while ignoring image noise.

3.5.2. EfficientNetBX

EfficientNetBX belongs to the EfficientNet family, which employs compound scaling to scale network depth d, width w simultaneously, and input resolution r using a global scaling coefficient ϕ:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi}.$$

Subject to the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2.$$

This model achieves state-of-the-art accuracy while maintaining computational efficiency. It benefits fine-grained classification tasks, such as differentiating between benign and malignant tumors.

3.5.3. VGG16

VGG16 is a deep CNN consisting of 13 convolutional layers and 3 fully connected layers. It uses small 3×3 filters and ReLU stands for Rectified linear unit (ReLU) activations. The basic form of each convolutional layer output is given by:

$$y = \max(0, W * x + b),$$

where $*$ denotes convolution, W is the kernel, and b is the bias. Despite its relatively simple design, VGG16 remains a strong baseline for visual tasks. However, it has a higher memory footprint compared with other modern architectures.

3.5.4. Xception

Extreme Inception, or Xception, is based on the idea that it is possible to completely separate the spatial and channel-wise correlations found in feature maps. The use of depthwise separable convolutions is:

$$y = (x * K_{\text{depthwise}}) * K_{\text{pointwise}}.$$

This approach significantly reduces the number of parameters and computations while maintaining accuracy. Xception's high representational capacity effectively models complex lesion boundaries in our setting.

3.5.5. Training protocol

The breast ultrasound dataset helped us optimize each model to its maximum potential. The last dense layer calculates a probability for each class using either a softmax activation (for tasks with three classes) or a sigmoid activation (for functions with two classes). We used the Adam optimizer to lower the cross-entropy loss. Batch normalization and dropout were also used to lessen the effects of overfitting.

We employed several model architectures to investigate the various trade-offs among the model's complexity, data processing speed, and accuracy in classifying both coarse-grained and fine-grained breast ultrasound data.

3.6. Ensemble learning with genetic algorithm weight optimization

Enhanced image classification capabilities are achieved by utilizing an ensemble learning model, which capitalizes on the benefits of two robust CNNs. Both models are initially supplied with weights that have been pre-trained using ImageNet, providing a solid foundation for extracting features. One of the most critical aspects of the combination is th weighted average of the results obtained from each model's global average pooling layers. This weighted average is an excellent method for combining the learned characteristics of both models, thereby capitalizing on their strengths. After feeding the combined features into a dense layer with 256 neurons and ReLU activation, a final thick layer with softmax activation is applied to classify the data into the desired classes. Compared with utilizing either model independently, this ensemble model tries to generate a more reliable and accurate classification. This is accomplished by merging the feature representations from both models.

An algorithm that utilizes genetics can automatically determine the optimal combination of weights for deep learning models, thereby eliminating the need for manual weight assignment. In this method, the weights are viewed as genes within a population of possible solutions. An accuracy evaluation is performed on each solution, representing a collection of weights, based on how well it performs on a validation set. After that, the genetic algorithm will iteratively evolve the population by picking the solutions that perform the best (with the highest accuracy), using genetic operators such as mutation and crossover to create new solutions, and then repeating the evaluation process. Throughout several generations, the algorithm will eventually arrive at a collection of weights that will maximize the precision of the ensemble model when applied to the validation data. The genetic algorithm uses the principles of natural selection to conduct an intelligent exploration of the weight space. Compared with manual weight assignment, it enables the identification of the ideal combination, resulting in improved classification performance. To optimize the weight combination for the ensemble model, the following Algorithm 1 illustrates the evolutionary process used. This method involves evaluating each individual's fitness, depending on the quality of their classification accuracy.

Algorithm 1 Genetic algorithm for optimal ensemble weight selection.

Initialization: Generate an initial population of N candidate solutions, where each solution consists of weights w_1 and w_2 for Model 1 and Model 2, such that $w_1 + w_2 = 1$, and $w_1, w_2 \in [0, 1]$.
for each generation $g = 1$ to G **do**
    for each individual i in the population **do**
        Apply w_1^i and w_2^i to the ensemble model.
        Evaluate on the validation set.
        Set fitness $f_i = \text{accuracy}(w_1^i, w_2^i)$.
    end for
    Selection: Choose the top k individuals with the highest fitness (accuracy) as parents.
    Crossover: Generate offspring by combining the parent weights

$$w^{\text{child}} = \frac{w^{\text{parent1}} + w^{\text{parent2}}}{2}.$$

    Mutation: With the small probability p, slightly perturb the offspring weights to explore new solutions.
    Replacement: Form the new generation by selecting the best individuals from current and new populations.
end for
Output: Return the individual with the highest fitness (best validation accuracy).

To enhance generalization and robustness, we developed an ensemble of ResNet50 and EfficientNet outputs:

$$y_{\text{ensemble}} = w_1 y_{\text{ResNet}} + w_2 y_{\text{EffNet}}, \quad w_1 + w_2 = 1.$$

A genetic algorithm was used to optimize the weights w_1, w_2 with:

$$\text{fitness} = \text{accuracy}.$$

3.6.1. Genetic algorithm configuration

We employ the optimization of ensemble weights on the probability simplex. We do this by encoding an unrestricted vector $\mathbf{z}$ and transforming it to $\mathbf{w} = \text{softmax}(\mathbf{z})$, ensuring that $w_m \geq 0$ and $\sum_m w_m = 1$. The fitness score is the macro-F1-score derived from the validation split, which is logical, considering the class imbalance and clinical significance. To resolve the connections, we utilize sensitivity at a consistently elevated specificity. The tournament sampling approach, with a size of t=3, is employed for selection; one-point crossover is applied with a probability of p_c; and mutation is executed by introducing a zero-mean Gaussian noise to $\mathbf{z}$ with a probability of p_m. Elitism retains the top $e\%$ candidates from each generation, while early stopping terminates the process when the fitness fails to improve for a certain number of consecutive generations. In the federated context, each client executes this genetic algorithm on their own machines to obtain $\widehat{\mathbf{w}}_k$. The server uses FedAvg to obtain the base models and, optionally, executes a brief global genetic algorithm with $\{\widehat{\mathbf{w}}_k\}$ to derive the broadcast weight vector $\mathbf{w}^\star$. The attributes and default configurations of the genetic algorithm employed in this study are presented in Table 1, outlining the key factors that guided the optimization process.

Table 1. Attributes and defaults of the genetic algorithm used in this study.

Attribute	Symbol/option	Default/range	Purpose/notes
Population size	N	50 (30–80)	Exploration capacity of candidate weight vectors
Generations	G	100 (50–150)	Convergence depth; early stopping may end sooner
Initialization	$\mathbf{z}_0$	Small $\mathcal{N}(0, 0.1)$ plus uniform seeds	Diversity; seeds include a near-uniform $\mathbf{w}$
Selection	Tournament	size $t = 3$	Stable selection pressure under class imbalance
Crossover	One-point	$p_c = 0.8$	Mix parental solutions while preserving the structure
Mutation	Gaussian on $\mathbf{z}$	$p_m = 0.1$; $\sigma = 0.05$	Escape local optima; operates prior to softmax
Elitism	Top %	$e = 5\%$	Keep the best candidates each generation
Constraint handling	softmax	$\mathbf{w} = \mathrm{softmax}(\mathbf{z})$	Enforces $w_m \geq 0$ and $\sum_m w_m = 1$
Fitness metric	$J(\mathbf{w})$	macro-F1	Matches clinical preference and class imbalance
Early stopping	patience	10 generations	Prevents overfitting and excess compute
Random seed	`Seed`	Fixed (report)	Reproducibility of GA runs
Local optimization	$\widehat{\mathbf{w}}_k$	per client	Adapts to site-specific distributions
Server fusion (option)	$\mathbf{w}^\star$	GA-global or barycenter	Seeds with $\{\widehat{\mathbf{w}}_k\}$; broadcasts the final weights

3.7. Federated learning integration

We employed the FedAvg method to implement FL using distributed datasets while maintaining patients' data privacy. Multiple decentralized clients can train a shared global model using this technology without sharing raw data. This kind of plan is beneficial in healthcare, where data centralization is impossible due to its sensitive nature and governing rules.

In our system, each client i was responsible for maintaining a local dataset of size n_i that was up to date. Clients trained their local models with weights w_t^i over several local epochs using either stochastic gradient descent (SGD) or the Adam optimizer. This training happened every time there was a communication round t. After the local training, the most recent weights were sent to the central server. The server then used a weighted average to create a new global model:

$$w_{t+1} = \sum_{i=1}^{K} \frac{n_i}{n} w_t^i \quad \text{where} \quad n = \sum_{i=1}^{K} n_i.$$

This aggregation formula ensures that customers with larger datasets will have a significant impact on the global update. The process will continue iteratively until it reaches a point of convergence or a

certain number of communication rounds have been completed. Algorithm 2 below shows how to train FedAvg. This method utilizes a central server to coordinate decentralized clients in building a global model without sharing raw data.

Algorithm 2 FedAvg algorithm.

Input: Number of clients K, local epochs E, number of rounds T, initial global model w_0.
Output: Final global model w_T
for $t = 0$ to $T - 1$ **do**
    Server broadcasts the global model w_t to all clients
    for each client $i \in \{1, \ldots, K\}$ **in parallel do**
        Initialize $w_t^i \leftarrow w_t$
        Train w_t^i on local dataset $\mathcal{D}_i$ for E epochs using SGD
        Send updated model w_t^i to server
    end for
    // Server aggregates updates:
    $w_{t+1} \leftarrow \sum_{i=1}^{K} \frac{n_i}{n} w_t^i$, where $n = \sum_{i=1}^{K} n_i$
end for
return w_T

All clients used the Adam optimizer with a learning rate of $\eta = 0.001$, a batch size of 32, and a local training epoch count of $E = 5$. There were 10 rounds of communication during the global training.

In TensorFlow federated (TFF), we employed techniques such as dataset caching, shuffling, and batch prefetching to manage the substantial computational load associated with processing large amounts of image data in federated settings. We also ensured the model's stability by adding protections against non-finite updates, such as Not a Numbers or infinite values.

In conclusion, FL is a method for training high-performance deep learning models in areas where privacy is crucial, such as breast cancer ultrasound imaging. It is both possible and scalable.

3.7.1. Privacy model and why FL is unique in healthcare

We employ a threat model that is both candid and attentive to the interests of clients and servers. In medical AI, it is crucial to consider several key issues, including model alterations that lead to gradient leakage, membership inference from global snapshots, the introduction of poisons or backdoors, and the surveillance of channels. In FL, only modifications to the model are communicated, excluding raw images or labels. This renders FL a distinctive method for fulfilling data minimization requirements. We utilize a technique referred to as "secure aggregation" to ensure that the server receives only an aggregate of client changes. We also allow users to select the "DP-SGD" mode, which applies layer-wise clipping with Gaussian noise and reports (ε, δ) when activated. The transport layer employs the Transport Layer Security (TLS) to provide provenance and auditability, while the clients authenticate updates through signatures. These design-stage constraints enhance privacy and security without imposing additional burdens on the testers. They are particularly advantageous for healthcare facilities with different locations, where data centralization is unfeasible.

3.7.2. Security by design: Threat model and risk mitigations

Rather than providing actual attack benchmarks, which necessitate additional permissions, data governance stages, and computational resources, we present documentation of the security posture embedded within our design. The danger model posits that the server is honest yet inquisitive, whereas the clients may likewise exhibit curiosity. The hazards encompass revealing gradients, identifying members, introducing poisoning/backdoors, and eavesdropping on the channel. A crucial component of our privacy-preserving solution is data localization, meaning that no unprocessed images or labels may exit the premises. Additional capabilities include secure aggregation to maintain the confidentiality of updates and an optional DP-SGD mode that employs clipping and Gaussian noise to prevent information leakage. TLS encrypts transmissions, while the clients authenticate updates to ensure traceability and source identification. Table 2 illustrates the correlation between dangers and our corresponding management strategies within the system.

Table 2. Threat-informed security checklist (design time).

Threat	Mitigation in our framework
Update inspection / gradient leakage	Secure aggregation; optional DP-SGD with (ε, δ) accounting; raw data remain on-site
Membership inference on global model	DP-SGD option; limited model exposure; audit logging of model exports
Poisoning / backdoor attempts	Client update signing; anomaly screening on update norms; cross-round consistency checks
Adversarial inputs at inference	Operator guidance for thresholding; empirical attack evaluation noted as future work
Channel eavesdropping	TLS for client–server traffic

3.8. Deep learning-based segmentation of malignant cases

Segmentation is the last and most crucial step in our pipeline. The purpose is to identify the edges of lesions in images diagnosed as malignant. Segmentation provides doctors with pixel-level accuracy, allowing them to examine the shape, area, and regularity of the lesion's boundaries. These are essential signs in cancer diagnosis. Classification, on the other hand, gives a rough diagnosis.

We utilized the segmentation models PyTorch (SMP) module to aid in building the U-Net architecture, which we employed for malicious lesion segmentation. This design has been effective in biomedical imaging due to its encoder-decoder structure, skip connections, and ability to preserve both semantic and spatial information.

The encoder is the central component of the U-Net network, which is why it is sometimes referred to as the backbone. Backbones are a type of CNN that have already been trained to find hierarchical features in the images sent to them. These features can capture low-level details, such as edges and textures, as well as the high-level semantic information necessary for identifying the correct object. The backbone in our system is essential for the quality of the feature representations sent to the decoder part of the network. Using pre-trained backbones trained on large datasets, such as ImageNet, can facilitate transfer learning, accelerate convergence, and make medical imaging tasks more manageable with less training data.

We examined various convolutional backbones, including ResNet50, VGG16, DenseNet201, EfficientNet-B7, MobileOne-S4, Xception, and timm-efficientnet. We used our breast ultrasound dataset to fine-tune all of them. We selected these backbones because they have been demonstrated to perform well in classification and segmentation tasks. These backbones have unique structural features, such as depth, skip connections, or compound scaling, which alter how the features are extracted and retained.

When compared with the other models tested, the timm-efficientnet and ResNet50 models both performed better. The timm-efficientnet library is a high-capacity EfficientNet version pretrained on ImageNet21k. It is built from the timm (PyTorch image models) library. It can achieve accurate multi-resolution features due to its compound scaling and more thorough design. This makes it ideal for delineating complex lesion borders in ultrasound images. In addition to providing strong spatial encoding through residual connections, ResNet50 also shows high accuracy across several evaluation measures.

We made a hybrid model by combining the results of ResNet50 and EfficientNet. This was done to make the segmentation methods even more stable and reliable. To be more specific, each model makes its unique probability mask. These masks are then combined using a weighted average:

$$M = \alpha M_{\text{ResNet}} + (1 - \alpha) M_{\text{EffNet}}.$$

This fusion method leverages the best aspects of both designs: ResNet's structural sensitivity and EfficientNet's fine-grained encoding. At the same time, it minimizes the effects of each model's limitations.

The final segmentation map is refined through a convolutional layer followed by a sigmoid activation function:

$$M_{\text{final}} = \sigma(W * M + b).$$

This produces a pixel-wise probability map for lesions' presence.

Visual inspection of segmentation maps (Figure 2) confirmed that the model could reliably find the edges of lesions. This was true even in challenging situations with low contrast or different textures. This segmentation stage is the last step in building our end-to-end AI architecture. It accurately detects malignant lesions, making automated breast ultrasound analysis more valuable and reliable in clinical settings.

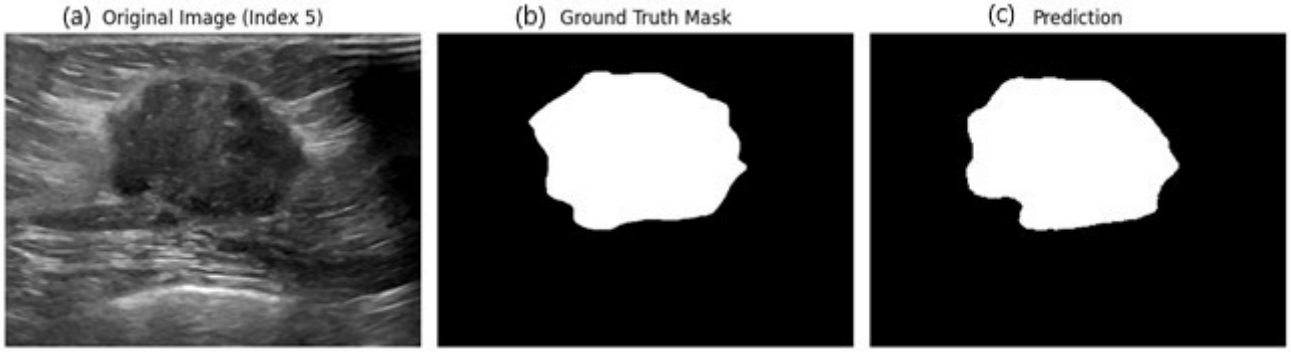

Figure 2. Qualitative segmentation example ((a) original ultrasound; (b) ground truth; (c) prediction). Shown as a prototype illustration.

4. Result and discussion

This section provides an overview of how we tested our proposed framework for classifying and segmenting breast cancer using ensemble deep learning and FL techniques. The evaluation is done using ultrasound imaging data, with a focus on two main classification tasks: (1) Finding normal and abnormal instances, and (2) sorting abnormal cases into groups that are either benign or malignant. We examine the performance of several deep learning models, including ResNet-50, EfficientNet-B7, VGG-16, and Xception. We also examine the performance of the ensemble model, which combines the outputs of these models using weighted averaging optimized by a GA. We also look at the FL setup to determine how to balance data's accuracy and privacy. We investigate the performance of U-Net-based architectures with different encoder backbones (such as ResNet50 and EfficientNet-L2) in identifying malignant lesions for segmentation. We use evaluation metrics like accuracy, precision, recall, F1-score, Mathews correlation coefficient (MCC), IoU, and the Dice coefficient to test the model's performance thoroughly. The results demonstrate that the proposed pipeline can be applied in clinical settings, as it shows how model fusion and decentralized training can enhance the robustness of classification and the accuracy of segmentation.

4.1. Dataset description

"Dataset-BUSI-with-GT" (breast ultrasound images with ground truth) is a standard dataset often used in research on classifying and segmenting breast cancer. This study's experimental evaluation is based on this dataset, available to the public at [*]. Al-Dhabyani et al. [76] made this dataset available, with 780 ultrasound images. We assigned these images into normal, benign, and malignant groups. A corresponding ground truth mask accompanies each image. This makes it easier to evaluate lesion boundaries on a pixel-by-pixel basis. This mask is used for tasks that involve separating benign and malignant groups.

This collection features black and white photos at various resolutions. These images depict various visual patterns, textures, and noise levels similar to those found in real-life clinical settings. There are three folders in the dataset, which are as follows:

- Normal/ – containing 133 images without tumor regions;
- Benign/ – containing 437 images and the corresponding binary masks;
- Malignant/ – containing 210 images and the corresponding binary masks.

The photos were resized to the exact resolution (128 * 128 or 224 * 224, depending on the model used), converted to RGB by duplicating channels, and then normalized to meet the preprocessing requirements of various deep learning architectures. This was done to utilize the data for training and testing purposes. We used a stratified method of 80:10:10 to divide the dataset into training, validation, and testing subsets, ensuring that there was no imbalance between the classes.

Because there are labelled masks and clinically relevant differences in lesions, this dataset is a good starting point for classification and segmentation research. Another helpful feature is that benchmarking works for centralized and FL situations.

Metadata limitation: The present study cannot conduct subgroup analyses based on density or size due to the lack of annotations concerning lesion size or breast density in the publicly available

[*]https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset/data

dataset employed. The aim is to do subgroup analyses in future cohorts with more extensive metadata. The segmentation example is presented to demonstrate the prototype's completeness, not to assert that the segmentation model has undergone comprehensive benchmarking. The primary objective is classification.

4.2. Evaluation metrics

To assess the models' performance, we employed a comprehensive set of evaluation metrics suitable for binary classification tasks. These the following.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.1}$$

Precision

$$Precision = \frac{TP}{TP + FP}. \tag{4.2}$$

Recall

$$Recall = \frac{TP}{TP + FN}. \tag{4.3}$$

F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{4.4}$$

MCC

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{4.5}$$

where:

- TP: True positives;
- TN: True negatives;
- FP: False positives;
- FN: False negatives.

The MCC returns a value between -1 and $+1$, where $+1$ indicates a perfect prediction, 0 represents random prediction, and -1 indicates total disagreement between the prediction and the observation.

IoU

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}, \tag{4.6}$$

where P stands for the predicted segmentation mask and G stands for the ground truth mask. The numerator in this equation shows how many pixels are shared by the prediction and the ground truth, and the denominator shows how many pixels are in either the prediction or the ground truth, p_o is the observed accuracy, and p_e is the expected accuracy by random chance.

These tools help us thoroughly understand the model's success, particularly concerning the aging class imbalance and ensuring strong detection performance in real-world environments.

4.3. Performance analysis of deep learning models for breast cancer classification

In this section, we discuss how well the deep learning classifiers perform with various types of models. The first step is to create baseline models and assess the performance of traditional algorithms before moving on to ensemble methods.

4.3.1. Classification of normal and pathological ultrasound images

This section presents the results of experiments that utilized various deep learning models to distinguish between normal and abnormal breast ultrasound images in a binary classification problem. Table 3 compares how well the different models classify things. This study uses accuracy, precision, recall, F1-score, area under the curve (AUC), balanced accuracy, and MCC as its metrics.

Table 3. Performance comparison of various deep learning models and basic ensemble learning for the classification of normal and abnormal breast lesions.

	Accuracy	Precision	Recall	F1-score	AUC	Balanced accuracy	MCC
Resnet50	**0.9067**	1.0000	0.4167	0.5882	0.9778	0.7083	0.6124
EfficientNetB7	**0.8400**	0.5000	0.4167	0.4545	0.8439	0.6687	0.3638
EfficientNetB0	0.8800	0.6667	0.5000	**0.5714**	0.8879	0.7262	0.5104
MobileNetV3Large	0.8733	0.8571	0.2500	0.3871	0.9229	0.6210	0.4207
InceptionV3	0.7200	0.3333	0.7500	0.4615	0.8042	0.7321	0.3546
VGG16	0.8733	0.8571	0.2500	0.3871	0.9378	0.6210	0.4207
Xception	0.8600	0.7143	0.2083	0.3226	0.8046	0.5962	0.3345
InceptionResNetV2	0.3800	0.2000	0.9583	0.3309	0.7464	0.6141	0.1978
Basic ensemble	**0.9400**	**1.0000**	**0.6250**	**0.7692**	**0.9911**	–	**0.7638**

The results of the classification tests for distinguishing between normal and abnormal breast ultrasound images reveal significant differences in the models' performance. These differences highlight the architectural strengths and weaknesses of the models when applied to medical imaging. ResNet50 achieves the highest overall accuracy (90.67%), the highest precision (1.000), and the highest MCC (0.6124) compared with the other models. It can accurately distinguish between the two groups because it has a high AUC score of 0.9778. However, its recall is still very low (0.4167), indicating that it prefers the regular class and may miss unusual cases.

EfficientNetB0 performs just as well as ResNet50, achieving an accuracy of 88.00%, a better recall of 0.5000, and a good MCC of 0.5104. Compared with its lighter version, EfficientNetB7, EfficientNetB7 does not perform as well. It has lower accuracy (84.00%) and MCC (0.3638), possibly because it was overfitted or not fine-tuned enough on the dataset. MobileNetV3Large and VGG16 have similar accuracy (87.33%) and precision (0.8571) scores. However, their low recall (0.2500) shows they are biased towards the majority class.

InceptionV3 and Xception have high recall values (0.7500 and 0.2083, respectively), but their precision and F1-scores are low, resulting in overprediction of the abnormal class and creation of false positives. InceptionResNetV2 has the highest recall of 0.9583 among all the models. This means that it can identify almost all unusual cases. Its accuracy, on the other hand, drops to 38.00%, which gives it the lowest MCC (0.1978) of all the models. This balance between sensitivity and specificity indicates a tendency to identify all samples as pathological.

Essential ensemble learning, which combines predictions from multiple models, achieves the best overall performance. The ensemble achieves impressive results, with an accuracy of 94.00%, perfect precision of 1.00, improved recall of 0.6250, and a strong F1-score of 0.7692. The AUC score of 0.9911 indicates that it can distinguish between things very effectively, and the MCC score of 0.7638

demonstrates that the predicted and actual labels agree very well. These results demonstrate that ensemble methods are beneficial because they combine the best features of different models to make them more robust and able to generalize.

In short, although each model has its own strengths and weaknesses regarding precision or recall, the ensemble approach is the only one that gives a more balanced and reliable performance. This demonstrates that ensemble learning is a suitable approach for our proposed framework and highlights the importance of balancing sensitivity and specificity in medical diagnostics, where even a single false negative could have severe clinical consequences.

4.3.2. Clinical interpretation (normal vs. abnormal)

The current confusion matrix indicates [TN=126, FP=0, FN=5, TP=19]. The GA-optimized ensemble (ResNet50: EffNetB0 = 0.7:0.3) has a precision of 1.000, a negative predictive value of around 0.962, a sensitivity of roughly 0.7917, a specificity of 1.000, an accuracy of approximately 0.9667, and an AUC of 0.9911. In screening scenarios when it is crucial to exclude disease, an operational point with a precision of 1.0 eliminates false positives. The drawback is that the recall is approximately 0.79, indicating that around 20.8 out of every 100 aberrant instances will be overlooked. To enhance memory for triage operations when the AUC approaches 1, we should adjust the thresholds to the left to get greater sensitivity.

4.3.3. Malignancy classification of detected abnormalities

This section aims to discuss the performance of deep learning models in the binary classification task of distinguishing between benign and malignant breast lesions using ultrasound images. Table 4 shows the evaluation results, which use various performance metrics to show how well each model works.

Table 4. Performance comparison of various deep learning models and basic ensemble learning for the classification of benign and malignant breast lesions.

	Accuracy	Precision	Recall	F1-score	AUC	Balanced accuracy	MCC
Resnet50	0.8333	0.9565	0.5238	0.6769	0.9048	0.7560	0.6247
EfficientNetB7	0.7619	0.7727	0.4048	0.5312	0.8455	0.6726	0.4287
EfficientNetB0	0.7857	0.7419	0.5476	0.6301	0.8260	0.7262	0.4951
MobileNetV3Large	0.8413	0.8667	0.6190	0.7222	0.9351	0.7857	0.6325
InceptionV3	0.6984	0.5588	0.4524	0.5000	0.7526	0.6369	0.2908
VGG16	0.8730	0.8611	0.7381	0.7949	0.9368	0.8393	0.7081
Xception	0.7937	0.6905	0.6905	0.6905	0.8231	0.7679	0.5357
InceptionResNetV2	0.3333	0.3333	1.0000	0.5000	0.5060	0.5000	0.0000
Ensemble learning	**0.9206**	**0.9444**	**0.8095**	**0.8718**	**0.9674**	–	**0.8199**

One of the most significant problems in medical imaging is determining the difference between benign and malignant breast lesions. This choice has substantial effects on the healthcare system. The tested models have various strengths in identifying malignant cases. Ensemble learning always gives the best results.

The VGG16 model performs the best among all the models, achieving an accuracy of 87.30%, an F1-score of 0.7949, and an MCC of 0.7081. These numbers indicate that it can accurately recognize both classes in a balanced manner, primarily due to its high recall of 0.7381, which is crucial for identifying as many cancers as possible. MobileNetV3Large also performs well, with a slightly lower accuracy (84.13%) but good generalization, as shown by a recall of 0.6190 and an MCC of 0.6325, indicating that the predictions are generally reliable. MobileNetV3Large also does a good job.

ResNet50 has a low recall (0.5238), but it achieves a high level of precision (0.9565), indicating that it generates few false positives. This is because it is well-known for its powerful spatial feature extraction capabilities. This demonstrates a careful approach to classification, leaning towards benign predictions, which could result in malignant cases going unnoticed. EfficientNetB0 also strikes a good balance, with an accuracy of 78.57%, a recall of 0.5476, and an MCC of 0.4951. EfficientNetB7 performs poorly, with the lowest F1-score among all EfficientNet variations. This could be because it is easy to fine-tune.

Interestingly, Xception has a symmetrical performance profile, with a precision and recall of 0.6905. This makes it a reliable choice for use in the real world, where consistency is essential in all classes. InceptionV3 performs well for many visual tasks, but not in this case. It does not perform an excellent job of distinguishing between different types of lesions, as evidenced by its low accuracy (69.84%), F1-score (0.5000), and MCC (0.2908). InceptionResNetV2 could recall samples 1000 perfectly, but it did so at the cost of a 33% accuracy and a zero MCC. This means that it incorrectly identified almost all samples as malignant, resulting in numerous false positives.

The ensemble learning method has the best overall performance, with an accuracy of 92.06%, a precision of 0.9444, a recall of 0.8095, an F1-score of 0.8718, and a AUC of 0.9674. This is the result that stands out the most. When diagnosing cancer, finding a balance between accuracy and memory is essential. This model combines the best aspects of various architectures, enhancing recall without compromising accuracy. Its MCC value of 0.8199 shows that the projected and actual classes are very similar.

Ensemble learning makes the classification problem much more reliable and accurate, even though different models capture different aspects. Ultimately, ensemble learning represents a significant step forward. To minimize errors during breast cancer screening, clinical AI pipelines must achieve both high sensitivity (recall) and specificity. These results demonstrate the critical importance of hybrid methods in this case.

4.3.4. Clinical interpretation (benign vs. malignant)

The GA-weighted ensemble exhibits an accuracy of approximately 0.9603, a negative predictive value of roughly 0.965, a sensitivity of approximately 0.9286, a specificity of about 0.9762, and an AUC of around 0.9674, as derived from the confusion matrix [TN = 82, FP = 2, FN = 3, TP = 39]. At this operational juncture, the expected number of overlooked malignant occurrences is around 100*(1 - 0.9286), equating to about 7.1 per 100 malignant cases. This represents an improved equilibrium for assisting in diagnostic determinations. Clinicians may choose varying criteria to optimize sensitivity and specificity, contingent upon the requisite number of biopsies, the prevalence of the ailment, and safety standards.

4.4. Performance analysis of weight selection for deep learning model fusion

We used a late-fusion method that combined predictions from ResNet50 and EfficientNet using a weighted average. This enhanced the classification's reliability and leveraged the strengths of various CNN architectures. We employed a GA to identify the optimal combinations of weights that yielded the best possible classification performance. The GA improves a multi-objective fitness function by determining accuracy and selecting the optimal weights. This function encompasses the model's ability to distinguish, be sensitive, and make balanced predictions.

4.4.1. Evaluation on normal vs. abnormal classification

This subsection presents the results of applying the genetic algorithm-guided weight selection method to combine ResNet50 and EfficientNetB0 for the binary classification task. The challenging part is distinguishing between normal and abnormal cases. Table 5 summarizes the evaluation metrics collected when different combinations of fusion weight measurements were used.

Table 5. Classification performance for normal vs. abnormal detection using different weight combinations of ResNet50 and EfficientNetB0.

ResNet weight	EfficientNet weight	Accuracy	Precision	Recall	F1-score	MCC
0.024	0.976	0.8733	0.7273	0.3333	0.4571	0.4353
0.401	0.599	0.9533	1.0000	0.7083	0.8293	0.8192
0.845	0.155	0.9200	1.0000	0.5000	0.6667	0.6757
0.881	0.119	0.9533	0.9048	0.7917	0.8444	0.8197
0.711	**0.289**	**0.9667**	**1.0000**	**0.7917**	**0.8837**	**0.8726**

The GA examined various combinations of ResNet and EfficientNet weights to determine which ones were normal and which were not. ResNet made up 71.1% of the ensemble output, and EfficientNet made up 28.9%. This mix yielded the best accuracy (96.67%). This setup resulted in an F1-score of 0.8837 and an MCC of 0.8726, which is excellent. It also had perfect precision (1.0000) and a high recall (0.7917). These results suggest a good balance between false positives and false negatives. This is particularly important for screening activities and situations where failing to identify atypical cases could be very harmful. It was confirmed that EfficientNet alone could not adequately capture all discriminative features, as a lower ResNet weight (for example, 0.024) resulted in a significant drop in recall and F1-score. High ResNet weights (like 0.845–0.881), on the other hand, improved memory and MCC, but they were at their best when those weights were balanced around the 70–30 range. This demonstrates that both models present valid yet distinct perspectives.

4.5. Evaluation on benign vs. malignant classification

This subsection summarizes the effectiveness of the weight fusion technique in addressing the benign versus malignant classification problem. It is based on GAs. Table 6 shows a complete analysis of how different weight distributions affect the classification results.

Table 6. Classification performance for benign vs. malignant detection using different weight combinations of ResNet50 and EfficientNetB0.

ResNet weight	EfficientNet weight	Accuracy	Precision	Recall	F1-score	MCC
0.02	0.98	0.8254	0.8846	0.5476	0.6765	0.5963
0.358	0.642	0.8889	0.9118	0.7381	0.8158	0.7459
0.878	0.122	0.8571	1.0000	0.5714	0.7273	0.6860
0.806	0.194	0.8730	0.8611	0.7381	0.7949	0.7081
0.700	**0.300**	**0.9603**	**0.9512**	**0.9286**	**0.9398**	**0.9103**

When classifying benign and malignant tumours, the best fusion again favoured a more substantial contribution from ResNet (70 %) and a minor contribution from EfficientNet (30 %). This resulted in an F1-score of 0.9398, a precision of 0.9512, a recall of 0.9286, and an accuracy of 96.03%. The MCC for this value of 0.9103 shows a strong agreement between the predictions and the actual data. This is better than any one model. Other combinations of weights that were examined revealed different trade-offs. For example, a ResNet weight of 0.358 provided a good balance (accuracy 88.89%, F1-score 0.8158), but a weight of 0.98 for EfficientNet resulted in a drop in recall and F1-Score. This demonstrates the importance of ResNet's robust spatial encoding in detecting cancers.

These results demonstrate that ResNet50 and EfficientNet possess complementary features, and that combining them via genetic optimization not only enhances performance but also addresses the issues inherent in separate networks. This is why choosing weights based on the GA is essential to our ensemble method. This is especially true for sensitive tasks like diagnosing cancer, where both false positives and false negatives have a lot of clinical weight. By using this method, we achieved better results in both binary classification stages. This indicates that it can also be applied in other medical imaging workflows.

4.6. Comparative analysis with state-of-the-art approaches

The results in Table 7 demonstrate that our proposed FL framework for breast cancer classification outperforms many other methods considered to be the best. In a FL environment, our method utilizes optimized weight selection and feature selection algorithms to incorporate deep learning classifiers. According to what we learned from existing centralized and hybrid frameworks, the model demonstrated good accuracy in classification, outperforming the results of other frameworks.

Table 7. Comprehensive evaluation: FL with optimized weight selection and classifier integration.

Dataset		
	Methods	Accuracy (%)
Proposed model	Ensemble learning based on weight selection	96
Anwar et al. [77]	Ensemble model based on CNN and random forest	94
Sahu et al. [78]	Hybrid framework	93.18
Sahu et al. [79]	Homogeneous ensemble	94.62
Bianca et al. [80]	CNN	86
Rai et al. [81]	Real+GAN	92

Compared with the ensemble CNN and random forest model of Anwar et al., [77] (94%) and the hybrid framework of Sahu et al., [78] (93.18%), our FL-based model does a much better job of making predictions. These two models both use CNN and random forest in different ways. It is especially impressive that Sahu et al. [79] used a homogeneous ensemble technique with an accuracy of 94.62%. On the other hand, Bianca et al. [80] got an accuracy of 86% by using only a CNN. These comparisons highlight the limitations of traditional pipelines, which rely on a single model, and underscore the importance of utilizing a diverse range of classifiers and adaptive weighting methods.

Rai et al. [81] developed a GAN-based synthetic data augmentation strategy that achieved 92% accuracy. This is a good start, but less accurate than our federated architecture. Unlike these centralized methods, our model operates in a way that ensures that people's privacy remains safe. It spreads the training process across fake customers without sharing private information, ensuring that it is accurate and adheres to modern healthcare data governance rules.

The performance improvement that was seen can be traced back to two main factors: (1) The use of a genetically optimized weighting scheme to combine multiple classifiers, like ResNet50 and EfficientNetB0, and (2) the use of a GA for weight selection to cut down on redundancy and computational overhead while keeping the ability to tell the difference between things. As a result, our model was better able to generalize across decentralized data divisions, which made learning more efficient and less prone to overfitting.

In conclusion, the results demonstrate that our federated ensemble learning architecture can achieve high classification accuracy while maintaining data privacy. This is a critical need for use in real-life medical settings. This new development makes it possible for breast cancer testing to be scalable and safe and done by healthcare institutions spread out over a large area.

4.7. Detection of breast cancer based on deep learning models

Finding and separating breast cancer lesions correctly is significant for making clinical decisions. This is because it allows radiologists to examine the tumor's shape, size, and edges, which helps them make a diagnosis, plan treatment, and monitor its progress. This study aimed to evaluate several deep learning models and ensemble methods to assess their ability to distinguish between cancerous and non-cancerous lesions in breast ultrasound images. We examined key performance metrics, including accuracy, precision, F1-score, recall, and IoU. Table 8 shows the results gathered for this work.

Table 8. Performance comparison of various deep learning models and ensemble learning methods for the breast cancer detection.

	Accuracy	Precision	F1-score	Recall	IoU
resnet50	0.9738	0.9806	0.9291	0.8828	0.8676
VGG16	0.9695	0.9824	0.9163	0.8585	0.8455
efficientnet-b7	0.9767	0.9697	0.9381	0.9084	0.8833
mobileone_s4	0.9797	0.9417	0.9481	0.9545	0.9012
timm-efficientnet-l2	0.9794	0.9635	0.9461	0.9293	0.8977
densenet201	0.9697	0.9573	0.9188	0.8833	0.8498
xception	0.9737	0.9636	0.9300	0.8986	0.8691
Ensemble learning	0.9742	0.9665	0.9311	0.8981	0.8710

All models did well overall, but MobileOne-S4 and EfficientNet-L2 (both from the timm library) stood out as the best. MobileOne-S4 had the highest accuracy (97.97%), F1-score (0.9481), and recall (0.9545). This demonstrates that it can identify most lesion pixels while maintaining a low rate of false negatives. It excels at defining lesion contours, as evidenced by its IoU of 0.9012, which demonstrates a substantial spatial overlap with the ground truth annotations.

Timm-EfficientNet-L2 comes in a close second with an accuracy of 97.94%, a high recall of 0.9293, and the second-best F1-score of 0.9461. It also has an outstanding IoU of 0.8977. On the basis of these results, we can conclude that EfficientNet-L2 can accurately capture small amounts of spatial and contextual data, which aids in accurate lesion boundary prediction. EfficientNet-B7 also generalizes significantly, with the third-highest recall (0.9084) and F1-score (0.9381). This suggests that compound-scaled architectures are effective for medical segmentation.

The accuracy of ResNet50 and Xception, which exceeds 97.3 %, their F1-scores, which are close to 0.93; and their recall, which exceeds 0.88, all indicate that they are effective at capturing lesion-specific patterns. The recall and IoU values for VGG16 and DenseNet201 are slightly lower than those of other neural networks (for example, for VGG16, recall = 0.8585 and IoU = 0.8455). This means that these neural networks may be unable to segment lesions or discern finer boundary details, even though they are highly accurate.

It is interesting to note that the ensemble learning strategy combines the results of many models to create a balanced performance. This gives an accuracy of 97.42%, an F1-score of 0.9311, a recall of 0.8981, and an IoU of 0.8710. The ensemble's ability to make the model more resilient and less likely to overfit specific patterns is demonstrated by the fact that it performs well on all metrics, even though it does not consistently outperform the best individual model on every parameter. Due to this, it is an excellent choice for use in real-world clinical workflows, where consistency is crucial.

MobileOne-S4 and EfficientNet-L2 are the best choices for detecting lesions, although other models are also very effective at segmentation. The ensemble model also appears to be helpful because it provides a solution that is both stable and can be applied in various situations. These results show how important it is to have a variety of models and a pleasing architectural design in the field of medical segmentation when it comes to diagnosing breast cancer. They also emphasize the importance of striking a balance between recall and precision to ensure that lesion mapping is accurate and trustworthy, as illustrated in Figure 3, which presents the segmentation results.

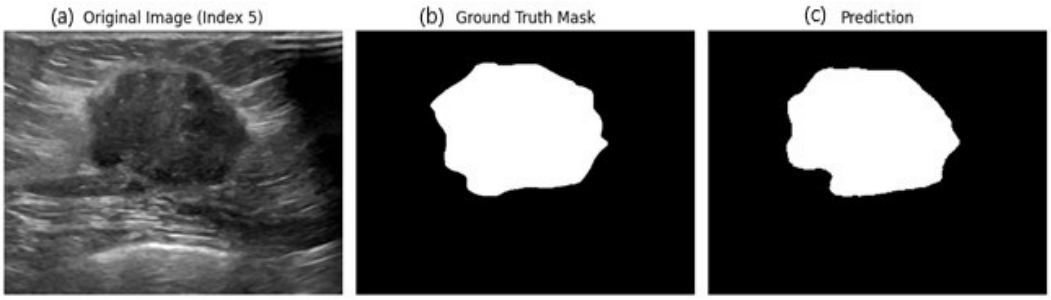

Figure 3. Breast cancer detection.

5. Conclusions

This study aims to develop a more comprehensive and multi-phase learning framework that will enhance breast cancer diagnosis using ultrasound imaging while maintaining patients' data privacy through FL. The framework combines advanced preprocessing, augmentation, classification, ensemble fusion, and segmentation components to make a single, therapeutically beneficial and technologically cutting-edge pipeline. We can demonstrate the ability to classify breast ultrasound images into normal, benign, and malignant classes with high accuracy using cutting-edge CNN architectures such as ResNet50, EfficientNetB7, VGG16, and Xception.

Using a genetic algorithm to optimize an ensemble learning technique helped us make our models stronger and more accurate at making predictions. This helped us overcome the limitations of using only one model. Additionally, using the FedAvg algorithm for FL enabled the training of models across simulated clients in a distributed manner without sharing sensitive patient data during the process. Our research shows that FL only slightly improves classification performance compared with regular centralized training. However, it enhances privacy and compliance with laws regarding medical data.

The segmentation part of our method provides accurate lesion boundaries for malignant cases, making our technology significantly more helpful in the clinic. This is important for both planning treatment and evaluating follow-up care. The combination of ResNet50 and EfficientNetB7's features for segmentation, followed by sigmoid activation and evaluation using Dice and IoU metrics, demonstrates the accuracy of our method.

Our tests showed that the suggested system could achieve an accuracy rate of 0.96 for the classification stage and 0.97 for the segmentation stage. This indicates that our technology performs exceptionally well and is highly reliable.

In conclusion, our architecture demonstrates that deep learning and federated computing can be effectively combined in a medical setting to achieve high accuracy, protect patients' data, and facilitate real-time clinical decision-making. This is important because medicine is a very delicate field. In the future, work will include field deployment on edge devices. This work represents the first step toward developing AI systems that are scalable, secure, and easy to understand in medical imaging.

Our design emphasizes privacy through construction via FL and secure aggregation, with an optional differentially private training mode. Empirical attack benchmarking is left for future work.

We intend to (i) establish cohorts enriched with metadata to facilitate subgroup analyses based on breast density and lesion size; (ii) implement personalized FL (such as FedBN/FedPer) to address the persistent shift in inter-client distribution; (iii) employ end-to-end DP-SGD with designated privacy budgets and utility-privacy curves; (iv) assess uncertainty and utilize calibration-aware triage thresholds; (v) adopt resource-aware deployment strategies (quantization/pruning) with latency and energy profiling on edge devices; and (vi) conduct a human-AI reader study to evaluate time-to-decision and sensitivity enhancement.

Author contributions

Karim Gasmi: Conceptualization, methodology, software, formal analysis, resources, writing – original draft preparation, visualization, funding acquisition; Ibtihel Ben Ltaifa: Conceptualization, formal analysis, data curation, writing – original draft preparation; Moez Krichen: Methodology,

writing – original draft preparation; Shahzad Ali: Software; Omer Hamid: Formal analysis, resources, data curation, data curation, writing – review and editing, visualization; Mohamed O. Altaieb: Validation, formal analysis, data curation, writing – review and editing; Lassaad Ben Ammar: Software, validation, writing – review and editing, visualization; Manel Mrabet: Validation, resources, writing – review and editing, visualization; Mahmood Mohamed: Validation, writing – review and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare that they have not used AI tools in the creation of this article.

Data availability

The data used in this study are openly accessible at the following link: `https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset/data`

Acknowledgments

This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under Grant No. (DGSSR-2025-FC-01067).

Conflict of interest

All authors declare that there are no competing interests.

References

1. M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, et al., Current and future burden of breast cancer: Global statistics for 2020 and 2040, *Breast*, **66** (2022), 15–23. https://doi.org/10.1016/j.breast.2022.08.010
2. M. Krichen, M. S. Abdalzaher, Performance enhancement of artificial intelligence: A survey, *J. Netw. Comput. Appl.*, **232** (2024), 104034. https://doi.org/10.1016/j.jnca.2024.104034
3. S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: A review, *Artif. Intell. Rev.*, **54** (2021), 137–178. https://doi.org/10.1007/s10462-020-09854-1
4. N. I. Yassin, S. Omran, E. M. El Houby, H. Allam, Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review, *Comput. Meth. Prog. Bio.*, **156** (2018), 25–45. https://doi.org/10.1016/j.cmpb.2017.12.012
5. K. Suzuki, Overview of deep learning in medical imaging, *Radiol. Phys. Technol.*, **10** (2017), 257–273. https://doi.org/10.1007/s12194-017-0406-5
6. Y. Qiu, S. Yan, R. R. Gundreddy, Y. Wang, S. Cheng, H. Liu, et al., A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology, *J. X-Rray Sci. Technol.*, **25** (2017), 751–763. https://doi.org/10.3233/xst-16226

7. M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, et al., Automated breast ultrasound lesions detection using convolutional neural networks, *IEEE J. Biomed. Health Inform.*, **22** (2018), 1218–1226. https://doi.org/10.1109/jbhi.2017.2731873
8. H. Ju, Y. Cui, Q. Su, L. Juan, B. Manavalan, CODENET: A deep learning model for COVID-19 detection, *Comput. Biol. Med.*, **171** (2024), 108229. https://doi.org/10.1016/j.compbiomed.2024.108229
9. H. Ghazouani, Multi-residual attention network for skin lesion classification, *Biomed. Signal Process. Control*, **103** (2025), 107449. https://doi.org/10.1016/j.bspc.2024.107449
10. K. Gasmi, A. Alyami, O. Hamid, M. O. Altaieb, O. R. Shahin, L. Ben Ammar, et al., Optimized hybrid deep learning framework for early detection of Alzheimer's disease using adaptive weight selection, *Diagnostics*, **14** (2024), 2779. https://doi.org/10.3390/diagnostics14242779
11. P. C. Gøtzsche, Beyond randomized controlled trials: Organized mammographic screening substantially reduces breast carcinoma mortality, *Cancer*, **94** (2002), 578. https://doi.org/10.1002/cncr.10224
12. A. Chong, S. P. Weinstein, E. S. McDonald, E. F. Conant, Digital breast tomosynthesis: Concepts and clinical practice, *Radiology*, **292** (2019), 1–14. https://doi.org/10.1148/radiol.2019180760
13. S. V. Sree, E. Y. K. Ng, R. U. Acharya, O. Faust, Breast imaging: A survey, *World J. Clin. Oncol.*, **2** (2011), 171–178. https://doi.org/10.5306/wjco.v2.i4.171
14. L. Fass, Imaging and cancer: A review, *Mol. Oncol.*, **2** (2008), 115–152. https://doi.org/10.1016/j.molonc.2008.04.001
15. K. M. Brindle, Molecular imaging using magnetic resonance: New tools for the development of tumour therapy, *Brit. J. Radiol.*, **76** (2003), S111–S117. https://doi.org/10.1259/bjr/50577981
16. M. Salhab, W. Al Sarakbi, K. Mokbel, The evolving role of the dynamic thermal analysis in the early detection of breast cancer, *Int. Semin. Surg. Oncol.*, **2** (2005), 8. https://doi.org/10.1186/1477-7800-2-8
17. D. Q. Zeebaree, A. M. Abdulazeez, D. A. Zebari, H. Haron, H. N. A. Hamed, Multi-level fusion in ultrasound for cancer detection based on uniform LBP features, *Comput. Mater. Contin.*, **66** (2021), 3363–3382. https://doi.org/10.32604/cmc.2021.013314
18. P. Pathak, A. S. Jalal, R. Rai, Breast cancer image classification: A review, *Curr. Med. Imaging*, **17** (2021), 720–740. https://doi.org/10.2174/0929867328666201228125208
19. M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, Breast cancer classification using machine learning, In: *2018 Electric electronics, computer science, biomedical engineerings' meeting (EBBT)*, Istanbul: IEEE, 2018, 1–4. https://doi.org/10.1109/ebbt.2018.8391453
20. B. S. Abunasser, M. R. J. Al-Hiealy, I. S. Zaqout, S. S. Abu-Naser, Convolution neural network for breast cancer detection and classification using deep learning, *Asian Pac. J. Cancer Prev.*, **24** (2023), 531–544. https://doi.org/10.31557/apjcp.2023.24.2.531
21. S. Murugan, B. M. Kumar, S. Amudha, Classification and prediction of breast cancer using linear regression, decision tree and random forest, In: *2017 International conference on current trends in computer, electrical, electronics and communication (CTCEEC)*, Mysore: IEEE, 2017, 763–766. https://doi.org/10.1109/ctceec.2017.8455058

22. P. Bhuvaneswari, A. B. Therese, Detection of cancer in lung with k-nn classification using genetic algorithm, *Proc. Mater. Sci.*, **10** (2015), 433–440. https://doi.org/10.1016/j.mspro.2015.06.077

23. V. J. Kadam, S. M. Jadhav, K. Vijayakumar, Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression, *J. Med. Syst.*, **43** (2019), 263. https://doi.org/10.1007/s10916-019-1397-z

24. Y. A. Hamad, K. Simonov, M. B. Naeem, Breast cancer detection and classification using artificial neural networks, In: *2018 1st Annual international conference on information and sciences (AiCIS)*, Fallujah: IEEE, 2018, 51–57. https://doi.org/10.1109/aicis.2018.00022

25. M. F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, *Expert Syst. Appl.*, **36** (2009), 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009

26. S. K. Prabhakar, H. Rajaguru, Performance analysis of breast cancer classification with softmax discriminant classifier and linear discriminant analysis, In: *Precision medicine powered by health and connected health*, 2017, 197–201. https://doi.org/10.1007/978-981-10-7419-6_33

27. S. ÖZŞEN, R. Ceylan, Comparison of AIS and fuzzy c-means clustering methods on the classification of breast cancer and diabetes datasets, *Turk. J. Elec. Eng. Comp. Sci.*, **22** (2014), 1241–1254. https://doi.org/10.3906/elk-1210-62

28. M. Kumar, S. Singhal, S. Shekhar, B. Sharma, G. Srivastava, Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning, *Sustainability*, **14** (2022), 13998. https://doi.org/10.3390/su142113998

29. M. Nasser, U. K. Yusof, Deep learning based methods for breast cancer diagnosis: A systematic review and future direction, *Diagnostics*, **13** (2023), 161. https://doi.org/10.3390/diagnostics13010161

30. V. Nemade, S. Pathak, A. K. Dubey, A systematic literature review of breast cancer diagnosis using machine intelligence techniques, *Arch Computat. Methods Eng.*, **29** (2022), 4401–4430. https://doi.org/10.1007/s11831-022-09738-3

31. C. Cong, X. Li, C. Zhang, J. Zhang, K. Sun, L. Liu, et al., MRI-based breast cancer classification and localization by multiparametric feature extraction and combination using deep learning, *J. Magn. Reson. Imaging*, **59** (2024), 148–161. https://doi.org/10.1002/jmri.28713

32. H. U. Rashid, T. Ibrikci, S. Paydaş, F. Binokay, U. Çevik, Analysis of breast cancer classification robustness with radiomics feature extraction and deep learning techniques, *Expert Syst.*, **39** (2022), e13018. https://doi.org/10.1111/exsy.13018

33. G. Altan, Breast cancer diagnosis using deep belief networks on ROI images, *Pamukkale Ünive. Müh. Bilim. Derg.*, **28** (2022), 286–291. https://doi.org/10.5505/pajes.2021.38668

34. N. M. ud din, R. A. Dar, M. Rasool, A. Assad, Breast cancer detection using deep learning: Datasets, methods, and challenges ahead, *Comput. Biol. Med.*, **149** (2022), 106073. https://doi.org/10.1016/j.compbiomed.2022.106073

35. X. Wang, I. Ahmad, D. Javeed, S. A. Zaidi, F. M. Alotaibi, M. E. Ghoneim, et al., Intelligent hybrid deep learning model for breast cancer detection, *Electronics*, **11** (2022), 2767. https://doi.org/10.3390/electronics11172767

36. Z. Jafari, E. Karami, Breast cancer detection in mammography images: A CNN-based approach with feature selection, *Information*, **14** (2023), 410. https://doi.org/10.20944/preprints202305.2209.v1

37. E. L. Omonigho, M. David, A. Adejo, S. Aliyu, Breast cancer: Tumor detection in mammogram images using modified alexnet deep convolution neural network, In: *2020 International conference in mathematics, computer engineering and computer science (ICMCECS)*, Ayobo: IEEE, 2020, 1–6. https://doi.org/10.1109/icmcecs47690.2020.240870

38. Y. Zhang, Y. L. Liu, K. Nie, J. Zhou, Z. Chen, J. H. Chen, et al., Deep learning-based automatic diagnosis of breast cancer on MRI using mask R-CNN for detection followed by ResNet50 for classification, *Acad. Radiol.*, **30** (2023), S161–S171. https://doi.org/10.1016/j.acra.2022.12.038

39. Q. A. Al-Haija, G. F. Manasra, Development of breast cancer detection model using transfer learning of residual neural network (resnet-50), *Am. J. Sci. Eng.*, **1** (2020), 30–39.

40. A. R. Bushara, R. V. Kumar, S. S. Kumar, An ensemble method for the detection and classification of lung cancer using computed tomography images utilizing a capsule network with visual geometry group, *Biomed. Signal Process. Control*, **85** (2023), 104930. https://doi.org/10.1016/j.bspc.2023.104930

41. R. Rajakumari, L. Kalaivani, Breast cancer detection and classification using deep CNN techniques, *Intell. Autom. Soft Comput.*, **32** (2022), 1089–1107. https://doi.org/10.32604/iasc.2022.020178

42. S. Khan, N. Islam, Z. Jan, I. U. Din, J. J. C. Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, *Pattern Recogn. Lett.*, **125** (2019), 1–6. https://doi.org/10.1016/j.patrec.2019.03.022

43. S. Al Garea, S. Das, Image segmentation methods: Overview, challenges, and future directions, In: *2024 Seventh international women in data science conference at Prince Sultan University (WiDS PSU)*, Riyadh: IEEE, 2024, 56–61. https://doi.org/10.1109/wids-psu61003.2024.00026

44. A. Abo-El-Rejal, S. E. Ayman, F. Aymen, Advances in breast cancer segmentation: A comprehensive review, *Acadlore Trans. Mach. Learn.*, **3** (2024), 70–83. https://doi.org/10.56578/ataiml030201

45. I. R. B. Godoy, R. P. Silva, T. C. Rodrigues, A. Y. Skaf, A. de C. Pochini, A. F. Yamada, Automatic MRI segmentation of pectoralis major muscle using deep learning, *Sci. Rep.*, **12** (2022), 5300. https://doi.org/10.1038/s41598-022-09280-z

46. E. Michael, H. Ma, H. Li, F. Kulwa, J. Li, Breast cancer segmentation methods: Current status and future potentials, *Biomed Res. Int.*, **2021** (2021), 9962109. https://doi.org/10.1155/2021/9962109

47. S. Gu, Y. Ji, Y. Chen, J. Wang, J. U. Kim, Study on breast mass segmentation in mammograms, In: *2015 3rd International conference on computer, information and application*, Yeosu: IEEE, 2015, 22–25. https://doi.org/10.1109/cia.2015.13

48. N. Karunanayake, S. Moodleah, S. S. Makhanov, Edge-driven multi-agent reinforcement learning: A novel approach to ultrasound breast tumor segmentation, *Diagnostics*, **13** (2023), 3611. https://doi.org/10.3390/diagnostics13243611

49. H. Su, F. Liu, Y. Xie, F. Xing, S. Meyyappan, L. Yang, Region segmentation in histopathological breast cancer images using deep convolutional neural network, In: *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, USA: IEEE, 2015, 55–58. https://doi.org/10.1109/isbi.2015.7163815

50. N. M. Zaitoun, M. J. Aqel, Survey on image segmentation techniques, *Procedia Comput. Sci.*, **65** (2015), 797–806. https://doi.org/10.1016/j.procs.2015.09.027

51. L. Yu, W. Min, S. Wang, Boundary-aware gradient operator network for medical image segmentation, *IEEE J. Biomed. Health Inform.*, **28** (2024), 4711–4723. https://doi.org/10.1109/jbhi.2024.3404273

52. P. Ma, H. Yuan, Y. Chen, H. Chen, G. Weng, Y. Liu, A Laplace operator-based active contour model with improved image edge detection performance, *Digit. Signal Process.*, **151** (2024), 104550. https://doi.org/10.1016/j.dsp.2024.104550

53. Z. Xu, X. Ji, M. Wang, X. Sun, Edge detection algorithm of medical image based on Canny operator, *J. Phys.: Conf. Ser.*, **1955** (2021), 012080. https://doi.org/10.1088/1742-6596/1955/1/012080

54. S. Jardim, J. António, C. Mora, Image thresholding approaches for medical image segmentation-short literature review, *Procedia Comput. Sci.*, **219** (2023), 1485–1492. https://doi.org/10.1016/j.procs.2023.01.439

55. N. S. Hassan, A. M. Abdulazeez, D. Q. Zeebaree, D. A. Hasan, Medical images breast cancer segmentation based on K-means clustering algorithm: A review, *Asian J. Res. Comput. Sci.*, **9** (2021), 23–28. https://doi.org/10.9734/ajrcos/2021/v9i130212

56. S. B. Shokouhi, A. Fooladivanda, N. Ahmadinejad, Computer-aided detection of breast lesions in DCE-MRI using region growing based on fuzzy C-means clustering and vesselness filter, *EURASIP J. Adv. Signal Process.*, **2017** (2017), 39. https://doi.org/10.1186/s13634-017-0476-x

57. C. Gallego-Ortiz, A. L. Martel, A graph-based lesion characterization and deep embedding approach for improved computer-aided diagnosis of nonmass breast MRI lesions, *Med. Image Anal.*, **51** (2019), 116–124. https://doi.org/10.1016/j.media.2018.10.011

58. S. Gu, Y. Chen, F. Sheng, T. Zhan, Y. Chen, A novel method for breast mass segmentation: from superpixel to subpixel segmentation, *Mach. Vis. Appl.*, **30** (2019), 1111–1122. https://doi.org/10.1007/s00138-019-01020-0

59. A. E. Ilesanmi, O. P. Idowu, S. S. Makhanov, Multiscale superpixel method for segmentation of breast ultrasound, *Comput. Biol. Med.*, **125** (2020), 103879. https://doi.org/10.1016/j.compbiomed.2020.103879

60. S. H. Abdulla, A. M. Sagheer, H. Veisi, Breast cancer segmentation using K-means clustering and optimized region-growing technique, *Bulletin Electr. Eng. Inf.*, **11** (2022), 158–167. https://doi.org/10.11591/eei.v11i1.3458

61. X. Shen, H. Ma, R. Liu, H. Li, J. He, X. Wu, Lesion segmentation in breast ultrasound images using the optimized marked watershed method, *Biomed. Eng. Online*, **20** (2021), 57. https://doi.org/10.21203/rs.3.rs-137797/v1

62. A. Kaur, M. Rashid, A. K. Bashir, S. A. Parah, Detection of breast cancer masses in mammogram images with watershed segmentation and machine learning approach, In: *Artificial intelligence for innovative healthcare informatics*, Cham: Springer, 2022, 35–60. https://doi.org/10.1007/978-3-030-96569-3_2

63. N. Ramadijanti, A. Barakbah, F. A. Husna, Automatic breast tumor segmentation using hierarchical k-means on mammogram, In: *2018 International electronics symposium on knowledge creation and intelligent computing (IES-KCIC)*, Bali: IEEE, 2018, 170–175. https://doi.org/10.1109/kcic.2018.8628467

64. S. U. Khan, N. Islam, Z. Jan, K. Haseeb, S. I. A. Shah, M. Hanif, A machine learning-based approach for the segmentation and classification of malignant cells in breast cytology images using gray level co-occurrence matrix (GLCM) and support vector machine (SVM), *Neural Comput. Applic.*, **34** (2022), 8365–8372. https://doi.org/10.1007/s00521-021-05697-1

65. J. Y. Lee, K. Lee, B. K. Seo, K. R. Cho, O. H. Woo, S. E. Song, et al., Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on MRI, *Eur. Radiol.*, **32**, (2022), 650–660. https://doi.org/10.1007/s00330-021-08146-8

66. L. Hirsch, Y. Huang, S. Luo, C. R. Saccarelli, R. Lo Gullo, I. D. Naranjo, et al., Radiologist-level performance by using deep learning for segmentation of breast cancers on MRI scans, *Radiol. Artif. Intell.*, **4** (2021), e200231. https://doi.org/10.1148/ryai.200231

67. L. Zhang, Z. Luo, R. Chai, D. Arefan, J. Sumkin, S. Wu, Deep-learning method for tumor segmentation in breast DCE-MRI, In: *Medical imaging 2019: Imaging informatics for healthcare, research, and applications*, **10954** (2019), 109540F. https://doi.org/10.1117/12.2513090

68. H. Sun, C. Li, B. Liu, Z. Liu, M. Wang, H. Zheng, et al., AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms, *Phys. Med. Biol.*, **65** (2020), 055005. https://doi.org/10.1088/1361-6560/ab5745

69. T. Shen, C. Gou, J. Wang, F. Y. Wang, Simultaneous segmentation and classification of mass region from mammograms using a mixed-supervision guided deep model, *IEEE Signal Process. Lett.*, **27** (2019), 196–200. https://doi.org/10.1109/lsp.2019.2963151

70. N. Saffari, H. A. Rashwan, M. Abdel-Nasser, V. K. Singh, M. Arenas, E. Mangina, et al., Fully automated breast density segmentation and classification using deep learning, *Diagnostics*, **10** (2020), 988. https://doi.org/10.3390/diagnostics10110988

71. M. S. Hossain, Microc alcification segmentation using modified U-net segmentation network from mammogram images, *J. King Saud Univ. Comput. Inf. Sci.*, **34** (2022), 86–94. https://doi.org/10.1016/j.jksuci.2019.10.014

72. Y. Xue, J. Zha, D. Pelusi, P. Chen, T. Luo, L. Zhen, et al., Neural architecture search with progressive evaluation and sub-population preservation, *IEEE Trans. Evol. Comput.*, **29** (2025), 1678–1691. https://doi.org/10.1109/tevc.2024.3393304

73. Y. Xue, X. Han, F. Neri, J. Qin, D. Pelusi, A gradient-guided evolutionary neural architecture search, *IEEE Trans. Neural Netw. Learn. Syst.*, **36** (2025), 4345–4357. https://doi.org/10.1109/tnnls.2024.3371432

74. P. Jiang, Y. Xue, F. Neri, Score predictor-assisted evolutionary neural architecture search, *IEEE Trans. Emerg. Top. Comput. Intell.*, 2025, 1–15. http://dx.doi.org/10.1109/TETCI.2025.3526179

75. O. N. Oyelade, A. E. Ezugwu, A bioinspired neural architecture search based convolutional neural network for breast cancer detection using histopathology images, *Sci. Rep.*, **11** (2021), 19940. https://doi.org/10.1038/s41598-021-98978-7

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

77. S. Anwar, A. Rahming, M. Fernander, O. Udedibor, S. Ali, Breast cancer diagnosing system: Using a rough set-ensemble classifier approach, In: *Pattern recognition. ICPR 2024 International workshops and challenges*, 2025, 22–35. https://doi.org/10.1007/978-3-031-88220-3_2

78. A. Sahu, P. K. Das, S. Meher, Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms, *Phys. Med.*, **114** (2023), 103138. https://doi.org/10.1016/j.ejmp.2023.103138

79. A. Sahu, P. K. Das, S. Meher, An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images, *Biomed. Signal Process. Control*, **87** (2024), 105377. https://doi.org/10.1016/j.bspc.2023.105377

80. B. Munteanu, A. Murariu, M. Nichitean, L. G. Pitac, L. Dioşan, Value of original and generated ultrasound data towards training robust classifiers for breast cancer identification, *Inf. Syst. Front.*, **27** (2025), 75–96. https://doi.org/10.1007/s10796-024-10499-6

81. H. M. Rai, S. Dashkevych, J. Yoo, Next-generation diagnostics: The impact of synthetic data generation on the detection of breast cancer from ultrasound imaging, *Mathematics*, **12** (2024), 2808. https://doi.org/10.3390/math12182808

AIMS Press