International Journal of Electrical and Computer Engineering (IJECE)
Vol. 15, No. 6, December 2025, pp. 5633~5646
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i6.pp5633-5646

Breast cancer detection using ensemble methods

Alaa Mohamed Ghazy[1], Hala Bahy Nafea[2], Fayez Wanis Zaki[2], Hanan Mohamed Amer[2]
[1]Department of Communications and Electronics Engineering, Mansoura Higher Institute of Engineering and Technology, Mansoura, Egypt
[2]Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Mansoura, Egypt

Article Info

Article history:

Received Mar 24, 2025
Revised Aug 22, 2025
Accepted Sep 14, 2025

Keywords:

Breast cancer
Convolutional neural network
External classifiers
Majority hard voting mechanism
Mammogram

ABSTRACT

Breast cancer (BC) is one of the most common cancers among women. This study's framework is divided into three phases. Firstly, a majority hard voting approach is used to apply an ensemble classification mechanism as a decision fusion technique on the level of convolutional neural networks (CNNs). Five pre-trained CNNs—visual geometry group 19 (VGG19), densely connected convolutional network 201 (DenseNet201), residual network 50 (ResNet50), mobile network version 2 (MobileNetV2), and inception version 3 (InceptionV3)—are evaluated, using a data splitting test ratio represents 30% of the total dataset. Secondly, the classification results of the five CNNs are compared to get the best-performance model. Then, seven state of art machine classifiers—decision tree (DT), histogram-based gradient boosting classifier (HGB), support vector machine (SVM), random forest (RF), logistic regression (LR), gradient boosting (GB), and extreme gradient boosting (XGB)—are used to improve system performance on the feature vector that was taken from this CNN model. Thirdly, to improve robustness, a majority hard voting technique is used at the external classifier level using the highest four classifiers selected based on their accuracy. Several experiments were conducted in this study, and the results showed that ResNet50 produced the best results in terms of precision and accuracy. The majority voting mechanism improves the system's accuracy to 99.85% through CNNs and to 100% through traditional classifiers.

Corresponding Author:

Alaa Ghazy
Department of Communications and Electronics Engineering, Mansoura Higher Institute of Engineering and Technology
Mansoura, Egypt
Email : alaamohamedghazy@gmail.com

1. INTRODUCTION

BC is the most common disease in the world, with 7.8 million people living with a diagnosis as of the end of 2020. It occurs in every country of the world in women at any age after puberty but with increasing rates in later life, according to the World Health Organization (WHO) [1]. Mammography is the most reliable and accurate screening technique and remains the gold standard for community breast cancer (BC) screening [2]. Currently, although it cannot be replaced, mammography is still utilized combined with MRI and ultrasound, particularly in cases where the density of breast tissue is significant. It is the major clinical testing Because it can be interpreted in several manners to provide additional information prior to diagnosis or detection, high precise, and can identify around 80%–90% of the danger of cancer. According to previous studies, early detection (the tumor is at least 20 mm in size) of breast cancer is crucial because it can contribute to up to a 40% decrease in mortality rate [3].

The goal of this research is to increase the precision of BC prediction models. Full article usually follows a standard structure: The related work is presented in section 2. Section 3 provides a detailed description of the framework's suggested methodology, employ deep learning algorithms, classification methods, and a majority voting system. Section 4 delves into dataset description, System specifications, Fine-tuned pre-trained convolutional neural networks (CNNs), data preprocessing, and the experimental results. It also includes the standards by which the models' performance is evaluated. The conclusion is shown in section 5. The principal contributions of the suggested work are:

a. Compare the accuracy and evaluation measure performance of individual pretrained CNNs then apply the majority hard voting to increase the benefit from features of each CNN and thus increase accuracy of the system.
b. Study the effect of using external machine learning classifiers on the performance of pretrained CNN model by applying seven machine learning classifiers to the feature vector extracted from best architecture out of the five pre-trained model from the first stage and compare its performance.
c. Studying the effect of applying two-level ensemble framework that combines deep learning models (CNNs) and external classifiers to ensure robustness and increase system performance.

2. RELATED WORK

Mammogram images were widely studied for BC detection and classification using various CNNs. Numerous research studies have reported significant results using different classifiers and deep learning techniques. A summary of the most significant related work is included in Table 1, showcasing diverse approaches applied to different datasets. Laghmati *et al.* [4] applied four traditional classifiers—artificial neural network (ANN), k-nearest neighbors (KNN), support vector machines (SVM), and decision tree (DT)—on the Mammographic Mass dataset, achieving a maximum accuracy of 84% with ANN. However, the study was limited by its reliance on classical models and relatively low accuracy. Nguyen *et al.* [5] utilized both supervised and unsupervised learning approaches with principal component analysis (PCA)—based feature selection. Although ensemble voting achieved around 90% accuracy, the study lacked integration with deep feature extractors and did not explore CNN architectures, which are known to perform better in image-based tasks.

Table 1. Show summary of related work

No.	Author	Year	Methodology	Results
1	Laghmati *et al.* [4]	2019	ANN, SVM, DT and KNN	With 84% accuracy, the ANN model produced the best results
2	Nguyen *et al.* [5]	2019	SVM, AdaBoost, LR and ensemble voting classifier	Ensemble voting classifiers achieved around 90% accuracy
3	Gopal *et al.* [6]	2021	RF, MLP and LR	The applied RF model achieved 95% accuracy
4	Bataineh [7]	2019	K-NN, NB, MLP, and SVM through WBCD dataset	The maximum accuracy of 99.12% was attained by MLP.
5	Mangukiya [8]	2022	RF, NB, SVM, DT, KNN, XGBoost, and AdaBoost through WBCD	XGBoost achieved 98.24% accuracy
6	Osman and Aljahdali [9]	2020	Ensemble boosting with RBF neural network	Ensemble boosting achieved 98.4% accuracy
7	Kumar *et al.* [10]	2019	12 classifiers including tree and lazy algorithms	Tree and lazy classifiers reached 99% accuracy
8	Mohamed *et al.* [11]	2022	CNN (U-Net) on DMR-IR dataset	Specificity = 98.67%, Sensitivity = 100%, and Accuracy = 99.33%
9	Singh *et al.* [12]	2023	Hybrid algorithms on WBCD dataset	Achieved 98.96% accuracy.
10	Chen *et al.* [13]	2023	XGB, RF, LR, and KNN	XGB achieved best accuracy of 97.4%.
11	Hamza and Mezl [14]	2024	CNN (U-Net++) with MobileNetV2 and InceptionV3 on ultrasound	MobileNetV2 achieved 96.58% accuracy; InceptionV3 achieved 72.80%.

Gopal *et al.* [6] employed RF, LR, and multilayer perceptron (MLP) on the WBCD dataset, reporting a maximum accuracy of 95%. However, their method did not consider combining deep learning with classical models, limiting its scalability and robustness on larger datasets. Bataineh [7] compared MLP, KNN, SVM, and Naïve Bayes (NB), where MLP achieved 96.70% accuracy. Despite this improvement, the study was constrained by the lack of ensemble strategies and absence of decision fusion mechanisms that could increase model reliability. Mangukiya [8] tested seven classifiers, reporting a peak accuracy of 98.14% using XGBoost. However, the lack of CNN-based feature extraction limits the model's applicability for high-dimensional image data. Osman and Aljahdali [9] used an ensemble boosting method combined with a radial basis function (RBF) neural network, reaching 98.4% accuracy. Yet, their approach did not incorporate

multiple CNNs or evaluate robustness through external classifiers. Kumar *et al.* [10] evaluated twelve classifiers and achieved up to 99% accuracy using tree-based models. While comprehensive, their study lacked a hybrid ensemble structure that combines CNN features with traditional classifiers. Mohamed *et al.* [11] applied a U-Net CNN to thermal images and achieved excellent performance (accuracy=99.33%, sensitivity = 100%). Nonetheless, the scope was limited to thermal imaging and did not extend to mammography or ensemble fusion strategies. Singh *et al.* [12] used a hybrid feature selection method that improved performance to 98.16%, but the study did not explore ensemble methods combining multiple CNNs with machine learning classifiers. Chen *et al.* [13] compared XGBoost, RF, LR, and KNN, with XGBoost performing best (97.4% accuracy). However, the absence of CNN-based feature extraction restricts its effectiveness for complex image data. Hamza and Mezl [14] employed a CNN (U-Net++) for breast region segmentation from ultrasound images and used MobileNetV2 and InceptionV3 for classification. Although MobileNetV2 achieved 96.58% accuracy, the study was limited by using single CNN models and lacked classifier-level ensemble mechanisms.

2.1. Identified gaps and motivation

While existing studies have made meaningful progress, most suffer from at least one of the following limitations: i) Used only single model, making them susceptible to overfitting and lack of generalization; ii) Limited integration of deep learning with traditional classifiers, reducing the potential for hybrid learning; iii) Lack of multi-level ensemble techniques, which could enhance robustness and accuracy; and iv) Use of non-mammographic datasets (thermal or ultrasound images), limiting applicability to mammography-based diagnosis. These limitations highlight the need for a more comprehensive, ensemble-based framework that integrates multiple CNNs and traditional classifiers through decision-level fusion. The proposed method addresses these gaps by:

a. Comparing and combining five different CNN models using a hard voting mechanism.
b. Feeding the best CNN's extracted features into multiple machine learning classifiers.
c. Applying a second ensemble at the classifier level to enhance system accuracy and robustness.

3. METHOD

This research proposes a robust hybrid ensemble foundation for BC binary classification with mammographic images. The methodology is structured into three stages, integrating the strengths of CNNs and classifiers together. The framework is evaluated using a publicly available mammography dataset. Figures 1 through 4 illustrate the workflow and components of the proposed system.

a. First stage: utilizing pre-trained CNNs for feature extraction

At the beginning, five pre-trained models—VGG19, DenseNet201, ResNet50, MobileNetV2, and InceptionV3—are used in mammography images to extract deep features as shown in Figure 1. These models are fine-tuned, using weights initialized from the ImageNet dataset. To enhance the adaptability of each model to the breast cancer classification task, custom classification heads are appended, consisting of Flatten layer, fully connected Dense layer using ReLU activation, Dropout layer to avoid overfitting, and final Soft max layer to perform binary classification.

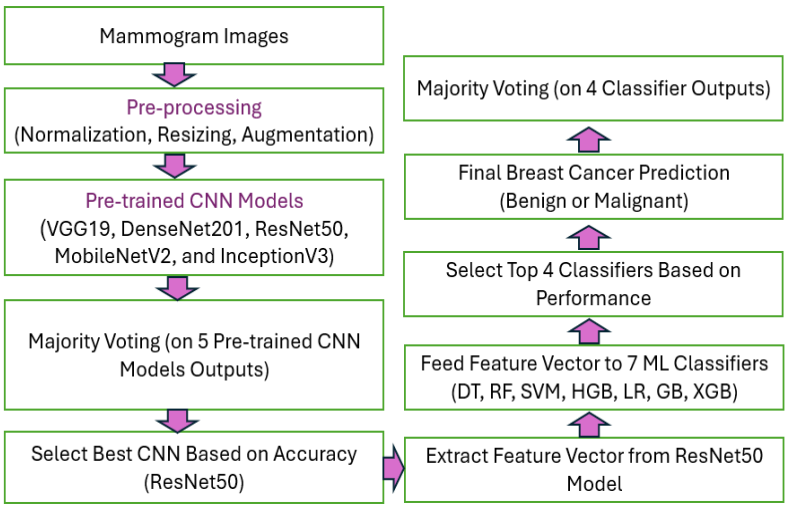

Figure 1. Prime mechanism of the proposed method

The dataset is split randomly into training set represents 70% and the testing set represents 30% of the total dataset. The Adam optimizer is used to fine- tune each model across 20 epochs and early stopping patience = 10 to reduce unnecessary computation and Minimize overfitting. The performance of each CNN is independently evaluated and compared to get the best-performance model. To enhance decision reliability and reduce model-specific biases, the outputs of all five CNNs are fused using a majority hard voting mechanism at the decision level as shown in Figure 2. This ensemble approach aggregates the predictions from individual models and selects the most frequent class as the final decision.

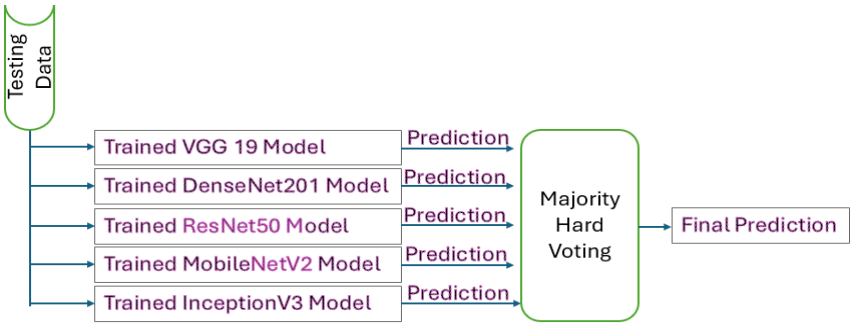

Figure 2. Majority-hard voting technique on the level of convolution neural networks

b. Second stage: applying seven classifiers through the best CNN

At second stage, the best-performing CNN model from the previous phase is used as a feature extractor. The feature vectors obtained from the penultimate layer of this model serve as input to seven classifiers: DT, LR, HGB, SVM, RF, GB, and XGB as shown Figure 3. Each classifier is trained on the extracted feature vectors using consistent hyperparameter tuning. This stage is designed to leverage the discriminative power of the CNN features and the classifiers abilities to distinguish between benign tumor and malignant tumor.

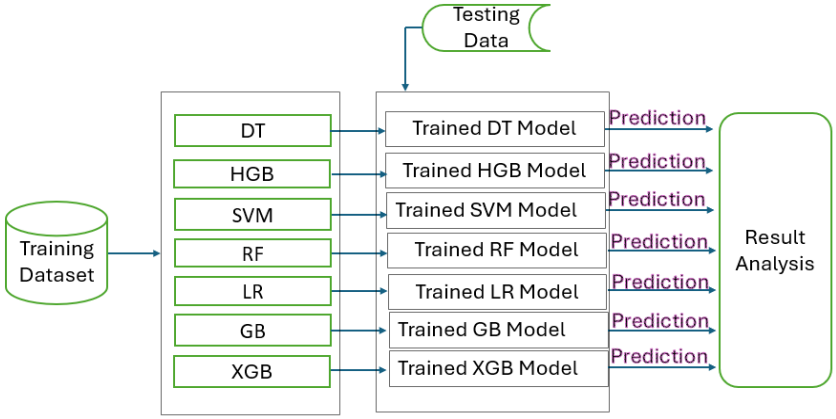

Figure 3. Seven state-of-the-art machine classifiers are used in the proposed method

c. Third stage: Using majority hard voting on the level of external classifiers

In the final stage, a majority hard voting mechanism is applied on the level of the external classifiers. The best four classifiers that achieved the highest accuracy performance are selected and integrated using a majority hard voting strategy as shown in Figure 4. This classifier-level ensemble aggregates the predictions of the most reliable models, aiming to reduce individual classifier variance and improve overall system robustness and stability.

The proposed three-stage framework comprising deep feature extraction, the incorporation of classifiers, and ensemble-based decision fusion strategically combines the benefits of deep learning and conventional methods. This hybrid strategy improves the system's ability to deliver high classification

accuracy, improved generalization, and increased robustness across diverse diagnostic conditions. Experimental findings validate the effectiveness of the framework, highlighting its promise as an effective and dependable method for BC detecting in medical imaging.

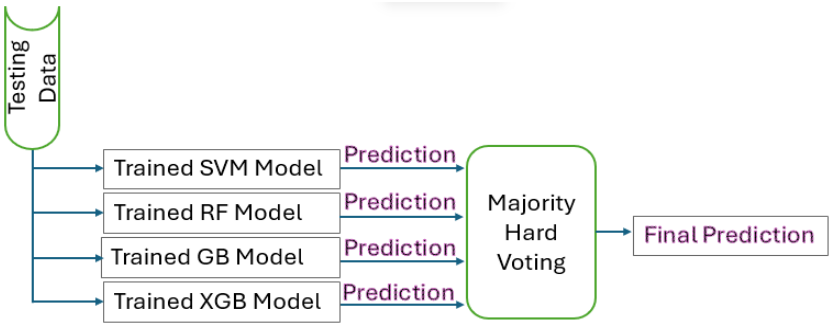

Figure 4. majority-hard voting mechanism on the level of external classifiers

3.1. The applied CNNs of proposed model

3.1.1. VGG19 architecture

It is a deep convolutional neural network architecture proposed by the Visual Geometry Group at the University of Oxford, introduced by Simonyan and Zisserman in 2014 [15]. It consists of 19 weight layers that use small 3×3 convolution filters consistently across all convolutional layers. The architecture follows a fixed pattern of convolutional layers then Rectified Linear Unit and max-pooling layers, which improves feature extraction while maintaining computational efficiency. This uniform architecture facilitates deeper networks without significantly increasing the number of parameters, making it well-suited for large-scale image recognition tasks [16].

3.1.2. DenseNet architecture

DenseNet, or densely connected convolutional network, was introduced by Huang *et al.* in 2017 [17]. By feed-forwardly connecting each layer to every other layer, this architecture improves gradient flow and information. Specifically, every layer receives feature maps from all preceding layers, promoting feature reuse and mitigating the problem of vanishing gradient. It is efficient in terms of parameter usage, as it avoids redundant feature learning and decreases the overall parameters in contrast to traditional CNNs. Its use of dense blocks and transition layers allows for compact, yet powerful, networks that perform exceptionally well on classification tasks [18].

3.1.3. ResNet50 architecture

It is a deep residual network introduced by He *et al.* [18], [19], which was developed at Microsoft Research Asia. The architecture is distinguished by its use of residual connections or "skip connections," which make it easier for gradients to move through the layers during the training of deep networks. These connections solving the problem of vanishing gradient, which commonly affects deep CNNs. ResNet50 consists of 50 layers and was pretrained on the ImageNet dataset to enhance its generalization capabilities.

3.1.4. MobileNet architecture

It is a lightweight deep convolutional neural network architecture designed by Google's Mobile Vision team, specifically tailored for efficient inference in mobile and embedded vision applications [20]. The model introduces Depthwise separable convolutions and linear bottlenecks to minimize model size and computing complexity while preserving competitive accuracy. The linear bottleneck layers reduce the dimensionality of feature maps and enhance non-linearity, enabling the model to operate efficiently on devices with limited processing resources.

3.1.5. InceptionNet architecture

It is also known as GoogLeNet, was introduced by Szegedy *et al.* [20], [21]. It employs factorized convolutions and inception modules to dramatically minimize over all parameters while retaining high accuracy in tasks of image classification. The architecture combines multiple convolutional filter sizes within a single module, which allows it to capture information at different scales. Due to its efficient design and powerful performance. It is pretrained on the ImageNet dataset and widely adopted in computer vision tasks.

3.2. The use of different classifiers through proposed model.

3.2.1. Random decision tree

It is supervised algorithms mainly used for the graphical representation of all the possible solutions [22]. It is characterized by the ability to identify and choose the most important attributes which are useful in the classification stage. It can also select the attributes which deliver the maximum information gain (IG) which is defined as:

$$IG = E(ParentNode) - AverageE(ChildNodes) \quad (1)$$

where *Entropy* (*E*) is defined as:

$$E = \sum_i^n -probi(\log_2 \quad probi) \quad (2)$$

and $Probi$ is the probability of class i.

3.2.2. Gradient boosting classifier

It is preferred for small samples and is considered an excellent model for regression and classification, particularly for tabular data [23]. It is characterized by easy implementation, low computational cost, and efficiency.

3.2.3. Hist gradient boosting classifier

It is an enhanced version of GBDT that creates a histogram of feature values during training while reducing training time and memory consumption and splitting the continuous variable into bins. It is characterized by speed with large number of samples. The utilization of histograms and better data structures is primarily responsible for this speed increase. The algorithm learns how to handle missing data during training, making the process more straightforward and efficient [24].

3.2.4. Support vector machine

It considers the most dependable algorithms based on statistical learning frameworks. It is regarded as a Decision plane-based model [25] that offers a solution for both regression and classification problems as well as for both linear and non-linear datasets. The basic idea of SVM was to separate different groups using hyperplanes. The two main issues with SVM are the correct selection of kernel function, and its parameters [26]. The kernel function allows SVMs to classify one-dimensional data in a two-dimensional approach. Typically, a linear kernel function is defined as follows:

$$k(x, x_i) = x \cdot x^T \quad (3)$$

And Polynomial kernel functions are defined as:

$$k(x, x_1) = (1 + x \cdot Tx_i)d \quad (4)$$

'd' is degree of kernel function.

3.2.5. Random decision forest

It builds many decision trees during the training phase and then generates classes for each. It is mainly used in classification and regression. By creating numerous decision trees from training data using bootstrapped samples with a small modification, the de-correlated tree via bagging. The impact of each prediction informs the final prediction. It can interpret irrelevant attributes and handle missing data [27].

3.2.6. Simple logistic regression model

The LR model is a popular choice for binary classifications [28]. It is believed that a linear combination of the input features equals the conditional probability of one of the two output classes. The classification model's logistic equation is as follows:

$$z_i = \ln\left(\frac{p_i}{1-p_i}\right) \quad (5)$$

where p represents probability that event i will occur.

3.2.7. Extreme gradient boosting

Extreme gradient boosting is a scalable machine learning system for tree boosting [29] which is implemented for supervised learning problems and developed specifically to boost model performance and computational effectiveness. It solves problems using minimal resources and incorporates a regularized model to prevent overfitting and is suited for classification problems.

3.3. Majority hard voting mechanism of proposed model

It is widely used through ensemble classification. It is also called plurality voting and used to improve the classification results. In this technique, the prediction of class label y performs via majority voting of each classifier C:

$$y = mode\{c1(x), c2(x), \dots, cn(x)\} \quad (6)$$

4. EXPERIMENTAL RESULTS

4.1. Data description

For experimentation, the dataset used in this study is the breast mammography dataset with masses introduced by Huang and Lin [30]. All images in this dataset on the Mendeley website were available in PNG format and sized to 227×227 pixels. The image datasets on this website were arranged into three main datasets: IN breast, MIAS, and DDSM datasets.

In this study, the proposed method was implemented on only the digital database for screening mammography (DDSM) datasets, which consisted of 2,188 mass images extracted from 1,319 cases before augmentation then the number of images reached 13,128 mass images after augmentation, arranged in two main folders. The benign folder comprised 5,970 images and the malignant folder comprised 7,158 images. This database isn't balanced. So, in this study, a randomly balanced subset of the DDSM dataset was selected, containing 1,600 mass images divided into 800 benign mass images and 800 malignant mass images in a balanced manner as illustrated in Table 2. A sample mammography mass image with benign and malignant tumors is shown in Figure 5.

Table 2. Number of mammography mass images used

Benign images	Malignant Images	Total
800	800	1,600

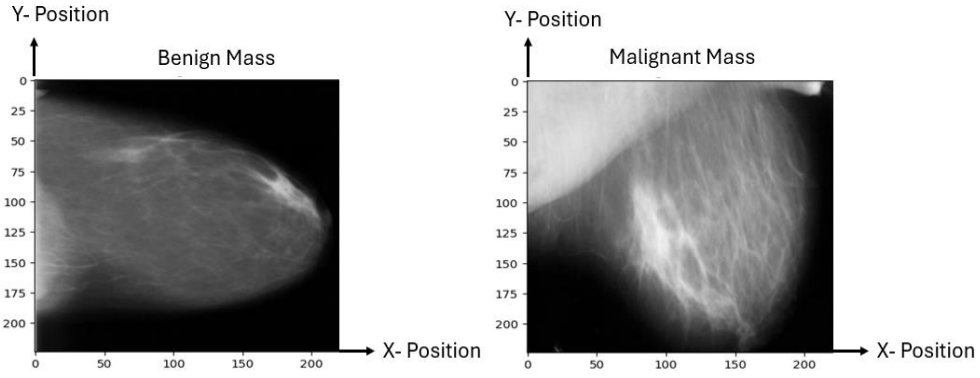

Figure 5. Benign and malignant mass tumor from DDSM dataset

4.1.1. Data splitting

The dataset is separated into random (training and testing) sets. The training dataset represents 70% and the testing dataset represents 30% of the total dataset. Table 3 illustrates the number of mammography mass images used in this study.

Table 3. Number of mammography mass images used through splitting ratio (70:30)

Splitting ratio	Trained images	Tested images	Total
Data splitting (70:30)	1120	480	1,600

4.2. System specifications

The experimental work is carried out using Python 3 in the Google Collaboratory with onlineT4 GPU using 64 GB of RAM and an Intel Core i7 processor are features of this laptop. In this case, the model is optimized using Adam optimizer, learning rate is 0.0001, minimum batch size is 23, and the maximum number of epochs is 20.

4.3. Fine-tuned the pre-trained (CNNs)

In order to maximize the advantages of transfer learning, every model utilized in this investigation was first trained on the ImageNet dataset. To enhance their ability to adapt to the task of BC classification, architectural modifications were made by adding some layers such as Flatten layer, Dense layer, a Dropout layer = 0.2, and Softmax output layer. These modified models were then fine-tuned on a dataset consisting of 1,600 images over 20 training epochs, with early stopping implemented using a patience value of 10 to avoid overfitting.

4.4. Data preprocessing

Normalization, scaling, and augmentation of the training images are all parts of preprocessing techniques. The technique of normalization involves transforming pixel values to fall within a predefined range, bounded between 0 and 1, to enhance model generalizability and streamline the training process. Then, the images are adjusted to a typical size of 224×224×3 pixels to confirm the required input size of all applied pre-trained models. To improve the diversity of training data and boost the model's performance and generalization ability, data augmentation is methodically applied to the training dataset. The applied augmentation function is an image data generator function including the following: shear range = 0.1, width shift range = 0.1, height shift range = 0.1, rotation range = 20 and zoom range = 0.1.

4.5. Performance evaluation measures

Performance evaluation metrics were computed based on the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{7}$$

$$\Pr e\,cision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1-score = 2 \cdot \frac{precision \cdot recall}{precision+recall} \tag{10}$$

And the area under the receiver operating characteristic (ROC) curve and AUC were calculated to evaluate the performance of BC classification.

4.6. Experimental results and discussion

In this part, the suggested BC classification system's performance is assessed using a data splitting ratio of 30% for testing and 70% for training. The system's design includes a three-stage framework to enhance system accuracy. Every step is essential to enhancing the model's capacity for prediction using a combination of conventional machine learning and deep learning methods. In the first stage, five pre-trained models—VGG19, MobileNetV2, DenseNet201, ResNet50, and InceptionV3—were trained and evaluated using data splitting ratio 70:30. Their individual classification performances are evaluated, after that a majority hard voting mechanism is applied to combine their predictions. This ensemble technique mitigates the disadvantages of any one CNN while utilizing the benefits of each model to increase overall accuracy.

The second step is to improve system performance by determining which CNN model performs best in terms of classification accuracy. Seven classifiers—DT, RF, LR, HGB, SVM, GB, and XGB—are fed feature vectors that are taken from the penultimate layer of this best-performing CNN model. Each classifier is trained and optimized to assess its ability to generalize and enhance detection accuracy.

In the third stage, the classification results of the seven classifiers are compared, and the four classifiers with the highest performance are selected. A majority hard voting mechanism is then applied to the outputs of these four classifiers, treating them as a decision-making committee. This ensemble strategy is designed to correct individual classifier errors and improve the final classification outcome. The integration of this voting mechanism demonstrates a significant enhancement in the robustness and accuracy of the overall system.

4.6.1. Experimental results of proposed model

while implementing the suggested classification system 30% of the data was used for testing and 70% for training. The experimental results are discussed in detail for each of the three main stages that comprise the framework. Each stage contributes to refining the classification process and collectively demonstrates the effectiveness of the suggested hybrid ensemble approach in BC detection.

a. The experimental result of the first stage

Results from the first step of the experiment are presented in Table 4, where the five pre-trained CNN models' classification performance is highlighted. These models—VGG19, DenseNet201, ResNet50, MobileNetV2, and InceptionV3—were evaluated individually. In addition to the individual performances, Table 4 also reports the overall classification accuracy achieved when the majority hard voting mechanism is applied to the combined outputs of these models. The strengths of each CNN are used in this ensemble technique to get a prediction result that is more reliable and accurate.

The ResNet50 model produced the best classification results among the five pre-trained CNNs, as indicated in Table 4. It recorded an accuracy and precision of 99.58%, along with an F1-score and recall of 99.58%, and ROC-AUC value of 99.98%. Furthermore, when the majority hard voting mechanism was applied to the outputs of all five CNN models, the system's overall classification accuracy improved to 99.85%, showing how ensemble learning can improve diagnostic reliability. Figures 6 to10 illustrate the five deep learning models' accuracy and loss during the training and testing steps over twenty epochs.

Table 4. The five pre-trained models' performance metrics and the majority hard vote's classification accuracy on the models' output

Trained models	Accuracy (%)	Precision (%)	Recall (%)	F1- Score (%)	Roc-AUC (%)	Time (Sec)
VGG 19	99.57	99.17	100	99.58	99.99	1350.29
DenseNet	97.50	98.30	96.66	97.47	99.71	484.72
ResNet50	99.58	99.58	99.58	99.58	99.98	438.18
MobileNet	95.62	94.33	97.08	95.68	99.12	333.65
InceptionNet	93.33	93.33	93.33	93.33	97.75	444.34
Majority hard voting based on the previously trained models	Increase the system's classification accuracy to 99.85%.					

Figure 6. Accuracy and loss during training and testing using VGG 19 pretrained CNN

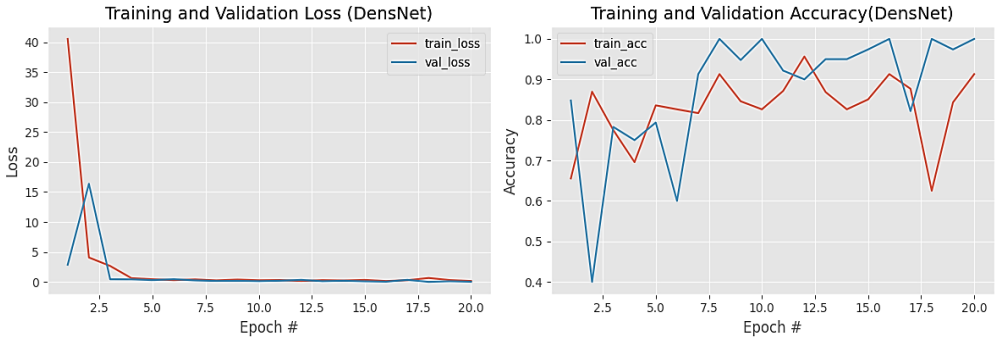

Figure 7. Accuracy and loss during training and testing using DenseNet pretrained CNN

Figure 8. Accuracy and loss during training and testing using ResNet50 pretrained CNN

Figure 9. Accuracy and loss during training and testing using MobileNet pretrained CNN

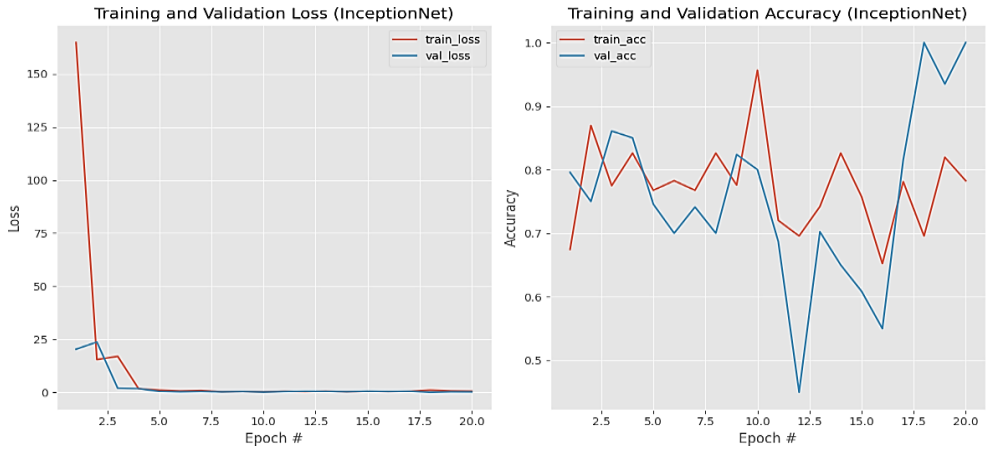

Figure 10. Accuracy and loss during training and testing using InceptionNet pretrained CNN

Figures 6 to 10 show comparison between (Training - Testing) accuracy and loss for the five deep learning models over twenty epochs. Figures 11 to 15 outline the confusion matrices acquired during experiments, where Classes "0" and "1" denote "benign" and "malignant," respectively. Figures 11 to 15 show comparison between the Confusion matrices through different five deep learning models in classifying 0 and 1 which refer to Benign and Malignant tumor in a testing dataset, respectively.

b. The experimental result of the second stage

The results of the second stage's experiment are shown in Table 5. where a preliminary analysis was conducted to compare the classification performance of the five proposed pre-trained models which led to the ResNet 50 achieving the highest classification result. Then, in the second stage of the suggested scheme, seven classifiers (DT, HGB, SVM, RF, LR, GB, and XGB) are applied to the feature vector extracted from ResNet 50 pre-trained model, and the classification performance of these machine learning classifiers is illustrated in Table 5.

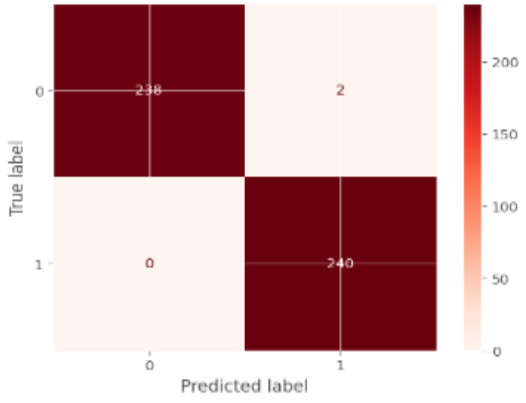

Figure 11. VGG19 model's confusion matrix

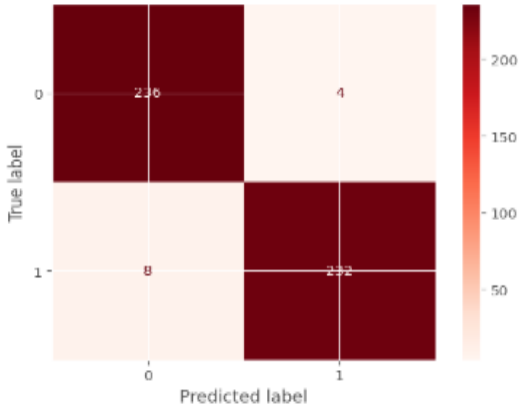

Figure 12. DenseNet model's conversion matrix

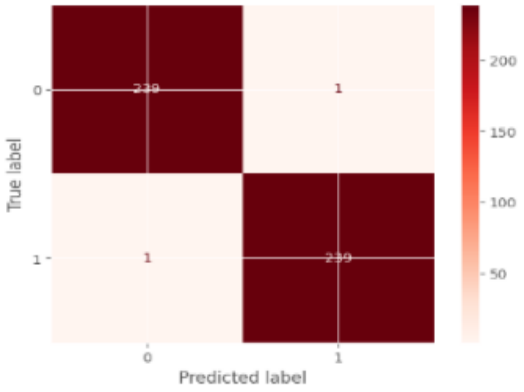

Figure 13. ResNet50 model's confusion matrix

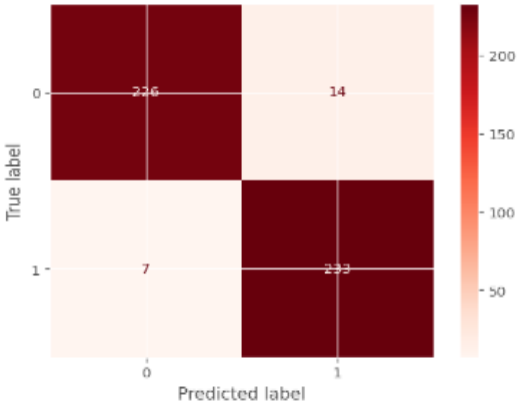

Figure 14. MobileNet model's confusion matrix

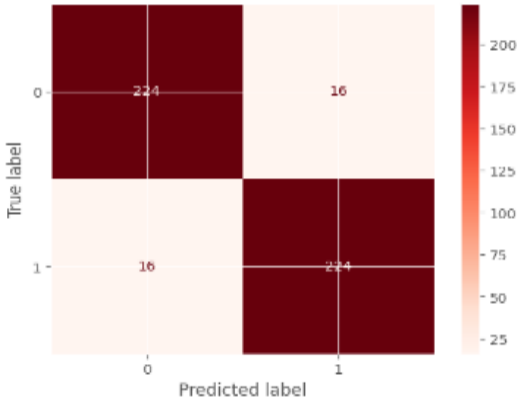

Figure 15. InceptionNet model's confusion matrix

Table 5. Shows the performance comparison of seven external classifiers with the pretrained ResNet 50 model

	Accuracy (%)	Precision (%)	Recall (%)	F1- Score (%)
(DT)	100	100	100	100
(HGB)	100	100	100	100
(SVM)	100	100	100	100
(RF)	100	100	100	100
(LR)	100	100	100	100
(GB)	100	100	100	100
(XGB)	100	100	100	100

As shown in Table 5, the performance of the external machine learning classifiers was remarkably high. All seven classifiers—DT, SVM, RF, LR, HGB, GB, and XGB— achieved perfect scores of 100% in F1-score, recall, accuracy, and precision. These results highlight the strength of the feature representations extracted from the best-performing CNN model and demonstrate the efficiency of traditional classifiers while diagnosing BC either benign or malignant when combined with deep feature extraction.

c. The experimental result of the third stage

The seven classifiers' classification performance was evaluated based on the feature vectors extracted from the ResNet50 model. These classifiers were compared to identify the top four models that achieved the highest classification results. Following this selection, a majority hard voting mechanism was applied at the output level of these four classifiers to improve the system's overall classification performance. The results of this ensemble approach are summarized in Table 6, demonstrating its effectiveness in increasing diagnostic accuracy and reliability.

As shown in Table 6, by applying the majority hard voting mechanism through the four best-performing classifiers, combined with the feature vectors obtained from the ResNet50 model, led to a substantial enhancement in classification accuracy. This ensemble strategy effectively combined the outputs of the selected classifiers, leveraging their individual strengths while minimizing the impact of potential misclassifications. As a result, the proposed system's classification accuracy was raised to a perfect score of 100%, demonstrating the strength and reliability of the hybrid ensemble approach in BC detection.

Table 6. Apply the Majority voting accuracy on the level of best performance external classifiers

ResNet50 pretrained model with	Accuracy
(SVM)	100%
(RF)	100%
(GB)	100%
(XGB)	100%
Majority voting on level of classifiers output	100%

5. CONCLUSION AND FUTURE WORK

Deep learning methods have significantly changed a number of industries in recent years, including image processing and the healthcare sector. These methods have been particularly helpful in the early diagnosis detection and classification of various types of cancer. By aiding medical professionals in highlighting regions of concern and improving diagnostic precision, deep learning has become an essential part of current healthcare technologies. This study focuses on enhancing binary classification of breast tumors by implementing a majority hard voting mechanism. To accomplish this, multiple pre-trained CNNs were utilized, including VGG19, DenseNet201, ResNet50, MobileNetV2, and InceptionV3, all optimized using the Adam optimizer. These models' performance was carefully assessed, and the results showed high classification metrics with 93.33% to 100% accuracy, precision, and recall values.

Among the evaluated models, ResNet50 achieved the highest accuracy of 99.58%, showing its ability to extract major features from mammography images. Moreover, applying majority hard voting mechanism enhanced system robustness, reaching 99.85% accuracy at the CNN level and a perfect 100% when applied to external classifiers. These findings validate the efficiency of ensemble techniques in medical imaging. For future work, the study proposes exploring alternative weighted voting strategies and extending the current binary classification to a multi-stage malignancy classification framework.

FUNDING INFORMATION

The authors declare that this research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Alaa Mohamed Ghazy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Hala Bahy Nafea		✓		✓		✓		✓	✓	✓		✓		
Fayez Wanis Zaki		✓	✓	✓				✓	✓	✓	✓	✓	✓	
Hanan Mohamed Amer	✓	✓		✓				✓	✓	✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Mendeley website at *https://data.mendeley.com/datasets/ywsbh3ndr8/5* [30].

REFERENCES

[1] WHO, "Cancer: fact sheet." 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer
[2] L. Shen, "End-to-end training for whole image breast cancer diagnosis using an all-convolutional design," *arXiv:1711.05775*, 2017.
[3] H. G. Welch, P. C. Prorok, A. J. O'Malley, and B. S. Kramer, "Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness," *New England Journal of Medicine*, vol. 375, no. 15, pp. 1438–1447, Oct. 2016, doi: 10.1056/NEJMoa1600249.
[4] S. Laghmati, A. Tmiri, and B. Cherradi, "Machine learning based system for prediction of breast cancer severity," in *2019 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct. 2019, pp. 1–5. doi: 10.1109/WINCOM47513.2019.8942575.
[5] Q. H. Nguyen *et al.*, "Breast cancer prediction using feature selection and ensemble voting," in *2019 International Conference on System Science and Engineering (ICSSE)*, Jul. 2019, pp. 250–254. doi: 10.1109/ICSSE.2019.8823106.
[6] V. N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning," *Measurement: Journal of the International Measurement Confederation*, vol. 178, p. 109442, 2021, doi: 10.1016/j.measurement.2021.109442.
[7] A. Al Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *International Journal of Machine Learning and Computing*, vol. 9, no. 3, pp. 248–254, 2019.
[8] M. Mangukiya, "Breast cancer detection with machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 2, pp. 141–145, 2022, doi: 10.22214/ijraset.2022.40204.
[9] A. H. Osman and H. M. A. Aljahdali, "An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model," *IEEE Access*, vol. 8, pp. 39165–39174, 2020, doi: 10.1109/ACCESS.2020.2976149.
[10] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, "Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 37, pp. 435–442, 2020, doi: 10.1007/978-981-15-0978-0_43.
[11] E. A. Mohamed, E. A. Rashed, T. Gaber, and O. Karam, "Deep learning model for fully automated breast cancer detection system from thermograms," *PLoS ONE*, vol. 17, no. 1 January 2022, p. e0262349, 2022, doi: 10.1371/journal.pone.0262349.
[12] L. K. Singh, M. Khanna, and R. Singh, "Artificial intelligence based medical decision support system for early and accurate breast cancer prediction," *Advances in Engineering Software*, vol. 175, p. 103338, 2023, doi: 10.1016/j.advengsoft.2022.103338.
[13] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification prediction of breast cancer based on machine learning," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, p. 6530719, 2023.
[14] A. Hamza and M. Mezl, "Deep learning-enhanced ultrasound analysis: Classifying breast tumors using segmentation and feature extraction," *IEEE Access*, vol. 13. pp. 83528–83541, 2025. doi: 10.1109/ACCESS.2025.3568588.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
[16] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.
[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 770–778.
[19] B. Deng *et al.*, "Application of ResNet50 convolution neural network for the extraction of optical parameters in scattering media," *arXiv preprint, arXiv:2404.16647*, 2024.
[20] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, vol. 07-12-June-2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.
[21] Z. Zhang, H. Tian, Z. Xu, Y. Bian, and J. Wu, "Application of a pyramid pooling Unet model with integrated attention mechanism and Inception module in pancreatic tumor segmentation," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 12, Dec. 2023, doi: 10.1002/acm2.14204.
[22] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.
[23] H. Aljamaan and A. Alazba, "Software defect prediction using tree-based ensembles," in *PROMISE 2020 - Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering, Co-located with ESEC/FSE 2020*, 2020, pp. 1–10. doi: 10.1145/3416508.3417114.
[24] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56–69, 2020.

[25] R. Jalloul, H. K. Chethan, and R. Alkhatib, "A review of machine learning techniques for the classification and detection of breast cancer from medical images," *Diagnostics*, vol. 13, no. 14, p. 2460, Jul. 2023, doi: 10.3390/diagnostics13142460.
[26] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications*, vol. 134, pp. 93–101, 2019, doi: 10.1016/j.eswa.2019.05.028.
[27] D. P. M. Abellana and D. M. Lao, "A new univariate feature selection algorithm based on the best–worst multi-attribute decision-making method," *Decision Analytics Journal*, vol. 7, p. 100240, 2023, doi: 10.1016/j.dajour.2023.100240.
[28] A. Ogunleye and Q. G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, 2020, doi: 10.1109/TCBB.2019.2911071.
[29] X. B. Kang, G. F. Lin, Y. J. Chen, F. Zhao, E. H. Zhang, and C. N. Jing, "Robust and secure zero-watermarking algorithm for color images based on majority voting pattern and hyper-chaotic encryption," *Multimedia Tools and Applications*, vol. 79, pp. 1169–1202, 2020, doi: 10.1007/s11042-019-08141-1.
[30] M. L. Huang and T. Y. Lin, "Dataset of breast mammography images with masses," *Data in Brief*, vol. 31, p. 105928, 2020, doi: 10.1016/j.dib.2020.105928.

BIOGRAPHIES OF AUTHORS

Alaa Mohamed Ghazy has received B.Sc. degree from Department of Communications and Electronics Engineering, Misr high institute for engineering and technology, Egypt. She is now an PHD student at Dept. of Electronics and Communications Eng. Faculty of Engineering, Mansoura University, Egypt. Her research interest is in deep learning applications in the medical field. She can be contacted by email: alaamohamedghazy@gmail.com.

Hala Bahy Nafea has received B.Sc., M.Sc., and Ph.D. degrees from Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Egypt. She is now an associate professor in Dept. of Electronics and Communications Eng., Faculty of Engineering, Mansoura University. Her main research interests are in Digital Communications, Mobile Communications and Communications Networks. She can be contacted by email: halabahyeldeen@mans.edu.eg.

Hanan Mohamed Amer received the B.Sc., M.Sc. and Ph.D. degrees from the Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, in 2007, 2011 and 2018, respectively. She is currently working as an Assistant Professor at the Electronics and Communications Department, Faculty of Engineering, Mansoura University. As well as she is the coordinator of the medical engineering program for postgraduate studies at Mansoura University. She has published more than 14 articles and supervised 35 postgraduate students at Mansoura University. She is interested in Artificial intelligent and digital image processing. She can be contacted by email: eng_hanan_2007@mans.edu.eg.

Fayez Wanis Zaki B.Sc., M.Sc., and Ph.D., is a professor at the Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Egypt. His main research interests are in Digital Communications, Mobile Communications, Communications Networks, Speech and Image Processing. He is with Dept. of Electronics and Communications Eng., Faculty of Engineering, Mansoura University since 1969. He received his Ph.D. from Liverpool University in 1982. He supervised several M.Sc. and Ph.D. theses. He is now a member of the professorship promotion committee in Egypt. He can be contacted by email: fwzaki@mans.edu.eg.