

Convolutional neural networks breast cancer classification using Palestinian mammogram dataset

Hanin Saadah¹, Amani Yousef Owda¹, Majdi Owda²

¹Department of Natural Engineering and Technology Sciences, Faculty of Graduate Studies, Arab American University, Ramallah, Palestine

²Faculty of Data Science, UNESCO Chair on Data Science for Sustainable Development, Arab American University, Ramallah, Palestine

Article Info

Article history:

Received Feb 29, 2024

Revised Aug 1, 2024

Accepted Aug 5, 2024

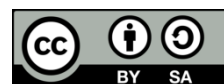
Keywords:

Breast cancer
Convolutional neural network
Image processing
Machine learning
Mammogram

ABSTRACT

Breast cancer is widespread across the globe. It's the primary cause of death in cancer fatalities. According to the Palestinian Ministry of Health annual report, it ranked as the third reported death of all reported cancer deaths in the West Bank. Mammogram screening is the most common technique to diagnose breast abnormalities, but there is a challenge in the lack of skilled experts able to accurately interpret mammograms. Machine learning plays an important role in medical image processing particularly in early detection when the treatment is less expensive and available. In this paper we proposed different convolutional neural network (CNN) models to detect breast abnormalities with promising results. Six CNN models were used in this research on a unique (first-hand) dataset collected from the Palestinian Ministry of Health. The models are VGG16, VGG19, DenseNet121, ResNet50, Xception, and EfficientNetB7. Consequently, DenseNet121 outperformed other models with 0.83 and 0.85 for testing accuracy and area under curve (AUC) respectively. As a future work, the outperformed model can be combined with other patient data like genetic information, medical history, and lifestyle factors to evaluate the risk of developing specific diseases. This would increase the survival rate and enable proactive measures.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Amani Yousef Owda
Department of Natural Engineering and Technology Sciences, Faculty of Graduate Studies
Arab American University
P600, Ramallah, Palestine
Email: amani.owda@aaup.edu

1. INTRODUCTION

Breast cancer is prevalent globally, and it's the first cause of cancer-related fatalities in women, particularly impacting those in low-and middle-income communities [1]. In the West Bank, breast cancer was the third reported death of all reported cancer deaths in 2021 as shown in Figure 1, and it accounted for 30% of deaths among women of reproductive age [2]. While breast cancer is typically diagnosed more frequently in women aged fifty years or older, there is a concerning increase in its occurrence among younger women. Currently, breast cancer ranks as the second leading cause of cancer-related deaths in women aged less than 40 globally [3]. Approximately 30% of breast cancer cases are preventable through modifiable risk factors like excess body weight, physical inactivity, and alcohol intake, and can be further reduced through mammography screening and advances in treatment [4]. Throughout a woman's life, the breast undergoes numerous transformations, spanning from infancy, adolescence, and motherhood including pregnancy and

breastfeeding, all the way to the menopausal stage. The breast is a glandular organ located on the chest wall of females. Though not anatomically categorized as part of the body's upper limb, it resides in the pectoral region, with its blood supply and lymphatic drainage primarily directed toward the armpit. It's a specialized accessory gland of the skin that secretes milk and exists in both males and females [5]. In males and immature females, their structure is similar.

There are many symptoms associated with breast cancer like shape, size, color, pain, or swelling in the breast or nipple. Nipple inversion turning inward or discharge other than milk, and redness or thickness in the underarm area. Symptoms can be physically obvious or tangible, while others may be subtle or intangible [6]. Many of these symptoms may occur at the same time. The cancer clinicians called this case symptom clusters [7]. A study in Jordan showed that there are five main symptom clusters among breast cancer women in Jordan [8]. The clusters are fatigue, pain, treatment side effects, psychological, nausea, and vomiting. On the other hand, breast cancer has many associated risk factors that affect increased risk, and these risk factors can be categorized into two categories; lifestyle and genetic [3], [9]. Arli *et al.* [10], conducted a cross-sectional study conducted in Turkey identified multiple births, a short breastfeeding period, overweight, low socioeconomic level, and low level of education as the most significant factors. Early detection is the cornerstone of controlling breast cancer and improving outcomes and survival. The World Health Organization (WHO) defined the early diagnosis of cancers as the detection of initial phases in women with symptoms, as this facilitates straightforward and cost-effective treatment, leading to elevated rates of recovery [11]. The Pan American Health Organization (PAHO) considered raising awareness through programs and education about breast cancer as the most important key element that improves outcomes, and women should be empowered to access cancer services timely [12].

Currently, there are multiple breast screening techniques available, such as mammography, ultrasound imaging, positron emission tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), and microwave imaging [13]. By far, studies have proven that mammography is the most sufficient evidence and effective in reducing mortality [14], [15]. Mammography is a recommended initial screening imaging method that uses a breast-specific X-ray imaging modality to produce images of the breast in various positions, helping to detect abnormalities and lesions [16]. A cross-sectional study was carried out in Palestine in 2016 to assess mammogram screening among women at risk [15]. The study revealed that 50% of women had undergone at least one mammogram, but only 21% had received timely mammograms. According to the 2021 health annual report released by the Palestinian Ministry of Health, 39.6% of all cases examined across the governorates, totaling 5,864, were found to be abnormal, with a total of 2,322 cases identified [2].

The breast imaging reporting and data system (BI-RADS) was introduced by the American College of Radiology (ACR) in 1993 [17]. It's a standard scale for mammogram reporting, aimed at enhancing communication among healthcare providers, minimizing ambiguity surrounding mammogram results, facilitating case management, and aiding in the monitoring of outcomes [18]. BI-RADS system is commonly used in mammograms, MRIs, and ultrasounds. It employs a numeric scale that ranges from 0 to 6, with each category denoting distinct levels of suspicion concerning breast cancer. Table 1 shows different BI-RADS categories and the assigned descriptions [19]–[23]. There are seven categories in the BI-RADS scale starting from zero to six. Each category has a special finding, management, and the percentage of these abnormalities being cancerous. The first category (zero) represents an incomplete assessment and no diagnosis can be extracted from the image. The incomplete assessment may be attributed to the mammogram device malfunctioning or the breast position and in this case, the woman is asked to repeat the screening. Categories of one to three most probably have benign or normal findings and have less than a 2% chance of being malignant. Furthermore, the chance that the abnormalities are cancerous increases as the scale increases from 4 to 6, and the probability approaches 100% in the sixth group, where the abnormalities have been diagnosed as malignant.

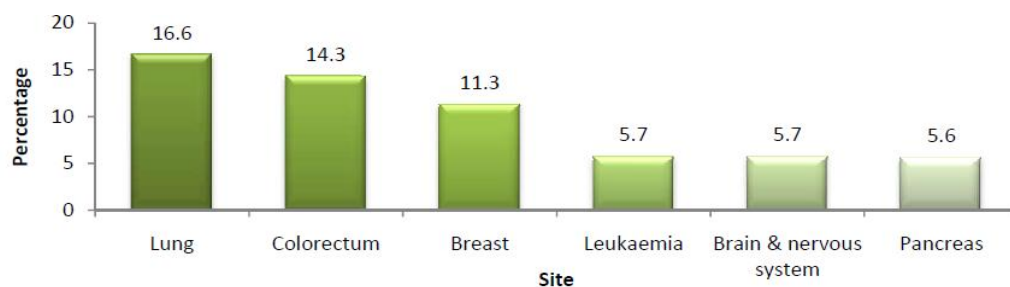


Figure 1. Proportional distribution of the most reported cancer deaths of West Bank in Palestine, 2021

Table 1. BI-RADS categories

Category	Findings/diagnosis	Management	Likelihood of developing breast cancer
0	Incomplete assessment	Additional imaging is required	-
1	Negative or normal findings	Routine follow-up every year	0%
2	Benign	Routine follow-up every year	0%
3	Probably benign	Routine follow-up at short intervals (every 6 months)	<=2%
4	Suspicious findings for malignancy	Biopsy should be considered	>2% to <=95%
5	Highly suggestive of malignancy	The doctor's decision and appropriate action should be taken. Biopsy is required	>=95%
6	Proven malignancy	Appropriate therapy/surgery	100%

2. PROBLEM STATEMENT

A significant volume of mammograms is gathered through an extensive mammography screening program and necessitates evaluation from proficient radiologists, who are qualified but overloaded with excessive workloads [24]. An important challenge in the field of mammography is the shortage of skilled domain experts capable of accurately interpreting mammograms. This scarcity of radiologists and clinicians specialized in breast imaging can lead to delays in diagnosis, increased workloads for existing experts, and potential errors in the interpretation process. The shortage can be attributed to various factors like inadequate training. By addressing the shortage of domain experts and implementing strategies to improve interpretation processes, healthcare systems can provide more timely and accurate diagnoses, ultimately improving patient outcomes in the realm of mammography. Therefore, there is an unmet need to develop machine learning (ML) models to assist radiologists with mammographic interpretation, and ML model development requires interdisciplinary research that integrates medical science and engineering [25].

3. LITERATURE REVIEW

ML plays a significant role in early detection across various fields, including healthcare, and has emerged as a powerful tool in mammography interpretation, revolutionizing the field of breast cancer screening and diagnosis [26]. By utilizing ML techniques, medical professionals can develop models that help detect diseases at their early stages, when intervention and treatment are most effective [27]. For example, in cancer detection, ML algorithms can analyze medical imaging data, such as mammograms or MRIs, to identify suspicious lesions or abnormalities that may require further investigation [28]. While classical ML has demonstrated its efficiency, the prevailing approach among researchers is deep learning (DL), which offers more potent techniques, particularly in the medical imaging domain, notably in mammography. Different studies and models were developed to detect breast cancer using DL techniques [29]–[31]. Researchers used many pre-trained CNN algorithms in this regard like AlexNet, VGG, ResNet, GoogleNet, and Inception [32]. Table 2 summarizes some of the studies conducted regarding breast cancer.

Table 2. Summary of studies in breast cancer

Reference	Dataset/input	Targeted classes	Algorithms	Outcome
[33]	Collected dataset of digital screening mammograms	Density and risk (dense, no dense, low risk, and high risk)	Logistic regression, ResNet18	AUC = 0.70
[34]	CBIS-DDSM, INbreast	ROI annotations	VGG16, ResNet50	AUC (DDSM) = 0.88 AUC (INbreast) = 0.95
[35]	15 microcalcifications features and 26 breast masses features	Breast lesions and microcalcifications	A stacked autoencoder (SAE) with n layers, SVM	Accuracy = 85.8%
[36]	INbreast	Malignant or benign lesions	Faster R-CNN	AUC = 0.95
[37]	MIAS, INbreast, ImageNet, and private databases	Normal and suspected ROI	(1) ConvNet+SVM (2) VGGNet16, (3) VGGNet19, (4) GoogLeNet, (5) MobileNetV2, (6) ResNet50, (7) DenseNet121 ConvNet	AUC: (1) 91.4, (2) 86.4, (3) 87.6, (4) 79.4, (5) 66.8, (6) 75.8, (7) 81.7
[38]	MIAS and private dataset	Benign and malignant ROI	ConvNet	AUC = 0.99

In previously mentioned studies, the datasets used are common public datasets like MIAS, DDSM, and INbreast. In this study, different pre-trained CNN models were implemented on a local first-hand dataset from the West Bank in Palestine. The models are: VGG16, VGG19, DenseNet121, Xception, ResNet50, and EfficientNetB7. The main task was to detect abnormalities in the breast and classify them into normal and abnormal. This contribution will add value to healthcare in clinical use since there is no such computer aided detection (CAD) system or decision support system (DSS) implemented in the mammogram departments at the Palestinian MoH to give the clinicians a second opinion which will improve the early detection of breast abnormality.

The following sections present the methodology for implementing the deep learning models to diagnose breast abnormalities and classify them into normal and abnormal. The methodology includes describing the collected dataset and how the images were pre-processed. The coming sections also illustrate some extracted insights from the data in the exploratory data analysis (EDA) stage. In addition, the implemented models in this study besides the performance measures are presented. Finally, the results will be presented and discussed in addition to the conclusion and future work.

4. METHOD

This section presents the methodology used in this study to detect breast abnormalities and classify them into normal and abnormal. Figure 2 shows the outline workflow for the methodology and Figure 3 illustrates the modelling approach. The proposed approach consists of six stages to address the research objective. The first stage talks about the dataset and how it's labeled while the second stage is about the pre-processing including the techniques used to prepare the mammogram images to be ready for the next stage. Additionally, stage three extracts the insights from the dataset and stage four includes splitting the dataset into subsets. In the modelling stage, the dataset is fed into six CNN models while in the last stage, the models are tested using different performance measures. All stages are presented in the following subsections.

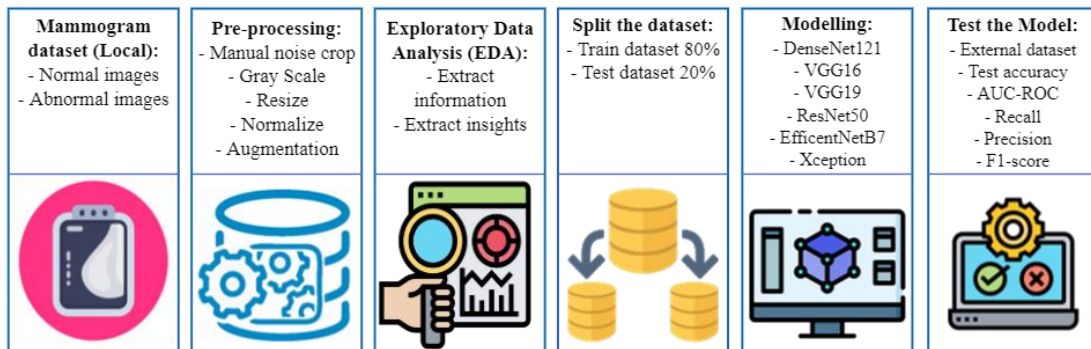


Figure 2. Research methodology conducted in this research

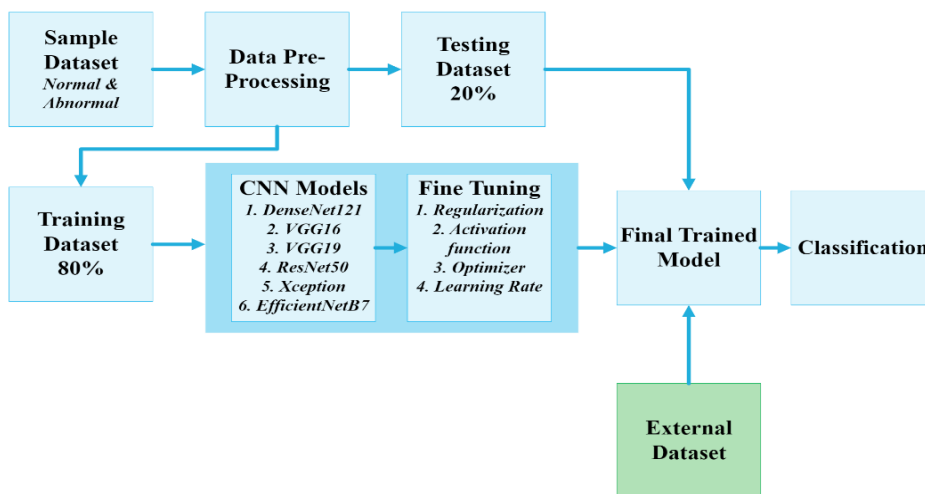


Figure 3. The modelling approach block diagram

4.1. Dataset

The first stage is data collection, the data was collected from the Bethlehem Mammography Center. It is one of the centers of the Palestinian Ministry of Health that possesses the necessary data that meets the conditions for conducting such studies. A total of 813 mammogram images were collected from 174 ladies who were screened at the center before July 2023.

The collected dataset included images taken from both breast sides. Each side is imaged in different orientations with two different views namely mediolateral oblique (MLO) and cranial caudal (CC). The MLO view is taken from the center of the chest outward, whereas the CC view is taken from above the breast. The MLO view offers a more extensive perspective of the upper-outer quadrant, providing an optimal visualization of the breast's lateral side. The images were converted by the center staff from DICOM to JPG format with a standard size of 4,710×5,844 pixels and were labeled using a BI-RADS score from 1 to 6 according to the doctor’s findings.

4.2. Pre-processing

In the second stage, the images were labeled as normal and abnormal instead of BI-RAD scores. Images of scores 1 to 3 were labeled as normal, and the images of scores 4 to 6 as abnormal. After that, a random sample was taken from the main dataset. The sample included 200 images categorized into two groups; 100 for the normal group, and 100 for the abnormal group. In addition, a small subset of 18 images (10 normal, 8 abnormal) were also taken from the original dataset to be fed into the trained models for external test purposes. The last dataset acts as the external dataset.

Moreover, the images in the random sample went through different techniques shown in Figure 4 to reduce the noise and clean the images to prepare them for the next stage. These techniques are: 1) Manually cropping text and pectoral muscle. 2) Convert images from RGB to Grayscale to reduce the dimensions and enhance computational time. 3) Resizing from the original size to 224×224 to be compatible with the input size of the pre-trained model used later. 4) Normalizing images to get a standard range from 0 to 1 to simplify the process. 5) Finally, augmentation was implemented on the random sample to reduce overfitting that might happen in the small dataset. The augmentation step resulted in creating 10 images from each image so the final number of images was 1,000 for normal and 1,000 for abnormal. The tools used in this stage and the next stages are Python and Jupyter Notebook and different libraries such as CV2, Keras, Tensorflow, and Sklearn. These libraries are commonly used in classification and prediction tasks.

As a result of this stage, 2,000 preprocessed images were created from the images in the random dataset; 10 images from each image. Figure 5 is a picture that shows a sample image from the random dataset before and after preprocessing. This picture includes 3 rows of images. The image located in the first row on the left is the original image before pre-processing. The image included annotations such as the patient's details, the view mode (MLO), other unwanted text, and the pectoral muscle. In addition, the image on the top right is after cropping the annotations, the pectoral muscle, and converting the color to grayscale. Finally, the ten images in the second and third rows are resized images into 224×224 and then normalized, and augmented. Moreover, the images in the second and third rows are the 10 images that were created from the original in the augmentation process by using techniques like rotation, width and height shift, shear, zoom, horizontal and vertical flip.

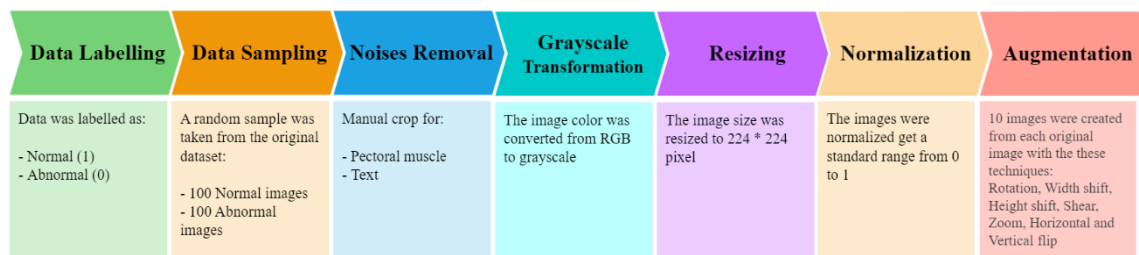


Figure 4. Data pre-processing methodology in this research

4.3. Exploratory data analysis

This stage helps in understanding the raw data and converting it to useful information and insights that can be used in making important decisions. Different visualizations have been extracted from the dataset used in this study. Seven columns were used to describe the data; BIRAD, side, year of birth, age in years, view, class, and age group. BIRAD is the global score for mammograms from 1 to 6, the side indicates whether the image is taken from a lady that has abnormalities on the left, right, or both sides of the breast.

The view indicates a CC or MLO view, class is the target column that says if the image is normal or abnormal, and finally age group categorizes the images into 6 groups which are: 30-39, 40-49, 50-59, 60-69, 70-79, 80-89. Figures 6 to 8 illustrate some of the most important insights.

Figure 6 is a chart to illustrate the distribution of screened images among the age group and the normality class. The majority of screened images are related to age group 50 to 59. This means that the largest group of women who have undergone a mammogram are between the ages of fifty and fifty-nine. This age group shows a close number of images between the normal and abnormal classes where there are 43 images in the abnormal and 41 in the normal. On the other hand, there are a small number of women in the sample who are under the age of thirty-nine and those who are over seventy. These age groups are shown more in the abnormal class than the normal. This means that the abnormalities are usually developed in women in older ages.

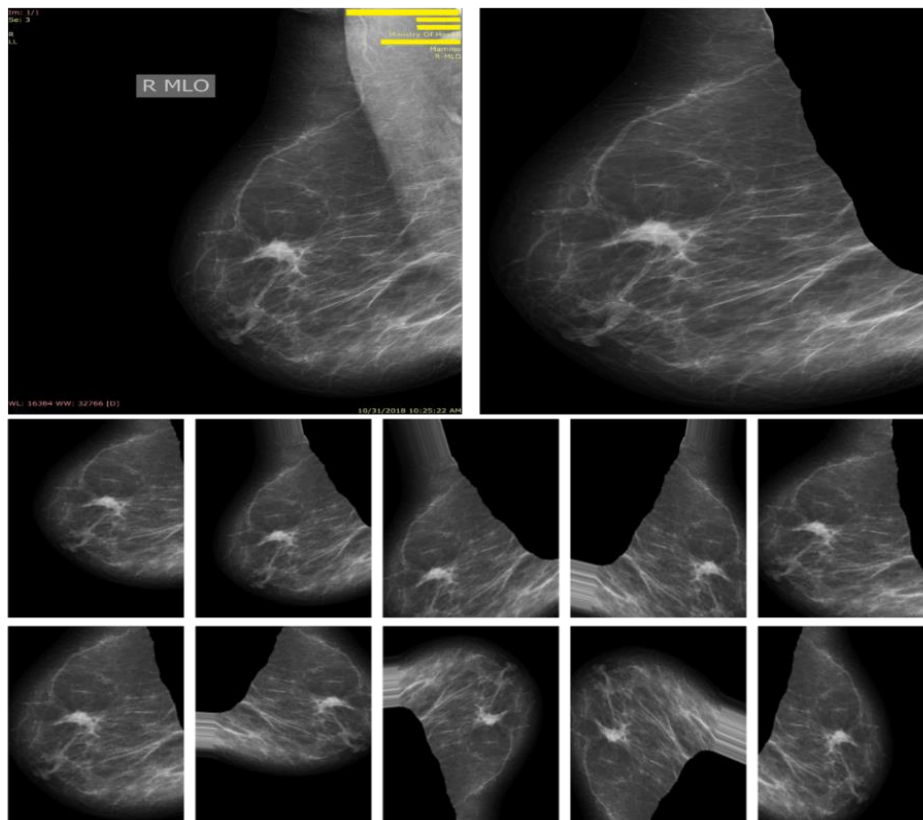


Figure 5. Mammogram image before and after pre-processing

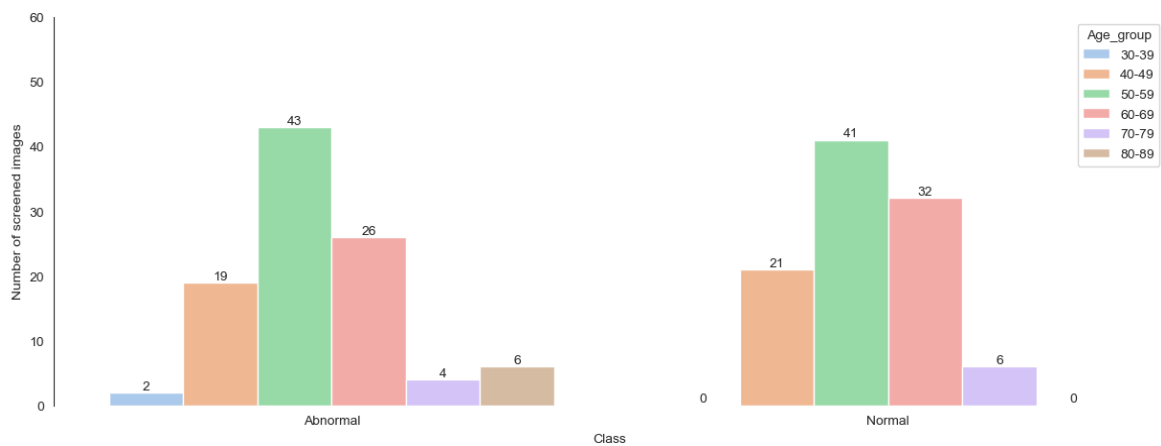


Figure 6. Number of screened images versus class and age group

Moreover, Figure 7 represents a chart of the number of screened images versus class and side. It can be noted that the largest number of images were taken from women with the possibility of cancer cells on the left side of the breast, and this group constitutes half of the sample, while the number of images taken from women with the possibility of cancer cells on the right side is less. Also, there are a good number of images taken from women suffering from possible cancer cells on both sides of the breast. Both classes show a close number of images taken on each side.

Similarly, the chart in Figure 8 shows the number of images taken from the CC and MLO views and how they are distributed between the normal and abnormal classes. In the abnormal sample, the numbers of the CC and MLO images are almost equal where there are 51 and 49 images for the CC and MLO respectively. Likewise, in the normal sample, there is a noticeable difference in the numbers, as the number of images in the CC view was higher and had 57 images while the MLO view had 43 images.

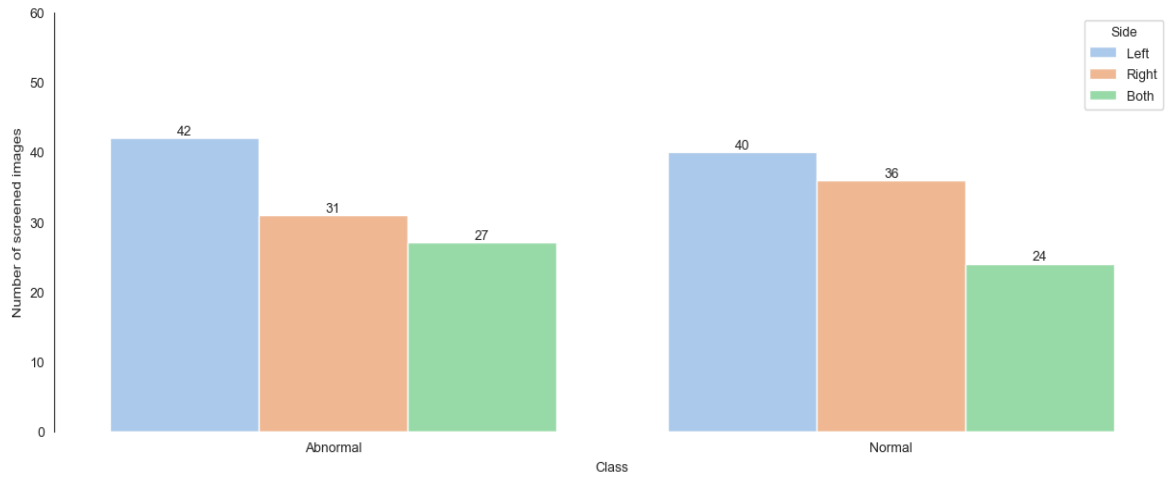


Figure 7. Number of screened images versus class and side

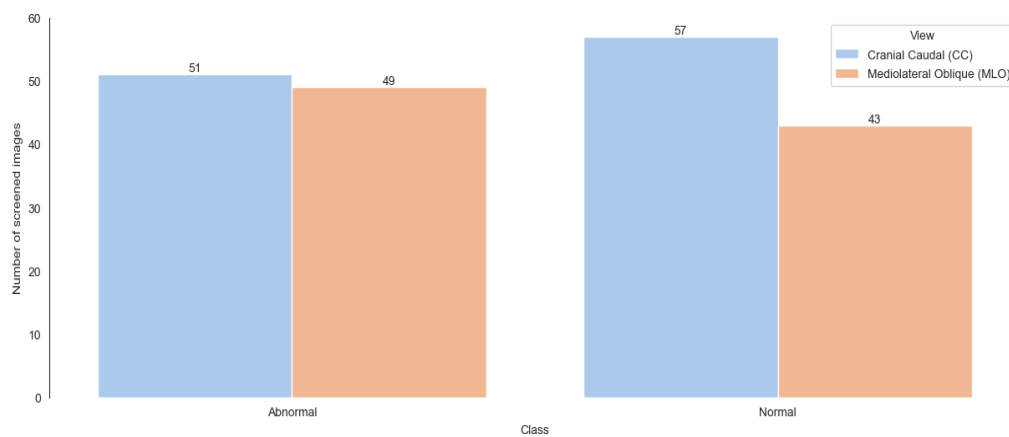


Figure 8. Number of screened images versus class and view

4.4. Data splitting

The process of splitting the data was in the fourth stage where the 2,000 images after augmentation were split into two subsets. The first subset is used to train the model and constitutes 80% of the dataset, and 20% for the testing subset which is used to test the real accuracy and performance of the model.

4.5. Modeling

In this paper, a transfer learning approach was followed and six DL pre-trained models were used. The training subset was the input for all models. The following subsections briefly present the models used in this research.

4.6.1. DenseNet121

Densely connected convolutional network (DenseNet) is one of the CNN models that achieved high performance in the literature compared to other pre-trained models. It's constructed from 120 convolutions and 4 AvgPool and includes four components which are connectivity, DenseBlocks, growth rate, and bottleneck layers. DenseNet takes the input from all previous layers' connected outputs to the dense block layer [39]. Figure 9 presents the schematic layout of the DenseNet [40]. The connectivity pattern is the main idea behind the DenseNet structure where the feature maps from previous layers are concatenated onto the inputs of future layers. Continually concatenating results in very deep inputs. There are many activation functions such as rectified linear unit (ReLU) are utilized to increase the non-linearity in the pooling layers when the feature maps are fed into the coming layers. In (1) illustrates the mathematical formula for this schema.

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (1)$$

Where l^{th} is the layer that receives the concatenated feature maps x_0, x_1, \dots, x_{l-1} from all previous layers. H_l is a composite function of operations like batch normalization (BN), a ReLU, and a 3×3 convolution (conv) that generates the K levels for the mapping features in the coming layer which determines the growth rate in the network. The growth rate K is calculated in (2).

$$k_l = k_0 + k * (l - 1) \quad (2)$$

Where k_0 is the number of channels in the input layer.

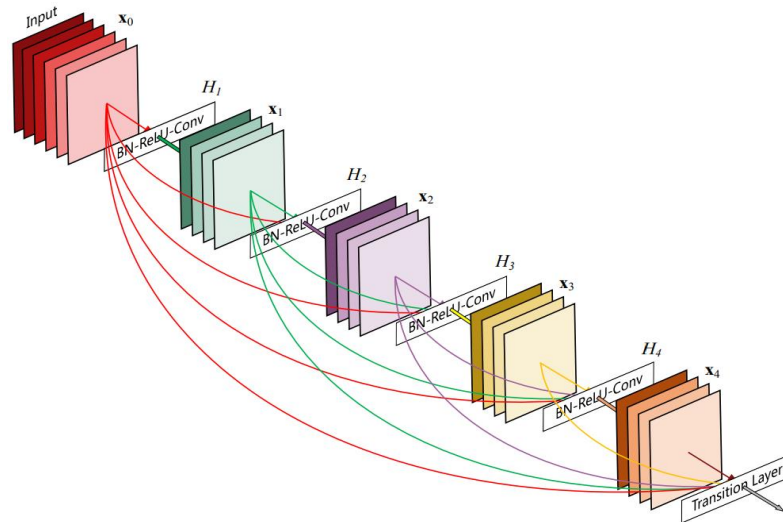


Figure 9. DenseNet schematic layout

4.6.2. VGG16 and VGG19

VGG stands for visual geometry group. VGG16 and VGG19 are also CNN models that are commonly used with image classification and detection. Their architecture is the same and the only difference is the number of layers. VGG16 is constructed from 13 convolutional layers and 3 fully connected layers while in VGG19 there are 16 convolutional layers and 3 fully connected [41]. The image shape should be 224×224 in both models.

4.6.3. ResNet50

Residual network; ResNet is a CNN model that is used to detect objects. It was the result of adding residual structure to the CNN to fix the gradient degradation and disappearance issue that frequently occurs in the training process [42]. ResNet50 contains 50 layers that are distributed over four types of layers which are convolutional, pooling, fully connected, and shortcut [43]. It can perform efficiently even with a large number of layers.

4.6.4. EfficientNetB7

EfficientNets is a family of models that obtained better efficiency and accuracy than the previous CNNs. EfficientNet-B7 in particular, achieved top-1 accuracy on ImageNet of all EfficientNet families [44].

4.6.5. Xception

The extreme inception; Xception was proposed by Google that utilizes CNN structure and relies on depthwise-separable convolution [45]. It was used in a lot of classification and recognition tasks with high accuracy. According to the official documentation, Xception architecture involves 36 convolutional layers that form the feature extraction base of the network [46].

4.7. Testing

The last stage in this research was the testing part where the testing subset was used as the input for the trained model to get the best model results in detecting breast abnormalities. Moreover, the trained models were implemented on the external dataset as well. To measure the models' performance, common ML performance metrics were used such as accuracy, recall, precision, and F1-score. These measures will be discussed in the next subsections.

4.7.1. Accuracy

Accuracy is the simplest and the most used measure in the classification studies. It measures how often the model correctly predicts the outcome. In general, it represents the ratio of the correct predictions of all predictions. The accuracy formula is shown in (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (3)$$

Where:

TP: true positive; when the model predicts the positive sample correctly.

TN: true negative; when the model predicts the negative sample correctly.

FP: false positive; when the model predicts the positive sample but it's negative in reality.

FN: false negative; when the model predicts the negative sample but it's positive in reality.

4.7.2. Recall

Recall is the metric that measures how often the model identifies "true positives" correctly from all ground truth positives. It's used to extract the ratio of true positives to all positives in the ground truth. Moreover, it represents type-II errors which occur when the model accepts the false null hypothesis (H_0). The formula is shown in (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

4.7.3. Precision

Precision is the ability of the model to predict the positive class correctly. It represents the ratio of "true positive" to all predicted positives. Additionally, it focuses on type-I error which occurs when the model rejects the true null hypothesis (H_0) by mistake. The formula is shown in (5).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

4.7.4. F1-score

The last measure is the F1-score. It's the harmonic mean of precision and recall. It's used in binary and multi-class classifications. The formula is shown in (6).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

4.8. Python code

In this study, Python language was used to achieve the goal of this study which is detecting breast abnormalities in mammogram images. Table 3 lists the detailed steps to reproduce the findings obtained in this study.

Table 3. Python functions

Step	Stage	The function used in Python
1	Import libraries	CV2, OS, Keras, numpy, tensorflow, matplotlib, sklearn
2	Load the dataset and set the saving location	data_directory = "your dataset directory" save_directory = "the directory of saved model"
3	Preprocessing	grayscale: cv2.cvtColor resize: cv2.resize normalize: img / 255.0 Augmentation: Image Data Generator ()
4	Split the dataset	def split_dataset (images, labels, test_size=0.2, random_state=42)
5	Modelling with L2 regularization	def create_efficientnetb7_model (input_shape= (224, 224, 3), num_classes=2, regularization_strength=0.01): base_model = EfficientNetB7 (weights= 'imagenet', include_top=False, input_shape=input_shape) model = models.Sequential ([])
6	Train the model	def train_model (model, train_images, train_labels, val_images, val_labels, epochs=20, batch_size=32): early_stopping = callbacks.EarlyStopping (patience=5, restore_best_weights=True) model.compile (tf.keras.optimizers.Adam (learning_rate=0.001), loss= 'sparse_categorical_crossentropy', metrics=['accuracy']) history = model.fit (train_images, train_labels, epochs=epochs, validation_data= (val_images, val_labels), batch_size=batch_size, callbacks=[early_stopping]) model.save ('trained_EfficientNetB7.keras') return model, history
7	Evaluation	cm = confusion_matrix (test_labels, predictions) display = Confusion Matrix Display (confusion_matrix=cm, display_labels=['Abnormal', 'Normal']) display.plot (cmap= 'Blues', values_format= 'd') plt.title ('Confusion Matrix2') plt.show()

5. RESULTS AND DISCUSSION

This study implemented a methodology approach to investigate the effects of the power and effectiveness of machine learning models in detecting breast abnormalities in local mammogram images while others focused on using the public commonly used datasets like MIAS and Inbreast. The approach used CNN deep learning models, namely DenseNet121, ResNet50, VGG16, VGG19, Xception, and EfficientNetB7. The input of these models was a training subset of size 1,600 was used to learn the models, while a testing subset of size 400 was used for testing and evaluation. The experiments in this approach were done using the same parameter values across all models. The model's input shape was 224×224 and the test subset size was 0.2 of the original dataset. In addition, the batch size was 32 with 20 epochs and Adam optimizer was deployed with a 0.001 learning rate. Also, the activation functions used in the models were ReLU and Softmax. Finally, the approach classified the images into normal and abnormal and was evaluated using different performance measures such as confusion matrix, accuracy, recall, precision, F1-score, and ROC curve. The results from this approach will be presented and discussed in this section.

Based on the experiment, we found that DenseNet121 outperformed other models in detecting abnormalities. Table 4 lists all obtained results from performance measures versus models in the testing subset and external dataset. It can be seen that the accuracies in the testing and external sets ranged between 0.83 and 0.50. In the testing subset, DenseNet121 outperformed all other models with the highest accuracy, AUC, recall, precision, and F1-score. The highest accuracy for detecting breast abnormalities obtained was 0.82 from DenseNet121, while EfficientNetB7 showed promising and close results of DenseNet121. EfficientNetB7 accuracy was 0.81 and got 0.80, 0.78, 0.82, and 0.80 as AUC, recall, precision, and F1-score respectively. Moreover, VGG16, VGG19, and Xception got accuracies of 0.74, 0.68, and 0.78 respectively. The performance in these models was better in detecting the abnormal class than the normal class. Furthermore, the lowest accuracy was 0.50 from ResNet50 where the model predicted all images as normal class. Likewise, a second experiment was conducted using the saved trained model from the first experiment on an external image as explained in the dataset section above. The results obtained from this experiment are shown in Table 4. The results here are very similar to the results obtained from the first experiment. In this case, DenseNet121 produced the best results in terms of performance and outperformed other models. DenseNet121 obtained accuracy, AUC, recall, precision, and F1-score of 0.83, 0.85, 0.70, 1.0, and 0.82

respectively. Unlike the first experiment, there was a notable difference between DenseNet121 and EfficientNetB7, where the accuracy reached 0.73 in the second. The accuracies of VGG16 and Xception were within a narrow range of each other, where they got 0.61 and 0.67 in order. Finally, the lowest accuracy was for the ResNet50 and VGG19 where they got an equal value of 0.56. VGG19’s performance was better in predicting the abnormal images while ResNet50 predicted all images as normal like in the first experiment. In both cases, ResNet50 acted like a random classifier. Overall, the performance in these models was better in detecting the abnormal class than the normal class.

Table 4. Performance results for the models on the testing and external datasets

Model vs performance measures	Testing subset					External dataset				
	Accuracy	ROC-AUC	Recall	Precision	F1-score	Accuracy	ROC-AUC	Recall	Precision	F1-score
DenseNet121	0.82	0.82	0.82	0.82	0.82	0.83	0.85	0.70	1.0	0.82
VGG16	0.74	0.74	0.64	0.80	0.71	0.61	0.62	0.50	0.71	0.59
VGG19	0.68	0.68	0.60	0.72	0.66	0.56	0.57	0.40	0.67	0.50
Xception	0.78	0.78	0.73	0.81	0.77	0.67	0.68	0.60	0.75	0.67
ResNet50	0.50	0.50	1.0	0.50	0.67	0.56	0.50	1.0	0.56	0.71
EfficientNetB7	0.81	0.80	0.78	0.82	0.80	0.73	0.74	0.60	0.86	0.71

In the same manner, Figure 10 shows the confusion matrix for the DenseNet121 in the first experiment in Figure 10(a) and 10(b) shows the second experiment. In Figure 10(a), DenseNet121 predicted 81.5% of all abnormal images correctly while misclassified 18.5% and classified them as normal. In contrast, 82.5% were classified correctly of all normal images but misclassified 17.5%. By the same token, in Figure 10(b), DenseNet121 succeeded in detecting all the abnormal images in the external dataset. On the other hand, DenseNet121 correctly detected 70% of the normal images and misclassified 30% of them. Based on the results from the two experiments above, it’s clear that the DenseNet121 is the best model out of the six models used in this research for medical image classification.

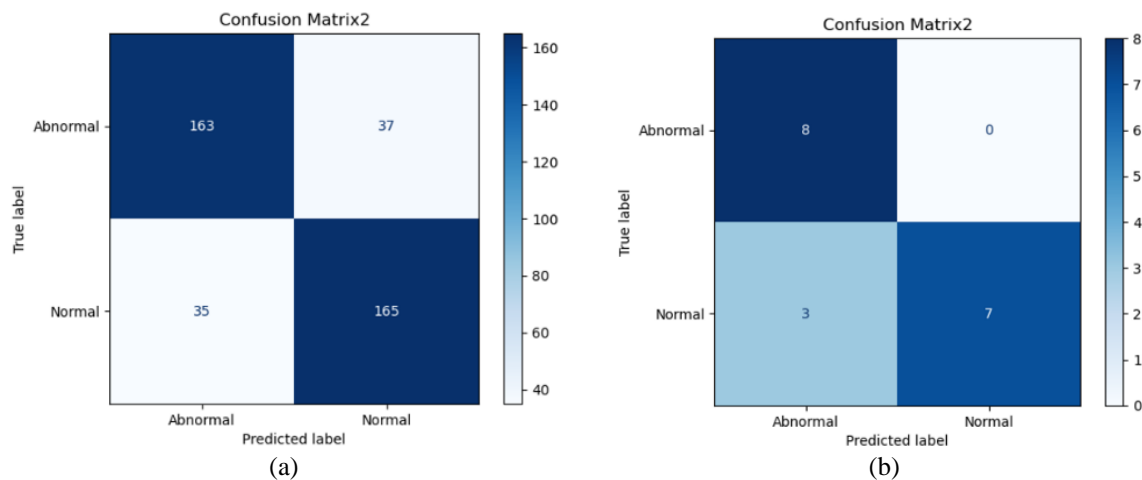


Figure 10. Confusion matrix results for DenseNet121 in (a) the first experiment and (b) second experiment

Based on the results derived from the first-hand dataset, the models succeeded in classifying the breast abnormalities efficiently with an accuracy of 0.83. These models can be generalized to be utilized in other tasks specifically, for detecting cancer diseases. In addition, it can be used as a decision support system in the mammogram centers at the Palestinian MoH for early detection of breast cancer. To the best of the author's knowledge, there are no such studies on mammogram images conducted in Palestine from the same perspective. Moreover, the results of this study are high compared to other studies using similar datasets; however, the ResNet50 showed low accuracy and this can be attributed to the small size dataset which is one of this study’s limitations.

6. CONCLUSION

Breast cancer is the top one common cancer in women globally and the primary cause of death in cancer deaths. It represented 30% of deaths among women of reproductive age in the West Bank in Palestine. Mammogram screening is the most common and efficient method to diagnose abnormalities in the breast. Early detection is very important to control cancer and improve the outcome because at this stage the treatment is cheap and effective. Machine learning plays a significant role in detecting breast cancer. Many studies proposed different CNN models to diagnose breast cancer with promising results. This study implements six pre-trained CNN models that obtained the best results in the state-of-the-art on a first-hand dataset collected from an MoH mammogram screening center in Palestine to investigate the ability of these models to interpret and detect abnormalities in the breast. The models are: DenseNet121, VGG16, VGG19, ResNet50, Xception, and EfficientNetB7. Based on the results obtained from this research, DenseNet121 achieved the highest testing accuracy of 0.83, and an AUC of 0.85, and it outperformed all other models. Based on the results, the presented work has proven its power and effectiveness in detecting breast suspected abnormalities. Additionally, the developed models in this study have obtained a high accuracy which could be implemented at the clinical level in the MoH mammogram centers in Palestine as a CAD. This CAD acts as a DSS to detect breast abnormalities at early stages when the treatment is affordable and available. In future work, these models will be used with other patient data such as genetic information, lifestyle factors, and medical history, to assess an individual's risk of developing certain diseases. This enables proactive measures, such as preventive screenings or lifestyle modifications, to be taken for individuals at higher risk.

ACKNOWLEDGEMENTS




Authors are grateful to the Arab American University in Palestine for their support. In addition, to the editors and reviewers for their constructive feedback on the manuscript.

REFERENCES




- [1] World Health Organization, "Global breast cancer initiative implementation framework: assessing, strengthening and scaling up of services for the early detection and management of breast cancer." World Health Organization, Geneva, p. 118, 2023, [Online]. Available: <https://www.who.int/publications/i/item/9789240065987>.
- [2] Ministry of Health (Palestine), "Health annual report palestine 2021," 2022. [Online]. Available: <https://ghdx.healthdata.org/record/palestine-health-annual-report-2021>.
- [3] J. W. Zhu, P. Charkhchi, S. Adekunte, and M. R. Akbari, "What is known about breast cancer in young women?," *Cancers*, vol. 15, no. 6, p. 1917, Mar. 2023, doi: 10.3390/cancers15061917.
- [4] A. N. Giaquinto *et al.*, "Breast cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, no. 6, pp. 524–541, Nov. 2022, doi: 10.3322/caac.21754.
- [5] R. Snell, "The upper limb," in *Clinical Anatomy By Region*, 9th ed., Lippincott Williams & Wilkins, 2011.
- [6] P. Kumari, S. Kumar, B. Shukla, and A. Dubey, "An overview on breast cancer," *International Journal of Medical and all body Health Research*, vol. 2, no. 3, pp. 59–65, 2022.
- [7] K. L. Kwekkeboom, "Cancer symptom cluster management," *Seminars in Oncology Nursing*, vol. 32, no. 4, pp. 373–382, Nov. 2016, doi: 10.1016/j.soncn.2016.08.004.
- [8] M. Al Qadire *et al.*, "Symptom clusters predictive of quality of life among jordanian women with breast cancer," *Seminars in Oncology Nursing*, vol. 37, no. 2, p. 151144, Apr. 2021, doi: 10.1016/j.soncn.2021.151144.
- [9] R. Martín-Payo, A. Martínez-Urquijo, E. Zabaleta-del-Olmo, and M. del Mar Fernandez-Alvarez, "Use a web-app to improve breast cancer risk factors and symptoms knowledge and adherence to healthy diet and physical activity in women without breast cancer diagnosis (Precam project)," *Cancer Causes & Control*, vol. 34, no. 2, pp. 113–122, Feb. 2023, doi: 10.1007/s10552-022-01647-x.
- [10] S. K. Arli, A. B. Bakan, and G. Aslan, "Distribution of cervical and breast cancer risk factors in women and their screening behaviours," *European Journal of Cancer Care*, vol. 28, no. 2, p. e12960, Mar. 2019, doi: 10.1111/ecc.12960.
- [11] O. Ginsburg *et al.*, "Breast cancer early detection: a phased approach to implementation," *Cancer*, vol. 126, no. S10, pp. 2379–2393, May 2020, doi: 10.1002/cncr.32887.
- [12] Pan American Health Organization, "Early detection: breast health awareness and early detection strategies." Pan American Health Organization, Geneva, pp. 4–5, 2016, [Online]. Available: <https://www.paho.org/hq/dmdocuments/2016/KNOWLEDGE-SUMMARY---EARLY-DETECTION.pdf>.
- [13] S. Bagchi, K. G. Tay, A. Huong, and S. K. Debnath, "Image processing and machine learning techniques used in computer-aided detection system for mammogram screening - a review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2336–2348, Jun. 2020, doi: 10.11591/ijece.v10i3.pp2336-2348.
- [14] World Health Organization, "Policy statement and recommended actions for early detection of breast cancer in the Eastern Mediterranean Region." World Health Organization, Cairo, 2016.
- [15] Z. Nazzal, H. Sholi, S. B. Sholi, M. B. Sholi, and R. Lahaseh, "Motivators and barriers to mammography screening uptake by female health-care workers in primary health-care centres: a cross-sectional study," *The Lancet*, vol. 391, p. S51, Feb. 2018, doi: 10.1016/S0140-6736(18)30417-3.
- [16] K. J. Geras, R. M. Mann, and L. Moy, "Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives," *Radiology*, vol. 293, no. 2, pp. 246–259, Nov. 2019, doi: 10.1148/radiol.2019182627.
- [17] D. A. Spak, J. S. Plaxco, L. Santiago, M. J. Dryden, and B. E. Dogan, "BI-RADS ® fifth edition: a summary of changes," *Diagnostic and Interventional Imaging*, vol. 98, no. 3, pp. 179–190, Mar. 2017, doi: 10.1016/j.diii.2017.01.001.

- [18] M. M. Eberl, C. H. Fox, S. B. Edge, C. A. Carter, and M. C. Mahoney, "BI-RADS classification for management of abnormal mammograms," *The Journal of the American Board of Family Medicine*, vol. 19, no. 2, pp. 161–164, Mar. 2006, doi: 10.3122/jabfm.19.2.161.
- [19] S. Obenauer, K. P. Hermann, and E. Grabbe, "Applications and literature review of the BI-RADS classification," *European Radiology*, vol. 15, no. 5, pp. 1027–1036, May 2005, doi: 10.1007/s00330-004-2593-9.
- [20] S. Prabhala, A. Srirambhatla, and S. Pasula, "Comparison of BIRADS lexicon to breast biopsy findings in low resource countries," *ScienceRise: Medical Science*, no. 4(49), pp. 55–60, Jul. 2022, doi: 10.15587/2519-4798.2022.262145.
- [21] M. Eghtedari, A. Chong, R. Rakow-Penner, and H. Ojeda-Fournier, "Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems," *American Journal of Roentgenology*, vol. 216, no. 4, pp. 860–873, Apr. 2021, doi: 10.2214/AJR.20.24894.
- [22] L. Falconi, M. Perez, W. Aguilar, and A. Conci, "Transfer learning and fine tuning in mammogram BI-RADS classification," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Jul. 2020, pp. 475–480, doi: 10.1109/CBMS49503.2020.00096.
- [23] I. Banerjee, S. Bozkurt, E. Alkim, H. Sagreiya, A. W. Kurian, and D. L. Rubin, "Automatic inference of BI-RADS final assessment categories from narrative mammography report findings," *Journal of Biomedical Informatics*, vol. 92, p. 103137, Apr. 2019, doi: 10.1016/j.jbi.2019.103137.
- [24] K.-J. Tsai *et al.*, "A high-performance deep neural network model for BI-RADS classification of screening mammography," *Sensors*, vol. 22, no. 3, p. 1160, Feb. 2022, doi: 10.3390/s22031160.
- [25] W. L. Bi *et al.*, "Artificial intelligence in cancer imaging: clinical challenges and applications," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 2, pp. 127–157, Mar. 2019, doi: 10.3322/caac.21552.
- [26] J. H. Yoon and E.-K. Kim, "Deep learning-based artificial intelligence for mammography," *Korean Journal of Radiology*, vol. 22, no. 8, pp. 1225–1239, 2021, doi: 10.3348/kjr.2020.1210.
- [27] World Health Organization, "Cancer control: knowledge into action: WHO guide for effective programmes," *World Health Organization*. World Health Organization, Geneva, p. 42, 2007, [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK195408/pdf/Bookshelf_NBK195408.pdf.
- [28] M. Tahmoresi, A. Afshar, B. B. Rad, K. B. Nowshath, and M. A. Bamiah, "Early detection of breast cancer using machine learning techniques," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 3–2, pp. 21–27, 2018.
- [29] S. J. Mambou, P. Maresova, O. Krejcar, A. Selamat, and K. Kuca, "Breast cancer detection using infrared thermal imaging and a deep learning model," *Sensors*, vol. 18, no. 9, p. 2799, Aug. 2018, doi: 10.3390/s18092799.
- [30] L. Wang, "Mammography with deep learning for breast cancer detection," *Frontiers in Oncology*, vol. 14, Feb. 2024, doi: 10.3389/fonc.2024.1281922.
- [31] M. Humayun, M. I. Khalil, S. N. Almuayqil, and N. Z. Jhanjhi, "Framework for detecting breast cancer risk presence using deep learning," *Electronics*, vol. 12, no. 2, p. 403, Jan. 2023, doi: 10.3390/electronics12020403.
- [32] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—a comparative study," *Journal of Imaging*, vol. 5, no. 3, p. 37, Mar. 2019, doi: 10.3390/jimaging5030037.
- [33] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, Jul. 2019, doi: 10.1148/radiol.2019182716.
- [34] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific Reports*, vol. 9, no. 1, p. 12495, Aug. 2019, doi: 10.1038/s41598-019-48995-4.
- [35] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li, "Discrimination of breast cancer with microcalcifications on mammography by deep learning," *Scientific Reports*, vol. 6, no. 1, p. 27327, Jun. 2016, doi: 10.1038/srep27327.
- [36] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Scientific Reports*, vol. 8, no. 1, p. 4165, Mar. 2018, doi: 10.1038/s41598-018-22437-z.
- [37] T. Mahmood, J. Li, Y. Pei, and F. Akhtar, "An Automated in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning," *Biology*, vol. 10, no. 9, p. 859, Sep. 2021, doi: 10.3390/biology10090859.
- [38] T. Mahmood, J. Li, Y. Pei, F. Akhtar, M. U. Rehman, and S. H. Wasti, "Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach," *PLOS ONE*, vol. 17, no. 1, p. e0263126, Jan. 2022, doi: 10.1371/journal.pone.0263126.
- [39] A. Ayad and M. E. Abdulmunim, "Detecting abnormal driving behavior using modified DenseNet," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 3, pp. 48–65, Jul. 2023, doi: 10.52866/ijcsm.2023.02.03.005.
- [40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Aug. 2017, vol. 2017-Janua, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
- [41] D. Hindarto, N. Afarini, and E. T. Esthi H, "Comparison efficacy of VGG16 and VGG19 insect classification models," *JIKO (Jurnal Informatika dan Komputer)*, vol. 6, no. 3, pp. 189–195, Dec. 2023, doi: 10.33387/jiko.v6i3.7008.
- [42] X. Du, L. Si, P. Li, and Z. Yun, "A method for detecting the quality of cotton seeds based on an improved ResNet50 model," *PLOS ONE*, vol. 18, no. 2, p. e0273057, Feb. 2023, doi: 10.1371/journal.pone.0273057.
- [43] W. Li *et al.*, "Machine learning model of ResNet50-ensemble voting for malignant–benign small pulmonary nodule classification on computed tomography images," *Cancers*, vol. 15, no. 22, p. 5417, Nov. 2023, doi: 10.3390/cancers15225417.
- [44] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019*, 2019, pp. 6105–6114.
- [45] S. D. Deb, A. Rahman, and R. K. Jha, "Breast cancer diagnosis using modified Xception and stacked generalization ensemble classifier," *Research on Biomedical Engineering*, vol. 39, no. 4, pp. 937–947, Oct. 2023, doi: 10.1007/s42600-023-00317-4.
- [46] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.




BIOGRAPHIES OF AUTHORS

Hanin Saadah    is currently a master's student at the Arab American University in Palestine in the Data Science and Business Analytics program. She completed her bachelor's degree in information and communications technology in 2016 from Al-Quds Open University. From 2018–2023, she worked with the World Health Organisation (WHO) in Palestine as a health system developer and implementer. She spent 5 years working closely with the care providers in the health field to understand their workflows, guidelines, and requirements and reflect them in a digital system to fulfill their needs. In 2023, she joined the University of Oslo (UiO) as an Implementation Specialist in the Middle East and North Africa. During her work with the WHO and UiO, she showed a strong interest in analyzing and exploring health data using machine learning, specifically medical images. She can be contacted at emails: h.saadah@student.aaup.edu or haninsaadah1187@gmail.com.



Dr. Amani Yousef Owda    assistant professor in Computer Engineering and Data Science in the Faculty of Graduate Studies at the Arab American University in Palestine. She worked as a head of department of Natural, Engineering, and Technology Sciences in the Faculty of Graduate Studies at Arab American University from 2022-2023. She worked as a research associate in the Faculty of Engineering at the University of Manchester from 2019-2020. In addition, she worked in the School of Engineering at Manchester Metropolitan University from 2015 – 2019. She worked at Birzeit University from 2007- 2011. She received her MSc. degree (Hons.) from The University of Manchester, UK in 2013, and her Ph.D. degree in Computer Engineering from Manchester Metropolitan University, UK in 2018. Since 2018, she leads research in multi-disciplinary fields with a focus on artificial intelligence, machine learning, decision support systems, image processing, medical applications of microwave and millimeter-wave imaging, security screening, and anomaly detection. She has published more than 43 articles in well reputable journals. She is a reviewer in many well-known Journals, and she is supervising M.Sc. and Ph.D. students. She can be contacted at emails: Amani.Owda@aaup.edu or amaniabubaha@gmail.com.



Dr. Majdi Owda    associate professor in Computer Science and Dean of Faculty in Data Science at the Arab American University in Palestine. In addition, he is a UNESCO Chair for Data Science for Sustainable Development. Worked as a head of the Department of Natural, Engineering, and Technology Sciences in the Faculty of Graduate Studies at Arab American University from 2020-2022. Worked in the School of Computing, Mathematics, and Digital Technology at Manchester Metropolitan University from 2009 to 2020. He gained a B.Sc. in Computer Science from the Arab American University in 2004, and an M.Sc. by research in Computer Science with distinction from Manchester Metropolitan University in 2005 and a Ph.D. in Computer Science in 2011. His main research interests are AI techniques for natural language interfaces to relational databases, data science, conversational informatics, conversational agents, knowledge trees, knowledge engineering, planning, information extraction, AI techniques for the help of society, web/data/text mining, digital forensics processes and frameworks, digital forensics artefacts, information retrieval from large data sources, internet of things frameworks, and internet of things digital forensics artefacts and security. He can be contacted at email: Majdi.Owda@aaup.edu.