ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20211221131838
Original scientific paper

A Decision-Making Tool for Early Detection of Breast Cancer on Mammographic Images

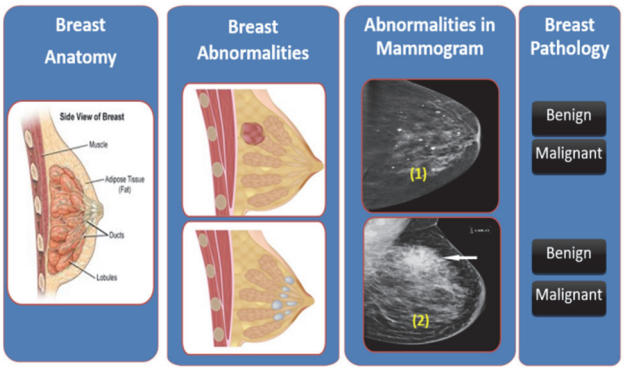
Duygu ÇELİK ERTUĞRUL*, Soona AHMED ABDULLAH

Abstract: Breast cancer is one of the most dangerous types of cancer in the world among females. In the medical industry, the early detection of a breast abnormality in a mammogram can significantly decrease the death rate caused by breast cancer. Therefore, researchers directed their focus and efforts to find better solutions. Whereas researchers earlier used semi-automatic algorithms of machine learning, recently the attention is redirected toward deep learning algorithms that automatically extract features. Therefore, in the research study, two pre-trained Convolutional Neural Network models, VGG16 and ResNet50, have been used and applied on mammogram images to classify their abnormalities in terms of (1) the Benign Calcification, (2) the Malignant Calcification, (3) the Benign Mass, and (4) the Malignant Mass. The mammographic images of the CBIS-DDSM dataset are used. In the training phase, various experiments are performed on ROI images to decide on the best model configuration and fine-tuning depth. The experimental results showed that the VGG16 model provided a remarkable advancement over the ResNet50 model; the accuracy obtained was 80.0% in the first model whereas the second model could classify with a 60.0% accuracy almost randomly. Apart from accuracy, the other performance metrics used in this study are precision, recall, F1-Score and AUC. Our evaluation, based on these performance metrics, shows that accurate detection effect is obtained from the two networks with VGG16 being the most accurate. Finally, a decision support tool is developed which classifies the full mammogram images based on the fine-tuned VGG16 architecture into Benign Calcification, Malignant Calcification, Benign Mass, and Malignant Mass.

Keywords: breast cancer; decision support systems; image classification; mammogram images; Resnet50; VGG16

1 INTRODUCTION

Early detection of breast abnormalities is a crucial process in extending the life span of patients in breast cancer [1]. Thus, radiologists can diagnose this disease with mammography medical imaging method, which is the most preferred procedure for detecting abnormalities in human breast tissues, due to the use of low-dose X-ray and its low cost [2]. Mammograms usually visualize the breast tissues in two viewpoints - Mediolateral Oblique (MLO) and Cranio Caudal (CC) - for a breast in search of the two main abnormalities calcification and masses. The (1) and (2) of Fig. 1 show two exemplary mammogram images in terms of MLO and CC.



(a) (b)

Figure 1 Breast abnormalities shapes in mammograms
(a) Breast anatomy and abnormalities; (b.1) A mammogram image is showing a calcification as small spots scattered around the breast tissues; (b.2) A mammogram image is containing a dense tissue mass.

Whereas the calcification can be observed as small white spotting with a small shape variation scattered in the breast tissues in the mammogram images, the masses can be shown as dense tissue with odd shapes compared to the normal tissue (Fig. 1). Even with modern screening technologies, the breast cancer mortality rate was 15.0% of the total cancer-related deaths in 2018. Because the classification of such abnormalities is still done by specialists in the hospital or clinical environments [1], therefore, the drive to develop computer-assisted clinical diagnostic tools to interpret medical images has recently motivated researchers in this field. Earlier researchers used image processing techniques for distinctive feature extraction in a semi-automatic fashion with machine learning methods [3]. The infeasibility in time and computation in those methods guided research works toward deep learning methodologies to automatically extract the discriminative features from raw images without the need for feature engineering. In this study, it is aimed to reduce the mortality rate in breast cancer by gaining the trust of the radiologist in imaging. In addition, it is aimed to reduce the 30.0% misdiagnosis rate resulting from traditional imaging and analysis method. This article's contributions are organized as follows: Section 2 gives background; Section 3 discusses similar research studies; Section 4 mentions the dataset used; Section 5 presents the major steps applied; Section 6 discusses the experimental studies and performance metrics used. While Section 7 provides experimental results, the Section 8 presents the tool developed followed by a conclusion of the research.

2 BACKGROUND

In literature, many researchers have applied machine learning methods, which involve preprocessing images, feature extraction, selection of features to reduce the features size, and finally a classification algorithm to achieve the expected result. Convolutional Neural Network (CNN) [4] and Transfer learning [5], which are well known and the most effective deep learning approaches currently in the medical area, are proven by various research to be superior to traditional methods.

CNN is categorized into three main layers: *convolutional layer, pooling layer* and *fully connected layer* that are organized in a way to build neural networks [4]. Fig. 2 shows the CNN layers and its functionalities on a mammography image. To produce high accuracy models using CNN requires to use massive datasets as well as high hardware specifications.

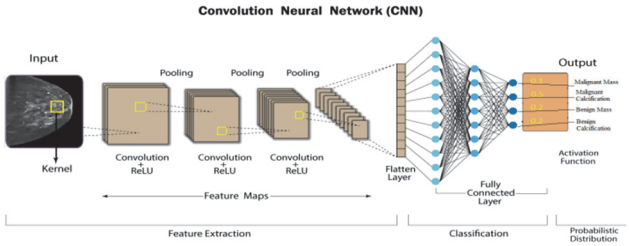


Figure 2 CNN concept

Instead, *Transfer Learning (TL)* concept was introduced [5] which was based on the principle that previously learned knowledge can be exceptionally implemented to solve new problems in a more efficient and effective manner. Therefore, it has been possible to use existing pre-trained models in other existing image datasets or in your own custom dataset. This often requires modifying and fine-tuning the pre-trained model to work with the new dataset. In TL, by using previous knowledge, higher success can be achieved with less training data and faster learning models can be obtained. Numerous medical imaging systems using TL techniques have been developed to support physicians in effective mammograms diagnosis, care, and follow-up examination [6-8]. In literature, TL is commonly used to distinguish malignant and benign breast cancer by fine-tuning multiple these pre-trained models. The most common pre-training models used for TL in mammogram ultrasound images are the VGG19, VGG16, AlexNet, and Inception-V3 models; VGG is the most extensively used, followed by AlexNet and Inception, which are the least common [9-11]. In our research study, two pre-trained models, VGG16 [12] and ResNet50 [13], are used on a set of mammogram images to classify their abnormalities in terms of: (1) the Benign Calcification, (2) the Malignant Calcification, (3) the Benign Mass, and (4) the Malignant Mass.

VGG16 is designed by Simonyan & Zisserman in 2014. It consists of 13 convolutional layers with 2 dense layers and an output layer. The VGG16 network has also 5 max-pooling layers. The model takes square RGB image of size 224. On the other hand, Residual Networks model, or ResNet50, is also a pre-trained CNN model developed by He et al. and is with 50 layers deep [13]. It is a complex CNN model which is trained on a huge dataset and consists of many layers; convolutional, pooling, residual (i.e., a stack of three convolution layer followed by a batch normalizer), dense, and an output. The model also takes square RGB image of size 224.

3 RELATED WORKS

The main components of a model on TL involves: (1) the input, (2) the feature extractor, (3) the classifier, and (4) the output. Therefore, the methodologies and challenges encountered in previous research studies on breast cancer are addressed and compared within four groups: (1) Type of Inputs Used, (2) Pre-processing Methods, (3) Classifiers, and (4) Output Labels.

(1) Focused Input Types: It has been observed that the input data fed to the classifier models applied by the researchers are in three forms; (i) a full mammogram image, (ii) extracted region of interest (ROI) from the mammogram images, and (iii) segmenting these mammograms into patches. Some research studies used the full image of mammograms as in [6, 14, 15], others used ROI images [16, 17] while some others used segmented patches of the full mammogram as in the research studies [18-20].

(2) Pre-processing Methods: In fact, this step is not mandatory in deep learning approaches, therefore, the pre-processing methods for the input images were not implemented by the researchers in [15-17, 20-23]. Whereas, other researchers [24-26] enhanced the mammogram images by the use of one or a combination of the methods such as normalizing, removing noises, or adjusting the contrast.

(3) Classification Methodologies: In 1995, models with shallow CNN architecture had been investigated in the research study [27]. Some recent research has also been applied on the same technique [28, 29]. The researchers in [6, 8, 14, 24, 29, 30] present more complicated CNN architectures. It is worth noticing that the researchers in the studies [31, 32] stated that deep CNN showed better than the former shallow layers; however [31] concluded that neither a shallow architecture nor an exaggerated deep architecture are useful in mammograms classification; perhaps a model with moderated depth can distinguish features better. Other researchers [7, 24, 33, 34] tried to improve the accuracy of classification using the TL approach by using a pre-trained model with general images of the "ImageNet" dataset [35]. The authors in [17, 24, 25, 28, 36, 33] fine-tuned their TL models. The most common CNN architectures used in TL studies for mammography classification are AlexNet, VGG16, and ResNet. AlexNet model is studied in [16, 18, 19, 21, 24, 32-34, 37] and the researchers in [7, 8, 15, 17, 19, 20, 22, 36] studied VGG16. In addition, the ResNet50 model is investigated in these research studies [13, 20, 21, 24, 25, 32, 33, 38].

(4) Output Labels: The main aim of the models used in the previous studies was performing either localization or classification of breast lesions. The studies [8, 16-18, 28, 30] aimed to locate and predicate the abnormalities while the researchers in [32, 39] were interested in lesion classifications applying CNN with probabilities for mass, calcification, and normal classes. Some studies [7, 24, 25, 33, 34, 36, 37] were interested in classifications of three or two classes of benign, malignant, and normal masses.

4 DATA SET USED

Curated Breast Imaging Subset of DDSM (CBIS-DDSM) is a curated subset of the Digital Database of Screening Mammography (DDSM) database which involves 2620 scanned mammograms which represent both MLO and CC views. According to DICOM metadata, the data appears to consist of 6671 patients.

However, it really consists of 1566 participants due to the original data set being structured so that each participant has more than one patient ID [40]. For instance, participant 00038 has 10 separate patients ID. The CBIS-DDSM dataset is divided into a train set and a test set with a ratio of 80:20 percentage of the whole dataset. Tab. 1 shows the number of samples in each partition and category.

The database is partitioned to 2864 ROI images for training set in which a validation set is generated by

holding out 20.0% of the former and 645 mammographic images as the test set. The original corps are arranged in separate folders for training and testing with two subfolders for each abnormality. While the files have been renamed with a pattern constructed from the first letter of abnormality type followed by the first letter of the pathology *type_sampleNumber_patientID* (e.g., cB_3_7 represents a calcification abnormality with benign pathology third sample and patient ID_7).

Table 1 CBIS-DDSM dataset taxonomy

DATASET (6671)							
Train set				Test set			
Full Images 2458		ROI Images 2864		Full Images 645		ROI Images 704	
Calcifications 1227	Masses 1231	Calcifications 1546	Masses 1318	Calcifications 284	Masses 361	Calcifications 326	Masses 378

In this study, it is used hold-out validation for the validation partition to keep the training dataset as big as possible. Our study also applies balanced filtration on the training dataset to assign equal weights on the importance of each sub-class. Train, validation, and test datasets are pre-processed using various methods, which are discussed in the next section.

5 METHODOLOGIES

Several stages were required to develop an anomaly detection tool to predict abnormalities on a breast image. This study performs various methods which are summarized in Fig. 3 and discussed in the next sections in details.

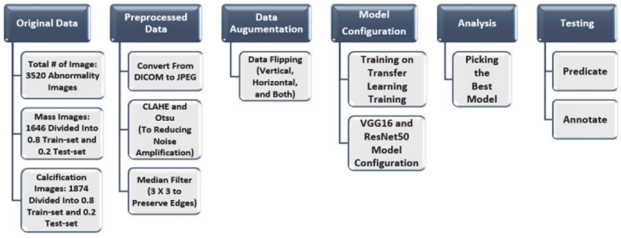


Figure 3 Classification workflow

5.1 Pre-Processing

Otsu segmentation [41] method is used to distinguish artifact objects from the breast region in the background area. Image quality has been improved using two filtering methods which are CLAHE and Median Filter [42]. The methods were applied to both the full mammogram images and the ROI images to reduce noise and preserve the edges for further extend. In this study, the dataset [40] has been resized to 224 × 224 square images and converted to RGB to match the input shape of the pre-trained models.

5.2 Data Augmentation

Data augmentation is one of the popular ways to increase dataset size to double or more of the original size along with assisting in preventing over-fitting problem [43] when training deep learning models with comparatively small dataset size. To avoid this, data augmentation approach has been used in many research studies [14, 18, 24, 25, 28, 34, 36]. Similarly, data augmentation is also applied in our study. Therefore, the transformations were made by flipping the images horizontally, vertically, and in both, also increased the size of the first data set used five times. Thus, it presents the lesions in different orientations.

5.3 Model Configuration

The overall architecture of CNN will not produce an efficient model without understanding some of the hyper-parameters used to adjust the training and learning tasks according to the problem in focus. For instance, a number of these parameters should be adjusted in the layers of the classifier model to overcome the main problems that face any model in the training phase such as over-fitting problem or weight vanishing between layers. One of the main hyper-parameters is the optimization algorithm that should be defined in the learning process which simply updates iteratively the attributes of the neural network such as weights and learning rate to find the best value to reduce the losses and provide the best accurate results. Another important hyper-parameter is the loss function that calculates the difference between the target output and the output produced from the model is specified and aimed to be minimized as much as possible. All these hyper-parameters along with the activation functions are essential to configure a suitable model accordingly. Details on the initialization of the hyper parameters in configuring the VGG16 and ResNet50 models in our study are discussed in detail in the next section.

5.4 Analysis

Research studies stated that the error rate in diagnosing abnormalities is estimated to be around 30.0% [44]. In addition, one survey study [45] reported that most of the legal cases against radiologists were due to their failure in interpreting the medical images. Therefore, the trials of model configuration try to reduce the error rate done by radiology experts and better predicate abnormalities. Whereas the previous step trains the pertained VGG16 and ResNet models with various configuration, the analysis step compares the returned result of each model and picks the most suitable classifier to be used later in the developed tool.

5.5 Testing

The final step represents the testing phase after developing our clinical diagnostic tool. When the diagnostic tool user picks an image sample from the test dataset, the selected classifier thus annotates the class label of the anomaly type. The class labels are presented in our tool as four labels: (1) the Benign Calcification, (2) the Malignant Calcification, (3) the Benign Mass, and (4) the Malignant Mass. Experimental studies conducted with the use of the tool are presented in detail in the next section.

6 EXPERIMENTAL STUDIES

A clinical diagnostic tool is developed using Deeplearning4j library and 8GB NVidia GTX 1660i hardware specification in this study. Although there is limited documentation available on the Deeplearning4j, it has been chosen because of its ability to support multiple GPUs and CPUs, the availability of pre-trained models, parallel computing, and most importantly, fast execution. In addition, the TL concept has been implemented by preserving the original structures of the VGG16 and ResNet50 models. The hyper-parameters of the models have been updated through various trials along the experiments done until the best configuration for the classification is reached. In this study, the experiments fine-tuned the original architectural composition of VGG16 and ResNet50 in different depths, either by fine-tuning only the last fully connected layer or the two last connected layers. The training experiments of this study are done with a batch size 64 at most for a maximum of 100 epochs using a learning rate that ranges from ($1e^{-4}$) to ($1e^{-7}$). When the CNN model is started using pre-trained weights, the feature learned during initial training is more precise, therefore a smaller learning rate is used with the default momentum value of 0.9. Moreover, the hyper-parameters used such as optimization function, loss function, weight initialization and more are used to fine tune the classifiers.

(1) Optimization Function: The output layer for both architectures, the VGG16 and ResNet50, are set to use Adam optimizer [46] as the optimization function instead of the Stochastic Gradient Descent (SGD) optimizer that was used in the original model as default optimization function. Adam optimizer is trying to find the optimal point of the cost function evaluated on a random mini batch of data. The step size in the Adam optimizer is invariant to the computed gradient and can be updated better as it navigates quickly through tiny gradients (saddles and ravines), unlike the SGD function. Adam optimizer is used to compute individual learning rates for different parameters to maximize the model's efficiency following Eq. (1):

$$w_t = w_{t-1} - n\frac{m_t}{\sqrt{v_t} + \in} \tag{1}$$

where: w: current model weight, n: step size, depending on the iteration, v_t, m_t: calculated estimators, $\in$: Epsilon. In each iteration, the weights of a feature are updated with the help of step size and computing the gradient to calculate the biased corrected mean estimator and variance estimator for a certain moment. The moment coefficient which represents the gradient percentage possessed in every iteration is equal to (0.9) due to the smaller values turn out to fluctuate a lot.

(2) Loss Function: Cross-Entropy [47] is applied as the validation loss function to measure the classification performance by minimizing the distance between two distributions of probabilities. The output of this function is a probability between 0 and 1 whereas the actual prediction is either 0 or 1. Thus, the output probability is tried to be as close as possible to the actual predictions. It is calculated for binary classification using Eq. (2):

$$L = -y\left(\log\left(p\right) + \left(1-y\right)\log\left(1-p\right)\right) \tag{2}$$

where, L: the validation loss, y: the actual labels, p: the predicted labels.

(3) Activation Functions: The typical activation functions applied in this study are: Sigmoid, SoftMax and Rectified Linear Unit (ReLU). ReLU function used in all convolutional layers, is a piecewise linear function that prunes the negative parts to zero and retains the positive parts directly, as applied in [40]. ReLU is calculated by using the formula, Eq. (3):

$$A_{i,j,k} = \max\left(Z_{i,j,k}, 0\right) \tag{3}$$

where, A: the activation function at the (i, j, k) position, Z: the input image at the position (i, j, k).

The activation functions, Sigmoid and SoftMax, are used in the output layer depending on the number of classified labels. Thus, in the models that classify two labels Sigmoid is applied, such as the model that classifies the abnormalities between mass and calcification. Sigmoid function is used to predict the probability on the output. It is calculated by Eq. (4):

$$f\left(x_i\right) = \frac{1}{1+e^{-x}} \tag{4}$$

In the other hand, the models that classify the ROIs to more than two classes use SoftMax as the output layer's activation function, such as the model that classifies the mammograms according to the abnormality as well as the pathology (Benign Calcification, Malignant Calcification, Benign Mass, and Malignant Mass). SoftMax activation function produces an output ranging from 0 to 1 by computing the probability distribution from a vector. It is calculated through Eq. (5):

$$f\left(x_i\right) = \frac{e\left(x_i\right)}{\sum_j e\left(x_i\right)} \tag{5}$$

(4) Weight Initialization: Weight initialization is set to Xavier [48], which is similar random initialization but turns out to perform much better. Xavier prevents layer activation output from vanishing through training by setting layers' weights to a random value chosen in between the uniform distribution boundary. The relationship between the input and output variance may dramatically vary according to the initialized weight, where the variance with values greater than 1 leads to exploding, the variance with values less than 1 vanishes the forward signal. Therefore, the chosen value of weights can be calculated by Eq. (6):

$$Uniform\ Distribution\ Boundary = \overset{+}{_-}\frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}} \tag{6}$$

where, n_i: the number of input connections to network, n_{i+1}: the number of output connections of the network.

(5) Weight Decay: Overwhelming the over-fitting problem in a model requires regularization to penalize the large weight features. One of regularization methods, L2, forces these weights to decay to zero [43]. Over-fitting has been suppressed via L2 regularization in our study. Parameter is used as (0.1) to penalize the large weighted features and prefer small ones in the network.

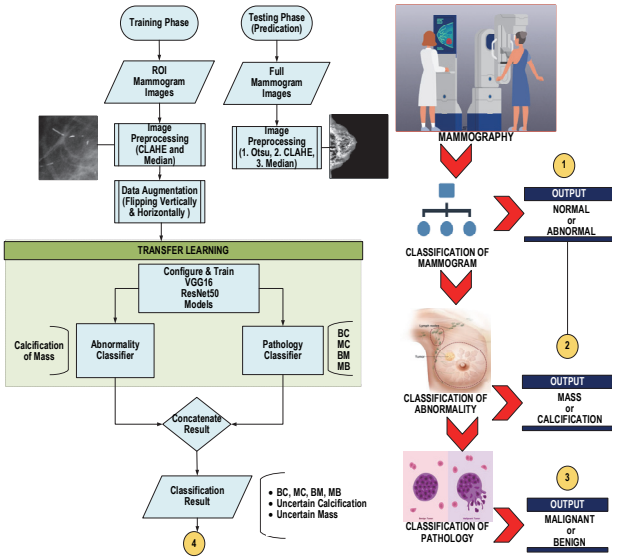


Figure 4 The structure of the proposed system for mammogram classification in four stages. (1) In this step, the main purpose of the proposed tool is to evaluate an input mammography as Normal or Abnormal. (2) Any abnormality detected in any case during (1), it is classified as either Mass or Calcification. (3) For further examination, the case is evaluated pathologically as Malignant or Benign. (4) All probability results are concatenated, then a prediction result is produced.

Fig. 4 illustrates the methodology applied during the training and prediction steps. The abnormality classifier of VGG16/ResNet50 classifies the mammograms into two classes (1-Calcification and 2-Mass) according to the abnormality presented in the mammogram, whereas the pathology classifier of VGG16/ResNet50 classifies mammograms into one of four classes (1-Benign Calcification, 2-Malignant Calcification, 3-Benign Mass, and 4-Malignant Mass) showing pathologies' probabilities of the mammogram. Although the system shows detailed probability results in both classifiers, if both classifiers agree on the same abnormality, it also compares the results of both classifiers to predicate its class label. As it is seen in Fig. 4, the training step uses the ROI images, unlike the prediction step which uses full mammogram images. Then, the input images of both phases are being pre-processed separately. Later, the images pre-processed in the training convey into batches of size 64 which are fed into both classifiers (abnormality and pathology) to extract the distinctive features after training the unfrozen layers of the pre-trained model. Finally, step (4) shows how both phases predicate the results for both classifiers giving the possible probability for each class label, then compares the outcomes. The comparison step, trained, occurs away if both classifiers did not state the same abnormality class; then the system indicates uncertain result that may require further medical investigations.

Evaluation Metrics Used: Typically, to quantify the performance of a model, a set of individual evaluation metrics is required such as (1) Accuracy, (2) $F1$ score, and (3) AUROC (Area Under Receiver Operating Curves).

(1) Accuracy: Accuracy is a suitable evaluation metric and is applied for both a binary and a multi-class classification problem. It simply calculates the portion of true values among the examined cases.

(2) *F*1 score: $F1$ score is a harmonic mean of precision and recall which is calculate through Eq. (7).

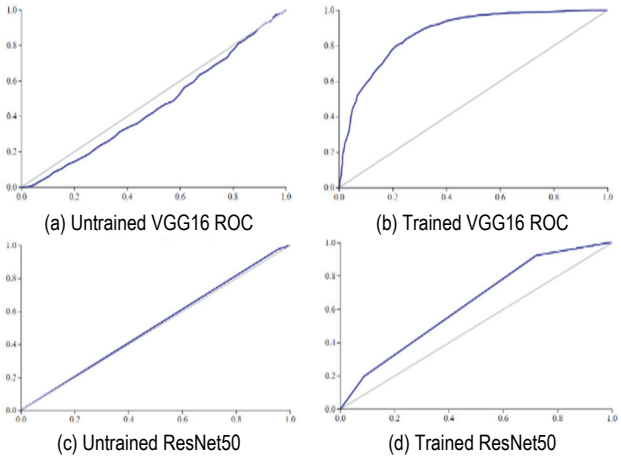
$$F1 = 2 \times \frac{recall \times Precision}{recall + Precision} \tag{7}$$

(3) AUROC: It measures how well a prediction is ranked rather than its actual value (i.e., scale-invariant) by calculating the area under the ROC curve. Next section discusses experimental results.

7 EXPERIMENTAL RESULTS

In the first stage of trials, the proper type of input on the pre-trained CNN models is decided. As second, suitable hyper-parameters for the models are investigated. These two steps confirmed that the larger batch size is the fastest in the computations. It also verified that batch size and learning rate impact on generalization gap between train and validation data. The combination of large batch size with small learning rates or vice versa is found to be a balanced trade-off between them according to the computation speed and the hardware specifications. Furthermore, when fixing the learning rate at a certain value and doubling the batch size the performance improved 1.00% only. In the aspect of epoch numbers, increasing it 5 times raised the performance of the CNN model lightly by 2.00% only and extended the training time. The performance of the training using pre-trained weights obtained from ImageNet [35] versus the performance of initializing the network randomly and training the samples from scratch achieved better in terms of accuracy and speed. The later classified the samples randomly with accuracy equal to 0.5 only and that is due to the small dataset size in comparison to ImageNet which contains hundreds of thousands of images for each class. Although data augmentation has been applied, yet the augmented data is still correlated to the original data. The nature of CBIS_DDSM [40] is very different in content compared to the ImageNet dataset. This allowed additional experiments of fine-tuning different deepness levels of the CNN model searching for performance improvement, either by training only the last fully connected layer or several layers. Fine-tuning the last output layer by changing the activation function and the number of outputs in both models acquired a superlative score over fine-tuning the last three layers in VGG16 "0" adding "3" more fully connected layers in the ResNet50.

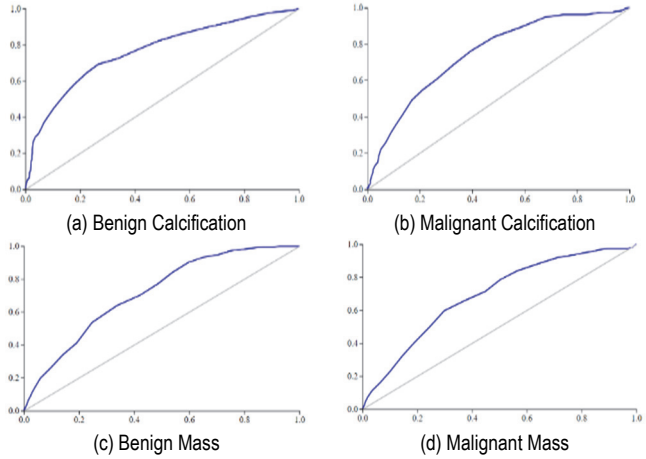
The differences in the ROC curve of the VGG16 model are presented in Fig. 5. The classification of the abnormality between mass and calcification was easier, while malignant and benign classification was more challenging, especially in mass anomaly. Fig. 6 shows the ROC curve for the four different classes. The AUC score for each was (Benign Calcification = 0.76%, Malignant Calcification = 0.74%, Benign Mass = 0.71% and lastly, Malignant Mass = 0.70%).



(a) Untrained VGG16 ROC (b) Trained VGG16 ROC

(c) Untrained ResNet50 (d) Trained ResNet50

Figure 5 ROC curve of VGG16 and ResNet50 [ROC: TPR/Recall (*y*) vs. FPR (*x*)]



(a) Benign Calcification (b) Malignant Calcification

(c) Benign Mass (d) Malignant Mass

Figure 6 VGG16 classifier ROC for all the labels [ROC: TPR/Recall (*y*) vs. FPR (*x*)]

In the training with the above-mentioned parameters, a higher accuracy was obtained in the VGG16 model compared to the ResNet50 model. Although the number of trainable parameters was much less than the ResNet50 model, the accuracy of the VGG16 is scored at 80.0%. The accuracy of ResNet50 is scored as 60.0%. The comparison results between VGG16 and ResNet50 models used with other similar studies are given in Tab. 2. Experimental studies revealed that while the trainable parameters of the VGG16 model were only 8194 in the proposed system, the original VGG16 model had 134268738 trainable parameters. Whereas the ResNet50 model has 25636712 total trainable parameters, only 4475906 parameters of them were trainable.

In summary, the findings from the experimental studies are given below:

- **Input**: Distinctive features can be observed more in ROI images. The reason of this, in the first trials, using a pre-trained model with a full image (with background cropped) first, only 34% accuracy was achieved without applying data augmentation. This unsatisfactory accuracy was due to the use of full image, where the exact region of abnormality could not be observed. Therefore, under the same conditions, the input data was treated as ROI images and its accuracy was slightly increased to 48%. This revealed that the amount of data was considered insufficient in trials and data augmentation was necessary.
- **Dataset Size**: For data augmentation, some transformations were applied by flipping the input images both horizontally and vertically, and the size of the initial dataset was increased 5 times and the lesions were presented to the system in different directions. After data augmentation by increasing the training dataset 5 times, the accuracy increased to 54%. After this, the experiments done were on searching for suitable model parameters.
- **Regularization**: Improving model performance and avoiding over fitting problem.
- **Batch Size**: The larger the batch size, the faster the calculation is based on hardware specifications.
- **Xavier**: To maintain the variance of activations and back-propagated gradients for all the layers.
- **Adam Optimizer**: The parameters of the output layer for both, architecture VGG16 and ResNet50, set to use Adam optimizer instead SGD that is used in the original model. Calculating learning rates for the different parameters using Adam maximizes the model efficiency.

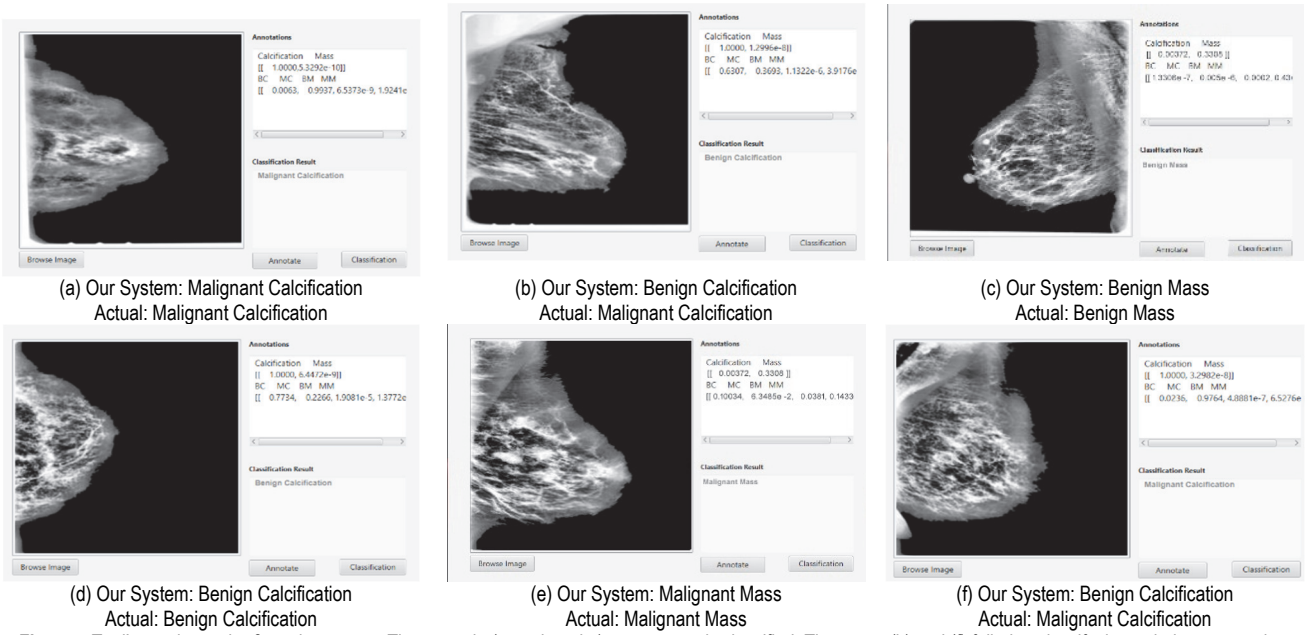
Table 2 Comparison of performance results of the system proposed with others

Researcher	Classification	Dataset	Image Type	Model	Acc.	AUC
Xavier [48]	Mass vs. Normal	CBIS-DDSM + INbreast	Patches	VGG16	NA	NA
				ResNet50	0.84	0.92
Perez [40]	MM vs. BM	CBIS-DDSM	ROI	VGG16	0.64	0.64
				ResNet50	0.84	0.84
Lazaros [17]	MM vs. BM, and Normal	CBIS-DDSM	ROI	VGG16	0.71	0.78
				ResNet50	0.74	0.80
LiShen [15]	MM, BM, MC, BC, and B*	CBIS-DDSM + INbreast	Patches	VGG16	NA	0.84
				ResNet50	NA	0.63
Our system	Mass vs. Calcification	CBIS-DDSM	Training ROI to Classify Full Img.	VGG16	0.80	0.87
				ResNet50	0.60	0.66
Our system	MM, BM, MC, BC	CBIS-DDSM	Training ROI to Classify Full Img.	VGG16	0.72	0.73
				ResNet50	0.60	0.64
Abbreviation: BC: Benign Calcification, MC: Malignant Calcification, BM: Benign Mass, MM: Malignant Mass, *B: Background.						

8 THE TOOL DEVELOPED & CASE STUDIES

A visual analysis tool is developed that provides clinical decision support to the healthcare professionals during the detection of breast cancer. The screenshots of the tool's returned results on six random mammograms are shown in Fig. 7. User can simply choose a mammogram of a patient using the "Browse Image". As soon as the pre-processing step for the selected mammogram is completed and successfully loaded, the classification button will be activated automatically. Thus, the button loads two VGG16 models to predicate the breast abnormality and

pathology. The final prediction of the classification is made by using the two most accurate VGG16 models achieved; an accurate VGG16 classifies the mammogram to mass or calcification. The other accurate VGG16 classifies that mammogram to the (1) Benign Calcification, (2) Malignant Calcification, (3) Benign Mass, and (4) Malignant Mass.



(a) Our System: Malignant Calcification
Actual: Malignant Calcification

(b) Our System: Benign Calcification
Actual: Malignant Calcification

(c) Our System: Benign Mass
Actual: Benign Mass

(d) Our System: Benign Calcification
Actual: Benign Calcification

(e) Our System: Malignant Mass
Actual: Malignant Mass

(f) Our System: Benign Calcification
Actual: Malignant Calcification

Figure 7 Tool's results on the 6 random cases. The cases in (a, c, d, and e) were correctly classified. The cases (b) and (f) failed to classify the pathology correctly.

9 DISCUSSION AND CONCLUSION

In this study, a classification tool has been proposed to assist the radiologist in diagnosing breast cancer with the least effort and time. The following are the key contributions of this research: (1) Mammograms show radiological indications that are readily detectable symptoms. As a result, deep learning-based methods can be used to automatically analyze mammograms, which significantly reduces analysis time. (2) To do this, a decision support tool is developed which classifies the full mammogram images based on the fine-tuned VGG16 architecture into Benign Calcification, Malignant Calcification, Benign Mass, and Malignant Mass. (3) Various evaluation metrics are used such as: Accuracy, Precision, Recall, *F*1 score, and AUROC (Area Under Receiver Operating Curve). As a result of the experimental studies, higher accuracy than human accuracy has been obtained. The pre-trained VGG16 model achieved 80.0% grading accuracy, and the ResNet-50 model achieved 60.0% accuracy. The robustness of the methodology used was based on the accuracy after comparing the results obtained with the results obtained by four other researchers. The results obtained from the training and test experiments indicated that slightly higher accuracy was achieved in classification. (4) Extensive comparative comparisons were performed to assess the effectiveness of the proposed system. (5) The main achievement of this study can be summarized in rejuvenating the use of the inhibited CNN models by data limitation in breast cancer imaging fields with the use of efficient transfer learning models. (6) To improve the generalization effectiveness of the proposed system and prevent overfitting, a different training protocol assisted by different combinations of training policies (e.g., data augmentation, hold-out validation) were used.

In future studies, we will concentrate on the application of the methods used in this paper to compress other highly regarded networks including but not limited to VGG16 and ResNet50. In addition, increasing the dataset size will be interesting to work with since each class requires at least 1000 samples to extract useful features more precisely. Moreover, the patch classifier in the system used the ROI of abnormal cases only. For the future, adding the normal patches will help not only to classify the abnormality type, perhaps it may provide comprehensive classifications. Another improvement on the decision tool is integrating segmentation function to detect the exact location of the abnormality.

10 REFERENCES

[1] American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. Retrieved from: www.cancer.org.

[2] Suri, J. S. & Rangayyan, R. M. (2006). Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. *SPIE*. https://doi.org/10.1117/3.651880

[3] Baldi, P., Brunak, S., & Stolovitzky, G. A. (2002). Bioinformatics: The machine learning approach. *Physics Today*, *55*(12), 57-58. https://doi.org/10.1063/1.1537915

[4] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology* (ICET), 1-6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

[5] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data, 3*(1), 1-40. https://doi.org/10.1186/s40537-016-0043-6

[6] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Sánchez, C. I. et al. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005

[7] Huynh, B. Q., Li, H., & Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, *3*(3), 034501. https://doi.org/10.1117/1.JMI.3.3.034501

[8] Chougrad, H., Zouaki, H., & Alheyane, O. (2020). Multi-label transfer learning for the early diagnosis of breast cancer. *Neurocomputing*, *392*,168-180. https://doi.org/10.1016/j.neucom.2019.01.112

[9] Huynh, B., Drukker, K., & Giger, M. J. M. P. (2016). MO-DE-207B-06: Computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks. *Medical physics*, *43*(6), 3705-3705. https://doi.org/10.1118/1.4957255

[10] Byra, M., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., & Andre, M. (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Medical physics*, *46*(2), 746-755. https://doi.org/10.1002/mp.13361

[11] Yap, M. H., Pons, G., Martí, J., Ganau, S., Sentis, M., Zwiggelaar, R., Marti, R. et al. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics, 22*(4), 1218-1226. https://doi.org/10.1109/JBHI.2017.2731873

[12] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE conference on computer vision and pattern recognition*, 770-778. https://doi.org/10.1109/CVPR.2016.90

[14] Lu, H. C., Loh, E. W., & Huang, S. C. (2019). The classification of mammogram using convolutional neural network with specific image preprocessing for breast cancer detection. *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 9-12. https://doi.org/10.1002/jmrs.385.

[15] Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, *9*(1), 1-12. https://doi.org/10.1038/s41598-019-48995-4

[16] Cai, H., Huang, Q., Rong, W., Song, Y., Li, J., Wang, J., Li, L. et al. (2019). Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. *Computational and mathematical methods in medicine*. https://doi.org/10.1155/2019/2717454

[17] Tsochatzidis, L., Costaridou, L., & Pratikakis, I. (2019). Deep learning for breast cancer diagnosis from mammograms - a comparative study. *Journal of Imaging*, *5*(3), 37. https://doi.org/10.3390/jimaging5030037

[18] Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2015, August). Convolutional neural networks for mammography mass lesion classification. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 797-800. https://doi.org/10.1109/EMBC.2015.7318482

[19] Dhungel, N., Carneiro, G., & Bradley, A. P. (2015). Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1-8. https://doi.org/10.1109/DICTA.2015.7371234

[20] Xi, P., Shu, C., & Goubran, R. A. (2018). Abnormality Detection in Mammography using Deep Convolutional Neural Networks. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 1-6. https://doi.org/10.1109/MeMeA.2018.8438639

[21] Yi, D., Sawyer, R. L., Cohn III, D., Dunnmon, J., Lam, C., Xiao, X., & Rubin, D. (2017). Optimizing and visualizing deep learning for benign/malignant classification in breast tumors.

[22] Carneiro, G., Nascimento, J., & Bradley, A. P. (2017). Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. *Deep learning for medical image analysis*, 321-339. https://doi.org/0.1016/B978-0-12-810408-8.00019-5

[23] Park, J., Phang, J., Shen, Y., Wu, N., Kim, S., Moy, L., Geras, K. J. et al. (2019). Screening Mammogram Classification with Prior Exams.

[24] Jiang, F., Liu, H., Yu, S., & Xie, Y. (2017). Breast mass lesion classification in mammograms by transfer learning. *Proceedings of the 5th International Conference on Bioinformatics and Computational Biology*, 59-62. https://doi.org/10.1145/3035012.3035022

[25] Wang, J., Ding, H., Bidgoli, F. A., Zhou, B., Iribarren, C., Molloi, S., & Baldi, P. (2017). Detecting cardiovascular disease from mammograms with deep learning. *IEEE transactions on medical imaging*, *36*(5), 1172-1181. https://doi.org/10.1109/TMI.2017.2655486

[26] Agarwal, R., Diaz, O., Lladó, X., & Martí, R. (2018). Mass detection in mammograms using pre-trained deep learning models. *14th International workshop on breast imaging (IWBI 2018)*, *10718*, 107181F. https://doi.org/10.1117/12.2317681

[27] Chan, H. P., Lo, S. C. B., Sahiner, B., Lam, K. L., & Helvie, M. A. (1995). Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network. *Medical physics*, *22*(10), 1555-1567. https://doi.org/10.1118/1.597428

[28] Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*, *7*, e6201. https://doi.org/10.7717/peerj.6201

[29] Ertosun, M. G. & Rubin, D. L. (2015). Probabilistic visual search for masses within mammography images using deep learning. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1310-1315. https://doi.org/10.1109/BIBM.2015.7359868

[30] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A. et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[31] Charan, S. G., Khan, M. J., & Khurshid, K. (2018). Breast cancer detection in mammograms using convolutional neural network. *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 1-5. https://doi.org/10.1109/ICOMET.2018.8346384

[32] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Todd Hurst, R., Kendall, C. B., Gotway, M. B., & Liang, J. (2017). On the necessity of fine-tuned convolutional neural networks for medical imaging. *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, 181-193. https://doi.org/10.1007/978-3-319-42999-1_11

[33] Suzuki, S., Zhang, X., Homma, N., Ichiji, K., Sugita, N., Kawasumi, Y., Yoshizawa, M. et al. (2016). Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 1382-1386. https://doi.org/10.1109/SICE.2016.7749265

[34] Gallego-Posada, J. D., Montoya-Zapata, D. A., Quintero-Montoya, O. L., & Montoya-Zapa, D. A. (2016). Detection and diagnosis of breast tumors using deep convolutional neural networks. *Conference Proceedings of XVII Latin American Conference in Automatic Control*, *17*.

[35] Samala, R. K., Chan, H. P., Hadjiiski, L. M., Helvie, M. A., Richter, C. D. et al. (2017). Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine & Biology*, *62*(23), 8894. https://doi.org/10.1088/1361-6560/aa93d4

[36] Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Karssemeijer, N. et al. (2017). Large scale deep learning forcomputer aided detection of mammographic lesions.*Medical image analysis*, *35*, 303-312. https://doi.org/10.1016/j.media.2016.07.007

[37] Jiao, Z., Gao, X., Wang, Y., & Li, J. (2018). A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognition*, *75*, 292-301. https://doi.org/10.1016/j.patcog.2017.07.008

[38] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int. J. of Computer Vision, 115*, 211-252. https://doi.org/10.1007/s11263-015-0816-y

[39] Shen, L. (2017). End-to-end training for whole image breast cancer diagnosis using an all-convolutional design. https://doi.org/10.1038/s41598-019-48995-4

[40] Falconi, L. G., Perez, M., Aguilar, W. G., & Conci, A. (2020). Transfer learning and fine tuning in breast mammogram abnormalities classification on CBIS-DDSM database. *Advances in Science, Technology and Engineering Systems*, *5*(2), 154-165. https://doi.org/10.25046/aj050220

[41] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, *9*(1), 62-66. https://doi.org/10.1109/TSMC.1979.4310076

[42] Sukassini, M. & Velmurugan, T. (2016). Noise removal using morphology and median filter methods in mammogram images. *The 3rd International Conference on Small and Medium Business*, 413-419.

[43] Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

[44] Schreer, I. (2009). Dense breast tissue as an important risk factor for breast cancer and implications for early detection. *Breast Care*, *4*(2),89-92. https://doi.org/10.1159/000211954

[45] Parvat, A., Chavan, J., Kadam, S., Dev, S., & Pathak, V. (2017). A survey of deep-learning frameworks. *2017 International Conference on Inventive Systems and Control (ICISC)*, 1-7. https://doi.org/10.1109/ICISC.2017.8068684

[46] Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization.

[47] Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning.

[48] Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *AISTATS*.

Contact information:

Duygu ÇELİK ERTUĞRUL
(Corresponding Author)
Department of Computer Engineering, Engineering Faculty,
Eastern Mediterranean University,
Famagusta, North Cyprus, via Mersin-10, Turkey
E-mail: duygu.celik@emu.edu.tr, duygucelik@msn.com

Soona AHMED ABDULLAH
Department of Computer Science and IT,
College of Science, Salahaddin University,
Erbil, Iraq