

# Hierarchical Pretrained Deep Learning Features for the Breast Cancer Classification

Abeer S. Alsheddi

Computer Science Department-Imam Muhammad bin Saud Islamic University, King Saud University, Riyadh, Saudi Arabia

**Abstract**—Breast cancer is a common and fatal disease among women worldwide. Accurately and early diagnosing of breast cancer plays a pivotal role in improving the prognosis of patients. Recently, advanced techniques of artificial intelligence and natural image classification have been used for the breast cancer image classification task and have become a hot topic for research in machine learning. This paper proposes a fully automatic computerized method for breast cancer classification using two well-established pretrained CNN models, namely VGG16 and ResNet50. Next, the feature extraction process is used to extract features in a hierarchical manner to train a support vector machine classifier. Evaluating the proposed model shows achieving 92% accuracy. In addition, this paper investigates the effect of different factors, highlights its findings, and provides future directions for the research to develop more advanced models.

**Keywords**—Feature extraction; CNN models; Pretrained models; breast cancer classification

## I. INTRODUCTION

Breast cancer is the most common form of cancer in women, and invasive ductal carcinoma (IDC) is the most common form of breast cancer [1]. It is frequently occurring and increasingly fatal. Usually, a biopsy is taken from patients and then a pathologist must decide whether they have breast cancer or not. Manual diagnosis from slides is time-consuming and the decision itself depends on the expertise of the pathologist and their equipment. Image Processing and Deep Learning can be used to create models that complement doctors by automating and speeding up diagnosis to save time and minimize errors in detecting breast cancer. This problem is not new these days, one of the earliest DNN applications was on breast cancer images [2].

1) *Research problem*: in this paper, the main goal is to classify breast cancer images in the form of images into binary classification: IDC and non-IDC. To formulate the problem, let  $x$  be a 2D image that belongs to  $R^{m \times n}$  where  $R$  is the space of 2D images with width  $m$  and height  $n$ , and let  $Y = \{0, 1\}$  where 0 indicates no breast cancer (non-IDC) and 1 indicates to indicating breast cancer was found in the image (IDC). Then the problem of the breast cancer classification is to model a mapping  $f$  from  $R^{m \times n}$  to  $Y$ , such that,

$$f(\cdot, \theta): R^{m \times n} \rightarrow Y \quad (1)$$

where any value of  $x$  will be mapped to  $y$ ,  $f(x, \theta) = y$ , for any  $x \in R^{m \times n}$  and  $y \in Y$ .

2) *Research objective*: the main contribution of the proposed model in this paper is combining between two tasks; First: using multiple pretrained models to extract features; second: using a feature hierarchy concept during extracting the features. To achieve this contribution, two pretrained VGG16 [3] and ResNet50 [4] models that have excellent classification performance for natural image classification in the Image Large Scale Visual Recognition Challenges, are used to extract activations of different five convolution layers from each model, ten layers in total. Next, the features are reduced using pooling operations to be  $6 \times 6 \times 5$  at each pretrained model and then concatenated them to result a  $6 \times 6 \times 10$  layer. Finally, the resulting features feed to the support-vector machines (SVM) classifier [5]. The results show that the combined feature hierarchy from two pretrained models gets 92% accuracy higher than using a pretrained model individually.

The rest of the paper is recognized as the following. Section II reviews some of the state-of-art in the problem. Next, Section III provides the proposed method and Section IV presents its results and discusses them. Finally, Section V concludes the main results and provides some future directions.

## II. RELATED WORK

This section reviews some of the research that works on the breast cancer classification problem and gives information about the used methods. Table I summarizes the methods of these related works. Most research recently uses different techniques to deal with features and then applies the classification on these features instead of classifying the whole image using ANN.

TABLE I. METHODS SUMMARIZATION IN THE RELATED WORK

Types	Methods	Research
Feature extraction from pretrained models	Fuses activations from FC of three pretrained models	[6]
	Fuses the results of the classifiers and extracted activations from each FC in pretrained model	[7]
Transfer parameters	Fine-tuned pre-trained models with using logistic regression classifier	[8]
	Fine-tuned modified AlexNet model	[9]
Unsupervised learning with segmentation step	Selecting features using GA algorithm and CNN model for classification.	[10]
	Clustering using Lloyd's algorithm [11] for and CNN model for classification.	[12]
Feature selection	Feature selection using mRMR algorithm with 4 classifier: SVM, Naïve Bays, Function tree and End Meta	[13]

The researchers usually try to find a proper feature representation of images to train their model. One major way is using a pretrained DCNN to extract an image activation as its features. Several pre-trained models were used for this process. The proposed model in [6] extracts features from fully connected layers of three models, namely GoogleNet [14], VGG [3], ResNet [4] models. Then classifier was trained on these features. The accuracies of these three models individually were 93.5%, 94.15%, and 94.35% respectively, while the combination between them achieves 97.525% accuracy. Thus, fusing the features from different models leads to better classification compared to extracting from a single model. The other research [7] uses three pre-trained models, namely AlexNet [15], VGG16 [3] and ResNet-18 [4] to extract features and then uses them to train three SVM classifiers, one classifier for each pretrained model. Instead of combining the features, the research fuses the three results from each classifier by calculating the average and combining the probabilities for fusion to obtain the final decision score. It measures the performance using the receiver operating characteristics curve (AUC) and it achieves 83.83% and 97.55% for two different datasets. However, for both research, [6] and [7], FC layers contain usually a high number of activations than convolution layers which will consume more computation cost.

In transfer parameters, the learning assumes that the two models share some parameters that can learn effectively. The research [8] analyzes different pretrained models VGG [3] and ResNet [4] considering all activations values of the convolution layers without considering the fully-connected layers and using the same strategy of the previous research [16] but without using one pre-trained model, namely AlexNet [15] as done [16]. Next, the logistic regression classifier is to decide the predicted class. As a result, a fine-tuned pre-trained VGG16 achieves the best performance at 92.60% accuracy. The other research [9] uses the same strategy as AlexNet [15]. The authors adapted the AlexNet with some modifications in its architecture related to the normalization process and type of the activation function. These modifications provide different proposed models. Then they applied fine-tuned processing on the models and achieved individual model ranges between 75% and 77%, while the combining model 84% accuracy.

The other research [10] uses unsupervised learning to implement its model. The proposed model is based on k-mean algorithm [17] and the probabilistic model (GMM). The proposed model first finds the region of interest (ROI) and then applied the feature selection using genetic algorithms (GA) [18]. Next, the model applies the CNN algorithm to find out better results. The resulted accuracy achieved 95.8%. The other research [12] also uses the segmentation step before classification using Lloyd's algorithm [11] for clustering and CNN for classification. A 96% accuracy was achieved by the proposed methods. However, in both the previous research, the authors did not mention exactly the proposed CNN model that was used.

The feature selection is also used as a preprocess of classification in a hybrid approach [13]. It uses a minimum redundancy feature selection (mRMR) algorithm [19] to effectively identify object properties and narrow down their relevance and then can predict breast cancer. The proposed

approach uses four classifiers SVM, Naïve Bays, Function tree and End Meta to find out the best performance. The result shows that SVM outperforms at 99% accuracy on average by combining it with MRMR algorithm. However, the feature selection process may not be enough for training with large datasets without dealing with deep learning models. This proposed approach is not the only research that claims the outperformance of the SVM classifier. A number of research early and recently [20] [21] [22] [23] [24] compare different classifiers and reach the same result, such as the study in [25] focuses to compare random forest and SVM classifiers for breast cancer classification and claimed that the highest accuracies 95% is for SVM.

On the other hand, some research not related to breast cancer classification uses a feature hierarchy to represent the images in CNNs. The research [26] studies real-world video sequences. It uses different hierarchical features of convolutional layers in CNNs to deal with features at early layers that keep more fine-grained spatial details and are useful for localization. It claimed that dealing with multiple layers of CNN features to get better performance for learning video features and visual tracking.

In the end, the result shows that the concept of transfer learning can be successfully applied to the breast classification domain. The activations of the source model can be used as features in the target model in the breast classification domain with less implementation cost, i.e., using pretrained models instead of training from scratch. Moreover, the combination of the features from different neural networks improves the accuracy of the classifiers.

In addition, a common way in the previous research using the activations of the fully connected (FC) layer as features. However, FC layers contain a higher number of activations than convolution layers which consume more computation cost. Also, extracting features from different layers leads to a better performance in learning video features. Moreover, most of the previous research used accuracy as a metric to evaluate their proposed models.

### III. PROPOSED METHOD

The idea of the proposed solution is image classification by extracting a feature hierarchy from pretrained CNN models and then feeding it into a classifier instead of using the whole images as inputs to that classifier. Extracting features from different layers of a single network is shown to lead to better performance in previous research working on learning video features [26].

To build the proposed model, two sub-models are constructed, one for extracting features and the other for classification. Writing  $f_m$  for the final proposed model that its form maintained under composition  $f_c$  and  $f_t$ ,  $f_c$  for the classifier, and  $f_t$  for the features by

$$f_m = f_c (f_t(x), y) \quad (2)$$

where  $x$  is called the input images and  $y$  is the true class of the input images with two possible values 0 or 1, where 0 indicates to no breast cancer (non-IDC) and 1 indicates to breast cancer was found in the image  $x$  (IDC). Exactly, each

input  $x_i$ , where  $x_i \in \mathbb{R}^{m \times n}$  will be mapped to  $y_i$ , where  $y_i \in \{0,1\}$ . Fig. 1 displays the general diagram of the proposed method. While each component of the composition will be described in the following subsections supported by described figures.

A. FT: Feature Extraction Phase

Feeding the whole images into a classifier needs to extract features manually an extremely time-consuming process and needs a strong knowledge of the domain. Also, converting 2D images to 1D vectors increases the number of trainable parameters exponentially and it significantly can increase the chance of overfitting especially if the size of a dataset is less than the number of learnable parameters. Thus, a CNN model is used in this proposed model for extracting the features.

Pretrained CNN model is decided to be used because the process of training networks with a large number of parameters is time- and resource-consuming. Thus, two pretrained models VGG16 and ResNet50 are used in this paper which they are used previously on a similar domain [6] [7].

1) VGG16 model is a type of CNN Architecture proposed by Visual Geometry Group (VGG), Oxford University [3]. Using VGG16 with 16 learnable layers regarding the depth which is larger than 8 layers in AlexNet [15], as an example, gives important for achieving high performance [15]. Moreover, VGG16 shows excellent classification performance for different previous works natural image classification in the Image Large Scale Visual Recognition Challenges [27] and for different previous works [6] [7] [28].

VGG model contains 16 learnable layers separated into five groups where each group ends with a pooling layer. In this proposed model, a pretrained VGG model is used with the input size differs from the default size in VGG16. The input size is equal to 50x50 pixels to be the same size as the input dataset and the three fully connected layers are removed. Fig 2 (a) is zoomed a portion of Fig. 1 that presents the feature extraction process in VGG16 model. The resulting sizes are presented on Table II after applying the following steps:

a) Creating a VGG16 model without the three fully connected layers due to the purpose of using VGG16 model as a features generator from its intermediate layers, not as a classifier.

b) Feeding the input into the VGG16 model and extracting feature maps at five intermediate layers. The layers are

c) The last layer (pooling layer) in each group, the red layer in Fig. 2 (a). Hence, five layers generate five different blocks of feature maps with different shapes.

d) To combine extracted feature maps from the previous step in a specific axis, the layers must have the same dimension on them. To unify the size to be the same 6x6 as the size of width and high, up/down sampling operations are applied. The down-sampling operation is applied to the first and the second extracted layers using the max pooling layer. The third layer is already having the same required size, so it does not need to change. The up-sampling operation is applied to the fourth and

the fifth extracted layers using the transpose convolutional layer that performs an inverse convolution operation. More detail about the values of their hyper-parameters is detailed in Section 4.1.

e) Each layer has a high number of channels which will increase the computation time. At the same time, the activated region of a channel is semantically meaningful and serves a similar role as the feature detectors to identify different features present in an image [29]. Thus, max pooling over the depth operation is applied to extract the maximum value of activation in a specific location (receptive field) among all channels and decrease the number of channels to only one channel. It is noteworthy that the utilized up-sampling method returns one channel by default. Thus, we can remove this step from the up-sampling layers.

f) Concatenating the five resulted from layers of the previous step on the depth axis to get one 6x6x5 layer. These feature maps will be concatenated with the resulting layer from the ResNet50 model which will be described in the next subsection.

2) ResNet50 model [4] consists of 48 convolution layers along with one max pooling and one average pool layer. The model has two types of connections: Identity connections between every two convolution layers and skip connections between some of them. The skip connections help to solve the vanishing gradient problem by allowing for the gradient to flow through these shortcut paths. Thus, it enables CNN models to get deeper and deeper without decreasing the accuracy by adding more layers to the network.

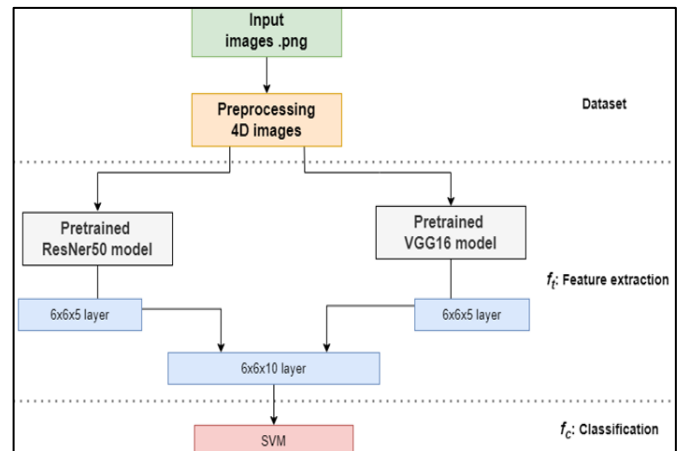


Fig. 1. Diagram of the proposed method

TABLE II. THE SIZES OF RESULTED FEATURES FROM VGG16 MODEL

Layer numbering	Extracted from Model	After up/down sampling	After depth pooling
1	25x25x64	6x6x64	6x6x1
2	12x12x128	6x6x128	6x6x1
3	6x6x256	-	6x6x1
4	3x3x512	6x6x1	6x6x1
5	1x1x512	6x6x1	6x6x1

In this proposed model, a pretrained ResNet model is constructed with the same VGG settings. Fig. 2(b) presents the feature extraction process and the resulting dimension is presented on Table III applying the following steps:

a) Creating a ResNet model without the fully connected layer due to the purpose of using ResNet model.

b) Feeding the input into the ResNet model and extracting feature maps at the last layer in each group. In total, there are five layers that generate five different blocks of feature maps.

c) The width and high are unified to be 6x6 using up/down sampling operations. The down-sampling operation is applied to the first three layers using the max pooling layer. The up-sampling operation is applied to the fourth and the fifth extracted layers using the transpose convolutional layer.

d) Decrease the number of channels to one channel using the max pooling over the depth operation.

e) Concatenating the five resulted from layers on depth axis to be one 6x6x5 layer.

3) *VGG and ResNet combination.* After building the two models separately, the resulting layer of each model is 6x6x5 layer. The last step in the extracting phase is to concatenate these two layers on the depth axis to be 6x6x10 layers as Fig. 1 shows.

TABLE III. THE SIZES OF RESULTED FEATURES FROM RESNET50 MODEL

Layer numbering	Extracted from Model	After up/down sampling	After depth pooling
1	13x13x64	6x6x64	6x6x1
2	13x13x256	6x6x256	6x6x1
3	7x7x512	6x6x512	6x6x1
4	4x4x1024	6x6x1	6x6x1
5	2x2x2048	6x6x1	6x6x1

### B. Classification Phase

The main goal of this paper is to classify breast cancer in the form of 2D images into binary classification: IDC and not IDC. The resulting features from the pretrained models along with the corresponding labels (i.e., IDC or non-IDC) are then used to train binary non-linear SVM classifier. The SVM classifier performs a good result in different works of breast cancer classification [19] [20] [21] [22] [23] [24] [25].

In SVM implementation, feature scaling is a crucial step because the methodology of SVM considers the distances among inputs to select the maximum decision boundary. This distance is surely different for non-scaled and scaled cases. Thus, the scaled step is applied using standardized features [30] with a mean equal to zero and standard deviation equal to one,

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

where  $x$  is the concatenated feature,  $\mu$  is the mean and  $\sigma$  is the standard deviation of these features. This step makes the features fall in a small range and leads to faster convergence in fewer iterations and then better performance [31].

## IV. RESULTS AND DISCUSSION

This section provides in detail the implementation of the proposed model and presents the results along with discussing it.

### A. Implementation

1) *Pretrained models construction.* The pretrained models are constructed using TensorFlow-Keras package with the same weights pretraining on ImageNet dataset [32] without changing or learning any weight. Thus, all convolution layers of the pretrained models are frozen. Moreover, the input size is equal to 50x50 pixels to be the same size as the input dataset. Although the default input size in the two pretrained models VGG16 and ResNet50 is 244x244 pixels, this proposed method discards the classification part with the fully connected layers to allow any input size. Table IV shows the value for each hyper-parameter in the construction.

2) *Extracting features phase.* To extract a feature hierarchy, ten temporary small models are constructed, five models for each pretrained model. Each small model is prepared to take the inputs equal to the input of the pretrained model and produce a block of feature maps as the output, which are used as features. Recall that the output layers of the small models are different regarding to producing five different layers in each pretrained model. Moreover, to unify the shape of the feature maps, different pooling layers in the Keras package are applied. In case of the layer size greater than 6x6, the down-sampling operation using the method *MaxPooling2D()* is applied, or in case of the layer size is less than 6x6, apply the transpose convolutional layer using the method *Conv2DTranspose()*. Moreover, the depth pooling operation is applied by the method *reduce-max()* to get a maximum element across a specific axis, here the depth. The assigned values for each parameter in the methods are presented on Table V. After unifying the shape of all feature maps, the method *concatenate()* is applied to concatenate all blocks of the feature maps among the depth axis.

3) *Classifier.* To can feed the features into the classifier, the array of the features must be reshaped to be 2D array with the number of inputs as the row and multiple of the 6x6x10 as the columns using the method *reshape()* in tensorflow package. The next step is applying the standardized features using the methods *StandardScaler()* and *transform()* in sklearn package.

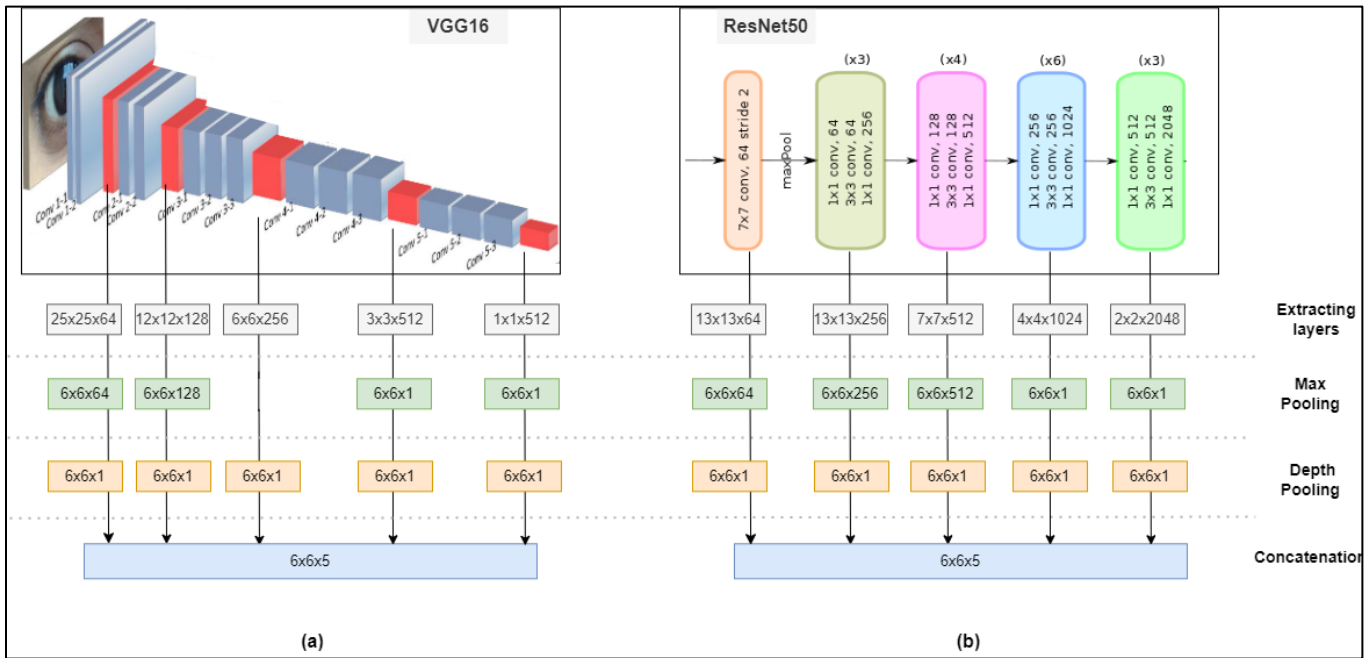


Fig. 2. Diagram of the pretrained (a) VGG16 (b) ResNet50 models.

After that, the non-linear SVM classifier is constructed from sklearn package using different methods. The method *SVC()* constructs the classifier by adjusting three hyper-parameters: regularization parameter *C* for giving a different level of regularization, the kernel parameter for enabling SVM to solve nonlinear classification problems when the inputs cannot be separated linearly, and the gamma parameter for considering as support vectors and therefore affect the decision region. When the value of gamma is low, the curve of the decision boundary is very low and thus the decision region is very broad and vice versa. Different values are assigned to these three parameters to estimate the best values. Table VI presents the suggested values for each parameter. The method *GridSearchCV()* helps to loop through the three parameters and fit SVM classifier on the training set to select the best values. The best value is optimized by the cross-validation splitting parameter *CV*.

To complete training SVM classifier, the number of maximum iterations is fixed to 30,000 iterations because the convergence warning appears due to convergence issues. The other solution to overcome this issue is using standardizing features that helps to reach the convergence state faster.

TABLE IV. THE HYPER-PARAMETERS OF CONSTRUCTING PRETRAINED MODELS

Parameters	Values
weights	imagenet
Layer trainable	Fasle (freeze)
Include top (FC)	False
Input size	50 x 50 pixcel
Parameters	Values

TABLE V. THE HYPER-PARAMETERS OF EXTRACTING FEATURE PHASE

Up/down sampling	Layers numbering	VGG16	ResNet50
Down	1	Pooling size= (4, 4)	Pooling size= (2, 2)
	2	Pooling size= (2, 2)	Pooling size= (2, 2)
	3	-	Pooling size= (2, 2) Strides = (1,1) with padding
Up	4	kernel_size = (2,2) Strides = (2,2)	kernel_size = (3,3) Strides = (1,1) with padding
	5	kernel_size = (6,6) Strides = (2,2)	kernel_size = (5,5) Strides = (1,1) with padding

TABLE VI. THE HYPER-PARAMETERS OF EXTRACTING SVM

Parameters	Values
Max iteration	30,000
C	[0.1,1, 5, 10]
kernel	Radial basis function (rbf), Polynomial kernel, sigmoid
gamma	[1,0.1,0.01,0.001]
Parameters	values

### B. Evaluation

1) *Cross-validation*. The cross-validation evaluates a classifier's performance by dividing the dataset into *k* parts. *K* is equal to 10 in this paper which is called 10-fold cross-validation. Thus, each image in this dataset will be used 9 times for training and once for testing. This validation then calculates the average between them to evaluate the classifier's performance. Thus, to evaluate the performance of the trained

classifiers, the cross-validation splitting parameter CV of the evaluation methods is assigned to 10 as 10-fold cross-validation.

2) *Performance metrics.* The Metric is accuracy as it is used in most of the previous works in Section II. Accuracy measures how many IDC and non-IDC images are classified correctly among all classifications. It shows overall how is the classifier classified correctly. Calculating the accuracy of the training set as an average over 10-cross-validation folds. Especially, the experiments are made in three cases for three SVM classifiers. Each SVM classifier is related to one of the following models: using only VGG16 model, only ResNet50, and using the combination of both models.

3) *Test platform.* The experiments are concurred using a personal laptop. However, the GPU in the laptop is not supported by python. Some of the tasks then run in long execution times and the memory of the laptop may not be enough. Thus, I have moved to use Google Colab Pro due to some commands could not be run using a free version of Google Colab.

4) *About the dataset.* The used dataset of the breast-cancer-image-classification is available in [33]. Fig. 3 shows the distribution of the dataset. The original images are for 279 patients with a small number of images scanned at 40x. However, overfitting is highly likely. Then, 50x50 patches were extracted including 198,738 negative examples (i.e., no breast cancer) and 78,786 positive examples (i.e., indicating breast cancer was found in the patch). Thus, the available dataset contains 277,524 patches in total. According to the figure, there is clearly an imbalance in the class data with over two times the number of negative data points than positive data points. However, in this work, the loading step, which loads the whole dataset into a programming notebook, has caused a crash multiple times after running the code in hours because of the available RAM space in Google Colab Pro. This leads to using a part of the dataset in the experiments with keeping the same percentage of imbalance in the class data.

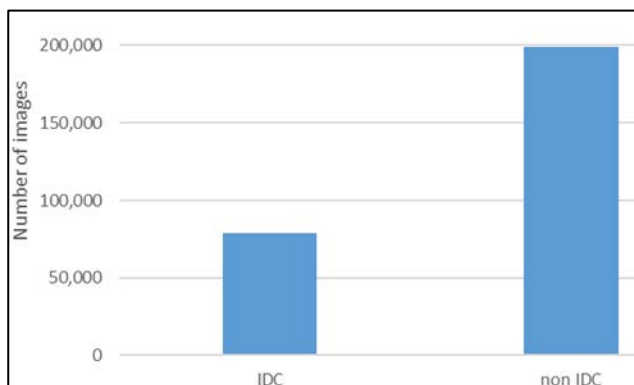


Fig. 3. The dataset distribution.

In this proposed method, to load and manipulate the images, the library image in Keras package is used. Then simple preprocessing is applied to stack all 50x50 images into 4D array to be able to deal with it in the implementation. Next,

the pixel values are normalized. The reason is that the pixel values can range from 0 to 256, where each number indicates a gray level value. The computation of large numeric values may get more difficult when sending these values through CNNs. We may lessen this by normalizing the numbers to a range of 0 to 1 by dividing the array by 255.

5) *Splitting dataset.* Train-Test split is a technique to evaluate the performance of the proposed model with giving 20% for the test set. The method *train-test-split()* in sklearn helps to split the images into training and test sets. The training set is used to train SVM classifier and then it calculates the accuracy of the training set as an average over 10-cross-validation. The training is also used to draw a learning curve. The test set is used to test the trained SVM classifier and then calculates the accuracy of the test set as an average over 10-cross-validation.

### C. Results

Different experiments were concurred to investigate the pretrained models and analyze the results trying to get a better performance. The best results are written down in this paper.

Table VII-A shows the overall comparison of the accuracy of the three cases. It shows the accuracy of training and test sets over the three classifiers from the three models. The values represent the mean of the accuracy which is the average value of accuracy among 10 different sizes of the training set along with their standard deviation values to represent the dispersion of accuracy values around the mean. The Table VII-B and Table VII-C represent values in the same manner as Table VII-A but investigate different effects. Table VII-B investigates the effect of swapping between the two steps of the methodology, Step 3 and Step 4. While Table VII-C investigates the effect of change the up-sampling methods. Table VIII shows the effect of applying standardization to the features on the accuracy of the combined SVM classifier, where the time is in seconds.

Fig. 4 shows the learning curve of the training and validation accuracy of the trained SVM classifier for varying numbers of training images. The x-axis shows the number of images that will be used to generate the learning curve. The y-axis shows the average of the accuracy values over 10 runs for each training subset size. The training and validation accuracies for different training set sizes in 10-cross-validation is measured to investigate influence of number of images on accuracy of the SVM classifier. Recall that the SVM classifier in this case is the combined SVM classifier.

### D. Discussion

Table VII-A shows both VGG16 and ResNet50 models give a satisfactory performance when using a feature hierarchy. The initialed experiments started without hierarchy, i.e., using only the last convolution layer as features but the result was lower. The result corresponds to the research [26], which confirms that the concept of feature hierarchy can be successfully applied to breast cancer classification. At the same time, the performance is better when the features from the two models are fused. This result shows the effect fusing of different pretrained models to get a better result than using each pretrained individually.

TABLE VII. THE ACCURACY OF THE THREE CLASSIFIERS FROM THE THREE MODELS

	Data sets	VGG16	ResNet50	Combination
A) Original	Training set	0.927±0.015	0.923±0.016	0.943±0.014
	Testing set	0.897±0.028	0.889±0.031	0.920±0.041
B) After swapping Step 3 & 4	Training set	0.923±0.013	0.923±0.015	0.943±0.021
	Testing set	0.892±0.030	0.890±0.036	0.908±0.033
C) After double operation	Training set	0.923±0.018	0.926±0.017	0.933±0.019
	Testing set	0.899±0.025	0.899±0.029	0.916±0.039

TABLE VIII. THE ACCURACY OF THE TWO CLASSIFIERS WITH/WITHOUT APPLYING THE STANDARDIZATION

Applied?	Data sets	Accuracy	Fit time	Fold time	Warning?
Yes	Training set	0.943±0.014	144.05 s	1054.24 s	No
	Testing set	0.920±0.041	-	128.61 s	No
No	Training set	0.897±0.011	186.02 s	1300.88 s	Yes
	Testing set	0.915±0.023	-	126.19 s	No

According to pretrained models, we can also observe from Table VII-A that the activations of pretrained model, that are trained on ImageNet dataset [32], can be used as features in the proposed model in the breast classification task with consuming less implementation cost, i.e., using pretrained models instead of training from scratch.

Moreover, the extracting steps shown on Fig. 2 describe applying up/down sampling (Step 3) before applying depth pooling (Step 4). However, both these two steps are related to unifying the shape of feature maps. Thus, if these two steps are swapped, i.e., applying up/down sampling after depth pooling, the result is almost the same with a small enhancement for the original case as shown on Table VII-B. One of the possible reasons is both operations work on getting the maximum value which will generate almost similar values in two directions.

Regarding up-sampling operations, different operations can be used other than the transpose convolutional layer. The other simple common type is using the method *UpSampling2D()* to double the dimensions of the input. After applying the simple double operation in the experiments, it gives almost the same result as Table VII-C shows with little enhancement for the original case. However, the key difference is in their learning. The simple double operation is a simple scaling up of the input without learning to achieve a less complicated in implementation. Whereas the transpose convolutional operation is a convolution operation whose kernel is learnt while learned the model to learn the best up-sampling for the task.

Turning to the normalization on Table VIII, the case of standardization (the top case) gives a faster result in fewer iterations. It can be considered as one of the solutions to overcome the warning of convergence issue that expresses that the estimation terminated early before reaching the convergence. Also, standardization achieves a better result in terms of accuracy. Especially for SVM classifier, the scaling helps to decrease the distances between inputs to select the maximum decision boundary.

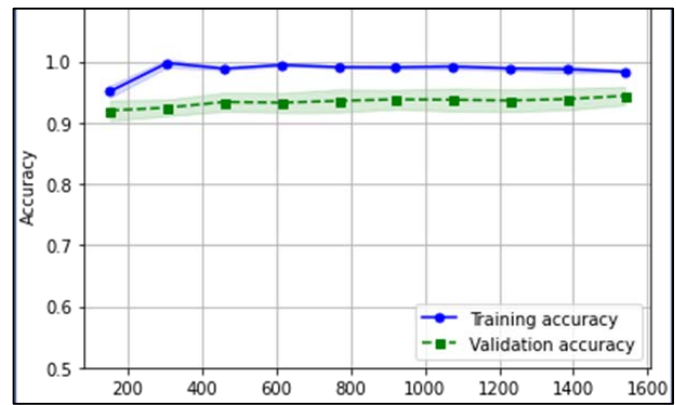


Fig. 4. The evaluation of the classifier over different examples.

The overall trend of Fig. 4 shows the effect of the number of images on the accuracy. The accuracy of the training set is higher than the validation set but with acceptable gap, i.e., the gap between them did not increase after a specific point to express about happening overfitting. One of the possible reasons is using the pretrained models with their learnable parameters, i.e. all layers are frozen and used the same weights. This helps to reduce the number of required images to reach the convergence.

Some other notes are appeared during the experiments. There are different possible sizes of the receptive field can be chosen to unify the shape of feature maps. When trying to unify the size to be 3x3, the resulting accuracy is almost the same in most cases, with little increase for 6x6 in other cases. However, more investigations in the future are better to be conducted on this size and other varied sizes to get adequate results about the effect of changing size on performance.

#### E. Comparing with the State-of-the-art

Ultimately, the results correspond to the research [26], which observes a good effect of using a feature hierarchy. While the proposed model in this paper uses the feature hierarchy for different domain, which is the breast cancer classification tasks.

At the same time, comparing to the other works provided in Section II, this proposed model seems to infer the same previous result [6] [7] about the effect fusing of different pretrained models to get a better result than using each pretrained individually. But in this paper, the combination is between a features hierarchy extracted from two models VGG16 and ResNet50.

#### V. CONCLUSION

IDC is the most common subtype of all breast cancers. Instead of manual diagnosis, it must find solutions to ease diagnostic burdens, especially in under-staffed laboratories and equipment. Thus, the goal of this paper is to classify breast cancer in the form of images into binary classification: IDC and not IDC. This paper proposes a CNN-based model for learning features of breast cancer images that combines two pretrained CNN models to extract a feature hierarchy and then feeds them into the SVM classifier. Besides, experimental results show that classification performance is higher in the

combined pretrained model and fusing the deep features from various layers from various pre-trained CNNs leads to better classification performance. In addition, other findings present the effect of some factors such as the normalization of training SVM classifier. However, those results are not the best results. It can be considered as a contribution, while the performance can be after the additional investigation in several factors, such as change the size of the receptive field of the features maps, number of pretrained models as well as other datasets with different pixel sizes may get another improvement.

In the future, this paper provides various recommendations that are expected to help in developing CNN models. First, combining other information along with the breast images during developing DNN models, such as changes in the breast shape and DNA sequences, may increase the accuracy of the classification. Second, the breast imaging modalities are better to consider during developing DNN models. Adopting new modalities of imaging may provide more accurate details, such as shear wave elastography (SWE) or magnetic resonance imaging (MRI). Third, while the enormous quantity of unlabeled photos is a valuable source of data, it cannot be used in supervised learning. Instead, the research can shift to training in an unsupervised manner, such as using clustering approaches. In the end, increasing research interest and rapid technological advancements creates a chance for researchers to continue to evolve models of breast cancer classification.

#### REFERENCES

- [1] Nahid and Y. Kong. "Involvement of machine learning for breast cancer image classification: a survey." Computational and mathematical methods in medicine 2017, 2017.
- [2] B Sahiner, H Chan, N Petrick, D Wei and M. Helvie. "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images". IEEE Trans Med Imaging, vol. 15, no. 5, pp. 598-610, 1996.
- [3] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.
- [4] H. Kaiming, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
- [5] C. Cortes, and V. Vapnik. "Support-vector networks." Machine learning, vol. 20, no. 3, pp. 273-297, 1995.
- [6] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. Rodrigues. "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning." Pattern Recognition Letters 125, pp: 1-6, 2019.
- [7] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Elling. "Skin lesion classification using hybrid deep neural networks." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1229-1233. IEEE, 2019.
- [8] R. Mehra, "Breast cancer histology images classification: Training from scratch or transfer learning?". ICT Express, vol. 4, no. 4, pp. 247-254, 2018.
- [9] R. Andrik, W. Scotney, J. Morrow, and H. Wang. "Breast mass classification in mammograms using ensemble convolutional neural networks." In 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services, pp. 1-6. IEEE, 2018.
- [10] S. Shamy, J. Dheebea, "A research on detection and classification of breast cancer using k-means GMM & CNN algorithms," International Journal of Engineering and Advanced Technology, vol. 8, no. 6S, pp. 501-505, 2019.
- [11] P. Lloyd "Least squares quantization in PCM", IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982.
- [12] V Sansya Vijayan, Lekshmy P L, "Deep learning based prediction of breast cancer in histopathological images," International Journal of Engineering Research & Technology, vol. 8, no. 07, pp.148-152, 2019.
- [13] M. Rathi and V. Pareek, "Hybrid approach to predict breast cancer using machine learning techniques," International Journal of Computer Science Engineering, vol. 5, no. 3, pp. 125-136, 2016.
- [14] S. Christian, W Zaremba, I. Sutskever, J. Bruna, D. Erhan, Ian Goodfellow, and R. Fergus. "Intriguing properties of neural networks." , arXiv preprint arXiv:1312.6199, 2013.
- [15] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.
- [16] K. Simonyan, A. Zisserman, "Very deep convolutional networks for largescale image recognition", in: International Conference on Learning Representations, pp. 1-14, 2015.
- [17] Lloyd, S. P."Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd., S. P, 1982.
- [18] D. E. Goldberg. "Genetic Algorithms in Search, Optimization and Machine Learning". New York: Addison Wesley, 1989.
- [19] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinforma Comput Biol, vol. 03, no. 02, pp. 185-205, 2005.
- [20] Lo C-S, Wang C-M. "Support vector machine for breast MR image classification". Comput Math Appl, vol. 64, pp :1153-1162, 2012.
- [21] J. Ren "ANN vs. SVM: which one performs better in classification of MCCs in mammogram imaging", Knowl-Based Syst, vol. 26, pp :144-153, 2012.
- [22] M. Tahmooreesi, A. Afshar, B. Bashari Rad, K. B. Nowshath, 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.
- [23] Ebru Aydindag Bayrak, Pinar Kirci, TolgaEnsari,"Comparison of machine learning methods for breast cancer diagnosis". 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp. 1-3,2019.
- [24] Ch. Shravya, K. Pravalika, ShaikSubhani, "Prediction of breast cancer using supervised machine learning techniques," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 6, pp. 1106-1110, 2019.
- [25] C. Aroef, Y. Rivan, and Z. Rustam. "Comparing random forest and support vector machines for breast cancer classification." Telkomnika 18, no. 2, pp. 815-821, 2020.
- [26] C. Ma, J. Huang, X. Yang and M. Yang, "Robust Visual Tracking via Hierarchical Convolutional Features," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 11, pp. 2709-2723, 2019.
- [27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015.
- [28] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks", Journal of Biomedical Informatics, vol. 86, pp. 25-32, 2018.
- [29] L. Liu, C. Shen, & A. van den Hengel. "Cross-convolutional-layer pooling for image recognition". IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 11, pp. 2305-2313, 2016.
- [30] Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning. 2nd ed. New York: springer. 2017.
- [31] Koo, KM., Cha, EY. "Image recognition performance enhancements using image normalization". Hum. Cent. Comput. Inf. Sci. vol. 7, 2017.
- [32] "About ImageNet", image-net, 2022. [Online]. Available: <https://www.image-net.org>. [Accessed: 01- Mar- 2022].
- [33] "Breast Histopathology Images", Kaggle, 2022. [Online]. Available: <https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images?resource=download>. [Accessed: 01- Mar- 2022].