Multidimensional Systems and Signal Processing (2021) 32:747–765
https://doi.org/10.1007/s11045-020-00756-7

Diagnosis of breast cancer based on modern mammography using hybrid transfer learning

Aditya Khamparia[1] · Subrato Bharati[2] · Prajoy Podder[2] · Deepak Gupta[3] · Ashish Khanna[3] · Thai Kim Phung[4] · Dang N. H. Thanh[4]

Received: 24 May 2020 / Revised: 9 December 2020 / Accepted: 19 December 2020 /
Published online: 11 January 2021

Abstract
Breast cancer is a common cancer in women. Early detection of breast cancer in particular and cancer, in general, can considerably increase the survival rate of women, and it can be much more effective. This paper mainly focuses on the transfer learning process to detect breast cancer. Modified VGG (MVGG) is proposed and implemented on datasets of 2D and 3D images of mammograms. Experimental results showed that the proposed hybrid transfer learning model (a fusion of MVGG and ImageNet) provides an accuracy of 94.3%. On the other hand, only the proposed MVGG architecture provides an accuracy of 89.8%. So, it is precisely stated that the proposed hybrid pre-trained network outperforms other compared Convolutional Neural Networks. The proposed architecture can be considered as an effective tool for radiologists to decrease the false negative and false positive rates. Therefore, the efficiency of mammography analysis will be improved.

Keywords Hybrid transfer learning · Medical image segmentation · Breast cancer · Mammography · 3D mammography · Convolutional neural networks

1 Introduction

Every year, 12% of women are diagnosed with breast cancer (McGuire et al. 2015). In the US alone, 40,000 women die of breast cancer annually (Bharati et al. 2018; Cancer.gov 2018). Evidence shows that early detection of breast cancer can significantly increase the survival rate of women.

✉ Dang N. H. Thanh
thanhdnh@ueh.edu.vn

[1] School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

[2] Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh

[3] Maharaja Agrasen Institute of Technology, Delhi, India

[4] School of Business Information Technology, University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam

Mammography, a special X-ray of the woman's breast, is one of the most common diagnostic tools for detecting breast cancer (Bharati et al. 2020a, b; Thanh and Surya 2019). A 3D mammography is an advanced model compared to mammography. A 3D mammogram uses multiple breast X-rays to create a 3D picture of the breast. A 3D mammogram is used for finding breast cancer in patients who have no signs or symptoms. It can also be used to investigate other issues on breasts, such as breast mass, pain, and nipple discharge (Kumar et al. 2020).

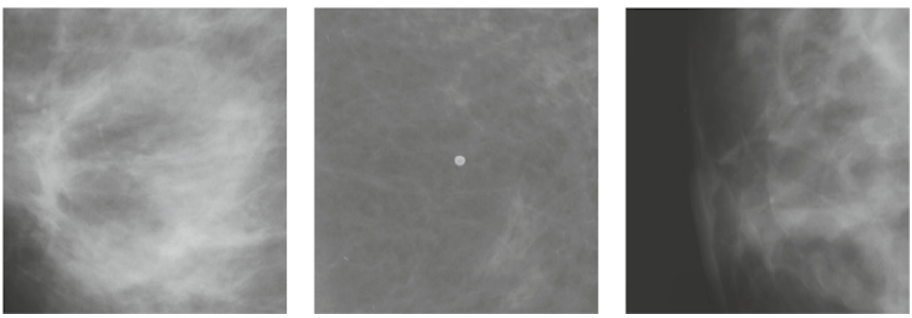
When screening breast cancer, 3D mammogram machines will create 3D images and standard 2D mammogram images (Clinic 2020). Studies showed that "Combining 3D mammograms with standard mammograms reduces the need for additional imaging and slightly increases the number of cancers detected during screening".

According to the mammography technique as well as the 3D mammography technique, one can show masses and even calcifications, which are precursors to breast cancer. However, correctly identifying these images can be challenging for radiologists. Moreover, time constraints in assessing the images often result in incorrect diagnosis with detrimental consequences. For instance, a false negative diagnosis, that a case is normal when it is, in fact, an early form of breast cancer, can decrease the chance of 5-year survival significantly.

A mammogram is a type of X-ray image of the breast. It can be captured by both mammography and/or 3D mammography. Doctors use the mammogram to identify the signs of breast cancer. So, it is capable of detecting calcifications, lumps, dimpling, etc. These are the common signs shown in the early stage of breast cancer. The mammograms of the digital database for screening mammography (DDSM) dataset is used in this paper. This dataset is available in a valid online repository (DDSM 2020) which is illustrated in the article of Lee et al. (2017). DDSM is a group of labeled mammographic images. This database is maintained by the research community (Heath et al. 1998). It is one of the largest datasets for studying breast cancer. Figure 1 shows some types of images of DDSM.

2620 instances are contained in the dataset. The instances are the mammograms of patients with masses, calcifications, etc. The labels for calcification and masses are specified in four categories.

(1) Benign.
(2) Malignant.
(3) Benign without a callback.
(4) Unproven.



(a) Lump or mass **(b)** Calcification **(c)** Normal

Fig. 1 Some types of images of DDSM

Table 1 Summary of the DDSM patch data set

	Malignant	Benign w/o callback	Benign	Unproven	Total
Calcification	797	539	800	16	2152
Mass	1075	179	1079	21	2354
Pathological cases	–	–	–	–	4506
Non-pathological cases	–	–	–	–	6207
Total number of patches	–	–	–	–	10,713

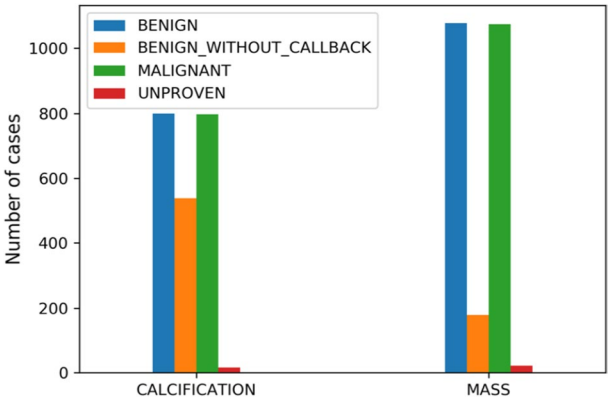


Fig. 2 Classes and labels of the DDSM dataset

Besides, the images have been categorized on a scale of 1–5, according to the BI-RADS. BI-RADS means breast imaging, reporting, and data system. BI-RADS can be considered the most effective tool to detect breast cancer. Score 5 shows that the mammogram results are very suspicious, and the probability of breast cancer is almost 95%. To simplify our analysis, patches are used instead of full images. It helps not only for efficient computation but also for better performance. Because feature detection becomes easier. 10,713 patches are contained in our dataset. Table 1 summarizes the statistics of the DDSM patch data set.

All images are separated at the abnormality level and full mammography as DICOM files where the whole mammography of breast images contains both CC and MLO visions of the breast mammograms. Moreover, abnormalities are depicted as binary mask breast images where the size of the images is the same as their related breast mammograms. The ROI of every abnormality is described in these mask breast images. Users can play out an element-wise choice of pixels inside an abnormality mask which was made for every mammogram. We have separated the images, including only abnormalities cropped for analysis of abnormalities. We have also separated the dataset like train, test, validation using python programming.

Exploring the data provides the plots shown in Figs. 2 and 3. Overall, there are more cases of masses than of calcification (in Fig. 2). The number of malignant and benign cases for calcification and masses seems to be the same. For both calcification and masses, a few cases have been categorized as 'unproven'. In our analysis, we decide to mark these as pathological as it is not clear if they can be considered healthy patients or not (the number is small and should not have a strong negative impact on our predictive power). As we aim

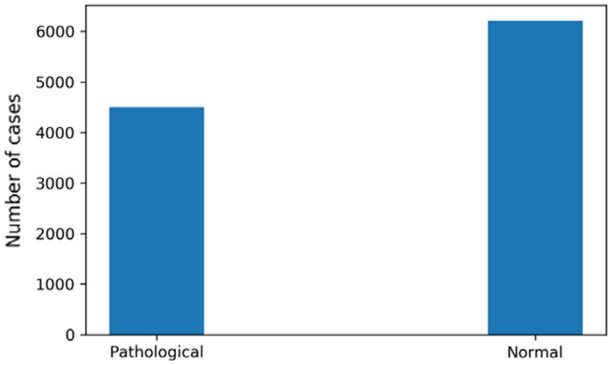


Fig. 3 Split of the patches based on pathology

to model a binary classifier, we label all mass and calcification patches as pathological. In total, we have 4506 pathological and 6027 non-pathological patches (Fig. 3).

In this paper, we have proposed MVGG based on VGG 16. VGG 16 is modified in our application by fine-tuning the feed-forward, dense layers in the end to just one layer with 32 nodes, followed immediately by an output layer with sigmoid activation and one node (for binary classification). Binary classification is necessary to predict breast cancer. Therefore, categorizing mass and calcification labels are categorized as 'pathological', and normal images are categorized as 'non-pathological'. VGG16 is designed initially to label up to 1000 classes and therefore have wide dense layers (4096 nodes). The width of these layers is cut down not to mix the information of the features at the time of passing from 4096 nodes to just one node in the output layer.

The significant contributions of the research work listed as follows:

- Modified VGG model has been proposed to diagnose breast cancer utilizing 2D and 3D images of mammograms.
- The proposed hybrid transfer learning model (a fusion of MVGG and ImageNet) provides an accuracy of 88.3% which surpasses existing machine learning models.
- The data augmentation and regularization approach enhance the breast cancer detection rate and improve the proposed system performance.

2 Literature reviews

There are several machine and deep learning approaches in health care systems. ML is conducted in various domains, including health care, disease detection, biomedical, etc. (Tiwari and Melucci 2019a). Some works (Tiwari and Melucci 2018, 2019b; Khamparia et al. 2020) related to binary and multi-class classifications using machine learning have been proposed, and they exhibited some performance matrix-like accuracy, recall, precision, F1 score, etc. (Tiwari and Melucci 2018). Some unsupervised algorithm was already utilized for the treatment of breast cancer, lung cancer, and coronavirus (Tiwari et al. 2020; Mondal et al. 2020). Therefore, we can use deep learning for the detection of diseases from image data. Conversely, the image fusion algorithm was conducted for medical images where the big data was efficiently utilized (Tan et al. 2020). Traditional deep learning techniques are used for detecting blood cells. The images of the dataset were 13 k and provided

results according to the performance matrix (Tiwari et al. 2018). Next, a hybrid method was offered by Reddy et al. (2020). They used hybrid deep belief networks (DBNs) and MRI images to detect glioblastoma tumors. The proposed method combined DBN with DTW to improve the efficiency of DBN. Thus, we use hybrid MVGG16 ImageNet for enhancing efficiency.

A deep learning-based system for the classification of the images of breast tissue is proposed in Rakhlin et al. (2018). For those images, 650×650 has been extracted with 400×400 pixels. Next, pre-trained VGG-16, InceptionV3, and ResNet-50 networks are conducted for the feature extraction. 10-fold cross-validation with LightGBM classifier has been driven to the classification and extraction of in-depth features. That technique gets an average accuracy of 87.2% across leave on out for breast cancer image classification (Rakhlin et al. 2018). In the other work (Kwok 2018), 4-DCNN architectures, i.e. InceptionResnetV2, InceptionV3, VGG19, and InceptionV4 have been used for the classification of images of breast cancer. The size of the images is 1495×1495 of 99 pixels. Various data augmentation systems have also been developed to increase the accuracy. In Vang et al. (2018), the ensemble-based architecture is proposed for multi-class image classification of breast cancer. Their conducted ensemble classifier involved; logistic regression, gradient boosting machine (GBM), majority voting to achieve the final prediction.

Moreover, the ensemble-based boosted neural network is also used for the diagnosis of lung cancer (Alzubi et al. 2019). The bagging algorithm is improved in the paper of Alzubi (2015). This algorithm cannot provide a good result for this complex dataset. Therefore, our proposed work will carry out for other complex image data. It can be used in IoT healthcare system and will be transmitted data securely according to the authors of Rani et al. (2019). Furthermore, the authors of Qian et al. (2020) conducted an unsupervised dictionary learning in an internet-based healthcare system for patient monitoring. They offered an ECG compression method where they measured EEG. This method has developed the dictionary continuously while the hidden pattern refined and occurred the dictionary. In Vahadane et al. (2015), the stain-normalization technique is applied to stain images for normalization where the achieving accuracy is 87.50%. Another research (Sarmiento and Fondón 2018) conducts a machine learning method where feature vectors are extracted from various characteristics i.e. texture, color, shape, etc. For the ten-fold cross-validation, SVM provides 79.2% accuracy. Lastly, the paper (Nawaz et al. 2018) uses a fine-tuned AlexNet for the automatic classification of breast cancer. They achieve an accuracy of 75.73% where the patch-wise dataset is used.

Our literature review covered three topics: (1) The state-of-art deep learning architectures for the task of binary classification of images, (2) Performance achieved in similar tasks as a benchmark for our algorithm, (3) Studies on physicians' performance to understand the clinical implications of such algorithms.

2.1 State of the art architectures

VGG network is presented by the author of Simonyan and Zisserman (2014). It is a simple model. It consists of a 13 layered CNN where 3×3 filters (Fig. 4) are used. VGG model has 2×2 max-pooling layers. The performance of multiple, smaller-sized kernels is comparatively better than a single larger-sized kernel because the increased depth of the VGG network can support the kernel to learn more complex features.

Secondly, Residual Network (ResNet) is considered. For image classification, ResNet is the most popular architecture. It is presented by the paper (He et al. 2016). Residual block

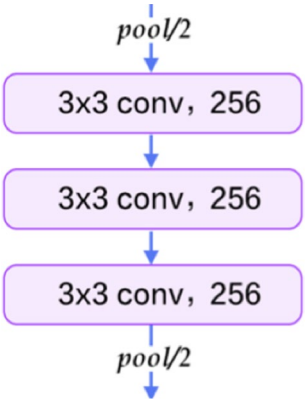


Fig. 4 The characteristic of 3×3 convolution layers of the VGG

can be considered a distinguishing feature in ResNet. (in Fig. 5). The residual block allows the residual network to achieve a depth of 152 layers. Vanishing gradients is a common problem in DCN. This problem can be moderated by the residual block. Because of vanishing gradients, the performance of ResNet can be degraded with the increase of depth.

Finally, MobileNets for mobile and embedded vision applications are proposed, which are based on a streamlined architecture that uses depth-wise separable convolutions to build light weight deep neural networks. This network is introduced by the authors of Howard et al. (2017). The core layer of MobileNet is depth-wise separable filters, named as depth-wise separable convolution. Finally, the width and resolution can be tuned to tradeoff between latency and accuracy. The purpose of using MobileNet in comparison to other architectures is that it has very little computation power to run or apply transfer learning to. This makes it a perfect fit for Mobile devices, embedded systems, and computers without GPU or low computational efficiency with compromising significantly with the accuracy of the results (in Fig. 6).

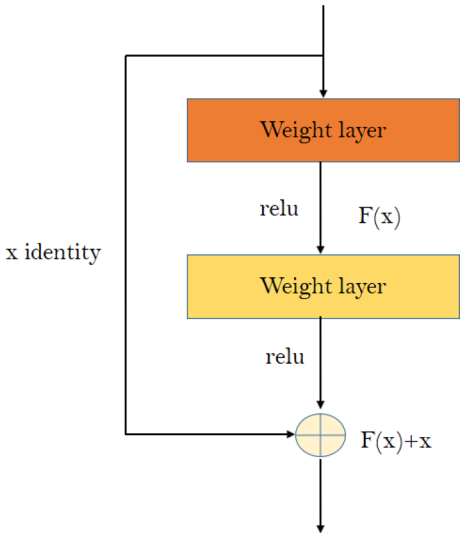


Fig. 5 The residual model of the ResNet architecture

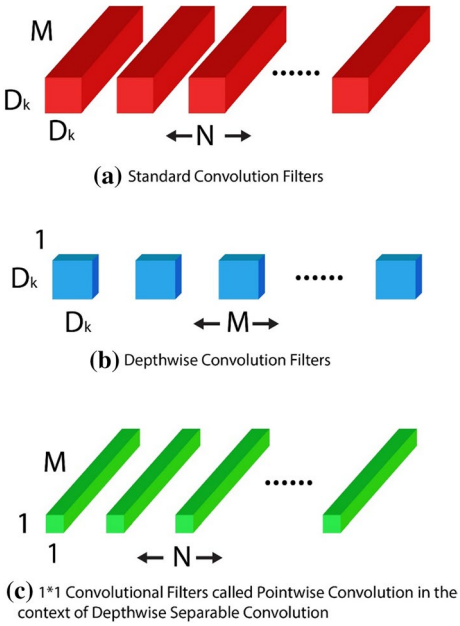


Fig. 6 MobileNet architecture

2.2 Prior art in breast cancer classification

To detect breast cancer, several algorithms and classification have been developed using different datasets. For instance, a paper published in 2015 obtained 85% accuracy for identifying images with a mass and also localizing 85% of masses in mammograms with an average false positive rate per image of 0.9 (Ertosun and Rubin 2015). In Shen (2017), developed an end-to-end training algorithm for a whole-image diagnosis. It deploys a simple convolutional design achieving a per-image AUC score of 0.88 on the DDSM dataset. We adopt this metric as the benchmark for our algorithm, in addition to an accuracy benchmark of 85%.

2.3 Physician performance

Several high-quality studies explored the performance of physicians diagnosing mammograms. A study was looked at radiologist performance on mammographs from 1192 patients (Rafferty et al. 2013). In a first study, 312 cases (48 cancer cases) were diagnosed by 12 radiologists who recorded if an abnormality that requires a callback was present. This resulted in a sensitivity of 65.5% and a specificity of 84.1%. In a second study, 312 cases (51 cancer cases) were analyzed by 15 radiologists. They obtained additional training and also reported the type and location of the lesion. It was in a sensitivity of 62.7% and a specificity of 86.2%. Another high-quality study compared different diagnosis methods such as mammography, ultrasonography (US), and physical examination (PE) using a data set of 27,825 screening sessions Kolb et al. (2002) and compared the results of the three diagnosis methods with the actual biopsy. The results showed a sensitivity of 77.6% and a specificity of 98.8%. However, these scores

Table 2 Comparison of risks for type 1 and 2 diagnostic errors

	Risks and costs of a diagnostic error
False-positive	Additional test: Costs and minimal-invasive biopsy Short-term distress/long-term risk of anxiety
False-negative	5-year survival rate is strongly impacted by later detection: Decreases from 93 to 72% from stage 3 to stage 2

Table 3 The survival rate of breast cancer stages

5 years overall survival by stage		
Stage	5-year overall survival (%)	Classification
0	100	In situ
1	100	Cancer formed
2	93	Lymph nodes
3	72	Locally advanced
4	22	Metastatic

were not achieved by radiologists using only mammograms and thus do not fit well for a benchmark for this task.

Most relevant as a benchmark for our analysis is the first study by the authors of Rafferty et al. (2013) as the 12 radiologists restricted themselves to binary classification, which is similar to our approach.

2.4 Clinical significance

Medical suggestions of the mammogram diagnosis are essential to improve an algorithm with Clinical significance. Radiologists have a considerably higher specificity than sensitivity. So, it means that the false-negative rate is higher than the false positive rate. Table 2 illustrates the Comparative analysis of risks for two types (1 and 2) of diagnostic errors.

False-positive diagnosis indicates that the radiologist judges a normal mammogram of malignant or benign type. As a result, that patient has to revisit the clinic, and in most cases, further testing through a biopsy is performed. Biopsy for breast cancer detection is minimally invasive, and only a small incision is needed. However, there is a range of evidence showing the psychological effects of such false positives. According to a study from 2000, it can lead to short-term distress as well as long-term anxiety (Aro et al. 2000). On the contrary, a false negative implies that a potentially cancerous case is misinterpreted as healthy. The consequences of this can be very severe because breast cancer when left untreated progresses in its stages and with each stage, a different 5-year life expectancy is associated (in Table 3 Cancer.gov 2018).

In general, it is more important to avoid false-negative over a false positive. A mammogram is first followed by sonography, and if this is positive as well, further testing is done via a minimally invasive biopsy. However, in the future, it would be great to differentiate

further and decrease the false positive in the three categories of BI-RADS. As of now, 98% of patients in this category have to come back every 6, 12, 24 months for a check-up, yet do not have breast cancer. This is a large burden.

Considering Table 2 and the two interviews, we conclude that a false negative error can have more severe consequences than a false positive error. Thus, we decide to design our algorithm to have a threshold that is more sensitive than specific.

3 Modeling

3.1 Data cleaning

Before building the model, we clean the data set to convert it into the appropriate form. We assign new, binary labels to the images by categorizing all the original mass and calcification labels as 'pathological', and the normal images as 'non-pathological'. Thus, the problem is decreased to binary classification. Next, we randomly divide the data set into train, validation, and test splits, in approximate proportions of 75:10:15, respectively. While doing so, we ensure that the splits are evenly balanced between the two classes (as evident in Fig. 7). Train test split method is adopted to check the validation of this work. For validation of this work, 75% of the total images are trained in the total dataset.

3.2 Performance evaluation

The performance of classification is evaluated by the following metrics: accuracy and AUC. AUC means the area under the ROC curve. These two parameters are widely-used for evaluating classification. Accuracy means the percentage of cases that the model classifies correctly. AUC implies the capability of the model to discriminate between the two classes.

3.3 Modeling process

The flow chart can easily describe the modeling process shown in Fig. 8.

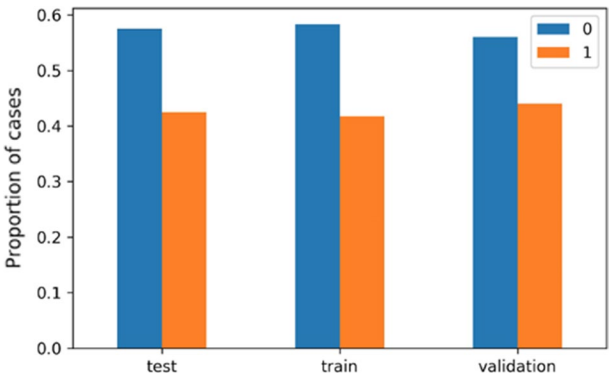


Fig. 7 Balance of classes in the train, validation, and test splits

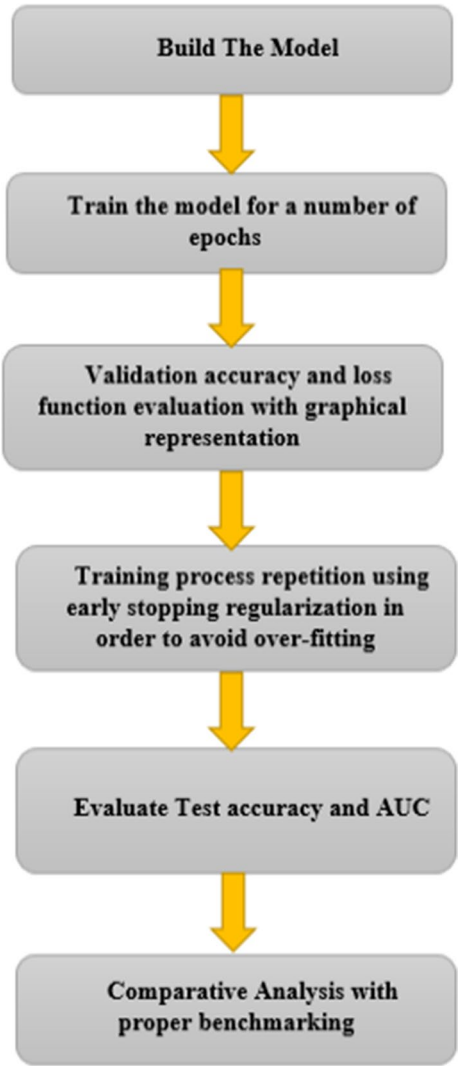


Fig. 8 Workflow diagram of the modeling process

3.4 Model building

Model building is the first step of model processing. It can be divided into four sub-stages:

1. Construction of the baseline model and performance evaluation.
2. Training of popular models with various architectures, and selection of best models.
3. Deployment of regularization and data augmentation methods to develop the performance. Choose the best model as the final model based on the performance.
4. Tuning of hyper-parameters on the final model to accomplish the desired level.

Table 4 describes the details of the architecture of the baseline (simple) model. There are two 2-dimensional convolution layers. There are 32 and 64 filters. There is a dense layer in the architecture having 32 nodes on the top. At the time of the performance

Table 4 Architecture of the baseline model

Layer (type)	Output shape	Parameters
Conv2d_3 (2D convolution layer)	(None, 254,254,32)	320
Conv2d_4 (2D convolution layer)	(None, 252,252,64)	18,496
Max_pooling2d_2	(None, 126, 126,64)	0
Flatten_2	(None, 1016064)	0
Dense_3	(None, 32)	32,514,080
Dense_4	(None, 1)	33

evaluation of the model on the test set, the accuracy of 75.9% is obtained. It is about 13% lower than the desired accuracy or desired benchmark.

In the next stage, three image classification models are implemented. The models are ALEXNET, VGG16, VGG19, MVGG, MobileNet, and ResNet50. The models are then modified by tuning the feed-forward, dense layers in the end to only one layer having 32 nodes. These models are designed initially to label up to 1000 classes and therefore have wide dense layers (4096 nodes).

Two additional methods have been considered after choosing the final model to observe if they improve performance. The two techniques are

(1) Data augmentation and
(2) Pre-training.

In the data augmentation stage, three operations are performed on the input images. The operations are:

(1) Flip the images along a horizontal axis
(2) Shift vertically/horizontally within a width range of 0.2
(3) Rotate randomly within a twenty-degree range.

The pre-training process includes initializing model parameters with values learned from a different data set, instead of random ones. The pre-training process not only can speed up learning but also achieve improved local optima in gradient optimization. In this paper, the best model is pre-trained using weights. ImageNet data set is trained in this case.

To tune on the best model, batch size and learning rates are varied to enhance the accuracy.

4 Results and interpretation

4.1 Performance of different models

The performance of different implemented models is shown in Table 5.

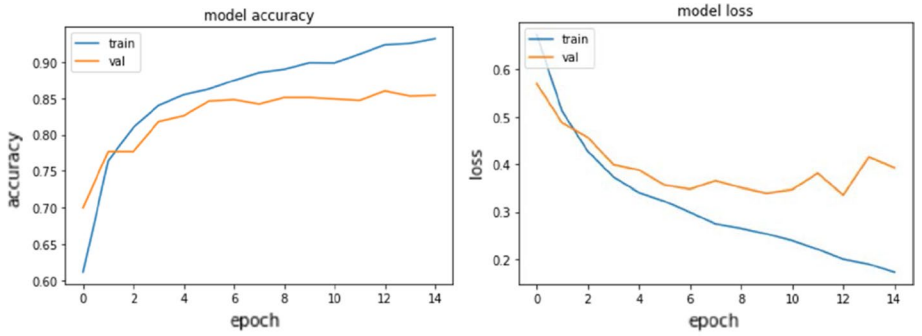
Table 5 depicts the comparison of various classification algorithms including ALEXNET, ResNet50, MobileNet, VGG16, VGG19, etc. where our final model hybrid MVGG16 ImageNet provides higher accuracy, precision, recall, and F1 score than another traditional algorithm for our DDSM dataset. We calculate our results for 15

Table 5 Performance evaluation of different models

Model	Size of batch	Special pre-processing	Precision (%)	Recall/sensitivity (%)	F1-score (%)	Accuracy (%)	Number of epochs
Simple model	32	No	74.2	75.1	74.7	75.9	15
ALEXNET	32	No	81.3	83.1	82.2	83.4	15
ResNet50	32	No	82.4	84.8	83.6	85.1	15
Mobile Net	32	No	85.1	85.9	85.5	87.2	15
VGG16	32	No	82.6	82.9	82.8	83.1	15
VGG19	32	No	82.8	82.2	82.5	82.5	15
MVGG16 (Modified VGG 16)	32	No	88.6	87.2	87.9	89.8	15
MVGG16 and Augmentation	32	Flips, shifts, rotations	90.5	92.2	91.3	92.8	15
Hybrid MVGG16 ImageNet (Final Model)	32	Pre-trained on ImageNet	93.5	93.7	93.6	94.3	15

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
dense_1 (Dense)	(None, 16)	16016

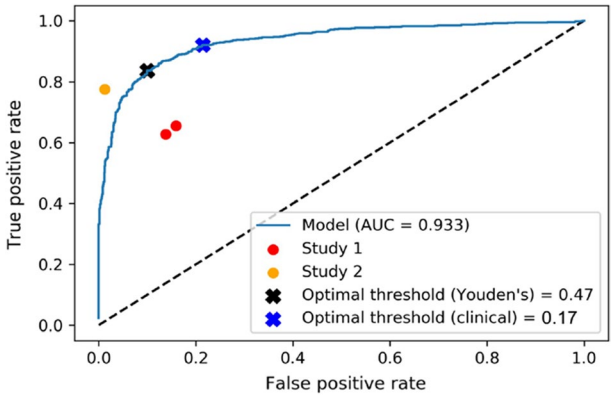
Fig. 9 Details of the architecture of MVGG



(a) Accuracy vs epoch **(b)** Loss vs epoch

Fig. 10 Training and validation of the best model

Fig. 11 ROC curve of the final model



epochs. Validation loss gets saturated after 15 epochs that why we considered these many numbers of epochs. It depends on how validation loss is behaving after each epoch. VGG 16 is modified in our application by fine-tuning the feed-forward, dense layers in the end to just one layer with 32 nodes, followed immediately by an output layer with sigmoid activation and one node (for binary classification). At the end of the model building process, we realize that the pre-trained modified VGG16 (MVGG 16) model outperforms all others in terms of accuracy. The architecture of the model is shown in Fig. 9. It produces an accuracy of 86.9% on the test set and an AUC of 0.933. This is better than our benchmark on both metrics. In Fig. 10, we can see that the model starts to strongly over fit after 6 epochs.

After testing three architectures, it is seen that Modified VGG16 outperforms both ResNet50 and MobileNet. MobileNet produces lower accuracy. The modified VGG16 model outperforms ResNet50 model. It can be assumed that this might be due to the features of the images and fixation of ResNet50's loss function on a higher local minimum.

It is also observed that the pre-training model provides better performance compared to the data augmentation. Because the initial weights might have enabled the model to find a better local minimum of the loss function during the gradient descent process. It can also be useful to run the model with extra resources and data augmentation for more epochs as the convergence is slow owing to the large size of the data.

4.2 ROC analysis

Figure 11 showed that the final modified proposed model has an AUC value of 93.3%. This is better than our benchmark of an AUC value of 88% from (Shen 2017). Additionally, our model also outperforms radiologists to classify mammograms as pathological or not. The benchmark of the first study on 12 radiologists on 312 cases with a sensitivity of 65.5% and a specificity of 84.1% (Rafferty et al. 2013) was surpassed by our model. For study 2, the physicians did use not only mammography but also other diagnostics, which could be a reason for better results.

After finalizing the algorithm, we estimate the mathematically optimal mode for the algorithm is, i.e. what the best threshold for the algorithm to declare a mammogram as either pathological or non-pathological. We compute the Youden's J statistic as follows:

$$J = maximum\, sensitivity(c) + specificity(c) - 1. \tag{1}$$

Youden's index is the probability of an informed decision (as opposed to a random guess) and considers all predictions. We have used it for setting optimal thresholds on medical breast cancer tests. In different words, this threshold minimizes the error rate of false positive and false negative, taking both as equally important. However, as written in chapter 3.4. from a clinical perspective, reducing false negatives is more important than false positives. Thus, we decided to weigh reducing false negatives twice as important as false positive and calculated the optimal threshold maximizing the cost function as 0.66 * *true positive rate* + 0.33 * (1-*false negative rate*). The clinically optimal threshold is 0.17, enabling us to further increase the false positive rate (thereby decreasing the false negative rate) by 10% while increasing the false positive rate by 15%.

Conclusively, this model with its well-performing accuracy as well as the estimated clinically relevant threshold would be well suited to sufficiently reduce errors, especially false negatives, in the clinical setting.

Several data mining algorithm is applied for cancer detection and classification (Shapiro et al. 1982; Bharati et al. 2019; Zhou et al. 2020; Celik et al. 2020; Benhammou et al. 2020; Kose and Alzubi 2020) using the dataset as a CSV file, but disease detection and classification using image dataset is a challenging task. To classify images into multiple categories such as benign, malignant (Hu et al. 2020; Bharati et al. 2018), and normal, our focus is to implement binary classification. This is because classifying a case as normal with higher confidence is more clinically relevant and immediately applicable than multinomial classification. The strength of the paper is the balance between breadth and depth in the scope. Testing various transfer learning models can enable us to recognize the best model for the task. On the other hand, there are some shortcomings with more time consuming and computing resources. So, we have tried to fine-tune our proposed models better, trying different hyper-parameters, and constructing our network.

Table 6 Comparative analysis with other existing works

Method	Accuracy (%)
Fusion of various deep CNN (Rakhlin et al. 2018)	87.20
Inception-Resnet-v2 (Kwok 2018)	79.00
Ensemble of LR, MV and GMV with refinement (Vang et al. 2018)	87.50
ALEXNET (Nawaz et al. 2018)	81.25
Quadratic SVM (Sarmiento and Fondón 2018)	79.20
Dense U-Net (Li et al. 2019)	78.38
CNN-GTD (Wang et al. 2019)	86.50
cGAN (Singh et al. 2020)	80
Proposed architecture	94.3

4.3 Comparison among the others work of accuracy

In the work of Rakhlin et al. (2018), the authors used the fusion methods of various deep CNN algorithms. They calculated sensitivity, AUC for two class and four class classification of breast cancer. Moreover, the authors of Kwok (2018) and Nawaz et al. (2018) proposed Inception-Resnet-v2 and ALEXNET, respectively for the same dataset. The using dataset of Rakhlin et al. (2018), Kwok (2018), Vang et al. (2018), Sarmiento and Fondón (2018), Nawaz et al. (2018) differs from our using dataset. Therefore, direct comparisons are not sufficient. Our adopted dataset is the same as the works of Li et al. (2019), Wang et al. (2019), Singh et al. (2020). The author of Li et al. (2019) proposed Dense U-Net algorithm that is not a traditional algorithm like DenseNet or U-Net. The obtained accuracy is 78.38%. Furthermore, the reference papers of Wang et al. (2019), Li et al. (2019) and Singh et al. (2020) also used our adopted DDSM dataset. They also proposed novel algorithms. The accuracy of these papers is less than our proposed method where the accuracy of 86.50% and 80% are obtained for CNN-GTD and cGAN, respectively.

Our proposed method provides higher accuracy than other methods presented in Table 5 for other transfer learning methods (Rakhlin et al. 2018; Kwok 2018; Vang et al. 2018; Sarmiento and Fondón 2018; Nawaz et al. 2018). It can also be explored from Table 6 that the architectures in Rakhlin et al. (2018), Kwok (2018), Vang et al. (2018), Nawaz et al. (2018), Sarmiento and Fondón (2018), Li et al. (2019), Wang et al. (2019) and Singh et al. (2020) provide an accuracy of 87.20%, 79.00%, 87.50%, 81.25%, 79.20%, 71%, 78.38%, 86.50%, and 80%, respectively, whereas our proposed model provides the accuracy of 94.3%.

5 Conclusion

The best performing architecture is an MVGG network which has been pre-trained on the ImageNet with an accuracy of 94.3% and the value of AUC is 93.3%. The clinical analysis shows that recall or sensitivity should be highlighted over specificity in the case of breast cancer. Thus, a clinical classification threshold is chosen, which is much lower than the mathematical threshold value. This algorithm will help to considerably decrease the false negative cases of mammograms. This will also increase the chances of 5-year survival.

There are different approaches we would like to follow in the future. Instead of binary classification, it will be interesting to make a categorical classification based on the BI-RADS scores. But in this situation, masses and calcification have to keep merged. We will also need additional, very detailed data set containing not only the BI-RADS scores but also other medical explanations.

Additionally, it will be interesting to integrate additional features into our algorithm. The tissue density of the woman plays a critical role in the breast cancer assessment. Obtaining this and adding it as a feature could potentially increase the accuracy of the algorithm significantly.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The data used for this research work is collected from DDSM (Digital Database for Screening Mammography) database and the authors do not violate the ethical statement.

References

Alzubi, J. A. (2015). Diversity based improved bagging algorithm. In *2015* (pp. 1–5).

Alzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., & Thaventhiran, C. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. *Applied Soft Computing, 80,* 579–591.

Aro, A. R., Absetz, S. P., van Elderen, T. M., van der Ploeg, E., & van der Kamp, L. J. T. (2000). False-positive findings in mammography screening induces short-term distress—Breast cancer-specific concern prevails longer. *European Journal of Cancer, 36*(9), 1089–1097.

Benhammou, Y., Achchab, B., Herrera, F., & Tabik, S. (2020). BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing, 375,* 9–24.

Bharati, S., Podder, P., & Mondal, M. R. H. (2020a). Artificial neural network based breast cancer screening: A comprehensive review. *International Journal of Computer Information Systems and Industrial Management Applications, 12,* 125–137.

Bharati, S., Podder, P., & Mondal, M. R. H. (2020b). Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*, *20,* 100391.

Bharati, S., Podder, P., & Paul, P. K. (2019). Lung cancer recognition and prediction according to random forest ensemble and RUSBoost algorithm using LIDC data. *International Journal of Hybrid Intelligent Systems, 15*(2), 91–100.

Bharati, S., Rahman, M. A., & Podder, P. (2018). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In *2018 4th International Conference on electrical engineering and information & communication technology (iCEEiCT), Dhaka, Bangladesh, 2018* (pp. 581–584). IEEE. https://doi.org/10.1109/ceeict.2018.8628084.

Cancer.gov. (2018). Cancer stat facts: Female breast cancer. Retrieved December 7, 2018, from, https://seer.cancer.gov/statfacts/html/breast.html.

Celik, Y., Talo, M., Yildirim, O., Karabatak, M., & Acharya, U. R. (2020). Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. *Pattern Recognition Letters*, *133*, 232–239.

Clinic, M. (2020). Retrieved December 7, 2020, from, https://www.mayoclinic.org/tests-procedures/3d-mammogram/about/pac-20438708.

DDSM. (2020). Retrieved December 7, 2020, from, http://www.eng.usf.edu/cvprg/Mammography/Database.html.

Ertosun, M. G., & Rubin, D. L. (2015) Probabilistic visual search for masses within mammography images using deep learning. In *2015* (pp. 1310–1315). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016* (pp. 770–778).

Heath, M., Bowyer, K., Kopans, D., KegelmeyerJr, P., Moore, R., Chang, K., et al. (1998). Current status of the digital database for screening mammography. *Digital Mammography, 13,* 457–460. https://doi.org/10.1007/978-94-011-5318-8_75.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

Hu, Q., Whitney, H. M., & Giger, M. L. (2020). A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. *Scientific Reports, 10*(1), 1–11.

Khamparia, A., Gupta, D., de Albuquerque, V. H. C., Sangaiah, A. K., & Jhaveri, R. H. (2020). Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. *The Journal of Supercomputing*, *76,* 1–19.

Kolb, T. M., Lichy, J., & Newhouse, J. H. (2002). Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: An analysis of 27,825 patient evaluations. *Radiology, 225*(1), 165–175.

Kose, U., & Alzubi, J. (2020). *Deep learning for cancer diagnosis*. Berlin: Springer.

Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N. H., & Verma, A. (2020). Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. In *Advances in data science and management* (pp. 435–442). Springer.

Kwok, S. (2018) Multiclass classification of breast cancer in whole-slide images. In *2018* (pp. 931–940). Springer.

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data, 4*(1), 170177. https://doi.org/10.1038/sdata.2017.177.

Li, S., Dong, M., Du, G., & Mu, X. (2019). Attention dense-u-net for automatic breast mass segmentation in digital mammogram. *IEEE Access, 7,* 59037–59047.

McGuire, A., Brown, J. A. L., Malone, C., McLaughlin, R., & Kerin, M. J. (2015). Effects of age on the detection and management of breast cancer. *Cancers, 7*(2), 908–929.

Mondal, M. R. H., Bharati, S., Podder, P., & Podder, P. (2020). Data analytics for novel coronavirus disease. *Informatics in Medicine Unlocked*, *20*, 100374.

Nawaz, W., Ahmed, S., Tahir, A., & Khan, H. A. (2018) Classification of breast cancer histology images using alexnet. In *2018* (pp. 869–876). Springer.

Qian, J., Tiwari, P., Gochhayat, S. P., & Pandey, H. M. (2020). A noble double-dictionary-based ECG compression technique for IoTH. *IEEE Internet of Things Journal, 7*(10), 10160–10170.

Rafferty, E. A., Park, J. M., Philpotts, L. E., Poplack, S. P., Sumkin, J. H., Halpern, E. F., et al. (2013). Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: Results of a multicenter, multireader trial. *Radiology, 266*(1), 104–113.

Rakhlin, A., Shvets, A., Iglovikov, V., & Kalinin, A. A. (2018). Deep convolutional neural networks for breast cancer histology image analysis. In *Paper presented at the international conference image analysis and recognition.*

Rani, S. S., Alzubi, J. A., Lakshmanaprabu, S. K., Gupta, D., & Manikandan, R. (2019). Optimal users based secure data transmission on the internet of healthcare things (IoHT) with lightweight block ciphers. *Multimedia Tools and Applications*, *79*, 1–20.

Reddy, A. V. N., Krishna, C. P., Mallick, P. K., Satapathy, S. K., Tiwari, P., Zymbler, M., et al. (2020). Analyzing MRI scans to detect glioblastoma tumor using hybrid deep belief networks. *Journal of Big Data, 7*(1), 1–17.

Sarmiento, A., & Fondón, I. (2018) Automatic breast cancer grading of histological images based on colour and texture descriptors. In *2018* (pp. 887–894). Springer.

Shapiro, S., Venet, W., Strax, P., Venet, L., & Roeser, R. (1982). Ten-to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute, 69*(2), 349–355.

Shen, L. (2017). End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv:1711.05775.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Singh, V. K., Rashwan, H. A., Romani, S., Akram, F., Pandey, N., Sarker, M. M. K., et al. (2020). Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Systems with Applications, 139,* 112855.

Tan, W., Tiwari, P., Pandey, H. M., Moreira, C., & Jaiswal, A. K. (2020). Multimodal medical image fusion algorithm in the era of big data. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-05173-2.

Thanh, D., & Surya, P. (2019). A review on CT and X-ray images denoising methods. *Informatica, 43*(2), 151–159.

Tiwari, P., & Melucci, M. (2018). Towards a quantum-inspired framework for binary classification. In *2018* (pp. 1815–1818).

Tiwari, P., & Melucci, M. (2019a). Binary classifier inspired by quantum theory. In *2019* (Vol. 33, pp. 10051–10052).

Tiwari, P., & Melucci, M. (2019b). Towards a quantum-inspired binary classifier. *IEEE Access, 7,* 42354–42372.

Tiwari, P., Qian, J., Li, Q., Wang, B., Gupta, D., Khanna, A., et al. (2018). Detection of subtype blood cells using deep learning. *Cognitive Systems Research, 52,* 1036–1044.

Tiwari, P., Uprety, S., Dehdashti, S., & Hossain, M. S. (2020). TermInformer: Unsupervised term mining and analysis in biomedical literature. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-05335-2.

Vahadane, A., Peng, T., Albarqouni, S., Baust, M., Steiger, K., Schlitter, A. M., et al. (2015) Structure-preserved color normalization for histological images. In *2015* (pp. 1012–1015). IEEE.

Vang, Y. S., Chen, Z., & Xie, X. (2018) Deep learning framework for multi-class breast cancer histology image classification. In *2018* (pp. 914–922). Springer.

Wang, Z., Li, M., Wang, H., Jiang, H., Yao, Y., Zhang, H., et al. (2019). Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. *IEEE Access, 7,* 105146–105158.

Zhou, L.-Q., Wu, X.-L., Huang, S.-Y., Wu, G.-G., Ye, H.-R., Wei, Q., et al. (2020). Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology, 294*(1), 19–28.