A MULTI-STAGE DEEP LEARNING TRAINING APPROACH FOR LESION DETECTION AND CLASSIFICATION IN MAMMOGRAMS

Siavash Salemi
AIIH Lab
ICT Research Institute
Faculty of Intelligent Systems
Engineering and Data Science
Persian Gulf University
7516913817 Bushehr, Iran
30yavash@gmail.com

Hamed Behzadi-Khormouji
Dep. of Computer Science
University of Antwerp, imec-IDLab
Belgium
Hamed.Behzadi
Khormouji@uantwerpen.be

Habib Rostami*
AIIH Lab
ICT Research Institute
Faculty of Intelligent Systems
Engineering and Data Science
Persian Gulf University
7516913817 Bushehr, Iran
habib@pgu.ac.ir (*Corr. Author)

Ahmad Keshavarz
IoT and Signal Processing
Research Group
ICT Research Institute
Faculty of Intelligent Systems
Engineering and Data Science
Persian Gulf University
7516913817 Bushehr, Iran
a.keshavarz@pgu.ac.ir

Yaser Keshavarz
Persian Gulf Nuclear Medicine
Research Center
Bushehr University of Medical Science,
Bushehr, Iran
ysrkvz@gmail.com

Yahya Tabesh
Dep. of Mathematical Sciences
Sharif University of Technology
Iran
tabesh@sharif.edu

Article history:
Received 09.08.2022, Accepted 25.09.2022

Abstract

Recently, Deep Convolutional Neural Networks (DCNNs) have opened their ways into various medical image processing practices such as Computer-Aided Diagnosis (CAD) systems. Despite significant developments in CAD systems based on deep models, designing an efficient model, as well as a training strategy to cope with the shortage of medical images have yet to be addressed. To address current challenges, this paper presents a model including a hybrid DCNN, which takes advantage of various feature maps of different deep models and an incremental training algorithm. Also, a weighting Test Time Augmentation strategy is presented. Besides, the proposed work develops the Mask-RCNN to not only detect mass and calcification in mammography images, but also to classify normal images. Moreover, this work aims to benefit from a radiology specialist to compare with the performance of the proposed method. Illustrating the region of interest to explain how the model makes decisions is the other aim of the study to cover existing challenges among the state-of-the-art research works. The wide range of conducted quantitative and qualitative experiments suggest that the proposed method can classify breast X-ray images of the INbreast dataset to normal, mass, and calcification classes with Accuracy 0.96, 0.98, and 0.97, respectively.

Key words

Deep Learning, Incremental Training, Mammography, Test Time Augmentation.

1 Introduction

Recently, machine learning has shown high efficiency in processing of sensory information, which allows machines to better handle complex data [Hosny et al., 2018]. Deep Learning is a branch of machine learning based on automatic feature extraction from data (e.g., image, video, voice, etc.) and processing them using a deep structure usually based on neural networks loosely inspired by the human brain. In the recent years, deep learning algorithms have been widely used in different image processing and machine learning tasks such as classification, identification and segmentation [LeCun et al., 2015]. Deep learning and especially Convolutional Neural Networks (CNNs) have opened their ways into medical image processing. A desirable result in using a deep learning algorithm not only depends on the architecture of the model, but the preprocessing and training algorithms also have important roles. For example, Kaggle Diabetic Retinopathy challenges different results obtained by different teams using the same architecture [Litjens et al., 2017].

The Deep Convolutional Neural Networks (DCNNs) can be used to build Computer-Aided Diagnosis (CAD) systems. Two types of Computer-Aided Diagnosis (CAD) systems are computer-aided detection (CADe) and computer-aided diagnosis (CADX) systems. In medical image processing systems, CADe systems serve as a second reader in assisting a radiologist to detect suspicious abnormalities, where subsequently patient management decisions are made by the radiologist [Esteva et al., 2019], [Geras et al., 2019], and [Krizhevsky et al., 2012]. Computer aided diagnosis (CADX) systems, however, assist to characterize abnormalities localized by a radiologist or a CADe system. A CADX system estimates the probabilistic class of the abnormality (e.g. benign or malignant) and the radiologist then decides about further evaluation.

Breast cancer can be detected through early diagnosis and screening strategies. Screening involves the systematic use of testing across an asymptomatic population to detect and treat cancer or pre-cancers. Machine Learning based CAD systems can help radiologists to detect and diagnose abnormalities such as masses, calcification and architecture distortion in time in mammograms; and provide an economical way to reduce the death rate among women with breast cancer [Tang et al., 2009]. Therefore, development of CAD systems for interpretation of mammograms has drawn the attention of both machine learning scientists and radiologists. The traditional approach to increase the diagnostics performance of mammography is double reading. Studies show that double reading of mammography increases the cancer detection rate (CDR) by 15% with no significant effect on the positive predictive value [Katzen and Dodelzon, 2018], [Thurfjell et al., 1994]. However, double reading is time-consuming, it is not a cost-efficient approach and it cannot be applied in many practices [Katzen and Dodelzon, 2018], [Posso et al., 2016]. In comparison with double reading, CAD systems reduce the workload of radiologists. However, they need more improvements to fulfill the requirements of routine clinical applications [Tang et al., 2009]. The diagnostic performance of AI based CAD systems, particularly those based on deep convolutional neural networks, can be increased through incorporating new data, whereby encourages the medical community to use them [Katzen and Dodelzon, 2018]. The aim of this research work is to propose an AI system, including a novel transfer learning, a hybrid convolutional neural network architecture, and a new evaluation method to classify and localize tumor lesions in breast X-ray images. The contribution of our work is as follows.

1. Proposing an incremental training algorithm, which instead of training a neural network monotonically, divides the training procedure to several superepochs in such a way that each superepoch includes several epochs.
2. Proposing a test time augmentation (TTA) evaluation strategy, in which a predicted label is considered as valid if there is an agreement between the predictions made on non-distorted and distorted version of the input images. The final decision is made in different than conventional TTA.
3. Designing a hybrid convolutional neural network which benefits of various feature maps of different convolutional neural network architectures. In this model, Mask RCNN is used in a unique way that not only classifies images into different lesions, but can also detect normal images.
4. We conduct comprehensive experiments on the proposed approach for lesion detection and classification in mammograms. These experiments, were conducted on two mammography datasets, namely CBIS-DDSM and INbreast.
5. The proposed method produces a localization on the region of interest as a visual explanation to show how the model makes decisions.

2 Related Work

In this section, first, we review the recent works that use CNN-based architectures for detecting breast cancer. Then, we discuss the weakness of the related works and explain how the proposed model addresses some of their weaknesses.

2.1 Breast cancer detection and classification based on deep models

In this subsection, the related works are overviewed and challenges of current works are investigated. Finally, we explain how the proposed approach addresses some of their weaknesses.

Xi [Xi et al., 2018] used AlexNet, VGG, GoogleNet and ResNet to classify mammography images and detect calcification and tumor in the images. They trained the network using patches of tumors and calcifications and then fed the whole image to the networks using transfer learning and a class activation map. They used a CBIS-DDSM dataset in the experiments. Al-Masni [Al-masni et al., 2018] developed a YOLO-based model to localize tumors and classify them into malignant and benign. They used 600 images from DDSM for the training of the network. Furthermore, they could detect masses existing over the pectorals or surrounded by the dense tissue considered as the most challenging cases in mammograms. Qiu [Qiu et al., 2017] proposed a network with five convolutional and two fully connected layers to detect tumors in mammography images. The network is pre-trained by ImageNet and trained by DDSM and MIAS datasets. They achieved better performance by substitution of SGD by parasitic metric learning [Jiao et al., 2018]. Singh [Singh et al., 2020] proposed a CNN model based on generative adversarial to segment breast cancer and create the binary masks (i.e. shapes of tumors). After that, a CNN is used to classify the shapes of tumors to four classes. Li [Li et al., 2020] proposed a Siamese-Faster-RCNN which detects masses in the bilateral mammography images. They aimed at solving the problem of separating single mass detection and bilateral comparison apart. For training the Siamese-Faster-RCNN, they just used the first four convolutional layers of VGG pre-trained by the ImageNet dataset and fine-tuned the remaining convolutional layers using the INbreast dataset. In another work, Runyu [Song et al., 2020] extracted multiple features from an improvement inception module, Gray-Level Co-occurrence Matrix (GLCM) and Histogram of Oriented Gradients (HOT) algorithms. Then, the features have been used to train a combined AI system including Support Vector Machine(SVM) and Extreme Gradient Boosting for classifying mammographic masses into three classes: normal, benign and cancer masses. The authors investigate the performance of their model in the cases of non-transfer learning and transfer learning from the ImageNet dataset. Transfer learning based on a dataset similar to the dataset under study is a novel and common approach which has been adopted in some recent works [Agarwal et al., 2020], [Behzadi-khormouji et al., 2020]. For example, Agarwal [Agarwal et al., 2020] used a transfer learning approach on a Faster RCNN network pre-trained on Hologic images to detect masses on small datasets such as INbreast. Finally, Ting Pang [Pang et al., 2020] conducted an extensive survey on published research works from 2015 until 2019 on detection and classification of breast cancer in mammography images using deep convolutional neural networks.

2.2 Challenges

According to the related works reviewed in the previous section, shortage of interpretation and explanation about proposed deep models and lack of collaboration between radiology experts and computer science specialists are among the common drawbacks of the current research works in this area. Also, they showed that the research works have focused on classifying and detecting mammography images into one or two lesions, including masses and calcification without classifying the normal images (i.e., the images which do not have any abnormalities and are belonged to the normal category). On the other hand, transferring the learned features from a CNN architecture pre-trained with the ImageNet dataset and fine-tuning it with the datasets under study is a common approach among them. The common transfer learning has had significant results in different applications. However, due to learning abstract features in the middle and upper layers of a CNN, these layers require more time to be fine-tuned with the dataset under study. This is different for lower layers which are responsible to learn low-level features such as edges that are common among different datasets. For example, in the case of medical images, the lesions are not appeared in the image with a clear and sharp edges. Therefore, the convolutional operation cannot detect medical concepts as well as those concepts appeared in the ImageNet dataset. Then, the upper convolutional layers require more time to learn the region of abnormalities [Behzadi-khormouji et al., 2020]. This problem cannot be overcome with the common transfer learning strategy where all or part of the layers are provided with the same opportunity to be fine-tuned. To cope with this problem, instead of training a neural network monotonically, we divide the training epochs to several superepochs in such a way that each superepoch subsumes several epochs. In each superepoch, the network is trained in some predefined number of epochs and at the end of the superepoch the weights of the network, which yield the best validation loss (or other metrics), are saved to be transferred to the next superepoch (it should be noted that the best weights may not be yielded from the last epoch of the superepoch). Furthermore, in contrast with current efforts, the proposed model not only detects mass and calcification lesions, but also modifies the MASK RCNN to classify the normal images. Moreover, the proposed model benefits from a hybrid architecture including different CNNs, which provide efficient proposals from images, thus increasing the performance of mammography classification in comparison to the related works. Besides, instead of using multiple CNNs independently to classify images based on voting among them, the present study proposes a Test Time Augmentation (TTA) approach which benefits only from one CNN architecture to make decisions based on voting from fed images from different distortions. More importantly, we had a radiologist evaluate the images and the performance of the proposed model. Finally, in order to explain the models, MASK RCNN localized the detected lesions from different distortions, which demonstrated the performance of the model.

3 Proposed Method

In this section, we propose a method to train a model to interpret mammogram images.

3.1 Training Algorithm

In this section, we present the proposed algorithm for training convolutional neural networks for mammogram interpretation. Then, we employ the algorithm to train the mammography neural network.

The general Idea is as follows. Instead of training a neural network monotonically, the training algorithm divides the training procedure to several superepochs in such a way that each superepoch includes several epochs. In each superepoch, the network is trained in several predefined number of epochs and at the end of the superepoch the weights of the network, which yield the best validation loss (or other metrics), are saved to be transferred to the next superepoch (it should be noted

```
Algorithm 1 $SuperTrain(model, dataset, NoSuperEpochs, NoEpochs, FreezRate[], LR[])$
  for $i = 1$ to $NoSuperEpochs$ , i += 1 do
    $BestWeights = Train(model, dataset, NoEpoch, LR[i])$;
    $model.load(BestWeights)$;
    $model.freezlayers(layers((i-1)*freezRate))$;
  end for
  return $Model$
```

```
Algorithm 2 $Train(model, dataset, NoEpochs, LR)$
  $model.compile()$
  $model.fit(dataset, NoEpoch, SGD(LR))$
  return $BestWeights$
```

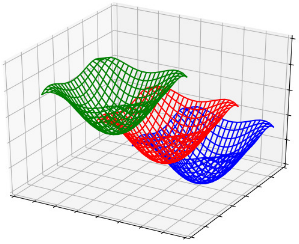


Figure 1. An overview of the fitness values of CNN in each superepoch. As can be seen, the CNN reaches a significant fitness value in the first super epoch (the green plot). As the transfers the weights of best classifier to the next superepochs, it is frequently fine-tuned with the dataset under study, which yields a classifier with a lower fitness value.

that the best weights may not be yielded from the last epoch of the superepoch).

At the beginning of the next superepoch, the best weights are transferred (loaded) into the network and some of the non-frozen bottom layers are frozen and the epochs of the superepoch run. The layers in different superepochs are frozen from the bottom up (i.e. layers before the classifier layers). Therefore, simple patterns are learnt in the first epochs and the corresponding layers are frozen. On the other side, the top layers corresponding to the sophisticated features receive more training epochs and have enough time to be trained.

In the early superepochs, larger learning rates are used to explore the weight space and in later superepochs, smaller learning rates are used to better exploit the vicinity of the best validation loss (or other metrics) found. Figure 1 illustrates an overview of the fitness values of CNN in each superepoch. As can be seen, the CNN reaches a significant fitness value in the first superepoch (the green plot). As the algorithm transfers the weights of the best classifier to the next superepoch, it is frequently fine-tuned with the dataset under study, which yields a classifier with a lower fitness value.

Algorithms 1 and 2 show the training method. Algorithm 1 is given a neural network model, the subject dataset and training hyper parameters as initial inputs. The hyper parameters include the number of superepochs (*”NoSuperEpochs”*), the number of epochs per superepoch (*”NoEpochs”*), the rate of layers that should be frozen in each superepoch (*”FreezRate”*) and an array of learning rates (*”LR[]”*). In superepoch *i*. *”LR[i]”* is used as the learning rate for the hyper parameter training procedure. In each superepoch, Algorithm 2 is called to be run as an ordinary training algorithm. The output of Algorithm 2 is the weights that yield the best validation loss. The incremental training approach is achieved by loading the returned weights to the model and freezing some layers and moving to the next superepoch.

3.2 Training the mammography model

According to the shortage of data mentioned in Section 2.1, we utilize advantage of two mammography datasets in our approach. First, as an intermediate transfer learning procedure, the model is pre-trained with the CBIS-DDSM dataset by a super train with superepochs $N = 8$, each of which includes 100 epochs in order to classify images as mass and calcification images. First the convolutional filters learn some patterns of mass and calcification such that whole network obtains a background knowledge about how mass and calcification lesions look like. The learning rates are 1e-3, 1e-4, 1e-4, 1e-4, 1e-4, 1e-5, 1e-5, 1e-6 in all of the eight superepochs, respectively. Then, the model is trained with INbreast with the same hyper-parameters in which network is provided with an additional opportunity to adjust its internal learned distributions with more and diverse images.

3.3 Test Time Augmentation

Deep neural networks are generally unstable against even small distortions on input images [Zheng et al., 2016]. Therefore, data augmentation is used to better train them. However, training with augmented data cannot solve all the problems, and marginally classified unseen images can be misclassified as a result of just a small noise. On the other side, the images that are classified with high confidence tolerate noises well without misclassification. The traditional approach to address this problem is using meta-classifiers, which employ more than one model to classify the input and then based on decisions made by all the models, the final decision is taken. The main disadvantage of this approach is that training some models imposes major time and resource requirements. On the other hand, Test-Time Augmentation (TTA), as an ensemble prediction solution for improving the model’s performance, creates multiple augmented copies of each image in the dataset. Then, the network, given several manipulated images of the same image, makes output probabilities for each class. In the next step, using the soft voting strategy (or similar strategy) the probabilities of each class are summed across

Table 1. Number of examples per class

Dataset/Classes	Mass			Calcification			Normal
	Train	Val	Test	Train	Val	Test	Test
CBIS_DDSM	2415	445	—	1199	212	—	—
INbreast	86	7	14	212	7	14	64

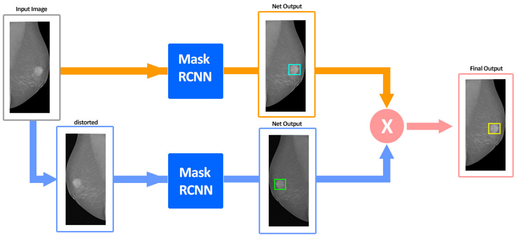


Figure 2. The proposed test time augmentation strategy. In the first round we feed the original image into the model and then we feed a distorted (e.g. flipped, rotated) version of it into the model. If the model classified both of them as mass, we will identify it as mass and if the model classified both of them as calcification, we will identify the image as calcification.

all the predictions and the class with maximum value is selected as the final predicted class. Our proposed TTA strategy, instead of using different models to classify the same image, inspires TTA strategy with some modifications such that not only does improve the network's performance, but also skills the network to classify normal images, in the case the networks, such as MASK-RCNN, cannot naturally detect normal images (i.e., images without any abnormality). The proposed strategy feeds just two versions of the same image (i.e., the original and manipulated ones) to one model and vote among the predicted output classes. In our problem, if a mass (or calcification) is classified by the model with high confidence, a distortion in the image should not change the decision of the model, otherwise the image is considered as the normal image. Therefore, we propose a scheme that includes double feeding of an image into the model (Figure 2). As can be seen in Figure 2, we feed the original image into the model in the first round, and then we feed a distorted (e.g. flipped, rotated) version of it into the model. If the model classifies both of them as mass, then we will identify it as mass; and if the model classified both of them as calcification, we will identify the image as calcification. Otherwise, we will consider the image as a normal one with no abnormality which shows the proposed strategy provides implicitly the network with the ability of classifying normal images. The experimental results (Section 4) show that the false positive rate reduces significantly. Also, as mentioned, Mask RCNN cannot naturally classify images with normal labels (i.e. images without any masks). However, in this research, Mask RCNN is used in a way that not only detects the mass and calcification images, but also classifies normal ones.

4 Results

4.1 Dataset

There are some public and private image datasets in the area of breast cancer [Xi et al., 2018]. In this paper, we used the CBIS-DDSM dataset to pre-train our models and the INbreast dataset to train and test them.

DDSM is one of the biggest datasets of mammography with 2620 cases including MLO and CC images from each breast (10180 images in total) with all types of findings [Moreira et al., 2012]. The dataset includes normal, benign and malign lesions. The dataset was published in 1997 and has some shortcomings. For instance, the region of interest annotations for abnormalities is not precise and only provides a general position of the lesions [Lee et al., 2017].

The CBIS-DDSM is an improved version of the DDSM with better RIO segmentation and data accessibility. The dataset includes 3061 mammography images from 1597 cases. It contains 1698 masses in 1592 images from 891 cases from CC and MLO views [Lee et al., 2017].

Another public dataset for mammograms is INbreast. It contains 410 full-field digital images from 115 cases including both CC and MLO views [Moreira et al., 2012]. The images of INbreast have been acquired between April 2008 and July 2010, are in the DICOM format. The dataset includes normal images with masses, calcifications, asymmetries, architectural distortions and images with multiple findings. The annotations of the dataset were made by a specialist and validated by another work [Moreira et al., 2012].

Of 410 INbreast mammograms, 67 images are normal without any masses or calcifications, 53 images include masses, 54 images include both masses and calcifications, 179 images include only calcifications and 57 images have other findings. In this paper, of 286 images which including masses or calcifications, 258 images were selected as the train data, 14 images were selected as the validation data, and 14 images were selected as test data. In addition, 30 normal images were added to the test data. It should be noted that, since MASK RCNN can be trained and validated just with segmented data, no normal images are used as train or validation data. It is noticeable that according to the three-part hold-out validation strategy, each dataset was randomly organized into three folders (train, validation, and test) and contained sub folders for each image category (i.e., normal, mass, and calcification). Table 1 summarizes the statistics related to each dataset used in this study. The Mass and Calcification images were used for training, validation and test procedures, but the normal images were just used in case of evaluating the performance of the models in test procedure.

4.2 Pre-processing

We used data augmentation to prepare the data which were to be fed into the models. Data augmentation is a technique to generate new samples from existing samples. Data augmentation may prove to be a solution [Krizhevsky et al., 2012], [He et al., 2020], [Heaton, 2017], [Roth et al., 2016], where training samples are not adequate and the trained model overfits.

The samples generated by data augmentation are the existing ones, but with different views. As a result, the model learns the patterns from different views. The operations of flipping, sharing, rotation and rescaling are applied in this research to augment the data. In order to overcome hardware limitations, the images were cropped from 2560×3328 pixels to 1024×1024.

4.3 Experiment setup

The model was implemented with the Python and Keras library [Gullì, 2017] on Tensorflow [Abadi et al., 2016] as its backend with CUDA 9 /cuDNN 7 (NVidia Corporation, Santa Clara, Calif) dependencies for graphics processing unit (GPU) acceleration. The model has been run on a computer with a Linux operating system (Ubuntu 18.04). The computer ran on an Intel Core(TM) i7-6850k CPU 3.60GHz processor, 32TB of hard disk space, 7889 MB of RAM, and a CUDA-enabled NVidia Titan 11 GB graphics processing unit (NVidia).

All statistical analyses were accomplished using statistic functions and sklearn packages of Python 3.5.2 and the Keras platform. To compare different experiments, ROC [Obuchowski, 2003] curves were computed on the same test dataset. Moreover, several experiments were designed to investigate the performance of the proposed method.

4.4 Efficiency of data augmentation

In order to evaluate the accuracy of the model, first all of the test images are augmented by flipping left and flipping right operations. Second, all of the applied augmentation related to the train data are applied on the test images. Finally, the augmented test image is fed to the model. The lesion recognized by the model is compared with the lesion which has been acquired by feeding the model with the real image. If both images show the same lesion, then it is considered as a correct estimation of the model.

4.5 Efficiency of different learning strategies

In order to investigate the efficiency of the proposed training algorithm, a variety of relevant training strategies has been examined. The details of each strategy is as follow.

1. **The First Strategy:** One-stage training using a constant learning rate. In this stage, the whole model was trained with the learning rate=1e-5 and epoch=1000. Also, DDSM and INbreast were used for pre-training and fine-tuning the model initialized with random weights, respectively.
2. **The Second Strategy:** One-stage training using a decreasing learning rate. In this scenario, all of the mask RCNN's weights were trained. The initial value of the weights and learning rate were set randomly and to 1e-3, respectively. To optimize the loss function efficiency, the learning rate was reduced using the coefficient 0.1 in each 200 epochs until it reached the value 1e-7. These settings were used for both pre-training and fine-tuning the model with the DDSM and INbreast datasets, respectively.
3. **The Third Strategy:** Multi-stage training from scratch. In this strategy, the weights of the network have been initialized randomly. Then, the network has been pre-trained using the DDSM dataset. Finally, it was trained by the proposed incremental training algorithm using the INbreast dataset. Each phase of training included 8 superepochs and 100 epochs. The initial value of the learning rate was set to 1e-3. To optimize loss function efficiency, the learning rates used for the superepoch 1 to 8 were $1e^{-3}$, $1e^{-4}$, $1e^{-4}$, $1e^{-4}$, $1e^{-4}$, $1e^{-5}$,$1e^{-5}$, and $1e^{-6}$, respectively. These values have been determined as a result of massive experiments.
4. **The Fourth Strategy:** Multi-stage training using a decreasing learning rate. In this strategy, the DDSM and INbreast datasets were used respectively for pre-training and fine-tuning a model pre-trained with the ImageNet dataset. In each phase of the training, the proposed incremental training algorithm was applied to the model. Each phase of training included 8 superepochs and 100 epochs. The initial value of the learning rate was set to 1e-3. To optimize loss the function efficiency, the learning rates used for the superepoch 1 to 8 were $1e^{-3}$, $1e^{-4}$, $1e^{-4}$, $1e^{-4}$, $1e^{-4}$, $1e^{-5}$, $1e^{-5}$, and $1e^{-6}$, respectively. These values have been determined as a result of massive experiments.

Table 2 shows the experimental results of the four strategies. The second, and third columns indicate the minimum loss function values of the train dataset, and the minimum loss function of the validation data of each strategy on the test data, respectively. As can be seen, the *Min Train Loss*, and *Min Validation Loss*, from low to high belong to Strategies 4 to 1, respectively.

The Figures 3 and 4 illustrate the train and validation loss function related to each strategy. Figures 3.A and 3.B demonstrate the validation and train loss function of the *"First Strategy"*, respectively. Figures 3.C and 4.D show the validation and train loss function of the *"Second Strategy"*, respectively. According to Figure 3.A, the classifier related to the epoch 893 in the *"First Strategy"* has the lowest validation loss of 0.829, while the figure related to *"Second Strategy"* shows that the lowest validation loss belongs to the classifier trained in the

Table 2. The minimum train loss, and minimum validation loss of each strategy

Strategy	Train Loss	Validation Loss
1	0.913	0.829
2	0.239	0.470
3	0.913	0.829
4	0.281	0.157

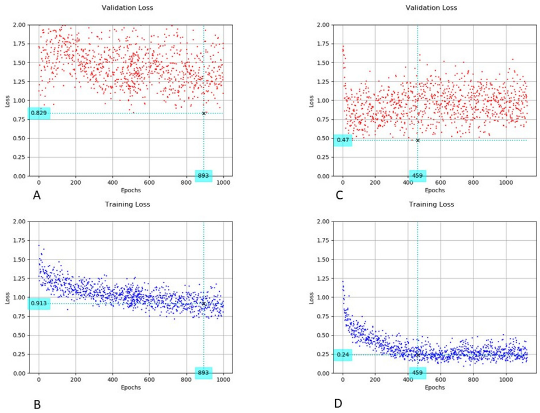


Figure 3. The validation and training loss functions of ”First and Second Strategies”. The Figures A and B illustrate the validation and training loss functions of the ”First Strategy”, while those of ”Second strategy” have been shown in Figures C and D. The ”First Strategy” shows that in the epoch 893, the lowest validation loss was 0.829, and the lowest validation loss in the ”Second Strategy” belonged to the epoch 459.

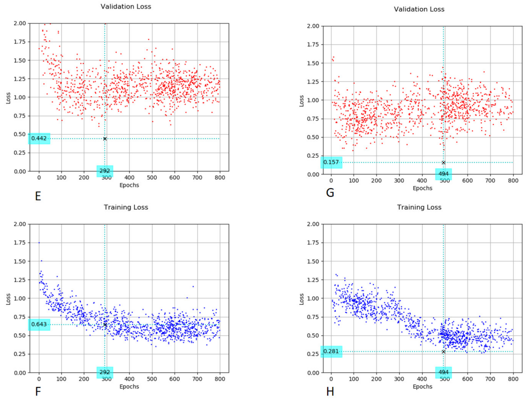
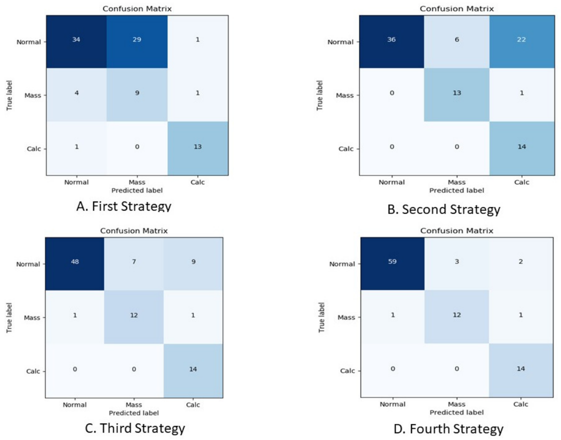


Figure 4. The validation and training loss functions of ”Third and Fourth Strategies”. The Figures E and F illustrate the validation and training loss functions of the ”Third Strategy”, while those of ”Forth Strategy” have been shown in Figures G and H. The ”Third Strategy” shows that in the epoch 292, the lowest validation loss was 0.442, and the lowest of validation loss in the ”Fourth Strategy” belonged to the epoch 494.

epoch 459 with a loss value of 0.470. According to Figure 4, the fourth strategy trained a classifier with the lowest validation loss function of 0.157 in epoch 494. Due to the better performance of the fourth strategy, it was considered as the training policy for the rest of experimental results.



A. First Strategy

B. Second Strategy

C. Third Strategy

D. Fourth Strategy

Figure 5. Figures A, B, C, and D illustrate the confusion matrix of the first, second, third and fourth strategies.

Figure 5 shows the confusion matrices of each training strategy. As can be seen, the *”Fourth Strategy”* correctly classified 59 images as the normal class, whereas the *”Third, Second and First strategies”* could correctly classify just 48, 36 and 34 images as the normal class, respectively. Also, the *”Second, Third, and Fourth Strategies”* had the same True Positive 14 for the class Calcification, which was higher than that of the *”First Strategy”*.

Table 3 demonstrates the Precision, Recall, F_1 and Accuracy factors of each strategy. These measures were calculated according to the test data which had not been used by model. Also, these factors were computed for each strategy using the sklearn package of Python 3.5.2; moreover, the precision, recall, F_1 and Accuracy factors were computed according [Zhu et al., 2010].

According to Table 3, *” Fourth, Third, Second and First Strategies”* had precision degrees of 0.92, 0.75, 0.56, and 0.53 for the normal class respectively, while the precision degrees of the Mass class from high to low belonged to the *”Second, Third, Fourth, and First Strategies”* with 0.92, 0.85, 0.85, and 0.64, respectively. Also, in the ”First Strategy”, the Calcification class had a precision degree of 0.92, while the *”Second, Third, and Fourth Strategies”* had a precision degree of 1. Moreover, the Recall related to the normal class from high to low belonged to the *”Second, Fourth, Third, and First Strategies”*, whereas that of the Mass class from high to low belonged to the *”Fourth, Second, Third and First Strategies”*, respectively. Also, the Recall related to class Calcification from high to low belonged to the ”First, Fourth, Third, and Second Strategies”.

Table 3. The Precision, Recall, F_1 , and Accuracy factors of each strategy

Strategy	Class	Precision	Recall	F_1	Accuracy
1	Normal	0.53	0.87	0.66	0.61
	Mass	0.64	0.23	0.34	0.63
	Calcification	0.92	0.86	0.89	0.96
2	Normal	0.56	1	0.72	0.69
	Mass	0.92	0.68	0.78	0.92
	Calcification	1	0.37	0.54	0.75
3	Normal	0.75	0.97	0.84	0.81
	Mass	0.85	0.63	0.72	0.90
	Calcification	1	0.58	0.73	0.89
4	Normal	0.92	0.98	0.95	0.93
	Mass	0.85	0.80	0.82	0.94
	Calcification	1	0.82	0.90	0.96

4.6 Effect of the test time augmentation method

Figure 6.A illustrates the confusion matrix of the model without the proposed TTA approach, and the results after applying the proposed approach are presented in Figure 6.C. In addition, Figures 6.B and 6.D show the receiver operating characteristics of the model without and with using the proposed test time augmentation strategy respectively. As can be seen, the area under the curve of the normal images increases from 0.84 to 0.943 by applying the proposed approach and at the same time, the area under the curve of the mass and calcification images show only a slight improvement.

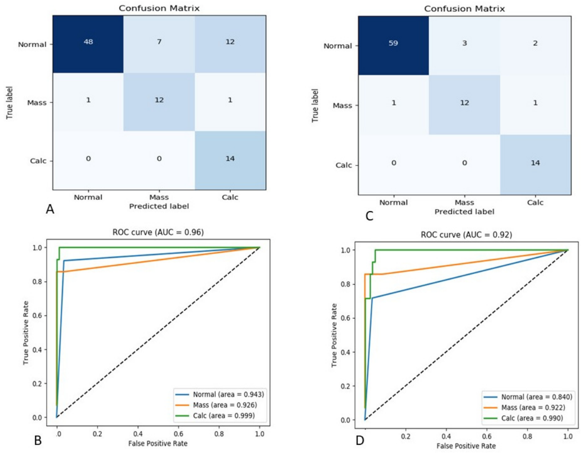


Figure 6. Effect of the test time augmentation method. Figures A and C show the confusion matrixes of the model without/with the proposed approach. Figures B and D show the receiver operating characteristics of the model without and with using the proposed approach, respectively. As can be seen, the area under the curve of the normal images increases from 0.84 to 0.943 by applying the proposed approach and at the same time, the area under the curve of the mass and calcification images show only a slight improvement.

4.7 A New hybrid MASK RCNN-CNN model

According to the previous experiments, the TTA approach improved the performance of the MASK RCNN model. In order to improve the proposed architecture, this section introduces a new hybrid architecture which improves the performance of the model. First, a set of six well-known pre-trained CNNs with the ImageNet dataset such as MobileNet, ResNet50, ResNetV2, VGG16, VGG19, and Xception were trained on abnormality and normal patches introduced in the mass and calcification images of the INbreast dataset. Then, two of the best models were selected to be combined with the TTA MASK RCNN to increase the performance of classification. According to Figure 7, Mask RCNN outputs the predicted label and an image including the region of interest. Then, the image is fed to a CNN model (in this case, the Xception model) to make another decision. In the end, the final decision will be taken by voting based on the three results.

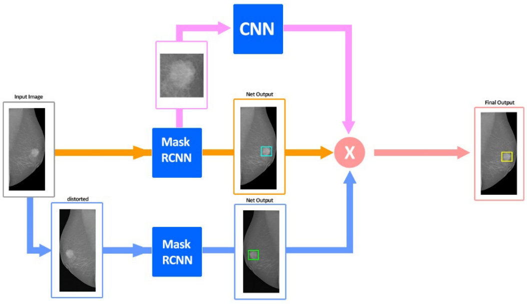


Figure 7. The hybrid MASK RCNN model. Mask RCNN outputs the predicted label and an image illustrating the region of interest. Then, the image is fed to a CNN model (in this case, the Exception model) to make another decision. In the end, the final decision will be taken by voting based on the three results..

Table 4 shows the Precision, Recall, F_1, and Accuracy measures of each CNN. According to the experimental results, the highest to the lowest *Accuracy* of the normal class belonged to ResNet50, Xception, ResNetV2, VGG16, VGG19, and MobileNet with the values of 0.94, 0.90, 0.80, 0.80, 0.41, and 0.38, respectively. Also, the same trend can be seen in the experimental results of the Mass class with the F_1 measures of 0.95, 0.93, 0.82, 0.80, 0.40, and 0.35, respectively, whereas these values for the Calcification class from the highest to the lowest belonged to VGG16, ResNet50, MobileNet, VGG19, ResNetV2, and Exception with the F_1 scores of 1, 0.98, 0.97, 0.96, 0.92, and 0.90, respectively. Finally, as ResNet50 and Exception had better performance with most of the factors in comparison to other models, they

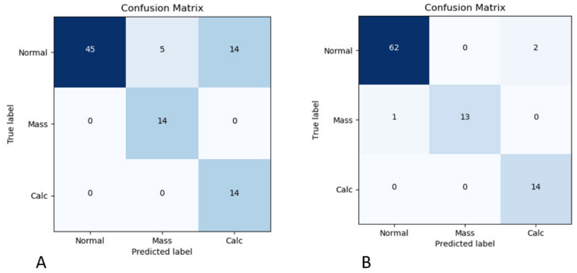


Figure 8. The confusion matrix of the Interpretation mammography image by a radiologist (A), and the best trained CNN model, test time augmentation MASK RCNN + Xception, (B).

Table 4. The Precision, Recall, F_1, and Accuracy factors of each CNN

CNN Model	Class	Precision	Recall	F_1	Accuracy
MobileNet	Normal	0.10	1	0.19	0.38
	Mass	1	0.19	0.32	0.35
	Calcification	0.85	1	0.92	0.97
VGG 16	Normal	0.71	1	0.83	0.8
	Mass	1	0.77	0.60	0.80
	Calcification	1	1	1	1
VGG 19	Normal	0.15	1	0.27	0.41
	Mass	1	0.20	0.33	0.40
	Calcification	0.85	0.92	0.88	0.96
ResNet50	Normal	0.92	1	0.95	0.94
	Mass	1	0.77	0.87	0.95
	Calcification	1	0.93	0.96	0.98
ResNetV2	0.75	0.96	0.84	0.80	
	Mass	1	0.50	0.66	0.82
	Calcification	0.85	0.85	0.85	0.95
Xception	0.89	0.96	0.92	0.90	
	Mass	1	0.70	0.82	0.93
	Calcification	0.85	0.92	0.88	0.93

Table 5. The Precision, Recall, F_1, and Accuracy factors of each hybrid structure

CNN Model	Class	Precision	Recall	F_1	Accuracy
TTA MASK RCNN + Xception	Normal	0.96	0.98	0.97	0.96
	Mass	0.92	1	0.96	0.98
	Calcification	1	0.87	0.93	0.97
TTA MASK RCNN + ResNet50	Normal	0.90	0.98	0.94	0.92
	Mass	0.93	0.93	0.96	0.97
	Calcification	1	0.77	0.87	0.95

are selected as the potential models for the hybrid architecture.

Table 5 shows the experimental results of two hybrid architectures. According to the results, the hybrid architecture including Xception model had Accuracy of 0.96, 0.98, and 0.97 for the Normal, Mass, and Calcification classes, respectively which were greater than those of the hybrid architecture including ResNet50.

4.8 Interpretation of the mammography images by a radiologist

In order to compare the accuracy of the proposed model with a radiologist, a radiologist with 12 years of experience was requested to classify the test images to three categories: Mass, Calcification and Normal. We provided him with 92 images. In 14 cases, he requested more images from different angels in order to make better diagnoses. In 19 cases he failed to recognize the correct label (Figure 8.A), while the best proposed model illustrated in Figure 8.B (TTA MASK RCNN + Xception) had lower failed detections.

5 Discussion

Machine learning is a branch of Artificial Intelligence (AI) which aims to provide computer systems with the ability to learn patterns and analyze the relationships between the data to perform tasks that normally require human intelligence [Kohli et al., 2017]. These tasks can be performed in two ways: supervised and unsupervised. In unsupervised tasks, the machine learning algorithms make decisions based on the similarity criteria among various categories, whereas the label assigned to each data in supervised problems is the main factor for making decisions by a machine learning algorithm [Choy et al., 2018]. In this study a hybrid convolutional neural network was designed in a supervised manner to classify mammography images as normal, calcification or mass.

5.1 The Performance of Proposed Model

One of the most significant advantages of the convolutional neural network is its great ability to provide various representations of high dimensional data such as images. MASK RCNN is an extension of the RCNN families, which generates the regions of interest using an alignment layer to classify and localize the objects inside the images. Despite significant achievements in proposing various pre-training strategies as well as CNN architectures, there are not enough images from various real-world applications such as medicine. This raises challenges how we can create efficient CNN architectures and train strategies for training such networks to learn appropriate patterns. In this study, we proposed a model which utilizes an incremental training strategy and a hybrid convolutional neural network to address such challenges.

From a training point of view, the majority of the state-of-the-art CNN models have been pre-trained by the ImageNet dataset. Since training such networks, which include millions of parameters, is extremely expensive, the transfer learning strategy is a common approach to address this problem. Because of the similarity between the features in lower views such as edges and points in different data, transferring the lower features to the new task can help to increase the learning speed and more importantly address shortage of data required for training the CNN model. However, transferring features from a network pre-trained with a dataset which is completely different from the one under study needs to be adjusted to the new task, especially in medical applications, where contexts of images are completely different from the ImageNet dataset. In this stage, we can address the mentioned issue to some extent by fine-tuning the model, especially the upper layers using the dataset under study [Becherer et al., 2019]. Shu [Shu, 2019] explains that fine-tuning a new classifier with a small dataset drastically increases the risk of overfitting and leads to poor generalization. Also, in a fine-tuning strategy, all the layers are updated at same time per fed data. In other words, all lower, intermediate, and upper layers have the same chance to be tuned with the target data. However, as intermediate and upper layers create an abstract of target data, they need more time to be tuned with the target data. As a result, we proposed a novel transfer-learning in which the number of updates skews to higher and more complex features (upper layers). According to the experimental result of Table 3, the ”Fourth Strategy” had better performance in comparison to the other learning strategy. This illustrates two important findings. First, providing the upper layers with more time for learning than the lower layers can lead a model to have higher performance than common transfer learning. Second, based on the experimental results of the ”Third and Second Strategies”, in cases where the model has millions of parameters and the dataset under study is incredibly small, applying the incremental training approach on a pre-trained network can deliver higher performance than applying it on a non-pre-trained network.

On the other hand, the Mask RCNN naturally cannot classify data as a class without any specific object (i.e. the normal class). In other words, this network is used in tasks where each class of dataset contains masks to determine the regions of interest which should be trained by the network. However, in this study, Mask RCNN was applied in a way that not only did detect different lesions, but also classified images without any lesions as the normal class. Moreover, in accordance with the experimental results of Figure 6, a new evaluation method based on distortion was proposed to increase the performance of the network in classifying normal images.

Recently, a variety of convolutional neural networks have been proposed with different architectures of convolutional layers to propose different presentations. In order to benefit from different representations for increasing the performance of mammography image classification, six well-known convolutional neural networks have been trained using the DDSM and INbreast datasets. Two of the models with higher performance were chosen to create a hybrid architecture. According to the experimental results in Table 5, the hybrid architecture with Mask RCNN-Xception had higher performance than the hybrid architecture with Mask RCNN-ResNet50. This showed that the former architecture (i.e., Mask RCNN-Xception) had provided better representations of images in comparison to other investigated architecture.

5.2 Visualization of the model’s perception

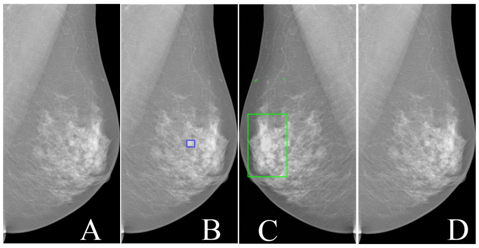


Figure 9. Image A is a Normal image. Images B and C (augmented Image B) show the different regions which were detected by the model. As a result Image D does not show any region and the image was correctly classified as a Normal image.

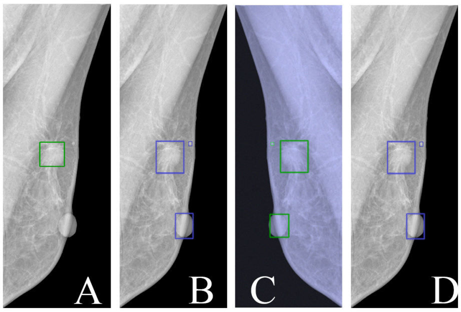


Figure 10. Image A with an area illustrating a mass lesion in the breast X-ray image. Image B shows that the model detected not only the correct region of interest, but also the wrong regions. Also, the model detected the same regions on the augmented image. As a result, Image D illustrates the mass area as well as other regions. Finally, this images has been classified as the Mass class.

Recently, deep convolutional neural networks have numerically shown that they can have promising results in a wide variety of tasks. However, in some critical fields such as medicine, reliance of the deep convolutional neural networks on appropriate features should be visually interpreted or explained. One of the significant advantages of Mask RCNN is identifying the region of image, based on which the model has made a decision. In this subsection, the visualization results of the proposed algorithm are explained. In Figure 9, Images B and C (augmented Image B) show the different regions which were detected by model. As a result, Image D does not show any region and the image was correctly classified as a normal image. Figure 9 shows that the model correctly classified Image B as the normal class, while it classified the augmented Image B (Image C) as the tumor class and detected a wrong area as a tumor lesion. As a result, Image D was classified as the normal class.

Figure 10 includes an Image A with an area illustrating a mass lesion in the breast X-ray image. Image B shows that the model detected not only the correct region of interest, but also the wrong regions. In addition, the model detected the same regions on the augmented Image B (i.e. image C). As a result, Image D illustrates the mass area as well as other regions. Finally, this image was classified as the Mass class. Figure 10 demonstrates a tumor image. In this experiment, the model detects the correct region of interest in both non-augmented and augmented images. As a result, image D shows the final detected region as a tumor lesion.

6 Conclusion

Deep Convolutional Neural Networks (DCNNs) have opened their ways into various medical image processing practices such as Computer-Aided Diagnosis (CAD) systems, especially in detecting and classifying abnormalities in mammograms such as masses, and calcification. Despite significant achievements of DCNNs in detecting lesions, especially in detecting and classifying breast lesions tasks, there are still some open challenges in the related works. For example, First, there is a shortage of decision making explanations of the proposed deep models. Also, lack of collaboration between radiology experts and computer science specialists is another weakness in the works. Moreover, lack of CNN models to classifying multiple breast lesions mass, calcification, and normal is common drawback in the recent works. Finally, a common transfer learning strategy to take advantage of a pre-trained CNN for the tasks such as detection and classification of breast images is other shortage in the related works. To cope with the mentioned issues efficiently, we used an incremental training strategy, which instead of training a neural network monotonically, divides the training epochs into several superepochs in such a way that each superepoch subsumes several epochs. MASK RCNN is the base CNN model used in this study. We modified MASK RCNN which not only detects mass and calcification lesions but also classifies the normal images. Moreover, the proposed model benefits from a hybrid architecture including different CNNs, which provide efficient proposals from images, thus increasing the performance of mammography classification in comparison to the related works. Besides, instead of using multiple CNNs independently to classify images based on voting among them, the present study proposed an incremental training approach that benefits only from one CNN architecture to make decisions based on voting from fed images from different distortions. In addition, we provided a set of visualizations of the model's outputs to explain how and based on what image features the proposed CNN makes its decisions. Finally and importantly, we had a radiologist who evaluated the images and the performance of the proposed model. Although, we designed a CNN model to cope efficiently with one of the main challenges in the related works (i.e., lack of clinical images), diversity of images was one of our limitations in the current study. Therefore, applying a across clinical adaption to increase the performance of the model on the wide range of sources of the clinical images from different hospitals is one of the follow-up work in this project.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Agarwal, R., Díaz, O., Yap, M. H., Lladó, X., and Martí, R. (2020). Deep learning for mass detection in full field digital mammograms. *Computers in Biology and Medicine*, **121**, pp. 103774.

Al-masni, M. A., Al-antari, M. A., Park, J.-M., Gi, G., Kim, T.-Y., Rivera, P., Valarezo, E., Choi, M.-T., Han, S.-M., and Kim, T.-S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. *Computer Methods and Programs in Biomedicine*, **157**, pp. 85–94.

Becherer, N., Pecarina, J. M., Nykl, S., and Hopkinson, K. M. (2019). Improving optimization of convolutional neural networks through parameter fine-tuning. *Neural Comput. Appl.*, **31** (8), pp. 3469–3479.

Behzadi-khormouji, H., Rostami, H., Salehi, S., Derakhshande-Rishehri, T., Masoumi, M., Salemi, S., Keshavarz, A., Gholamrezanezhad, A., Assadi, M., and Batouli, A. (2020). Deep learning, reusable and problem-based architectures for detection of consolidation on chest x-ray images. *Computer Methods and Programs in Biomedicine*, **185**, pp. 105162.

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A., Pianykh, O., Geis, J., Pandharipande, P., Brink, J., and Dreyer, K. (2018). Current applications and future impact of machine learning in radiology. *Radiology*, **288** (2), pp. 318–328. Publisher Copyright: © RSNA, 2018.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M. A., Chou, K., Cui, C., Corrado, G. S., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, **25**, pp. 24–29.

Geras, K. J., Mann, R. M., and Moy, L. (2019). Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology*, p. 182627.

Gullì, A. (2017). Deep learning with keras : implement neural networks with keras on theano and tensorflow.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2020). Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, pp. 386–397.

Heaton, J. (2017). Ian goodfellow, yoshua bengio, and aaron courville: Deep learning. *Genetic Programming and Evolvable Machines*, **19**, pp. 305–307.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, **18**, pp. 500–510.

Jiao, Z., Gao, X., Wang, Y., and Li, J. (2018). A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognit.*, **75**, pp. 292–301.

Katzen, J. T. and Dodelzon, K. (2018). A review of computer aided detection in mammography. *Clinical imaging*, **52**, pp. 305–309.

Kohli, M. D., Prevedello, L. M., Filice, R. W., and Geis, J. R. (2017). Implementing machine learning in radiology practice and research. *AJR. American journal of roentgenology*, **208 4**, pp. 754–760.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, **60**, pp. 84 – 90.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**, pp. 436–444.

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., and Rubin, D. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, **4**.

Li, Y., Zhang, L., Chen, H., and Cheng, L.-J. (2020). Mass detection in mammograms by bilateral analysis using convolution neural network. *Computer methods and programs in biomedicine*, **195**, pp. 105518.

Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, **42**, pp. 60–88.

Moreira, I., Amaral, I., Domingues, I., Cardoso, A. J. O., Cardoso, M. J., and Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic radiology*, **19 2**, pp. 236–48.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology*, **229 1**, pp. 3–8.

Pang, T., Wong, J. H. D., Ng, W. L., and Chan, C. S. (2020). Deep learning radiomics in breast cancer with different modalities: Overview and future. *Expert Syst. Appl.*, **158**, pp. 113501.

Posso, M., Carles, M., Rué, M., Puig, T., and Bonfill, X. (2016). Cost-effectiveness of double reading versus single reading of mammograms in a breast cancer screening programme. *PLoS ONE*, **11**.

Qiu, Y., Yan, S., Gundreddy, R. R., Wang, Y., Cheng, S., Liu, H., and Zheng, B. (2017). A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. *Journal of X-ray science and technology*, **25 5**, pp. 751–763.

Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K. M., Kim, L., and Summers, R. M. (2016). Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Transactions on Medical Imaging*, **35**, pp. 1170–1181.

Shu, M. (2019). Deep learning for image classification on very small datasets using transfer learning. Master's thesis, Iowa State University.

Singh, V. K., Rashwan, H. A., Romaní, S., Akram, F., Pandey, N., Sarker, M. M. K., Saleh, A., Arenas, M., Arquez, M., Puig, D., and Torrents-Barrena, J. (2020). Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Syst. Appl.*, **139**.

Song, R., Li, T., and Wang, Y. (2020). Mammographic classification based on xgboost and dcnn with multi features. *IEEE Access*, **8**, pp. 75011–75021.

Tang, J., Rangayyan, R. M., Xu, J., El-Naqa, I., and Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *IEEE Transactions on Information Technology in Biomedicine*, **13**, pp. 236–251.

Thurfjell, E., Lernevall, K. A., and Taube, A. (1994). Benefit of independent double reading in a population-based mammography screening program. *Radiology*, **191 1**, pp. 241–4.

Xi, P., Shu, C., and Goubran, R. A. (2018). Abnormality detection in mammography using deep convolutional neural networks. *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6.

Zheng, S., Song, Y., Leung, T., and Goodfellow, I. J. (2016). Improving the robustness of deep neural networks via stability training. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4480–4488.

Zhu, W., Zeng, N. F., and Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. In *NESUG proceedings: health care and life sciences*.