Fine-grained mammography object detection with multiple feature fusion transfer learning

Ri-Gui Zhou[a,b], Shihao Lv[a,b,*], Ding Yuan[c], ShengJun Xiong[c]

[a] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[b] *Research Center of Intelligent Information Processing and Quantum Intelligent Computing, Shanghai 201306, China*
[c] *HT-NOVA Co. Ltd, Building 39, Yard 12, Tianzhu Comprehensive Bonded Zone, Shunyi District, Beijing, 101318, China*

Abstract

As we all know, breast cancer is one of the important causes that endanger women's health. And screening mammography has become a technology that can effectively control breast cancer. It is necessary to build mammography image recognition systems. Previous solutions for mammography image recognition are usually based on hand-crafted features methods and use but limited in specific situations. Also, some methods utilize simple deep learning network, which can not achieve a good recognition effect. In this paper, we propose a deep learning based approach with multiple feature fusion transfer learning strategy. Firstly, we obtain the training data from an open data set called DDSM images. Then we employ data augment methods, and train a novel deep neural network to detect lesions and conduct the object detection job. Pre-trained model is used to initialize the network and help extract the basic features. Furthermore, we propose a fusion method that makes use of multiple transfer learning models in inference. Importantly, we take a strategy applied by hash learning in the deep network is cited to enhance the generalization ability of the model and solve the challenge of high-dimensional calculation in deep learning. In the end, regression analysis to analyze the object position. The experimental results prove that our method achieves high accuracy on the mammography image object detection and inspection task.

Keywords: CNNs, Random Forest, Learning to Hash, DensNet, Feature Fusion

1. Introduction

With the rapid development of deep learning, the medical influence community has also begun to try to apply this technology to improve the accuracy of cancer screening. In the United States, breast cancer has become the second highest death rate among all cancers. And screening mammography has become an effective way to reduce mortality. According to a study by the Breast Cancer Surveillance Consortium in 2009, the overall sensitivity of digital screening mammography in the U.S. is 84.4% and the overall specificity is 90.8% Oeffinger et al. (2015). Since the 1990s,

*Ri-Gui Zhou, Shihao Lv, are with the College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China and the Research Center of Intelligent Information Processing and Quantum Intelligent Computing, Shanghai 201306, China.

Email addresses: rgzhou@shmtu.edu.cn (Ri-Gui Zhou), lvshihao@stu.shmtu.edu.cn (Shihao Lv)

the Computer-Assisted Detection and Diagnosis (CAD) software has been applied in clinical diagnosis, which can effectively help radiologists improve the efficiency of cancer detection. However, according to the actual situation of the application, this system is not as envisaged, the effect is not obvious, and there is no substantial progress in the upgrade and improvement of the system. In recent years, deep learning has achieved better results in visual object recognition and detection, attracting more attention Lecun et al. (2015). There is much more interest in deep learning to assist radiologists and improve the accuracy of screening mammography Jamieson et al. (2012); Zhu et al. (2019); Shen (2017).

Early detection of subclinical breast cancer on screening mammography is challenging as an image detection task because the tumors themselves occupy only a small portion of the image of the entire breast. For example, a full-field digital mammography (FFDM) image is typically 4000 $\times$ 3000 pixels while a cancerous region of interest (ROI) can be as small as 50 $\times$ 50 pixels. If ROI annotations were widely available in mammography databases then established object detection and classification methods such as the region-based convolutional neural network (R-CNN) and its variants could be readily applied Girshick et al. (2014); Girshick (2015); Ren et al. (2017). However, approaches that require ROI annotations often cannot be transferred to large mammography databases that lack ROI annotations, which are laborious and costly to assemble Dai et al. (2016). Thus, it is essential to leverage both the few fully annotated datasets, as well as larger datasets labeled with only the cancer status of each image to improve the accuracy of breast cancer classification algorithms.

Pre-training is an effective way to train the network. For example, use hierarchical pre-training to initialize the weight parameters of the DBN with hidden layers and then fine-tune them. It is easy to find that pre-training can improve training speed and recognition accuracy. Another common method of training, first in large databases (such as ImageNet, and then fine-tune the model to complete other tasks Russakovsky et al. (2015); Li et al. (2020). Although certain tasks may be independent of the initial training data set, but the weight of the model parameters have been initialized weight, original features may be identified, which are easily used for other tasks. This can often save training time and improve the performance of the model He et al. (2016); Moreira and Fechine (2018).

Inspired by Lin et al. (2020), so and so. to use the feature extraction capabilities of neural networks, researchers have recently proposed a new mammography recognition method. However, we think that the underlying neural network capabilities to extratct image features can be better utilized. In this study, different deep neural network models are merged to jointly propose a mammography image recognition method that can learn fine-grained mammography by fusing multiple functions. In addition, the hash decoder method is used to simplify the classification calculation of complex high-dimensional feature vectors, and the results of different classifiers are merged at the end. In summary, our work can be shown as follows:

1. Build a deep convolutional neural network based on well-performed network structures, and design a transfer learning strategy to improve the representation power of the extracted features.

2. To further utilize the feature representation ability of CNNs, we propose a method to fuse extracted features from multiple trained models with similar topological structure to further improve the classification accuracy.

3. The strategy applied by hash learning in the deep network is cited to enhance the generalization ability of the model and solve the challenge of high-dimensional calculation in deep learning.

4. The designed feature fusion model is used in the feature extraction process of Faster RCNN, and a deep hash learning strategy is introduced in the classification stage.

This article is organized as follows. Section 2 details of the network structure and the proposed method. Section 3 elaborates the dataset used in our work. Section 4 shows experimental results demonstrate, and given the discussion. Section 5 makes the conclusion and future work are provided in the end.

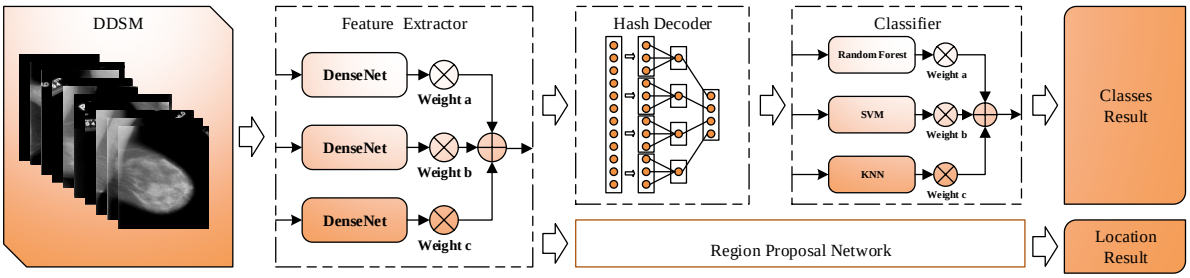

Figure 1: Overview of the proposed approach.

2. Methodology

Architecture of our method can be divided into 4 modules, as shown in Figure 1. Image Feature Extractor, which contains three pretrained models of DenseNet, which are trained on ImageNet task is used as the backbone of our network structure. Benefited from pre-trained models, the network gains the ability to extract basic features from images before training. Designing three different models has aimed for diverse features. To make use of different representation power of each model, fusing the feature from each model with weight add method. Hash Decoder, wherein the image object is randomly divided into a plurality of branches, each bit corresponds to a hash. By this step, the model can not only improve the generalization ability of the model, and may simplify the calculation of high-dimensional features, to facilitate the work of the classifier. Using the third module for classification, we also used three classifiers, including SVM (Support Vector Machine), KNN (K Nearest Neighbor) and RF (Random Forest). This design of classifications is to help improve the accuracy of model predictions. The fourth module is a linear regression analysis model, which has been utilied the RPN(Region Proposal Network) proposed by Faster RCNN. To do this, it is aimed at marking the specific location of the lesion in mammography.

2.1. Network design

Hand-crafted features have been widely used in the mammography image classification tasks. While feature-based approach these handmade perform well, they are always too complicated to build and constraints may under certain circumstances. With the development of deep learning, deep convolutional neural networks can extract intricate features from raw data Szegedy et al. (2016); Zeiler and Fergus (2014). After AlexNet great success in ImageNet competition, the potential of deep CNNs are gradually discovered by researchers, many deep network architectures with good performance on image classification tasks such as VGGNet, GoogLeNet, ResNet, DenseNet

and etc, network based on these well-performed architectures are naturally applied to mammography image recognition task. We build a fusion network made of four modules which are feature extractor(Explained in detail in Section 2.1.1), hash decoder(Explained in detail in Section 2.1.2), classifier(Explained in detail in Section 2.1.3) and regression model(Explained in detail in Section 2.1.4).

2.1.1. Feature Extractor Module

In the modules, building it with a truncated DenseNet without the fully connected layers Huang et al. (2017). The backbone of DenseNet consists of 3 dense blocks, each ciphertext block is composed of several compact layers. The number of dense layers in each dense block increases while the dense blocks go deeper. Dense block structure is shown in Figure 2.

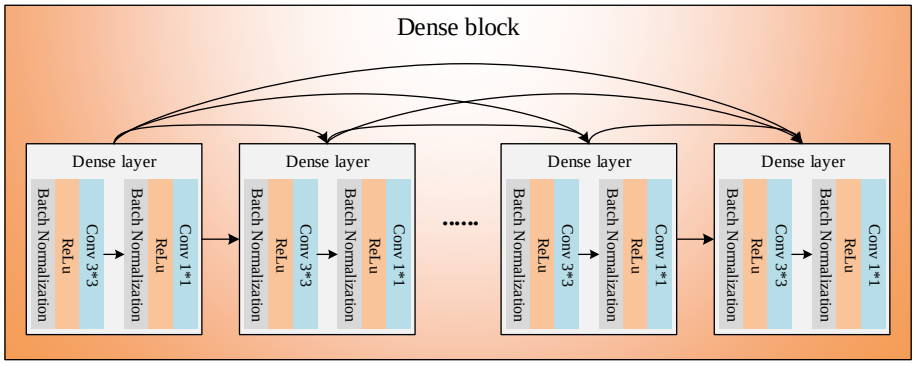

Figure 2: The architecture of DenseNet.

As DenseNet, different from the normal output of each dense layer are fed to all subsequent dense layer in the same block are connected through intensive, this is achieved by operating in series. By this way, from beginning to end global information may traverse dense blocks, each a dense layer may obtain additional knowledge from previous knowledge. For a dense layer, which comprises a base layer 2, the base layer of each batch normalized by layer, and two RELU activation function convolution layers. Srivastava et al. (2014); Ioffe and Szegedy (2015); Kingma and Ba (2015). Further, a feature extractor to help positioning the input feature the most abundant block portion, and a transitional layer and a batch standardization convolution layers, followed by the largest pool layers, which are inserted into each two between dense blocks. It is set for down sampling the output feature maps of each dense block. With the development of the network, the magnitude of the output characteristics of figure decreases, while the number of dimensions will increase the output characteristic of figure Huang et al. (2017). An example of the output feature maps after each dense block is shown in Figure 3.

Usually, most CNN based methods on mammography image classification tasks use a single network to extract features. However, we found some work by different ways to complete this work. It attempts to explore the potential of fusion of features extracted from different models. Wherein the fusion tag in two models with predictive input image, based on a AlexNet, on the other GoogleNet. Compared with the single network method, it utilizes different functions with higher precision. Inspired this proposed a method has the same structure by fusing, through a number of different models of transfer of learning strategy training to improve classification accuracy. Note

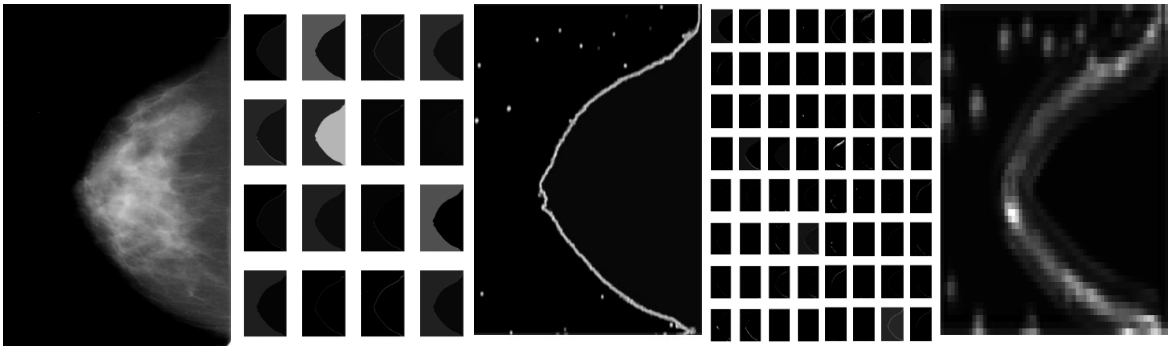

Figure 3: The feature of the fusion of extractors' result.

shallow layer tends to extract the basic characteristics, the deeper layer tends to abstract feature extraction, leading to the conclusion, the fixed parameters model different layers having different features will focus on the complexity of the training phase. Cireşan et al. (2012). In the inference phase, the model will be trained together to take advantage of the different features of each model extraction capability.

Different from it, train 3 models with the same structure, but freeze different layers of each model while training concurrently. There are two reasons why we select 3 same DenseNet based model to be fused. Firstly, we have tested the single-model method for this task, among which DenseNet performs best. Secondly, we have tried fusing different trained models such as Inception, ResNet, DenseNet structures together with frozen parameters, fusing 3 DenseNet based models is still the better way. We think the reason is that when fusing models in different topological structures, the relationship between features extracted from 3 models is relatively weak and we cannot assure that they are learning different knowledge from training images. Since the structures of different networks varies greatly, models may concentrate on similar features with different performance according to the depth of network. By training 4 similar structure models with different frozen layers, each model extracts features in different levels based on the number of frozen parameters, we can ensure that every model concentrates on different features of the input image. Meanwhile, in the experiment we found the false positive and false negative instances of each model are quite different, which proves that they are concentrating on different kinds of features. Since the 3 models keep a similar network structure, Their relationship extracted features stronger, and better integration features.

As figure 4 shows, training 3 models with parameters in different block frozen in parallel. Each row represents one model, the dense block in bright means that the parameters in this block are frozen during the training phase while the dark block means that parameters are activated. we fuse the feature of multiple transfer learning models with a weighted sum operation. The weights of 3 models are set manually in the experiments. With the advantages of functional diversity, the transfer learning mode fusion method achieves high accuracy in mammography image recognition tasks.

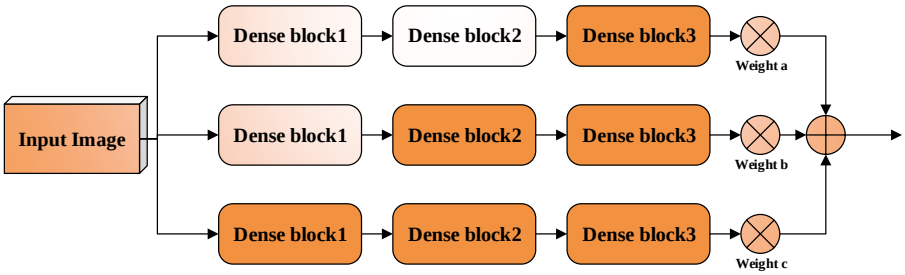

Figure 4: The module of feature through such steps, the model can not only improve the generalization ability of the model, but also simplify the calculation of high-dimensional features to facilitate the work of the classifier extractors' fusion.

2.1.2. Hash Decoder Module

After obtaining the intermediate image features from feature extractor module, we propose a hash decoder module, the map image features to the hash code. As seen in the Figure 5, the proposed hash decoder module first randomly divides the feature vector into several slices of equal lengthLai et al. (2015). Each slice is then connected by a full layer, then, is in the range [0,1] to the output limit value of the activation function s-shaped, and the threshold function is mapped to a segment of a size to encourage bit output binary hash. After that, connect the output hash bits as a code. A hashing decoder module may be a simple alternative to complete the connection layer, wherein the layer of the intermediate image mapped to the input vector, the vectors then converted to the activation function. Compared with this choice, the key idea of the overall strategy is to try to reduce redundancy between the hash bits. Several recent studies have advocated the use of the hash code has fewer redundant bits.

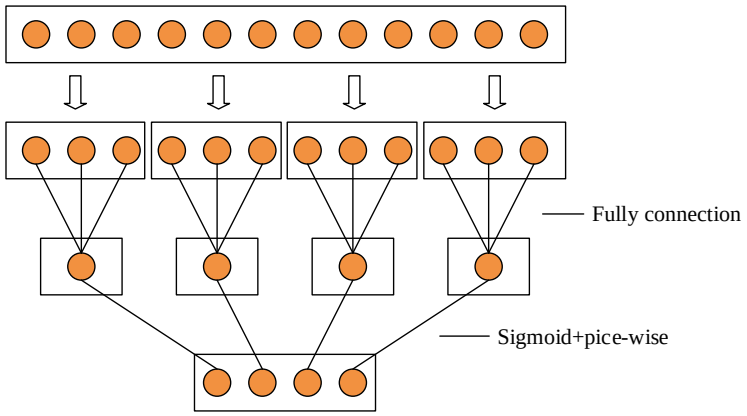

Figure 5: The module of the hash decoder.

To encourage hash decoder output binary code, we used activation function, then the fragmentation threshold function. Given a 50-dimensional slice $x^{(i)}(i = 1, 2, ...q)$, 50 pairs of output is defined as a fully connected layers

$$fc_i(x^{(i)}) = W_i x^{(i)} \tag{1}$$

with W_i being the weight matrix.

Given $c = fc_i(x^{(i)})$, the sigmoid function is defined by

$$sigmoid(c) = \frac{1}{1 + e^{-\beta c}} \tag{2}$$

where β is a hyper-parameter.

The piece-wise threshold function is to encourage binary outputs. Specifically, for an input variable $s = sigmoid(c) \in [0, 1]$, this piece-wise function is defined by

$$y = \begin{cases} 0, & s < 0.5 - \epsilon \\ s, & 0.5 - \epsilon \leq 0.5 + \epsilon \\ 1, & s > 0.5 + \epsilon \end{cases},$$

where ϵ is a small positive hyper-parameter.

This fragmentation threshold function similar to a hard-coded behavior, and encourages the use of binary output in the training. Specifically, if the outputs from the sigmoid function are in $[0, 0.5 - \epsilon]$ or $[0.5 + \epsilon, 1]$, they are truncated to be 0 or 1, respectively. Note that in prediction, the proposed deep architecture only generates approximate (real-value) hash codes for input images, where these approximate codes are converted to binary codes by quantization. With the proposed piece-wise threshold function, some of the values in the approximate hash codes are already zeros or ones. Hence, less errors may be introduced by the quantization step.

2.1.3. Classifier Module

In the classifier module, building it with a fusion structure by SVM, KNN, Random Forest classic methods. For each method, it predicates with the output of the hash decoder, then the three results weight add, as the result of class information. There is a reason why we select 3 different classifiers to be fused. These three algorithms target different classification types, so the results obtained for different feature information will also be different. The biggest advantage of doing this is to improve the robustness of the model, and at the same time it also improves the accuracy of model recognition.

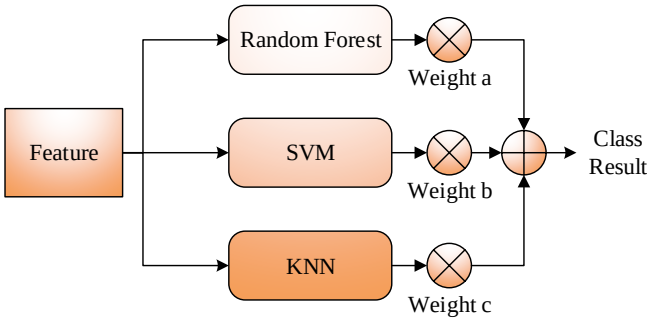

Figure 6: The module of the fusion of classifier.

2.1.4. Regression module

Proposed area network (the RPN) receiving (any size) image as an input, and outputs a set of rectangular objects proposed are each offer it has an objective rating. To generate the proposed region, a slide on a small network last layer output shared. The network is fully connected to the input of the conversion characteristics of FIG n × n window space. Each sliding window is mapped to a low dimensional vectors. This vector is fed into two fully connected layers at the box regression layer and the box classification layer. As shown in Figure 7

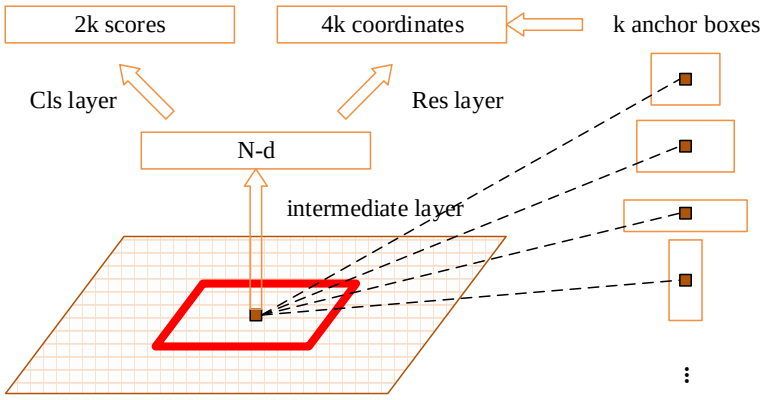

Figure 7: The module of the region proposal network.

Note that, since the small network operated sliding window, the whole connection layer at all spatial positions shared. The architecture is naturally implemented with the n×n CONV layer followed by two brothers 1×1 CONV layerRen et al. (2017). And the loss function is defined by

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{3}$$

In order to perform regression, we use the following 4 coordinate parameterization::

$$t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a, t_w = log(w/w_a), t_h = log(h/h_a), \tag{4}$$

$$t_x^* = (x^* - x_a^*)/w_a, t_y^* = (y^* - y_a^*)/h_a, t_w^* = log(w^*/w_a), t_h^* = log(h^*/h_a) \tag{5}$$

This can be thought of as bounding-box regression from an anchor box to a nearby ground-truth box.

2.2. Transfer learning

Aims to transfer learning between the source and target domains related to knowledge transfer, which is a useful tool for machine learning, can have a positive impact on the field due to insufficient training data and difficult to apply. Transfer learning methods can be divided into four categories: case-based, based on the mapping, network-based, learning against depth migration Pan and Yang (2010); Tan et al. (2018). Most network-based approach used in the convolution neural network. It is achieved by transferring the network structure and pre-trained parameters of the source

domain into a part of a deep convolutional neural network used in the target domain. In our work, we constructed a feature extractor according to the following in Section 2.1.1, then we initialize the network parameters with the parameters of a DenseNet model that were pre-trained on the ImageNet dataset.

In our experiments, adjusting the size of the input image to one piece, going that dataset; freezing, which consists of leaving the parameters in shallow part of the pretrained model unchanged and training only the rest part of the network, which can make use of the basic feature extracting ability of pretrained model. Freezing layers or not in training phase depends on the number of category and quantity variance between source and target dataset. Compared to other datasets, the DDSM dataset we used are much smaller. Moreover, most images in both datasets are natural images, they are similar in a way, hence we freeze some of the layers during training. In order to explore the effect of fusing different feature extractors, we try different freezing strategy on the networks, the details are described in Section 2.1.1.

3. Data

3.1. DDSM

Most computer aided diagnosis (in the CADx) and detection (CADe) algorithm breast cancer was evaluated in a private data set or subset unspecified public databases breast X-ray imaging. Lee et al. (2017); Gao et al. (2019) This leads to performance or replicate previous results can not be directly comparable methods. Choukroun et al. (2017) In order to address this major challenge, we need to publish digital breast X-ray screening database (DDSM) updated and standardized version, in order to assess breast X-ray photography in the future CADx and CADe systems (also sometimes referred to as CAD) study. Zhu et al. (2017) The CBIS-DDSM (planning DDSM mammography subset), including decompressing video data selected by trained mammographers curated updated segmentation quality, bounding boxes, and pathological diagnosis of training data, which looks like a modern computer vision data set. The data set included 753 cases and 891 calcified mass cases, be able to provide analysis and decision support system size of data sets in ammography. As shown in Figure 8

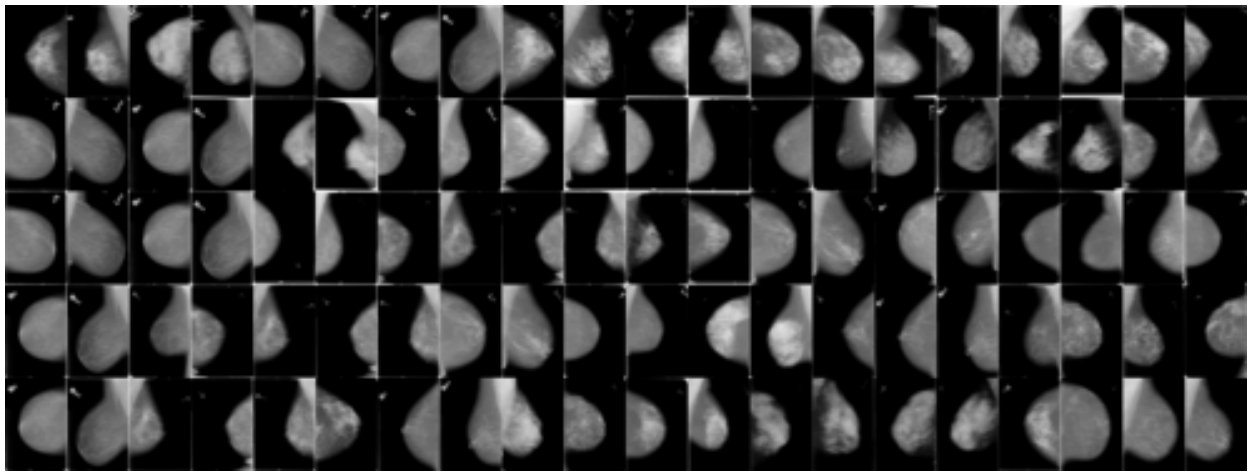

Figure 8: Examples in the dataset

3.2. Augmentation

Compared to the image in the object detection field other common data set (e.g., Pascal Voc), DDSM data set is relatively small. M. Heath, K. Bowyer, D. Kopans and Jr. (2001) Since small data sets are prone to reduce the model effect, to maintain the representation power of model and avoid overfitting in some ways, several data augmentation methods are employed. We don't augment data with the crop, translate or scale operations to avoid changing the original label of images. Instead, we randomly rotate and shear the original images between -10° and 10° , then flip them horizontally or vertically and fill the margin with white pixels.

4. Experiments

4.1. Training strategies

The approach of this study is implemented by the open source framework PyTorch. Abadi et al. (2016) The pre-trained models for initializing the network parameters are acquired from torchvision. Paszke et al. (2017) Models are trained and tested on a graphic workstation with an Intel Core i9-7980XE CPU(2.60 GHz, 18 core) and a GeForce GTX 1080Ti GPU. As described in Section 3, we choose a dataset DDSM to train and test our proposed network in the experiments.

In the training phase, we use a stochastic gradient descent method with a batch size of 32, Ioffe and Szegedy (2015) a momentum of 0.9 and a weight attenuates of 0.0005 similar to optimize our models for 100 epochs. As shown in Figure 4, we propose a method fusing 3 models with similar structure and freezing different layers called Freeze1 to Freeze3 from top to bottom. The accuracy curves on the validation dataset of 3 models in the training phase are shown in Figure 9. Due to the warmup strategy, Glorot et al. (2011) the accuracy grows slowly in the first 5 epochs. A relatively large batch size reduces the noise in the gradient, and an increased learning rate can make a larger progress along the opposite of the gradient direction. Liu and Lapata (2020) Since our network structure is based on DenseNet, the learning rate is initialized to 0.01 and lowered by 10 times every 20 epochs in the experiments. Jia et al. (2014)

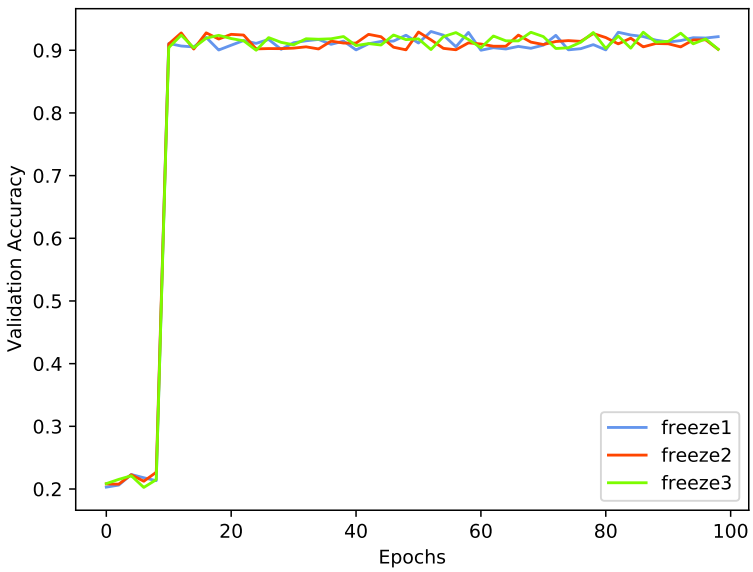

Figure 9: Validation accuracy.

Although the parameters of the network were initialized with the model trained by ImageNet at the beginning of the training phase, and the model is far from the final solution, we adopt a warmup strategy by using a very small learning rate 0.00001 in the first 5 epochs and then switch to the initial learning rate when the training process is stable. Besides, for every 5 epochs, we compare the performance between the current model and the best model of previous epochs. If the previous one performs better, the parameters in the former model will be loaded into the current model before the next epoch starts.

4.2. Results

For mammography image recognition, DDSM dataset contains 753 cases of calcifications, mammography provides the ability to analyze decision support system for data set size. We also compared the results on the DDSM dataset with other methods including CNN feature-based methods. As shown in Table 1, the proposed method achieved high accuracy close to other methods without obvious advantages on the DDSM dataset. We also trained 3 models concentrating on different kind of features by freezing different parameters during training. This method performs better on a fine-grained classification task. Our method achieved high accuracy on this task, significantly higher than other methods.

Table 1: Performance comparing to different methods

Method	Accuracy(%)
SIFT-BoVW	89.05
VGGNet	90.01
AGNet	91.28
Inception	93.48
DenseNet	93.70
Our	94.68

4.3. Statistical analysis of fusion method

In the previous sections, we have proposed a method which fuses 3 models with different layers' parameters frozen to improve performance, particularly in Section 2.1.1 we mention that different models concentrate on different aspects of input images. To evaluate the diversity of 3 models, we conduct statistical analysis with a series of ablation studies.

Firstly, we try to fuse any 2 models of 3 models with different frozen parameters and calculate their accuracies on the test set comparing to a single model. Details of the results are listed in Table 2. F1–F4 represent model Freeze1 to Freeze4 respectively and & means fusing two models together, for instance, F1&F2 means fusing model Freeze1 and Freeze2. According to the results, we find that fusing any of 2 models always performs better than a single one for each category. It can be concluded that models with different layers' parameters frozen focus on different features of images in different categories, which leads to the fused network model get better performance.

In addition, when testing each single network's performance, we count the numbers of instances that only one model recognizes them wrongly while the other three predict correctly. As shown in columns from 3 to 6 in Table 3, 3 models don't give the same predicting result of every image in each category. For each category, there are dozens of images which recognized wrongly by only one model, and the other 2 models can recognize them better.

Table 2: Results of fusing any 2 models

Category	F1(%)	F2(%)	F3(%)	F1&F2(%)	F1&F3(%)	F2&F3(%)
1	90.55	89.60	92.06	92.30	93.05	93.05
2	89.03	91.13	93.37	92.08	96.02	97.06
3	90.37	93.02	91.07	93.51	93.60	94.05
4	88.70	88.67	89.06	92.07	93.05	94.08
5	89.50	89.60	90.05	92.30	94.70	95.05
6	92.06	89.09	94.06	91.03	94.03	97.04
7	91.05	89.60	92.06	92.30	93.05	96.02

Table 3: Comparison of number of images recognized wrongly by only one model

Category	Total	Freeze1	Freeze2	Freeze3	All
1	1172	58	44	31	0
2	1029	24	29	36	0
3	1105	36	61	44	0
4	859	48	41	56	0
5	786	76	87	36	0
6	907	15	29	41	0
7	937	80	29	64	0

And It can be observed that none of the images are wrongly recognized by all the 3 models at the same time, from the last column in Table 3. In other words, the 3 models pay attention to different aspects of the images, and each model has its advantage and drawback. By integrating them and fusing their extracted diversified features together, the ensemble model learns more representative features.

4.4. Analysis the hash decoder

In this work, one of the innovation points, is to employ a hash decoder to reduce the difficulty of computing the high dimenision parameters. Natural alternative hash decoder is coupled with the connection layer after an S-shaped layer limits the output value in [0,1]. To investigate the effectiveness of the hash decoder. Implement and evaluate a single deep network-based DenseNet, connecting with its alternative choice which is a fully network. In the end, we employe a single classifier based SVM. To do this, to prove our hash decoder is effective. As can be seen from Table 4, the results of the proposed method outperform the competitor with the alternative. For example, the architecture with hash decoder achieves the accuracy of 0.581 with 48 bits, which indicates an improvement of 19.7% over the FC alternative. The root cause may be improved, as compared with FC Alternatively, redundancy between each lower output from the hash hash code decoder.

4.5. Comparing to the single classifier

In order to study the effectiveness of the classifier fusion, the present study a single DenseNet, because it can remain invariant feature vectors and expand the impact of differences in classification

Table 4: Comparison results of the hash decoder and fully connection

Method(MAP)	12 bits	24 bits	32 bits	48 bits
FC	0.877	0.896	0.909	0.912
Ours	0.899	0.914	0.925	0.923

of the test results. The specific experimental results are shown in the Table 5. From the data in the table, we can see that the difference between different classifiers is more obvious, and the effect of random forest is better, but the result of the fused classifier is the best. The reason for this is that the principle of different classifiers is different and the sensitivity of features is different. It is obvious that the results of the fusion classifier can make the accuracy more reliable.

Table 5: Comparison results of the classifers

Classifier	Accuracy(%)
k-NN	83.50
SVM	85.70
RandomForest	87.09
Ours	91.49

4.6. Comparing to single network methods

There have been approaches with single network structures applied on DDSM, well-performed structures like VGGNet, Inception, ResNet and etc are used as the backbones for constructing the networks of mammography image recognition approaches. To figure out which backbone is most suitable for this task, the experiments are conducted to test the performance of the network with different backbones. Figure 10 shows the accuracy of several backbones we tried on the validation set during 100 training epochs. In the first 5 epochs, the accuracy grows slowly because of the warmup strategy mentioned in Section 4.1, and the learning rate is set to a relatively small value before the training process is stable. We found the accuracy of DenseNet is slightly higher than other mod- els in the experiment. Hence it is used as one of the backbones.

The proposed transfer learning model fusion method is shown in Figure 4, 3 models are trained in parallel with different frozen parameters. weight of each model is set manually according to the singles model's accuracy. The accuracy of each single model with different frozen parameters is shown in Table 7. For instance, model Freeze1 represents the top row in Figure 4, which is the model with the most frozen parameters. In our work, we test different weights for each model according to their accuracies on the validation set, as shown in Figure 9. Lastly, the weights for 3 models are set to [0.4, 0.4, 0.3].

4.7. Comparing different network fusion methods

We have proposed a method of fusing 3 models with different network structure but different frozen parameters. This idea is inspired by AGNet, a method enriching extracted features by combining 2 networks with a weighted sum. Due to the different topological structures, the relationship between features they extracted is relatively weak and we cannot assure that they have learned knowledge of different aspects from the training images.

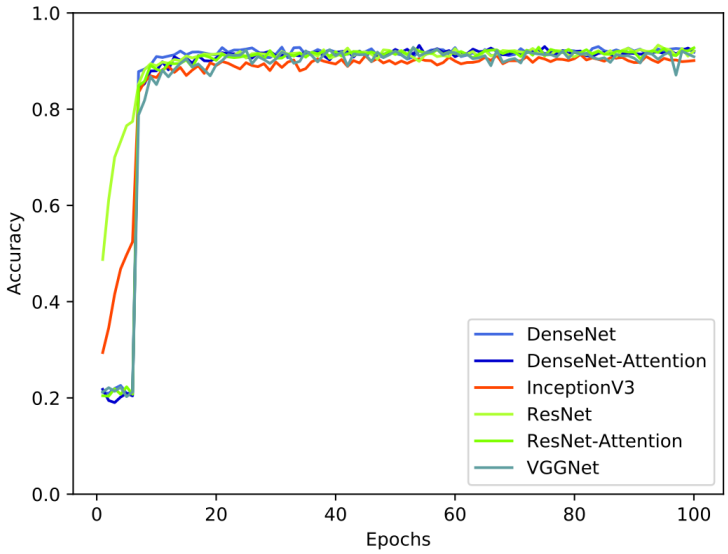

Figure 10: Validation accuracy of different backbone.

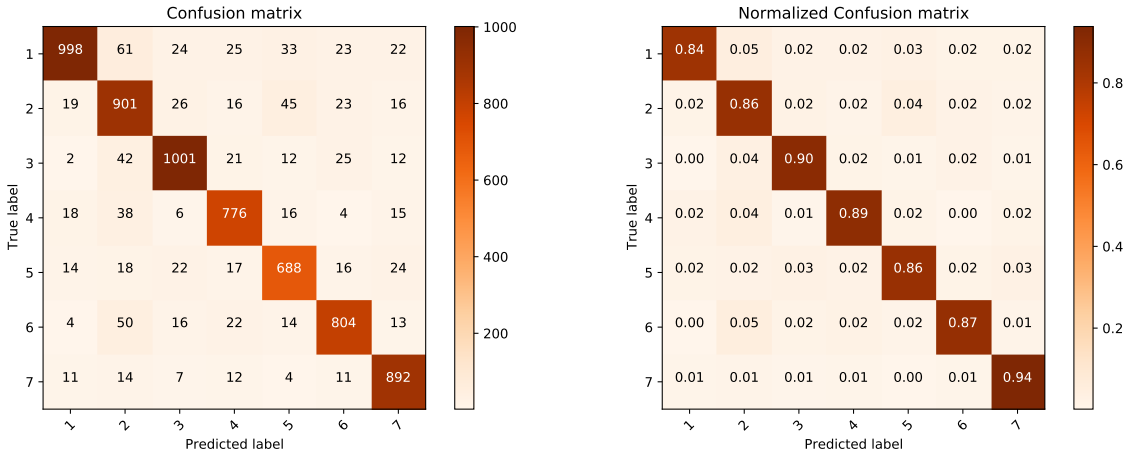

Figure 11: Confusion matrices produced by the first frozen models on the test set

Table 6: Comparing with the single network based methods on DDSM dataset

Method	Accuracy(%)	Average inference time (ms)
VGGNet based	90.09	6.062
InceptionV3 based	92.16	2.396
ResNet50 based	93.08	3.006
DenseNet based	92.67	3.762
Our	94.48	15.052

In the experiments, we test the methods of fusing models in different network structures and on the DDSM dataset. As shown in Figure 11 to Figure 13, among the 3 models we trained, Freeze3 achieves the best result on the validation set during the training phase. Therefore, we try

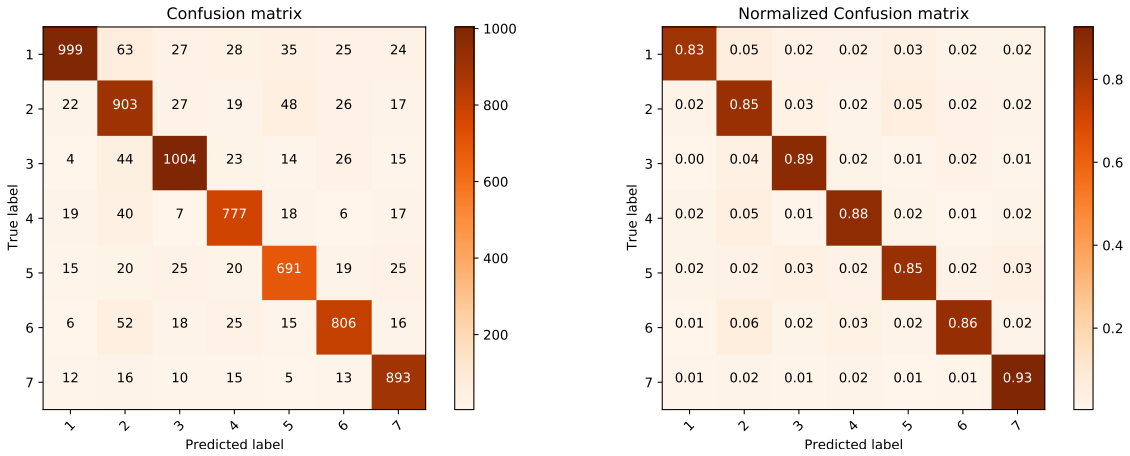

Figure 12: Confusion matrices produced by the second frozen models on the test set

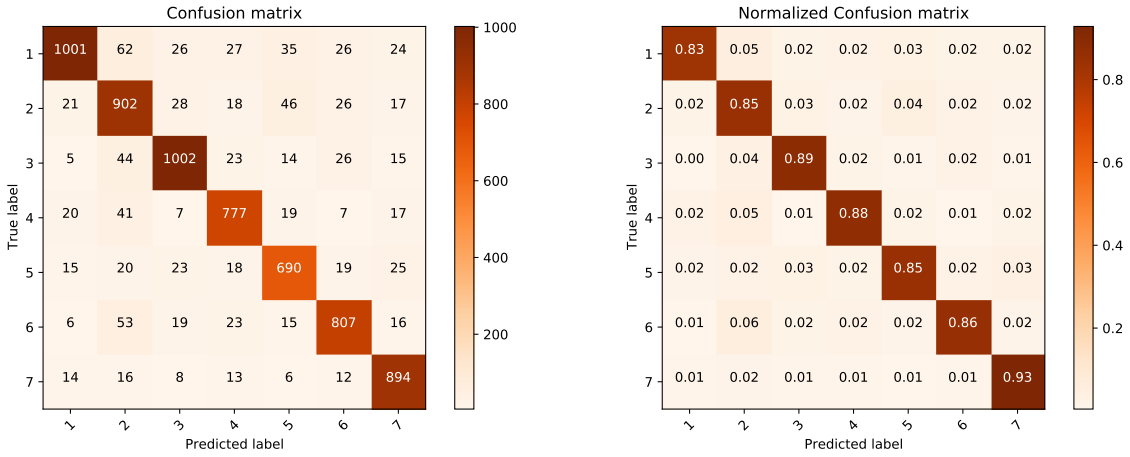

Figure 13: Confusion matrices produced by the thirty frozen models on the test set

Table 7: Accuracies of models with different parameters frozen

Models	Freeze params	Test accuracy(%)
Freeze1	4.76M	93.01
Freeze2	3.50M	92.36
Freeze3	0.95M	91.88
Fusion	-	94.48

to fuse it with other models in different network structure such as VGGNet, Inception, ResNet to check whether the fusion models in different network structure helps improve the performance or not. The parameters of these models are initialized by the corresponding pre-trained model on ImageNet and their shallow layers are frozen while training. According to the results listed in Table 6, compared to the single network-based method, it has higher accuracy, but fusing 3 models still performs better. In this way, the network extracts more rich features and the classifier

produces better recognition results.

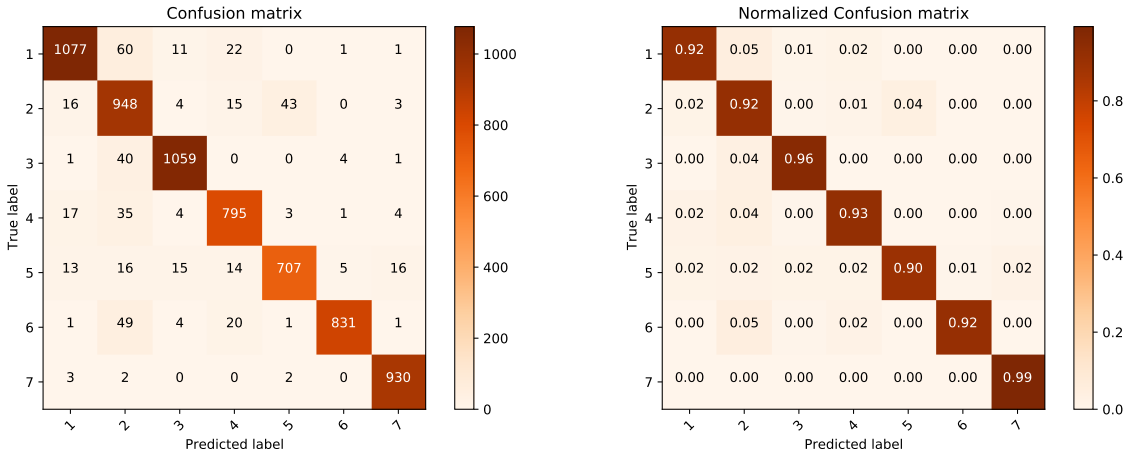

Figure 14: Confusion matrices produced by the fusion model on the test set.

Figure 14 shows the classification results of each class with confusion matrix. The accuracies of 7 classes are higher than 92%. Because the features are learned from the training set and there are some overlapping image parts between classes, which contain a large images are more likely be wrongly classified.

4.8. Comparing other object detection methods

For the target detection module of this work, it is mainly based on the Faster RCNN framework, and the innovation of this study is also the innovation of the classification stage. For the object detection module, we made a simple comparison with the original framework model in terms of time and accuracy. The specific experimental results are shown in Table 8. It is obvious from the data in the table that our method greatly improves the calculation time of the model without losing accuracy. Combining with our model framework, the reason can be clearly found because we have introduced a hash recoder module.

Table 8: Comparing to the region RCNN

Models	mAP(%)	time(ms)
Fast RCNN	60.50	534
Faster RCNN	70.03	325
Ours	75.01	105

5. Conclusion

In this paper, a deep convolutional neural network for mammography images object detection task, which employing the strategy of combining multiple model fusion transfer and transfer learning to improve the classification accuracy. Taking advantage of multiple models' representation power, the proposal achieves a higher accuracy than merely a single model. Meanwhile, the strategy applied by hash learning in the deep network is cited to enhance the generalization ability of

the model and solve the challenge of high-dimensional calculation in deep learning. In the experiments, it's found that classifying images based on CNN relies heavily on the quantity and breadth of the training data set. Our network is still not efficient enough to cover every situation of mammography diagnosed. In the future, we plan to collect a larger data set with more detailed labels and build a more effective architecture to further improve the performance of the mammography image object detection.

6. Acknowledgements

This work is supported by the National Key R&D Plan under Grant No. 2018YFC1200203; Shanghai Science and Technology Project in 2020 under Grant No. 20040501500.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016.

Choukroun, Y., Bakalo, R., Ben-Ari, R., Askelrod-Ballin, A., Barkan, E., Kisilev, P., 2017. Mammogram classification and abnormality detection from nonlocal labels using deep multiple instance neural network. In: VCBM 2017 - Eurographics Workshop on Visual Computing for Biology and Medicine.

Cireşan, D. C., Giusti, A., Gambardella, L. M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-FCN: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems.

Gao, X., Braden, B., Taylor, S., Pang, W., 2019. Towards real-time detection of squamous pre-cancers from oesophageal endoscopic videos. In: Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. In: Journal of Machine Learning Research.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q., 2017. Densely connected convolutional networks. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: 32nd International Conference on Machine Learning, ICML 2015.

Jamieson, A. R., Drukker, K., Giger, M. L., 2012. Breast image feature learning with adaptive deconvolutional networks. In: Medical Imaging 2012: Computer-Aided Diagnosis.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia.

Kingma, D. P., Ba, J. L., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Lai, H., Pan, Y., Liu, Y., Yan, S., 2015. Simultaneous feature learning and hash coding with deep neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning.

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

Li, Y., Zhou, R., Xu, R., Luo, J., Jiang, S., 2020. A quantum mechanics-based framework for eeg signal feature extraction and classification. IEEE Transactions on Emerging Topics in Computing, 1–1.

Lin, X., Qin, F., Peng, Y., Shao, Y., 2020. Fine-grained pornographic image recognition with multiple feature fusion transfer learning. International Journal of Machine Learning and Cybernetics.

Liu, Y., Lapata, M., 2020. Text summarization with pretrained encoders. In: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference.

M. Heath, K. Bowyer, D. Kopans, R. M., Jr., P. K., 2001. The Digital Database for Screening Mammography. In: the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., Medical Physics Publishing, 2001.

Moreira, D. C., Fechine, J. M., 2018. A Machine Learning-based Forensic Discriminator of Pornographic and Bikini Images. In: Proceedings of the International Joint Conference on Neural Networks.

Oeffinger, K. C., Fontham, E. T., Etzioni, R., Herzig, A., Michaelson, J. S., Shih, Y. C. T., Walter, L. C., Church, T. R., Flowers, C. R., LaMonte, S. J., Wolf, A. M., DeSantis, C., Lortet-Tieulent, J., Andrews, K., Manassaram-Baptiste, D., Saslow, D., Smith, R. A., Brawley, O. W., Wender, R., 2015. Breast cancer screening for women at average risk: 2015 Guideline update from the American cancer society.

Pan, S. J., Yang, Q., 2010. A survey on transfer learning.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Facebook, Z. D., Research, A. I., Lin, Z., Desmaison, A., Antiga, L., Srl, O., Lerer, A., 2017. Automatic differentiation in PyTorch. In: Advances in Neural Information Processing Systems.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision.

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

Zhu, W., Lou, Q., Vang, Y. S., Xie, X., 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

Zhu, Z., Albadawy, E., Saha, A., Zhang, J., Harowicz, M. R., Mazurowski, M. A., 2019. Deep learning for identifying radiogenomic associations in breast cancer. Computers in Biology and Medicine.