Received: 14 February 2020 | Revised: 9 August 2020 | Accepted: 24 August 2020

IET Image Processing

DOI: 10.1049/ipr2.12035

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Classification of breast mass in two-view mammograms via deep learning

Hua Li[1] | Jing Niu[1] | Dengao Li[2,3] | Chen Zhang[1]

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, China

[2] College of Data Science, Taiyuan University of Technology, Taiyuan, China

[3] Shanxi Engineering Technology Research Center for Spatial Information Network, Taiyuan, China

Correspondence
Dengao Li, College of Data Science, Taiyuan University of Technology, Taiyuan, China.
Email: lidengao@tyut.edu.cn

Abstract
Breast cancer is the second deadliest cancer among women. Mammography is an important method for physicians to diagnose breast cancer. The main purpose of this study is to use deep learning to automatically classify breast masses in mammograms into benign and malignant. This study proposes a two-view mammograms classification model consisting of convolutional neural network (CNN) and recurrent neural network (RNN), which is used to classify benign and malignant breast masses. The model is composed of two branch networks, and two modified ResNet are used to extract breast-mass features of mammograms from craniocaudal (CC) view and mediolateral oblique (MLO) view, respectively. In order to effectively utilise the spatial relationship of the two-view mammograms, gate recurrent unit (GRU) structures of RNN is used to fuse the features of the breast mass from the two-view. The digital database for screening mammography (DDSM) be used for training and testing our model. The experimental results show that the classification accuracy, recall and area under curve (AUC) of our method reach 0.947, 0.941 and 0.968, respectively. Compared with previous studies, our method has significantly improved the performance of benign and malignant classification.

1 | INTRODUCTION

Cancer is a worldwide public problem. Among cancer cases in women, breast cancer has the highest incidence [1]. According to the statistics from the American Cancer Society, by 2020, there will be about 276,480 new cases of breast cancer in women, accounting for 30% of new cases of cancer in women [2]. If it can be detected early in the onset of breast cancer, the patient's five-year survival rate will increase by 70% compared to advanced cancer [3]. Therefore, the early detection and treatment of breast cancer is extremely important for patients.

Mammography has become the most widely used and effective detection method for breast cancer because of its low cost and satisfying medical requirements [4]. Physicians get the diagnosis result mainly through the analysis of mammography, but the result is easily affected by the subjective experience and fatigue of physicians. In addition, because the features of breast masses are not obvious in the early stage, even for experienced physicians, it is still a challenging work to diagnose by mammograms. So it is very necessary to use computer aided diagnosis (CAD) system to help physicians make a diagnosis. Relevant research shows that a reliable CAD system can help physicians make correct judgments and effectively reduce the burden of patients [5].

The traditional method for breast mass classification is based on pattern recognition. First, features are extracted from mammograms manually, and then the extracted features are input into machine learning classifier for classification [6]. Although the traditional pattern recognition method has made some achievements in mammograms classification, this method relies on the artificially designed characteristics of researchers and lack the ability of autonomous learning. Convolutional neural network (CNN) is a method that can effectively overcome this shortcoming. It can automatically select and extract features from images and has achieved excellent performance in the field of natural image analysis. Therefore, CNN has attracted the attention of many researchers and they have tried to apply it to the analysis and diagnosis of medical images, such as lung CT image [7], brain MRI image [8], and thyroid ultrasound image [9]. In the field of medical image analysis, some researchers have started using CNN to diagnose breast masses [10]. We can divide these studies into two parts: Classification studies based on the whole mammograms and breast-mass patches.

At the beginning of the study, the researchers tried to apply CNN directly to the classification of the whole mammogram. Zhang et al. [11] evaluated the classification performance of two classic CNN models, AlexNet and ResNet50, on the whole mammograms. They use two strategies of data augmentation and transfer learning to improve the classification performance of the model. Similarly, Wang et al. [12] compared the classification performance of AlexNet, VGG16 and ResNet50 in the whole mammogram and conducted experiments on three currently popular public databases. In order to accelerate the convergence rate of the CNN model and improve the classification performance of the CNN model, they used a pre-trained network as a feature extraction network. Li et al. [13] used the Inception structure to construct a new CNN model DenseNet II to classify benign and malignant mammogram. The advantage of the Inception structure is that it contains multiple scale convolution kernels, which can pay attention to the information of different scales of the image. Agnes et al. [14] also adopted the multi-scale convolution strategy to realise the classification of the whole mammogram. They used three different convolution kernels in each convolution layer to extract deep features, so that the network can pay attention to a wider range of image information. However, the size of the whole mammogram is usually 5000 × 3000 pixels. The input image of CNN is generally with small size. Large size mammograms directly resize into small size mammograms and will lose many useful features. Smaller breast masses may even become invisible, severely limiting the classification performance of the model. In addition, some researchers are exploring the research on segmentation of the whole mammograms. One way is to segment the pectoral muscle and breast region [15]. This method can segment the breast region from the image, reducing the interference of the image background and pectoral muscles on the network feature extraction performance. Another way is to automatically segment the lesion area from the whole image [16], which provides more accurate feature information for further classification of the lesion area.

In order to improve the problem of using the whole mammograms to classify benign and malignant breast cancer, some researchers cropped the mass patches from the whole mammograms, and use the mass patches to classify breast cancer. The classification performance is improved by using different training strategies and integrating different network models. Arora et al. [17] proposed a two-stage classification system. In the first stage, five parallel CNN structures such as GoogleNet, ResNet18, and Inception are used to extract features from breast-mass patches, and the five extracted feature vectors are concatenated into one feature vector. In the second stage, they trained a neural network to classify mammograms. The main work of Sun et al. [18] is to compare the classification performance of three different networks for breast-mass patches. In addition, they also compared the classification results of the random and the pre-trained initialisation weights of the CNN model. The experimental results show that the pre-trained ResNet50 has achieved the best classification results in the DDSM database. Chougrad et al. [19] hope to improve the classification performance of breast images through fine-tuning of the network. The experiment found that Inception v3, which only fine-tuned two convolutional blocks, achieved the best results in the breast masses classification task. Some researchers have found that using a single image patch will ignore some useful information and limit the classification performance. So, they hope to overcome this difficulty by extracting multiple image patches.

Compared with the whole mammogram, more mass features can be extracted from the mass patch, which improves the network performance. However, the characteristic information contained in a single image is often limited. Some researchers try to improve classification performance by extracting features from multiple image patches. Lotter et al. [20] cropped two different-scale patches of the lesion area from the whole mammograms. Two ResNets with the same structure are used to extract the features of the two mammogram patches, and the extracted features are fused to achieve the classification of mammograms. Li et al. [21] proposed a two-path neural network model, one path is used to extract the features of breast lesion patches, the other path is used to extract the features of segmentation mask maps, and finally the features extracted by the two paths are connected to achieve the classification of lesions.

In clinical practice, mammography usually has two views called axial and lateral. They are called the craniocaudal (CC) projection and the mediolateral oblique (MLO) projection. Usually, the lesion area will appear in two different mammograms at the same time, but the features displayed are slightly different. Physicians usually need to combine two mammograms of the same breast to make a judgment about the lesion. However, most researchers only use a single-view mammogram to classify breast cancer, and it is often difficult to achieve a good response to the true classification results. Focusing on mammograms from two-view at the same time can extract more lesion features, which helps to improve the classification performance [22].

In recent years, the research of recurrent neural network (RNN) has provided more space for the improvement of CNN. RNN is also an important branch of deep learning research field [23], which has significant advantages in dealing with sequence data, and is widely used in video, signal, and text data. Recently, some researchers have tried to combine CNN and RNN for image classification. Moitra et al. [7] proposed a method combining CNN and RNN to automatically classify lung cancer images. Li et al. [24] used a model combining CNN and RNN to analyse brain MRI images. They used the model to analyse the features of the left and right hippocampal image patches to diagnose Alzheimer's disease. We have noticed that analysing images with spatial relationships through RNN can effectively improve the feature extraction ability and improve the classification performance of the model.

In this study, we made improvements to address the current research problems. We propose a two-view neural network (TV-NN) model to improve the performance of breast mass classification. Our contributions are the following:

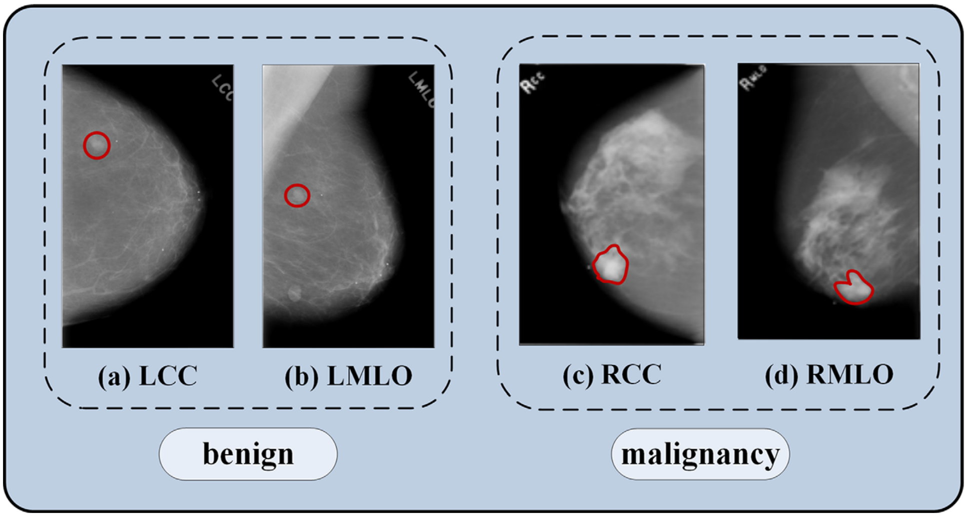


FIGURE 1 Benign and malignant mammograms of breast cancer. (a) and (b) are benign, (c) and (d) are malignant, two views are craniocaudal (CC) and mediolateral oblique (MLO), L and R represent left and right, respectively. The red circle shows the location of the breast mass

1. We combined a deep separable convolution and a residual block to propose a based classification CNN (BC-CNN), which can effectively reduce network parameters and increase network speed. We used two-path BC-CNN to extract features of mammograms from CC view and MLO view, respectively;
2. we combined CNN and RNN to analyse two-view mammograms. The BC-CNN model is used to extract two-view mammograms features, and the features are fused through the RNN's gate recurrent unit (GRU) model to achieve effective use of spatial features;
3. we verified the proposed network pre-training method and proved the effectiveness of this pre-training method;
4. we used the DDSM database to verify the TV-NN model and obtained good results. Finally, the area under the ROC curve (AUC) was 0.968, the accuracy was 0.947, and the recall was 0.941.

2 | DATABASE

The mammography images used in this article are collected from the digital database for screening mammography (DDSM) [25] database. The database can provide material for the research of computer aided diagnostic system, and at present DDSM is widely used in the research field. DDSM database is divided into four types of data: Cancer, normal, benign and benign without callback. There are 2620 cases available in 43 volumes, including 12 normal, 15 malignant, 14 benign and two undiagnosed volumes.

Each case of DDSM contains mammograms from two different views of the left and right mammograms (left CC (LCC), left MLO (LMLO), right CC (RCC), and right MLO (RMLO)). The location, shape, margin, benign and malignant of the lesion is also provided. Figure 1 shows the images of benign and malignant breast mass from two-view on the left and right sides.

In the DDSM database, the physicians' diagnosis information is contained in an OVERLAY file. The diagnosis information includes the lesion type, breast density, and the chain code of the lesion area segmentation. We converted the original LJPEG format to PNG format, and extracted the coordinates of the outermost boundary point of the lesion area from the annotation file.

2.1 | Image normalisation

Normalisation [26] not only can accelerate the convergence speed and increase the accuracy of the model, but also alleviate the scattered feature distribution in the deep network to some extent. It makes the training of deep network easier and more stable, so that the training can use a large learning rate. At present, standardisation has become the standard of neural network training.

In general, we will normalise the characteristics of the input sample to make the data normally distributed (mean 0, standard deviation 1)

$$\hat{x} = \frac{x - E(x)}{\sqrt{\mathrm{Var}(x)}} \tag{1}$$

where $E(x)$ is the mean of samples and $\sqrt{\mathrm{Var}(x)}$ is the standard deviation of samples.

2.2 | Image patch extraction

In the DDSM database, each case contains a detailed label by the physicians, which contains the coordinates of the boundary area of the mass. According to the physicians' annotation, we constructed a rectangle by attaching the outermost point of the mass boundary, and we used the centre of the rectangle as our

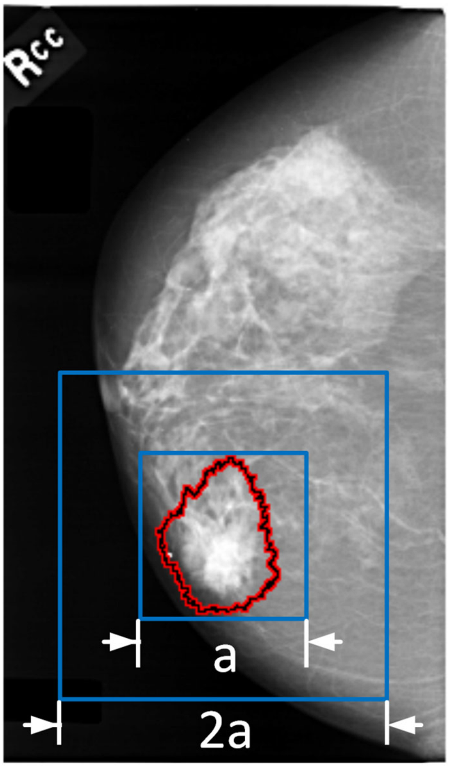


FIGURE 2 Image patch extraction. The longest side of the smallest rectangle containing the breast mass is a, centre is point o, we cut the area containing the breast mass with a square centred on the point o and side length 2a

extraction centre. In order to ensure that the shape of the mass plaque image does not change when it is input into the network, we set a square crop area. Figure 2 shows the patch extraction strategy of the mass. To enable the CNN to pay attention to the information around the mass, we set the side length of the extracted area to twice the length of the longer side of the rectangle. Because the pixel values of the regions of interest that are manually extracted are different, in order to facilitate the training of the network, we resize the pixel values of the image patches to 512 × 512.

2.3 | Data division

The DDSM database contains a total of 891 cases of masses, including 970 masses (522 benign and 448 malignant masses). We divide all the masses into 10 subsets according to the proportion, each of which contains 97 masses. Among them, the first eight subsets contain 52 benign and 45 malignant masses. The ninth and 10th subsets contained 53 benign and 44 malignant masses, respectively.

2.4 | Data augmentation

Deep learning requires a lot of data to ensure accuracy and prevent overfitting. Data augmentation was been given to increase the number of data in the case of few mammograms. Because of the positional correlation between CC and MLO, we only consider the use of flip and translation strategies for data enhancement. We enhance the data of the 10 subsets according to the following strategies, and the enhanced data still belongs to the original subset.

2.4.1 | Flip

We flip the original images according to the following image flipping strategies: (a) The CC image is flipped up and down, and the MLO image remains unchanged; (b) the CC image is flipped left and right, while the MLO image is also flipped left and right. The flipped images are then extracted according to the image patch extraction strategy in Section 2.2. Here, we consider the flipped mass image patch as a new mass. By flipping, we expand the data by three times.

2.4.2 | Translation

In order to achieve the robustness of position, we translate the image extraction area up, down, left and right by 10% to obtain more patches. At the same time, we randomly combine the five patches extracted from the CC image and the five patches extracted from the MLO image to form five pairs of two-view mammograms. Therefore, by translation, we expand the data by five times.

It should be noted that the nature of mass has not changed after data augmentation. Through the above two strategies, the data can be expanded to 15 times of the original. A total of 14,550 pairs of two-view mammogram patches were obtained, of which 7830 were benign and 6720 were malignant.

3 | METHOD

We proposed a TV-NN classification model. First, feature extraction of two-view mammograms (CC and MLO) on the same side was carried out on BC-CNN, and then the features extracted were fused by RNN. Finally, images were classified into benign or malignant ones. Figure 3 is the specific process structure of TV-NN.

The proposed method is divided into the following steps:

a. Two breast mammograms from two-view were input into BC-CNN, respectively, and the features were extracted.
b. In order to fuse the features of the two-view mammograms, the features of BC-CNN extracted from the two-view mammograms were, respectively, input into the recursive neural network.
c. The fused features of RNN were Input into the softmax layer for classification.

3.1 | Basic structure of CNN

In this study, residual block was combined with deep separable convolution to apply a new network structure called the inverted

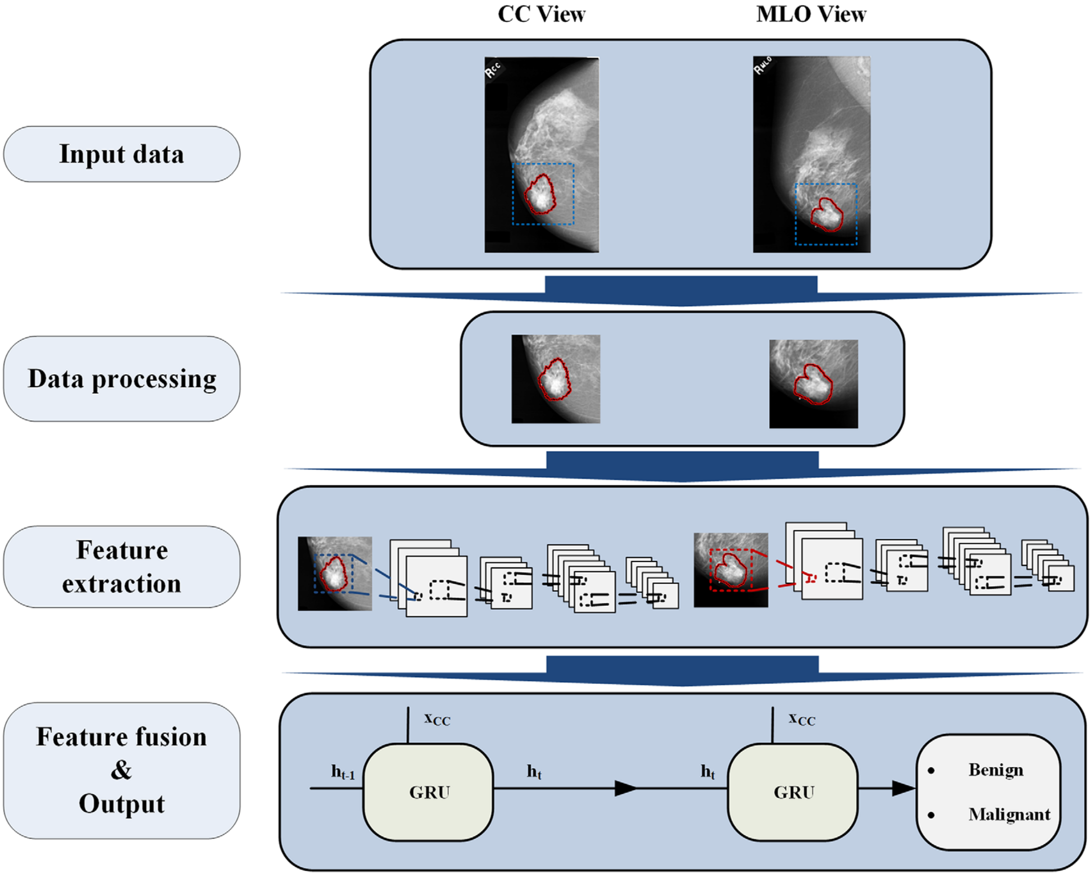


FIGURE 3 The overall architecture of the two-view neural network (TV-NN). The first step is data preprocessing, the breast masses were extracted from the whole mammograms from two views. The second step is feature extraction, based classification convolutional neural network (BC-CNN) extracts the characteristics of breast masses (BC-CNN is our proposed classification-based convolutional neural network). The third step is feature fusion, input of the extracted breast masses features into gate recurrent unit (GRU) structure for feature fusion

residual block. A new network structure based on BC-CNN is proposed.

3.1.1 | Residual block

As the plain network gets deeper and deeper, the classification results become worse, and the gradient disappears [27], leading to slower network convergence and worse classification accuracy. ResNet proposed a residual learning [28] method to improve this situation. ResNet solved the problem of disappearance of gradient return by introducing cross-layer linkage, making to train very deep CNN become simple. The comparison between traditional network connections and cross-layer residual connections is shown in Figure 4.

The rectified linear (ReLU) activation function is used after the convolution operation to increase the non-linearity of the model. If the input is less than zero, the ReLU outputs zero. Else, the output equals the input. The formula of ReLU function is as follows:

$$R\ (x) = \max\left(0, x\right) \tag{2}$$

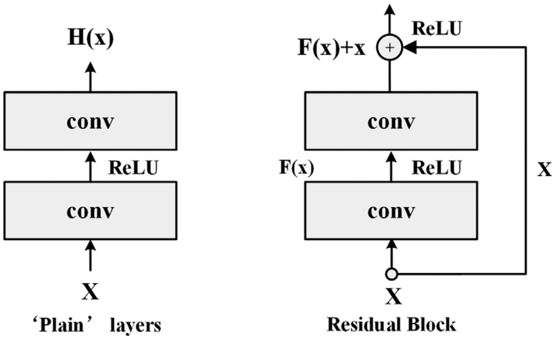


FIGURE 4 'Plain' layers (left) and residual block (right). Compared to ordinary network connections, the residual block adds cross-layer linkage

3.1.2 | Depthwise separable convolution

Depthwise separable convolution [29] is divided into depthwise and pointwise convolutions. Each convolution kernel of the standard convolution performs the convolution with the data of all channels. Convolution kernel of the depthwise separable convolution performs the convolution only with one channel.

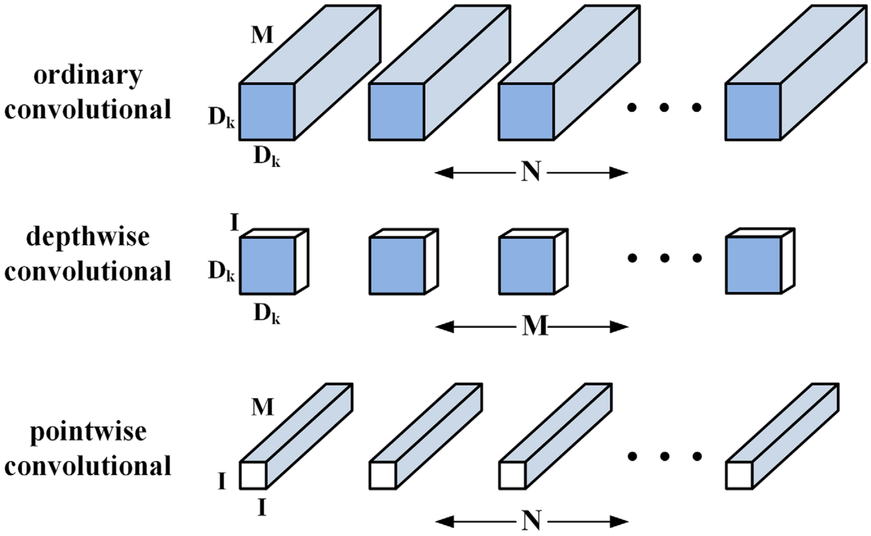


FIGURE 5 Ordinary convolution and depthwise separable convolution. Depth separable convolution solves the traditional convolution integral into a depthwise and a pointwise convolutions (1 × 1 convolutional)

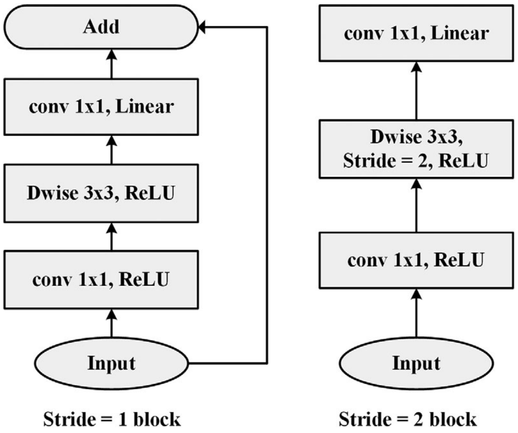


FIGURE 6 Inverted residual block. On the left is a block with a stride of 1 connected by a shortcut, on the right is the block with a stride of 2 to perform down sampling operation

Pointwise convolution is correlated with feature maps obtained by depthwise convolution. The detailed structure of depthwise separable is shown in Figure 5.

3.1.3 | Inverted residual block network structure

In this study, depthwise separable convolution is applied to residual structure instead of standard convolution operation in residual structure. The use of 1 × 1 convolutions first reduces and then increases the dimension of the feature map, and extracts the features by a 3 × 3 convolution. In order to avoid large information loss caused by ReLU to tensors with few channels, the linear layer is used to replace the ReLU non-linear operation. Figure 6 shows the network architecture of the residual block. Tensors with a small number of channels will lose information using the ReLU function. So, the linear layer is used to replace the non-linear operation of ReLU.

3.1.4 | Classification-based CNN

In this section, a BC-CNN was constructed using the inverted residual block mentioned above. Table 1 shows the concrete structure of the BC-CNN.

We input MLO and CC mammograms into two BC-CNNs and removed the last full connected layer. We noticed that the size of the extracted feature map was 8 × 8 × 1024. At this time, we used the global average operation to simplify the image feature into a one-dimensional vector, which was input into the recurrent neural network.

3.2 | Feature fusion based on GRU

There is a correlation between the two views of mammograms, and each view combines information from the previous view. The RNN has a unique structure which makes it very suitable for handling information related to time or space [30]. The features of two-view mammograms are spatially related. RNN consists of three parts: Input, hidden and output units. The emergence of Long Short-Term Memory (LSTM) [31] solves the problem of gradient disappearance and gradient explosion of RNN. The GRU [32] merges the forget gate and the inputs into an update gate, while the network loops back and forth only the output as a memory state. Inputs and outputs of GRU are simpler than LSTM.

Features from two-view mammograms are fused by two GRU modules. All GRU modules share the same parameters. The GRU function is shown in Figure 7. The classification results are obtained by using the softmax activation function. The formula is as follows:

$$\text{softmax}\,(x)_i = \frac{\exp\,(x_i)}{\sum_{j=1}^{n} \exp\left(x_j\right)} \tag{3}$$

where $i = 1, \ldots, n$.

$$r_t = \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{4}$$

TABLE 1 The description for BC-CNN architecture

Type/stride	Filter shape	Input size	Output size	Inverted residual block
Conv1/2	3 × 3 × 32	512 × 512 × 3	256 × 256 × 32	–
Conv2/1	3 × 3 × 16	256 × 256 × 32	256 × 256 × 16	1
Conv3/2	3 × 3 × 16	256 × 256 × 16	128 × 128 × 16	1
Conv4/1	3 × 3 × 24	128 × 128 × 16	128 × 128 × 24	2
Conv5/2	3 × 3 × 24	128 × 128 × 24	64 × 64 × 24	1
Conv6/1	3 × 3 × 32	64 × 64 × 24	64 × 64 × 32	2
Conv7/2	3 × 3 × 32	64 × 64 × 32	32 × 32 × 32	1
Conv8/1	3 × 3 × 64	32 × 32 × 32	32 × 32 × 64	2
Conv9/1	3 × 3 × 96	32 × 32 × 64	32 × 32 × 96	3
Conv10/2	3 × 3 × 96	32 × 32 × 96	16 × 16 × 96	1
Conv11/1	3 × 3 × 160	16 × 16 × 96	16 × 16 × 160	1
Conv13/1	3 × 3 × 320	16 × 16 × 160	16 × 16 × 320	1
Conv14/2	3 × 3 × 320	16 × 16 × 320	8 × 8 × 320	1
Conv15/1	1 × 1 × 1280	8 × 8 × 320	8 × 8 × 1024	–
FC		8 × 8 × 1024	1 × 1 × 1024	–
softmax				

Notes: The network uses convolution with step size of 2 to replace the maximum pooling layer for subsampling.

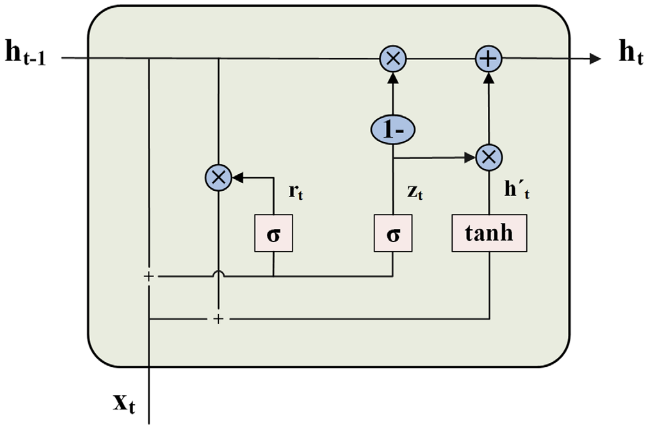


FIGURE 7 Working principle of GRU

$$z_t = \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right) \tag{5}$$

$$h'_t = \tanh\left(W'_t \cdot \left[r_t * h_{t-1}, x_t\right]\right) \tag{6}$$

$$h_t = \left(1 - z_t\right) * h_{t-1} + z_t * h'_t \tag{7}$$

Where [] represents the connection of two vectors and $*$ represents the product of matrices.

3.3 | Model-training strategy

Training of the proposed TV-NN network model includes the pre-training of BC-CNN and the fine-tuning of GRU network for specific classification task. In our implementation, the BC-CNN model which extracts the CC image patches features is pre-trained with enhanced MLO image patches. Similarly, the BC-CNN model which extracts the MLO image patches features is pre-trained with enhanced CC image patches. Then, the pre-trained BC-CNN models are fine-tuned with the enhanced mammogram patches from CC and MLO, respectively. The softmax function is used to connect the full connection layer to the category label of the output. We fixed the parameters of all convolutional layers, pooling layers and full-connected layer in BC-CNN. Finally, the GRU parameters are fine-tuned by using the two-view mammograms.

4 | EXPERIMENTAL RESULTS AND DISCUSSION

In this study, we experiment with the pre-processed DDSM database. The method is verified by contrast experiments, and the evaluation method is given to further analyse the experimental results. All experiments were carried out using Python3.6 in the PyTorch framework. The training and testing were done on a PC equipped with a 16 GB core i7 CPU and two NVIDIA Titan X GPUs, using the Ubantu18.04 system.

4.1 | Experimental setup

We use the method of k-fold cross-validation [33] to carry out our experiments. Each experiment selects k-1 subsets as the training set and the remaining one as the testing set. The experiment was carried out k times, and the corresponding accuracy rate was obtained for each experiment. The average of the accuracy of k experiments was calculated as the final accuracy. In the experiment, Adam optimiser was used to update parameters, and the cross-entropy function was used to calculate the error. The learning rate was changed to one tenth of the original one for every 100 epochs iterated, and the initial learning rate was 0.001. The resize of the original picture after input into the network was 512 × 512, a total of 300 epochs.

4.2 | Evaluation of metrics

For quantitatively analysing the experimental results, we used several commonly used evaluation indicators, accuracy, recall and receiver operating characteristic (ROC) curve. The evaluation indexes are represented by true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

The following three definitions of evaluation indicators are given:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{8}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

The horizontal coordinate of ROC curve is false positive case rate (FPR) and the vertical coordinate is true case rate (TPR) where FPR $= \frac{\text{TP}}{\text{TP+FN}}$, TPR $= \frac{\text{FP}}{\text{TN+FP}}$.

4.3 | Comparative experiment and result analysis

In this section, we set up different comparative experiments to verify the feasibility of our proposed method. Ten-fold cross-validation is used to evaluate the proposed model, and we have divided all data into 10 subsets previously. We compared and analysed different network models, pre-training methods and different fusion strategies.

4.3.1 | Comparative analysis of different network models

We used single-view and mix-view to analyse and evaluate the performance of different networks. We tested three different network structures, and Table 2 gives the relevant experimental results. We fine-tuned the original VGGNet and ResNet, changed the output of the network from 1000 to two classification, and compared with our proposed network BC-CNN in mammograms classification. Compared with the previous network, VGGNet uses a smaller convolution filter and deepens the network to 16 layers. However, with the deepening of the network, the gradient disappears. ResNet proposed a residual learning unit to improve the problem and make the network reach 152 layers.

TABLE 2 Classification performance in single-view based on different network model

		Evaluation Metrics	
View	**Framework**	**Accuracy**	**Recall**
CC	VGG	0. 877	0.868
	ResNet	0.887	0.892
	BC-CNN	**0.891**	**0.897**
MLO	VGG	0.871	0.861
	ResNet	0.882	0.889
	BC-CNN	**0.889**	**0.893**
Mix View	VGG	0.881	0.869
	ResNet	0.891	0.892
	BC-CNN	**0.892**	**0.900**

For VGG, ResNet and BC-CNN, we all used Gaussian random distribution to initialise the parameters. All network models are experimented with the same settings. In the experiment, each network was trained using breast-mass patches of CC and MLO views. In addition, in order to verify that the two-view mammograms can more effectively improve the classification performance of the networks, we mixed the CC and MLO views' breast-mass patches to train the networks, so that the networks can classify both two-view mammogram patches. From Table 2, we can see that in the single-view image classification task, compared with VGG and ResNet, the BC-CNN has advantages in accuracy and recall, so BC-CNN has better classification performance. Although the network trained with mix-view is superior to the single-view network, the difference is not significant.

4.3.2 | Analysis of the results of network pre-training

A major difficulty in medical image classification is that the amount of data is small, and training a network from scratch is often difficult to achieve good results. Related research shows that the initialisation of parameters through pre-training of the network can effectively overcome this problem [34]. Kooi et al. [35] used the symmetry of the breast to construct a two-path CNN to extract the features of bilateral image patches, and realised the classification of benign and malignant lesions of the breast. They proposed that the contralateral images can be used to pre-train the network in future study, which is a novel way to pre-train the network.

In this study, we proposed a two-path feature extraction network, and we used a multi-view network pre-trained strategy in our study. The strategy can be summarised as follows:

1. We use the mammograms from CC view to pre-train a BC-CNN network, and use the mammograms from MLO view to fine-tune the network. The fine-tuned network is used to extract the breast-mass features from the MLO view.

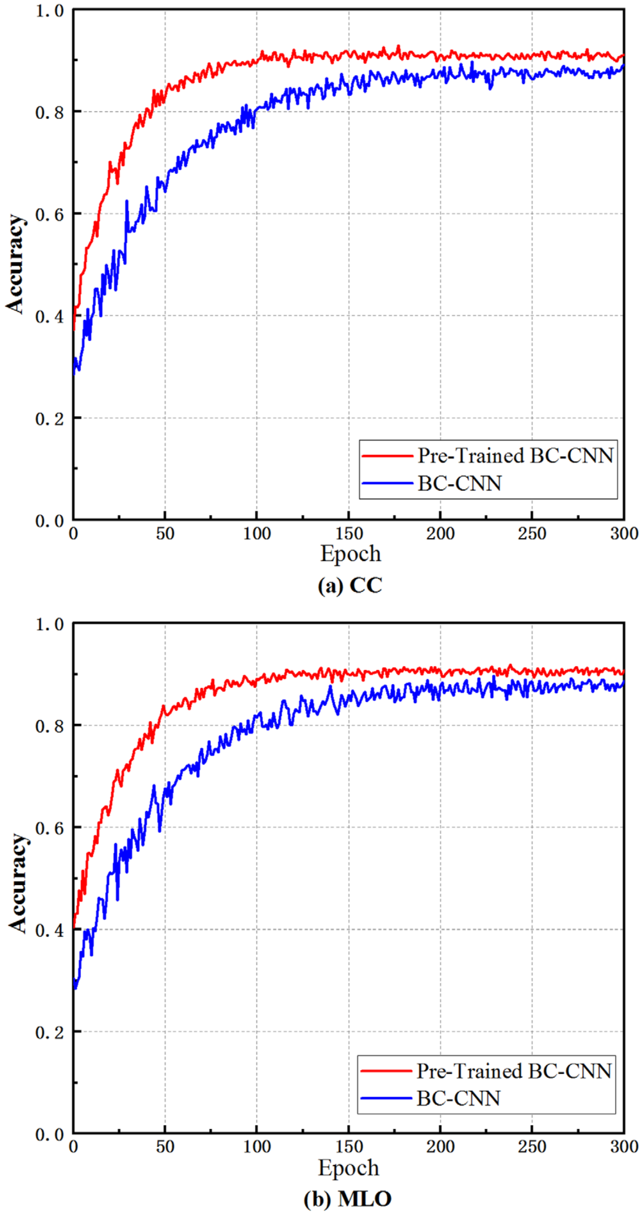
TABLE 3 Results of random initialisation model and pre-training model based on single-view image

		Evaluation Metrics	
Method	View	Accuracy	Recall
BC-CNN	CC	0.891	0.897
	MLO	0.889	0.893
Pre-trained BC-CNN	CC	0.910	0.922
	MLO	0.908	0.920

Notes: BC-CNN in the table means that Gaussian random distribution is used to initialise weights, and pre-trained BC-CNN means that pre-trained model parameters are used to initialise weights.

2. Similarly, we use the mammograms from MLO view to pre-train another BC-CNN network, and use the mammograms from CC view to fine-tune the network. The fine-tuned network is used to extract the breast-mass features from the CC view.

All experiments used the same data and experiment settings in this section. Table 3 shows the results of benign and malignant classification experiments for networks without pre-trained and pre-trained networks. Regardless of the CC or the MLO view, the classification performance of the pre-trained network is better than that without pre-trained network. Figure 8 shows the accuracy curve of the pre-trained network model and the network model without pre-trained in the training process. The classification accuracy curve of the CC view breast-mass patches is given in Figure 8(a), and the accurate classification curve of the MLO view breast-mass patches is given in Figure 8(b). It can be seen from Figure 8 that the pre-trained model has better performance than the model without pre-trained.



(a) CC

(b) MLO

FIGURE 8 Random initialisation model (blue) versus pre-training model (red) based on a single perspective. (a) The accuracy curve from the CC view, (b) the accuracy curve from the MLO view

4.3.3 | Comparison of different image feature fusion strategies based on two-view mammograms

According to the correlation characteristics of mammograms from two views, we combined the features of the two views to classify breast mass, and proposed a method for fusing the features of two-view mammograms. The extracted features of the two-view mammograms are fused through the GRU module. We used the pre-trained BC-CNN as the feature extraction network for breast-mass patches.

The proposed classification model TV-NN used two BC-CNNs to extract image features, and used GRU for feature fusion. The experiments in this section follow the experimental settings in Section 4.1 to train the model. Table 4 and Figure 9 show the classification results of each fold in the 10-fold cross-validation of the model. Among them, Figure 9 shows variation in the results of each fold. It can be seen from 10 experiments that the accuracy and recall of the model classification fluctuate between 0.937–0.954 and 0.934–0.949, respectively. Finally, average accuracy and average recall of the model are 0.947 and 0.941, respectively, and the standard deviations of the accuracy and recall are 0.005 and 0.004, respectively. These results show that the model we proposed is stable and reliable.

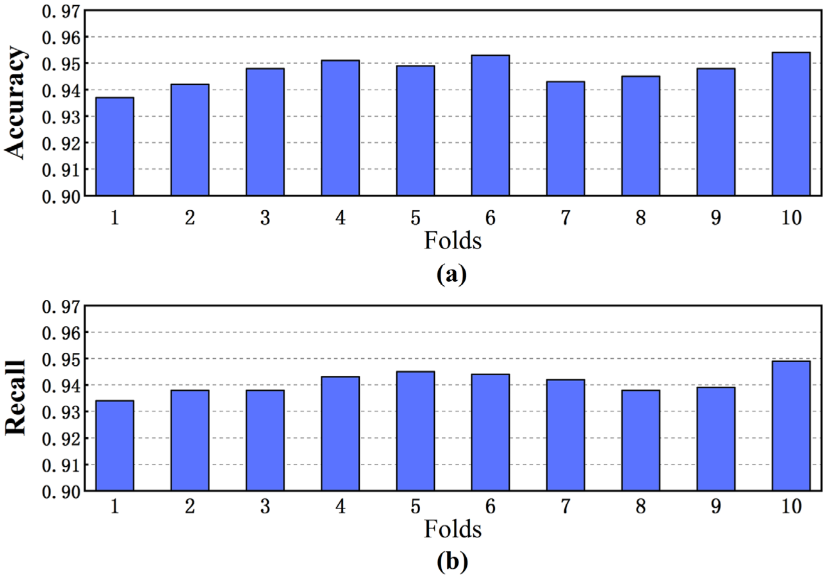
In recent years, some researchers are also exploring effective multi-view feature fusion methods. In this section, we compared our proposed strategy for feature fusion through GRU module with the two commonly used feature fusion strategies, (a) and (b), as shown in Figure 10 with (c) as our proposed fusion strategy through GRU module.

a. Average score: Using two pre-trained BC-CNN branches to extract image features from CC and MLO views. The last

TABLE 4 Benign and malignant classification results of 10-fold cross-validation based on TV-NN

Folds	1	2	3	4	5	6	7	8	9	10	Ave
Accuracy	0.937	0.942	0.948	0.951	0.949	0.953	0.943	0.945	0.948	0.954	0.947
Recall	0.934	0.938	0.938	0.943	0.945	0.944	0.942	0.938	0.939	0.949	0.941

FIGURE 9 The classification results of TV-NN model in 10-fold cross-validation



fully connected layer of each branch outputs the benign and malignant scores through a softmax function. We add the scores of the corresponding categories from the two networks and calculate the average as the final benign and malignant scores.

b. Fully connected layer: Using the fully connected layer to connect the last fully connected layer of two CNNs to obtain a new feature vector. Using the pre-trained BC-CNN as a feature extraction network.

Table 5 gives the experimental results of the three fusion methods. In order to verify that use of two-view mammograms can improve the model's classification performance, we added the single-view classification network for comparison. Both the single-view and the two-view classification networks use pre-trained BC-CNN as the feature extraction network. It can be seen from Table 5 that the classification performance of the three different two-view networks is better than the single-view network. Therefore, the classification performance of the network can be improved by two-view mammograms. The accuracy and recall of the feature fusion method we proposed using the GRU module are 0.947 and 0.941, respectively. Among the three fusion methods, our proposed method has better classification performance.

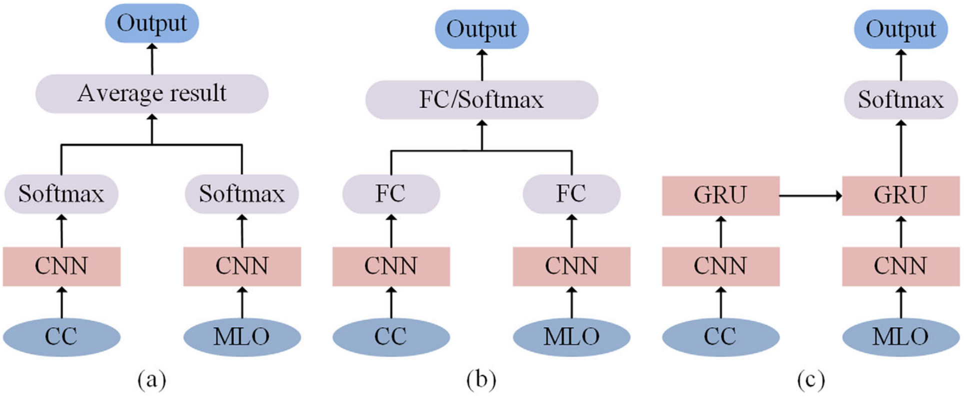


FIGURE 10 Three different ways of feature fusion (a) average the classification weights score of the two views, (b) features are fused through the fully connected layer, (c) the features extracted from CNN are fused through GRU

TABLE 5 Comparison of single-view and different two-view feature fusion methods

				Evaluation Metrics	
View		CNN Model	Fusion strategy	Accuracy	Recall
Single-view	CC	Pre-trained BC-CNN	–	0.910	0.922
	MLO	Pre-trained BC-CNN	–	0.908	0.920
Two-view		Pre-trained BC-CNN	Average weights score	0.919	0.925
		Pre-trained BC-CNN	Fully connected layer	0.923	0. 927
		Pre-trained BC-CNN	**GRU (ours)**	**0.947**	**0.941**

4.4 | Analysis of classification results of breast masses based on TV-NN

Through the comparative experiment in Section 4.3, we have the following conclusions. In terms of network structure, we combined depthwise separable convolution with residual block to propose BC-CNN. Compared with ResNet and commonly used VGGNet, our proposed network structure reduced the calculation parameters and improved classification performance in single-view classification tasks. We verified the feasibility of the pre-training method. Compared with the randomly initialised model, the proposed pre-training method can improve the classification performance of the model. Finally, we compared the feature fusion method proposed by us with the other two commonly used feature fusion methods. The experimental results show that our proposed way of using GRU fusion features has better performance in mammograms classification task.

In this section, we compared our proposed pre-trained two-view classification network TV-NN with four other two-view classification networks. The four classification networks are the BC-CNN network trained with mixed-view mammograms, the TV-NN classification network without pre-training, and two multi-view classification networks using the two feature fusion methods mentioned in Section 4.3.3. We used ROC curves to compare their performance. It can be seen from Figure 11 that the TV-NN model without pre-training obtains better classification performance than the BC-CNN using mixed mammograms.

In addition, the ROC curves of the three pre-trained two-view classification networks are higher than the TV-NN model without pre-training. And it can be seen that the classification performance of our proposed pre-trained TV-NN is significantly higher than the other four classification networks. Therefore, the proposed method can effectively improve the classification performance of mammograms.

Considering that in the k-fold cross-validation experiment, different values of k may affect the classification results of the model, we therefore chose different k values (k = 3, 5, 6, 8, 10, 12) to experiment with the proposed TV-NN model. Figure 12 shows the accuracy and recall rate of the k-fold cross-validation experiment of six different k values. It can be seen from the experimental results that when the k value is 10, the evaluation index of the model is optimal. And when the value of k changes, the overall performance of the model changes relatively smoothly, which shows that the model we proposed is stable and has good generalisation. Figure 13 shows the accuracy curve of training and validation in k-fold cross-validation.

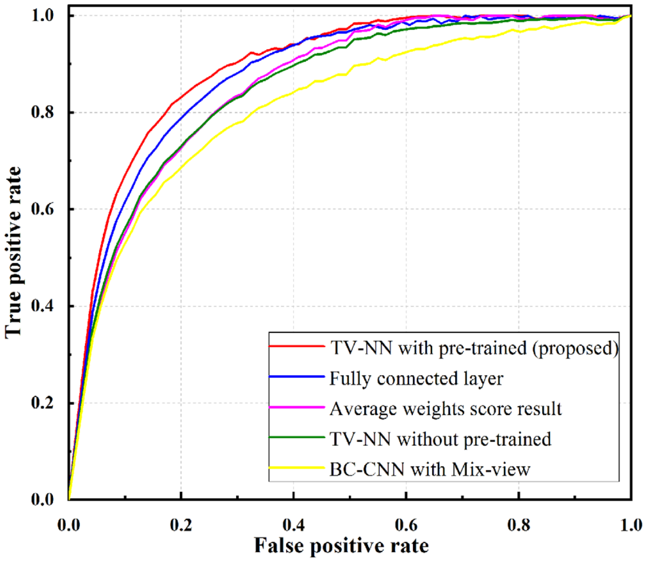


FIGURE 11 Receiver operating characteristic (ROC) curves of different classifiers

4.5 | Comparisons with state-of-the-art mammograms classification methods

In this study, we constructed a TV-NN to classify mammograms. In order to find the correspondence between the two-view mammograms, the network used the GRU module to fuse the features of the two-view mammograms. In addition, we also verified the effectiveness of a multi-view network pre-training strategy. In order to prove the advancement of our proposed method, we compare our proposed method with the method of classifying mammograms using deep learning in recent years.

Some researchers use different CNN models and different fine-tuning strategies in the mammograms classification task to improve the classification performance of the network. Ragab et al. [36] manually cut the area of interest, and used an AlexNet

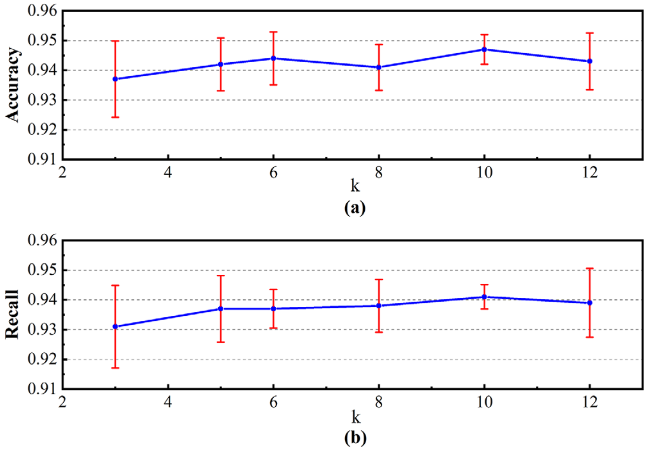


FIGURE 12 The classification performance of TV-NN using k-fold (k = 3, 5, 6, 8, 9, 10, 12) cross-validation. The figure shows the mean and standard deviation

based on fine adjustment to extract features of mammograms. They replaced the last fully connected layer in the network with an SVM structure and used SVM as a classifier to classify mammograms. Aboutalib et al. [37] explored the impact of four different training strategies on mammograms classification. They used the AlexNet as a feature extraction network and classify images through the softmax function. However, AlexNet is an earlier CNN model and its classification accuracy is not as good as some new CNN models. Lenin et al. [38] compared four currently popular network models which are MobileNet, ResNet50, Inception V3 and NasNet. ResNet50 achieved the best classification results in the classification task of mammograms. Recently, some researchers have tried to use feature fusion to make the network focus on a wider range of image features and improve the classification performance. Arora et al. [17] combined the five CNNs of AlexNet, VGG16, GoogLeNet, ResNet18, and Inception to build a holistic model to extract image features. They concatenated the feature maps extracted from different CNN models into one feature map. However, it increased the number of network parameters and limited the classification performance of the network. Khan et al. [39] cropped four image patches from bilateral mammograms and classified the mammograms in three stages. Although the method proposed by them has achieved good results in the classification of normal and abnormal images, this method limits the performance of benign and malignant classification. The big reason is that most lesions appear only in one breast. Carniero et al. [40] extracted the features of the original images from two perspectives, mass segmentation images and microcalcification segmentation mask images, and finally input the multi-view features into a CNN model for breast cancer risk prediction. Carneiro et al. [41] used whole mammograms from two views and corresponding segmentation maps to classify breast cancer. They used a network pre-trained with a large visual dataset to extract image features, and explored the effect of different fine-tuned layers on classification performance. Due to the obvious difference between medical and natural images, a network pre-trained with natural images cannot extract features from medical images well. Gao et al. [42] proposed a shallow-deep CNN (SD-CNN) to classify mammograms. Shallow CNN is used to synthesise the reconstructed image, and deep CNN is used to extract the features of the image. The features of the reconstructed image and the original image are combined for benign and malignant classification of breast images.

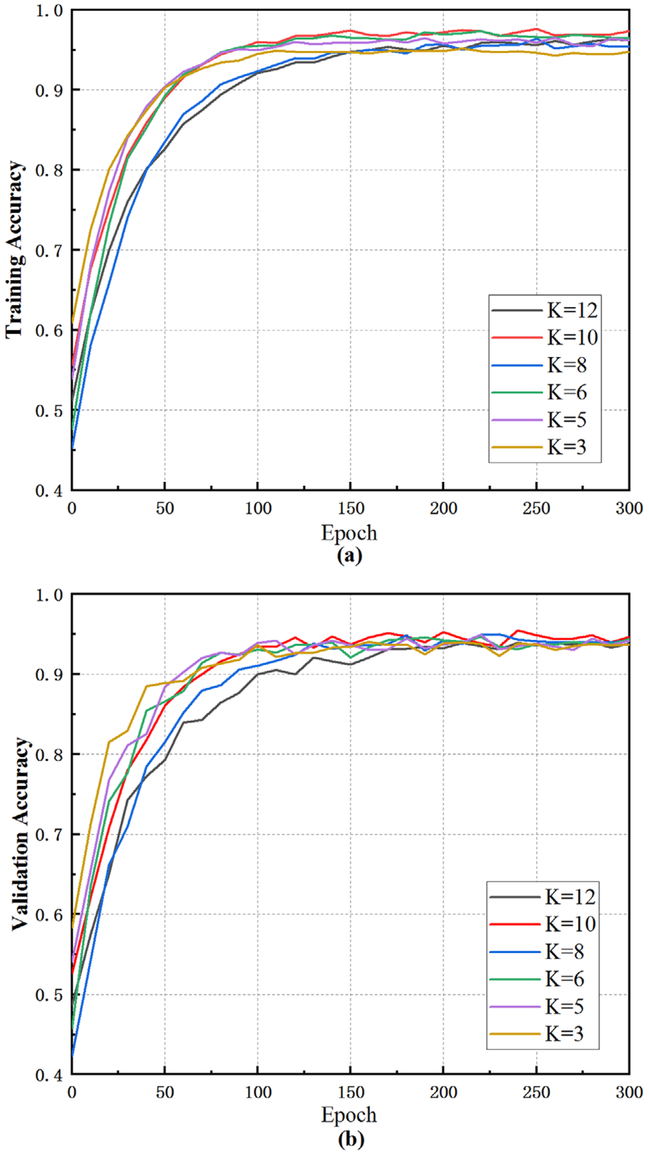


FIGURE 13 Classifier's accuracy curve of training and validation results. (Each curve is generated by taking the average of the result of k-fold cross-validation)

We compared our proposed method with the above methods, and the comparison results are shown in Table 6. Compared with the single-view image classification network, our proposed two-view feature fusion classification network can pay attention to the features of images from different views at the same time. And compared with the simple feature fusion method, our proposed GRU-based feature fusion method has obvious advantages in classification performance. It can be

TABLE 6 Comparisons with state-of-the-art mammograms classification methods

		Evaluation Metrics		
Method	**Dataset**	**Accuracy**	**Recall**	**AUC**
Ragab et al. [34]	DDSM	0.872	0.763	0.940
Aboutalib et al. [35]	FFDM + DDSM	–	–	0.780
Lenin et al. [36]	CBIS – DDSM	0.784	–	–
Arora et al. [15]	CBIS – DDSM	0.880	0.910	0.880
Khan et al. [37]	CBIS – DDSM + MIAS	0.776	0.818	0.920
Carniero et al. [38]	INbreast + DDSM	–	–	0.910
Carneiro et al. [39]	DDSM	–	0.940	0.910
Gao et al. [40]	INbreast	0.840	–	0.870
Our proposed	**DDSM**	**0.947**	**0.941**	**0.968**

TABLE 7 List of abbreviations

Abbreviation	Full name
CNN	Convolutional neural network
RNN	Recurrent neural network
CC	Craniocaudal
MLO	Mediolateral oblique
GRU	Gate recurrent unit
DDSM	Digital database for screening mammography
CAD	Computer aided diagnosis
TV-NN	Two-view neural network
BC-CNN	Based classification convolutional neural network
ROC	Receiver operating characteristic
AUC	Area under curve

concluded from the comparison in Table 6 that compared with the more advanced research, the proposed method has a significant improvement in overall classification performance.

5 | CONCLUSION

This study proposed a new architecture TV-NN for the benign and malignant classification of mammograms. In particular, the BC-CNN proposed is a new CNN structure used to extract features of images. Deep features of two-view mammograms are extracted and features of the two images are fused by GRU based on spatial correlation between different views. Experimental results show that our proposed method has stability and generalisation, and has achieved good classification performance on the DDSM database. The idea is to help physicians diagnose breast cancer as benign or malignant, thereby saving medical resources. Based on the experimental results, TV-NN is superior to the existing research methods, and its accuracy, recall and AUC are 0.947, 0.941 and 0.968, respectively.

Although our study has achieved good results, there is still room for improvement. Our proposed method is semi-automated and requires manual cutting of the breast-mass area. In the future study, we hope that the location of the breast mass can be automatically determined by a computer-aided system. In addition, we believe that a large amount of data can help improve model performance. We will use more data from different databases to improve the classification performance and generalisation ability of the model.

REFERENCES

1. DeSantis, C.E., et al.: Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J. Clin. 67(6), 439–448 (2017)
2. Siegel, R.L., et al.: Cancer statistics, 2020. CA Cancer J. Clin. 70(1), 7–30 (2020)
3. Cronin, K.A., et al.: Annual report to the nation on the status of cancer, Part I: National Cancer Statistics. Am. Cancer Soc. 124(13), 2785–2800 (2018)
4. Pedro, R.W.D., et al.: Is mass classification in mammograms a solved problem?—A critical review over the last 20 years. Expert Syst. Appl. 119, 90–1139 (2018)
5. Kooi, T., et al.: Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. 35, 303–312 (2017)
6. Tsochatzidis, L., et al.: Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. Pattern Recognit. 71, 106–117 (2017)
7. Moitra, D., Mandal, R.K.L: Automated AJCC (7th edition) staging of non-small cell lung cancer (NSCLC) using deep convolutional neural network (CNN) and recurrent neural network (RNN). Health Inf. Sci. Syst. 7(1), 14 (2019)
8. Talo, M, et al.: Application of deep transfer learning for automated brain abnormality classification using MR images. Cognit. Syst. Res. 54, 176–188 (2019)
9. Poudel, P., et al.: Patch based texture classification of thyroid ultrasound images using convolutional neural network. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, pp. 5828–5831 (2019)
10. Soffer, S., et al.: Convolutional neural networks for radiologic images: A radiologist's guide. Radiology 290(3), 590–606(2019)
11. Zhang, X.F., et al.: Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. IEEE Trans. Nanobiosci. 17(3), 237–242 (2018)
12. Wang, X., et al.: Inconsistent performance of deep learning models on mammogram classification. J. Am. Coll. Radiol. 17(6), 796–803 (2020)
13. Li, H., et al.: Benign and malignant classification of mammogram images based on deep learning. Biomed. Signal Process. 51, 347–354 (2019)
14. Agnes, S.A., et al.: Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). J. Med. Syst. 44(1), 30 (2020)
15. Maghsoudi, O.H., et al.: Automatic breast segmentation in digital mammography using a convolutional neural network. In: 15th International Workshop on Breast Imaging (IWBI2020), Leuven, Belgium (2020)
16. Sun, H., et al.: AUNet: Attention-guided sense-upsampling networks for breast mass segmentation in whole mammograms. Phys. Med. Biol. 65,(5), 055005 (2020)
17. Arora, R., et al.: Deep feature-based automatic classification of mammograms. Med. Biol. Eng. Comput. 58(6), 1199–1211 (2020)
18. Sun, W., et al.: Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Comput. Med. Imaging Graphics 57, 4–9 (2017)
19. Chougrad, H., et al.: Deep convolutional neural networks for breast cancer screening. Comput. Methods Programs Biomed. 157, 19–30 (2018)
20. Lotter, W., et al.: A multi-scale cnn and curriculum learning strategy for mammogram classification. In: Proceedings of the 3rd MICCAI International Workshop on Deep Learning in Medical Image Analysis (DLMIA), Quebec, Canada, pp. 169–177 (2017)

21. Li, H., et al.: A deep dual-path network for improved mammogram image processing. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, England, pp. 1224–1228 (2019)
22. Velikova, M., et al.: Improved mammographic CAD performance using multi-view information: A Bayesian network framework. Phys. Med. Biol. 54(5), 1131–1147 (2009)
23. Zhang, Q., et al.: A survey on deep learning for big data. Inf. Fusion 42, 146–157 (2018)
24. Li, F., Liu, M.: A hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease. J. Neurosci. Methods 323, 108–118 (2019)
25. Heath, M., et al.: Current status of the digital database for screening mammography. In: Digital Mammography, pp. 457–460. Dordrecht Springer, Netherlands (1998)
26. Zhou, S., et al.: Multi-view image denoising using convolutional neural network. Sensors 19(11), 2597 (2019)
27. Liu, Z., et al.: Retinal vessel segmentation using densely connected convolution neural network with colorful fundus images. J. Med. Imaging Health Inf. 8(6), 1300–1307 (2018)
28. Zhong, Z., et al.: Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework. IEEE Trans. Geosci. Remote Sens. 56(2), 847–85 (2018)
29. Ting, F.F., et al.: Convolutional neural network improvement for breast cancer classification. Expert Syst. Appl. 20, 103–115 (2019)
30. Mou, L., et al.: Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 55(7), 3639–3655 (2017)
31. Graves, A.: Supervised sequence labelling with recurrent neural networks. pp. 37–45. Springer, Heidelberg (2012)
32. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. Statistics (2014). arXiv:1406.1078v1
33. Zhang, Y., et al.: Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. IEEE Access 4, 8375–8385 (2016)
34. Christodoulidis, S., et al.: Multisource transfer learning with convolutional neural networks for lung pattern analysis. IEEE J. Biomed. Health Inf. 21(1), 76–84 (2017)
35. Kooi, T., Karssemeijer, N.: Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. J. Med. Imaging 4(4), 044501 (2017)
36. Ragab, D.A., et al.: Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 7, e6201 (2019)
37. Aboutalib, S.S., et al.: Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin. Cancer Res. 24(23), 5902–5909 (2018)
38. Falconí, L.G., et al.: Transfer learning in breast mammogram abnormalities classification with Mobilenet and Nasnet. In: Proceedings of the 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), Osijek, Croatia, pp. 109–114. (2019)
39. Khan, H.N., et al.: Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access 7, 165724–165733 (2019)
40. Carneiro, G., et al.: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 652–660. Springer, Cham (2015)
41. Carneiro, G., et al.: Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans. Med. Imaging 36(11), 2355–2365 (2017)
42. Gao, F., et al.: SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis. Comput. Med. Imaging Graphics 70, 53–62 (2018)

How to cite this article: Li H, Niu J, Li D, Zhang C. Classification of breast mass in two-view mammograms via deep learning. *IET Image Process*. 2021;15:454–467. https://doi.org/10.1049/ipr2.12035