BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 66, No. 6, 2018
DOI: 10.24425/bpas.2018.125931

Comparative study on the classification methods for breast cancer diagnosis

Y. QIU[1], G. ZHOU[1]*, Q. ZHAO[1,2] and A. CICHOCKI[3,4,5]

[1] School of Automation, Guangdong University of Technology, Guangzhou, China.
[2] Tensor Learning Unit, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan.
[3] Skolkovo Institute of Science and Technology (SKOLTECH), 143026 Moscow, Russia.
[4] System Research Institute, Polish Academy of Sciences, Warsaw 00-901, Poland.
[5] Hangzhou Dianzi University, College of Computer Science, Hangzhou 310018, China.

Abstract. Digital mammography is one of the most widely used approaches for breast cancer diagnosis. Many researchers have demonstrated the superiority of machine learning methods in breast cancer diagnosis using different mammography databases. Since these methods often have different pros and cons, which may confuse doctors and researchers, an elaborate comparison and examination among them is urgently needed for practical breast cancer diagnosis. In this study, we conducted a comprehensive comparative study of the state-of-the-art machine learning methods that are promising in breast cancer diagnosis. For this purpose we analyze the largest mammography diagnosis database: Digital Database for Screening Mammography (DDSM). We considered various approaches for feature extraction including principal component analysis (PCA), nonnegative matrix factorization (NMF), spatial-temporal discriminant analysis (STDA) and those for classification including linear discriminant analysis (LDA), random forests (RaF), k-nearest neighbors (kNN), as well as deep learning methods including convolutional neural networks (CNN) and stacked sparse autoencoder (SSAE). This paper can serve as a guideline and useful clues for doctors who are going to select machine learning methods for their breast cancer computer-aided diagnosis (CAD) systems as well for researchers interested in developing more reliable and efficient methods for breast cancer diagnosis.

Key words: breast cancer, mammography, DDSM, comparative study, deep learning.

1. Introduction

Breast cancer is the most common disease and also the leading cause of cancer death among women. In 2013, it accounted for 29% of all new cancer cases among women all over the world [1]. According to the latest cancer statistics in 2018, this rate is expected to rise to 30% [2]. In the treatment of breast cancer, early detection plays a key role. Mammography is the most commonly-used technology for early detection due to its low cost and wide availability [3]. However, it is often not easy to accurately identify the breast cancer from the mammograms due to the difficulty of mammograms interpretation [4]. Under such circumstances, computer-aided diagnosis (CAD) systems are necessary for presenting second suggestions for doctors in mammography-based breast cancer detection. A CAD system in general consists of three steps, namely, cropping the region of interest (ROIs) from the mammograms, extracting features from the ROIs, and detecting abnormality or malignancy based on the extracted features.

Digital mammography is one of the most widely used approaches for breast cancer diagnosis. Many researchers have demonstrated the superiority of machine learning methods in breast cancer diagnosis using different mammography databases. Since these methods often have different pros and cons, which may confuse doctors and researchers, an elaborate comparison and examination among them is urgently needed for practical breast cancer diagnosis. In this study, we conducted a comprehensive comparative study of the state-of-theart machine learning methods that are promising in breast cancer diagnosis. For this purpose we analyze the largest mammography diagnosis database: Digital Database for Screening Mammography (DDSM). We considered various approaches for feature extraction including principal component analysis (PCA), nonnegative matrix factorization (NMF), spatial-temporal discriminant analysis (STDA) and those for classification including linear discriminant analysis (LDA), random forests (RaF), k-nearest neighbors (kNN), as well as deep learning methods including convolutional neural networks (CNN) and stacked sparse autoencoder (SSAE). This paper can serve as a guideline and useful clues for doctors who are going to select machine learning methods for their breast cancer computer-aided diagnosis (CAD) systems as well for researchers interested in developing more reliable and efficient methods for breast cancer diagnosis.

In the past decades, machine learning methods have demonstrated their great potential in breast cancer diagnosis. Curvelet level moments (CLM) method was shown to achieve the accuracy of 91.27% and 81.35% for the Mammographic Image Analysis Society (mini-MIAS) database respectively on abnor-

*e-mail: guoxu.zhou@qq.com

Manuscript submitted 2018-05-12, initially accepted for publication 2018-05-29, published in December 2018.

mality and malignancy detection and the accuracy of 86.46% and 60.43% for Digital Database for Screening Mammography (DDSM) database [4]. Deep Belief Networks (DBN) reached a classification accuracy of 99.68% for the Wisconsin Breast Cancer Dataset (WBCD) [5]. Convolutional neural network (CNN) has already showed its powerful ability for classification in various data sets including ImageNet 2012 [6] and is widely utilized to solve various pattern recognition tasks such as speech recognition [7], objection detection [8], image classification [6], and etc. It has also demonstrated its effectiveness of detection in histopathological images and mammgrams [9–12]. The fusion machine learning models showed their efficiency in three databases, namely, Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) [9]. Stacked sparse autoencoder (SSAE), a representative learning method, achieved a diagnosis accuracy of 88.84% in breast cancer histopathological images [13]. In [14] authors presented a CAD system for mammographic masses which used a mutual information-based template matching scheme, and achieved accuracy up to 83% for DDSM database.

Many existing methods have presented promising and often satisfactory performance on some specific mammography databases. However, comparisons of these state-of-art methods for an unified large database have not been investigated till now. In this study, in order to better understand the characteristics of these methods for mammography diagnosis, we made a comprehensive comparative study of the state-of-art methods using DDSM database, the largest existing mammography diagnosis database [15]. We manually cropped ROIs which contain the mass tumor or suspicious texture from the mammograms at first. And then significant features of ROIs were extracted by applying state-of-art feature extraction methods including principal component analysis (PCA), nonnegative matrix factorization (NMF) and spatial-temporal discriminant analysis (STDA). Then, popular classification methods including linear discriminant analysis (LDA), random forests (RaF) and k-nearest neighbors (kNN) were employed to classify the extracted features into the normal or abnormal category and benign or malignant category. In deep learning methods: CNN and SSAE, discriminative features were extracted by the optimized multi-layered neural networks structures, and then the softmax function in the last layer was implemented to classify the features. Finally, we analyzed and ranked the performance of different feature extraction and classification methods according to their average classification accuracy. The acronyms of the methods used in this study are represented in Table 1 and the processing procedure from mammograms to classifications is illustrated in Fig. 1.

Table 1
The methods for comparison in this study.

Acronyms	Description
CNN	convolutional neural network
SSAE	stacked sparse autoencoder
PCA	principal component analysis
NMF	nonnegative matrix factorization
STDA	spatial-temporal discriminant analysis
LDA	linear discriminant analysis
RaF	random forests
kNN	k-nearest neighbors

2. Materials and methods

2.1. Data set. DDSM is a database resource which is available for the mammograms analysis research community, containing the cases from Massachusetts General Hospital,Wake Forest University School of Medicine, Sacred Heart Hospital and Washington University of St.Louis [15], and has already been widely explored in mammograms analysis. This database consists of 2620 cases, and each case contains four standard views (i.e., left, right, front and back) of full mammograms with the labeled location of ROIs. The labels of cases are divided into three types: normal, benignant and malignant.

2.2. ROIs extraction. Since the original mammograms in DDSM database contain muscles and background areas, we need to crop the ROIs from mammograms before our funda-

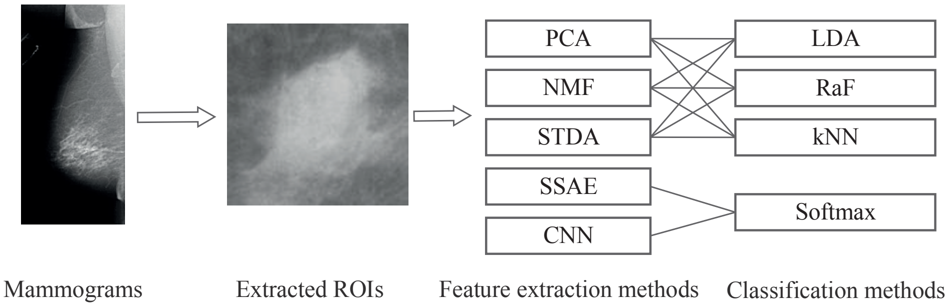

Fig. 1. A flowchart of processing from mammograms to classifications. The ROIs are cropped manually from mammograms based on the marks. And then the compared feature extraction methods are implemented on the ROIs. Finally, the compared classification methods classify the extracted features into normal or abnormal, or into benign or malignant

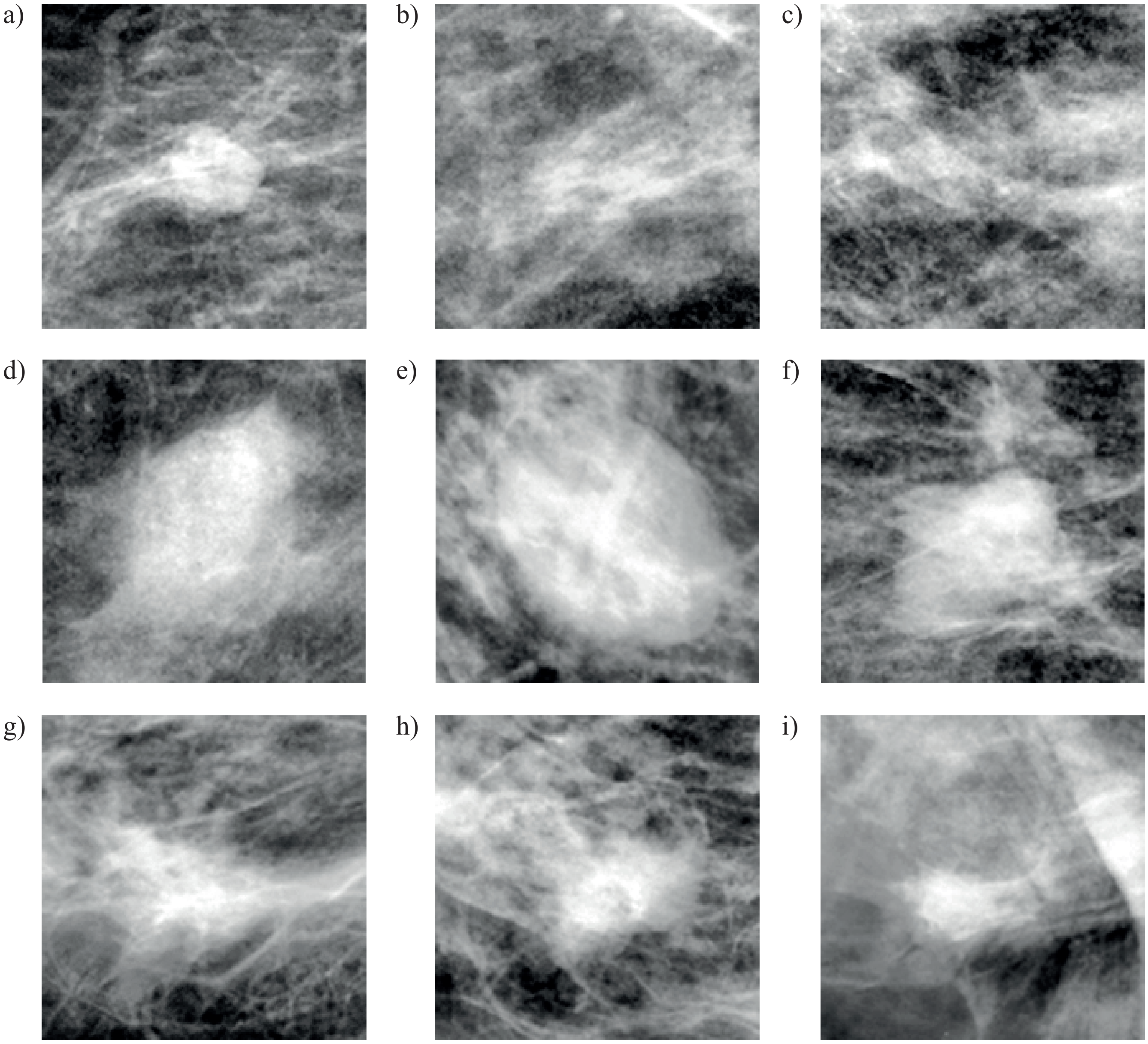

Fig. 2. A sample of ROIs cropped from DDSM database. Fig. a–c are normal cases, Fig. d–f are benignant cases, Fig. g–i are malignant cases

mental analysis. In this study, to guarantee the success of cutting off the mass tumor area, the center of abnormal or suspicious ROIs is the location marked manually and the resolution is 128×128 pixels. The extracted ROIs consist of 11 218 images, which include 9215 normal cases, 888 benign cases and 1115 malignant cases. Figure 2 exhibits nine examples of ROIs cropped from DDSM mammograms, where Fig. 2a to c show the normal cases, Fig. 2d to f represent the benign cases and Fig. 2g to i reveal malign cases.

2.3. Dimensionality reduction and feature extraction methods. Principal component analysis (PCA) is a technique that reduces the dimensionality of data set with minimum loss of variance information [16–18]. Given a set of input vectors x_i represented in vectorized form ROIs $\left(\boldsymbol{x} \in \mathbb{R}^D, D = H \times W\right)$ $\left(i = 1, 2, \ldots, N\right)$, where H and W are height and width of ROIs respectively. First, the mean and covariance can be estimated as follows

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}, \tag{1}$$

$$\Sigma = \frac{1}{N-1} \sum_{i \in N} \left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)^T. \tag{2}$$

Then, the eigenvectors with the largest eigenvalues of the covariance matrix Σ are called the principal components of the dataset. Essentially, PCA learns a linear transformation $\boldsymbol{h}_i = \boldsymbol{W}^T x_i$ where $\boldsymbol{W}$ are just these leading eigenvectors. Consequently, while the output vectors $\boldsymbol{h}_i$ are often of significantly lower dimensionality than the original data, the variance information is largely preserved.

Nonnegative matrix factorization (NMF) is an alternative efficient way to find low rank representation of nonnegative data [19]. For a given $n \times m$ nonnegative data matrix $\boldsymbol{V}$ represented ROIs, where m is the number of data samples, NMF attempts to find $n \times r$ nonnegative factor matrix $\boldsymbol{W}$ and an $r \times m$ nonnegative factor matrix $\boldsymbol{H}$ so that

$$\boldsymbol{V} \approx \boldsymbol{WH}. \tag{3}$$

The columns of $\boldsymbol{W}$ are called basic vectors for the linear combination, and the columns of $\boldsymbol{H}$ are weighted components [20]. The objective of the NMF is to find the optimal nonnegative $\boldsymbol{W}$ and $\boldsymbol{H}$ by solving e.g. the following minimization problem

$$\arg \min_{\boldsymbol{W} \geq 0, \boldsymbol{H} \geq 0} \left\| \boldsymbol{V} - \boldsymbol{WH} \right\|_F^2. \tag{4}$$

The nonnegativity constraints of $\boldsymbol{W}$ and $\boldsymbol{H}$ demonstrate that the combinations of $\boldsymbol{W}$ and $\boldsymbol{H}$ are only additive. Therefore, the columns of $\boldsymbol{W}$ can be seen as different parts of images that form a full image with corresponding coefficients given by $\boldsymbol{H}$. These important features of NMF are called parts-based representations [21]. It is possible to obtain two nonnegative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ by solving (4), where each columns of matrix $\boldsymbol{W}$ contain basis vectors, and each columns of matrix $\boldsymbol{H}$ contain the weight vectors. As to PCA, matrix $\boldsymbol{W}$ is the eigenvectors matrix and matrix $\boldsymbol{H}$ is the eigenprojections matrix [22].

Spatial-temporal discriminant analysis (STDA) is an alternative effective approach to reduce the dimension of multiway data samples by incorporating discriminative analysis [23–25]. Whereas PCA and NMF are designed for vector data, STDA presents a novel idea that can handle high dimensional tensor data. In this paper, STDA learns two projection matrices from original ROIs instead of vectorized features, which allows us to reduce the dimensionality of ROIs more effectively [23]. The between-class scatter matrix and within-class scatter matrix can be computed as follows

$$\boldsymbol{S}_B^j = \sum_{k \in K} N_k \left(\boldsymbol{Y}_k^j - \bar{\boldsymbol{Y}}^j\right)\left(\boldsymbol{Y}_k^j - \bar{\boldsymbol{Y}}^j\right)^T, \tag{5}$$

$$\boldsymbol{S}_W^j = \sum_{k \in K} \sum_{i \in \prod_k} \left(\boldsymbol{Y}_{k,i}^j - \bar{\boldsymbol{Y}}^j\right)\left(\boldsymbol{Y}_{k,i}^j - \bar{\boldsymbol{Y}}^j\right)^T, \tag{6}$$

where N_k is the number of kth class, $\boldsymbol{Y}_{k,i}^j$ is projection of ith sample in class k, $\bar{\boldsymbol{Y}}_k^j$ is the mean of $\bar{\boldsymbol{Y}}_{k,i}$ over i, $\bar{\boldsymbol{Y}}^j$ is the mean of all the projected samples, the symbol j denotes the jth way of samples, and in our case $j \in \{1, 2\}$. Therefore, the projection matrix for the jth way is obtained by solving the following optimization problem

$$\tilde{\boldsymbol{W}}_j = \arg \max_{\boldsymbol{w}_j} \frac{tr\left(\boldsymbol{W}_j^T \boldsymbol{S}_B^j \boldsymbol{W}_j\right)}{tr\left(\boldsymbol{W}_j^T \boldsymbol{S}_w^j \boldsymbol{W}_j\right)}. \tag{7}$$

Problem (7) is a generalized eigenvalue problem

$$\boldsymbol{S}_B^j \boldsymbol{W}_j = \boldsymbol{S}_W^j \boldsymbol{W}_j \boldsymbol{\Lambda}_j, \tag{8}$$

and the projection matrix $\tilde{\boldsymbol{W}}_j$ contains the L eigenvectors associated with the largest L generalized eigenvalues.

2.4. Classification methods. Linear discriminant analysis (LDA) is a basic but quite efficient supervised machine learning method for classification, also known as Fisher discriminant analysis (FDA) [23]. Supposed we are given a set of data vectors $\boldsymbol{x} \in \mathbb{R}^N$, and N equals the multiplication of ROIs' height and width, whereby the mean and covariance matrices of each class can be estimated as [26]

$$\mu_k = \frac{1}{m_k} \sum_{\boldsymbol{x}_k \in \prod_k} \boldsymbol{x}_k, \tag{9}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{k}} = \frac{1}{m_k} \sum_{\boldsymbol{x}_k \in \prod_k} (\boldsymbol{x} - \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)^T, \tag{10}$$

and $\prod_k$ indicates kth class in database, m_k represents the number of samples in class $\prod_k$.

In this work, ROIs are divided into normal and abnormal cases, where abnormal cases contain malignant and benign subclasses. The corresponding covariance matrix is computed by

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{m_k}{m} \boldsymbol{\Sigma}_k, \tag{11}$$

where m denotes the number of samples, K represents the number of classes. Finally the projection matrix can be computed as [23]

$$\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{12}$$

With the transformation matrix $\boldsymbol{W}$, LDA transforms the input samples onto the lower-dimensional vector space. In this vector space, the ratio of between-class distance to the within-class distance will be maximized, so that guarantee maximal discrimination [26].

Random forests (RaF) is an ensemble learning method [27], composed of many unpruned decision trees. These decision trees grow by the bootstrap samples of the data independently. Each node is generated by choosing the best split among all of the samples. Therefore an orthogonal hyperplane is conducted to split the samples. RaF is robust with respect to noise due to it splitting each node by randomly selecting features. In this case, decision of each sample is made by voting between these trees.

RaF method for classification can be summarized as follows:

a. Draw bootstrap samples and grow a corresponding unpruned classification tree.
b. At each node, randomly sample a subset of features to split.
c. Choose the tree with the maximum votes as prediction.

Another fundamental classification method is kNN, which is especially efficient when no prior knowledge about the distribution of input data is available [28]. In the feature spaces, for each training sample $\boldsymbol{x}$, points belonging to the same class will form a subspace. For a test sample, the nearest k points will decide its class. Usually, the distance is measured by the Euclidean distance defined as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_F^2. \tag{13}$$

Finally, the predicted class of each sample depends on the class that is most common among the nearest k samples. kNN is able to classify samples without any assumptions about the characteristics of them, therefore the training process's cost of kNN is relatively easy and fast.

2.5. Deep learning methods. Convolutional neural networks (CNN) were first introduced in [29], and achieved promising performance for hand-written digits at that time. They regained their popularity due to the competitive results for ImageNet 2012 competition [6]. CNN is a multi-layer artificial neural network consisting of convolutional layers, pooling layers and fully connected layers. The jth output feature maps $\boldsymbol{S}_j$ of convolutional layers are computed as

$$\boldsymbol{S}_j = f\left(\sum_{i=1}^{N} \boldsymbol{I}_i * \boldsymbol{K}_{i,j} + b_j\right), \tag{14}$$

where $\boldsymbol{I}_i$ and $\boldsymbol{S}_j$ are the ith input feature maps and the jth output feature maps respectively, $*$ is the convolution operation and b_j is the jth bias added to each element of the convolution output, the convolutional kernel matrices $\boldsymbol{K}$ are small square matrices, working over the input feature maps as filters. The sum of convolution output and the bias is activated by taking element-wise nonlinear activation function f. Then the maximum pooling is adopted for dimensionality reduction after activation process. As illustrated in Fig. 3, with 10 layers convolution and max pooling, the last feature maps are vectorized and then constructed three fully-connected layers. In order to reduce the test errors and avoid overfitting, dropout technique is exploited to the covolutional and fully-connected layers with dropout rate p [6, 30]. With labels information incorporated, the cross-entropy error is minimized by adopting stochastic gradient descent (SGD) algorithm.

Autoencoder (AE), an alternative unsupervised deep learning method, can generate a new representation from the input vector $\boldsymbol{x}$ by defining an encoder function f_θ. For every input vectorized sample $\hat{\boldsymbol{x}}^{(m)}$ from data set $\boldsymbol{X} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$, the representation has the form of

$$\boldsymbol{h}^{(m)} = f_\theta\left(\boldsymbol{x}^{(m)}\right), \tag{15}$$

where $\boldsymbol{h}^{(m)}$ is the representation of the mth input $\boldsymbol{x}^{(m)}$. AE can make a reconstruction $\hat{\boldsymbol{x}} = g_\theta(\boldsymbol{h})$ from representation $\boldsymbol{h}$ by im-

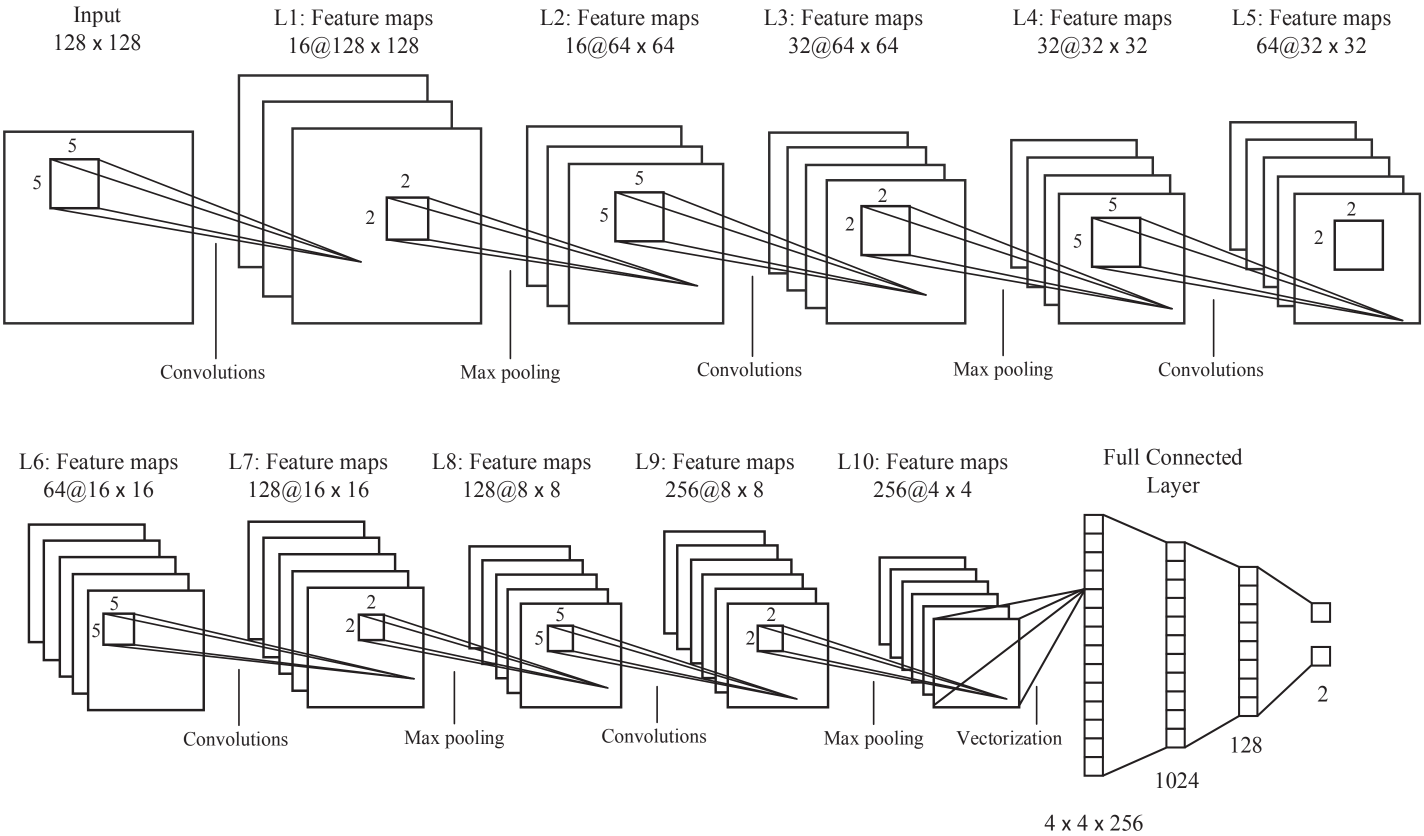

Fig. 3. The optimized detailed architecture of the CNN model evaluated in this study. The connection between L5 and L6 was the Max pooling operation, which hasn't been shown in the figure due to space issues

plementing the decoder function g_θ. In order to evaluate the quality of reconstruction, a reconstruction error $L(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is used to measure the similarity between input $\boldsymbol{x}$ and its reconstruction $\hat{\boldsymbol{x}}$. An optimal representation defined can be found by minimizing the reconstruction error

$$\mathscr{J}_{AE}(\boldsymbol{\theta}) = \sum_m L\left(\boldsymbol{x}^m, g_\theta\left(f_\theta\left(\boldsymbol{x}^{(m)}\right)\right)\right). \tag{16}$$

In order to extract more abstract features from ROIs, the network is stacked and trained by the greedy-wise strategy [31] (see Fig. 4). Sparsity constraints could provide a high-dimensional representation that improve the likelihood so that ROIs categories would be more separable [32]. A deep sparse architecture called Stacked Sparse AutoEncoder (SSAE) can be built, for which the objective function is defined as

$$\mathscr{J}_{SSAE}(\boldsymbol{\theta}) = \mathscr{J}_{AE}(\boldsymbol{\theta}) + \beta \sum_i KL(\rho \| \hat{\rho}_i), \tag{17}$$

where $\hat{\rho}_i$ is the activation value of the ith neuron, ρ is the desired value and KL means Kullback-Leibler divergence, a function to measure the difference between two probabilities. Two layers stacked autoencoder is illustrated in Fig. 4. After feed forward greedy layer-wise training, the output layer computes the loss of the full network with the label corresponding to the input vector $\boldsymbol{x}$.

For the purpose of classification, the network is finetuned with labels information. The last hidden layer of SSAE is linked to the softmax classifier with cross-entropy error function, and finally backpropagation (BP) strategy and SGD algorithm is exploiting to solve the gradients and optimize the parameters respectively.

3. Experimental setup

In clinical diagnosis, doctors usually first determine whether the patient contains mass tumors and then eventually determine whether the mass tumors are benign or malignant based on mammograms. At the same time the CAD systems may provide the second suggestions for doctors. Therefore, in this study, we mainly carry out two experiments. One is to detect whether the ROIs contain mass tumors and the other is to classify the mass tumor ROIs into two categories benign or malignant. In order to prove the effectiveness of each method, experiments are implemented with fivefold cross-validation. The ROIs were divided into five groups followed the standard routine for each category, and then every time one group was selected as the testing data set and the rest four groups were used as the training data set. The accuracy of diagnosis is an essential criterion of the effectiveness of the method because it reflects whether a method successfully distinguish between two different categories.

In order to compare the performance of different methods, we utilized the dimensionality reduction and classification methods, especially CNN, SSAE, PCA, NMF, STDA, LDA, RaF and kNN to extract significant features of ROIs and then distinguish the extracted features between the normal and abnormal case or benign and malignant cases. Note that the features extracted by deep learning methods: CNN and SSAE were classified by the softmax function at the last layer and they do not need to employ feature extraction techniques. After the

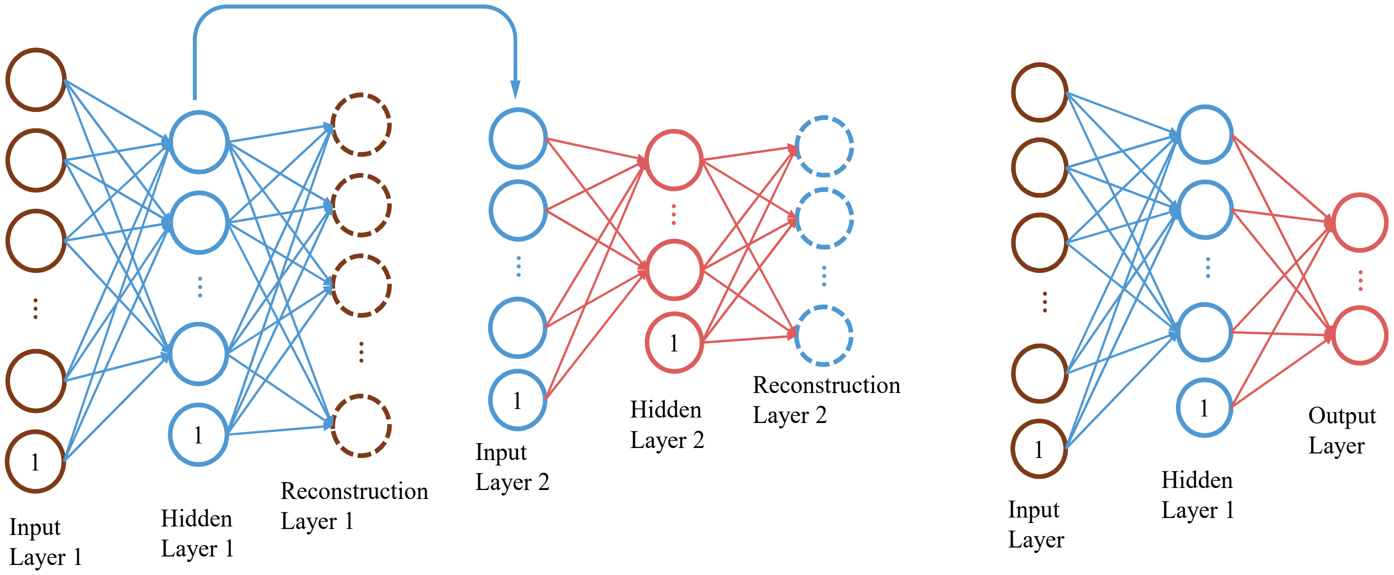

Fig. 4. Left: greedy layer-wise training. The hidden layer 1 and reconstruction layer 1 are the lower dimension representation and reconstruction of input layer 1 respectively. And then the hidden layer 1 will be the input layer 2 for training an even lower dimension representation. Right: two layer stacked autoencoder. Parameters and neurons in stacked autoencoder are pre-trained on the left

Table 2
Comparison of performance of various combinations of dimensionality reduction and classification methods for abnormality detection in DDSM case

Method Comparison	Number of Fold					
	1	2	3	4	5	Ave-acc
PCA + LDA	88.7	88.9	87.6	87.8	88.4	88.3
PCA + RaF	82.3	82.3	82.4	82.5	82.4	82.3
PCA + kNN	86.3	87.3	87.2	86.4	86.3	86.7
NMF + LDA	89.4	88.7	88.4	89.2	85.2	88.7
NMF + RaF	83.0	82.9	82.5	82.8	88.0	82.8
NMF + kNN	86.3	86.8	86.7	86.5	85.4	86.3
STDA + LDA	81.1	83.3	81.2	81.4	81.7	81.8
STDA + RaF	87.0	87.7	86.8	86.2	86.5	86.8
STDA + kNN	85.3	83.6	83.1	84.3	84.6	84.0
CNN	**93.8**	**92.7**	**93.2**	**92.4**	**92.8**	**93.0**
SSAE	92.7	91.5	91.3	91.2	90.8	91.5

Table 3
Comparison of performance of the same combination of dimensionality reduction and classification methods for malignancy detection in DDSM database

Method Comparison	Number of Fold					
	1	2	3	4	5	Ave-acc
PCA + LDA	56.6	56.1	58.6	50.5	52.8	54.9
PCA + RaF	55.6	52.1	55.9	56.3	55.6	55.1
PCA + kNN	53.1	48.9	54.1	53.8	50.3	52.0
NMF + LDA	56.6	56.6	55.4	57.3	56.8	56.5
NMF + RaF	56.4	55.9	55.6	56.3	52.0	55.2
NMF + kNN	54.1	52.9	56.6	51.0	50.3	53.0
STDA + LDA	59.1	56.4	53.9	56.3	56.3	56.4
STDA + RaF	58.1	53.1	57.9	55.8	54.3	55.8
STDA + kNN	53.4	50.4	53.6	52.5	48.8	51.7
CNN	**63.4**	**62.8**	**62.6**	**61.5**	**62.4**	**62.5**
SSAE	56.2	57.5	58.9	56.9	55.8	57.1

combinations of different dimensionality reduction and classification methods, our experiments provided a total of 11 classification methods (see Table 2 and Table 3).

Some important hyperparameters may greatly influence the performance of the compared methods, therefore it is necessary to give detailed descriptions for them. The optimal structure of CNN is in Fig. 3, we apply our convolutions with 5×5 filters and a stride of one, a zero padding of size two and dropout rate $p = 0.5$. Three layers SSAE with 16384 vectorized inputs, 4096 and 512 hidden features, two class outputs (denoted by 16384–4096–512–2), $\beta = 10^{-2}$ and $\rho =$ is constructed. For PCA, the number of components is selected based on the cumulative percentage of total variation. In this study, the cumulative percentage is set to be 90%. As for NMF, the optimal rank is selected to be 30. The required number of eigenvectors of STDA is set to be 10. As regards to kNN, k is chosen as 5. And concerning about RaF, the choice of k is equals to the square root of the number of samples.

All of the experiments were carried out in MATLAB 2016a on a PC equipped with an Intel i7–5830k, 128GB of RAM and four NVIDIA Titan X (Maxwell) Graphics Processor Units.

4. Results and Discussion

Table 2 and Table 3 demonstrate the accuracy for abnormality and malignancy classification respectively for the various combinations of dimensionality reduction and classification technique in fivefold cross-validation experiments.

From Table 2 and Table 3, we can see that CNN yields the highest average accuracy of 93.0% and 62.5% for abnormality and malignancy classification respectively, while SSAE is second only to CNN. Unlike PCA, NMF and STDA, deep learning methods CNN and SSAE can be stacked to form a deeper and more abstract architectures to generate a higher-level and well separable features which provides high efficiency for abnormality and malignancy detection for large amounts of ROIs. However, fitting a deep learning model with a large number of parameters is very time-consuming. Moreover, such models can be easy to be over-fitting in case when the number of training samples are insufficient. Therefore, in order to achieve reliable results in clinical breast cancer prediction, we should also consider alternatives methods to CNN and SSAE. Table 2 and Table 3 have demonstrated that NMF + LDA method has achieved 88.7% and 56.5% classification accuracy in two classification tasks respectively. It is probably because that the part-based features of ROIs learned by NMF are more distinguishable for LDA classifier.

As Table 3 showed that the average accuracy of malignancy classification of all methods are relatively low. This is probably because we only used the intensity features of ROIs in this study. Therefore, in order to improve the malignancy classification performance, some more features such as the shape and margin of mass could be incorporated in the feature.

5. Conclusions

Most of the recent studies have discussed different methods in various mammography databases. However, till now comparison of these state-of-art methods in one large database was not provided. This study has conducted a comprehensive comparative study of different state-of-art machine learning methods employed for DDSM database – the largest publicly available mammography database. The extensive computer simulations results have shown that deep learning methods could learn lower dimensional and higher-level features of ROIs, and therefore achieve the best classification accuracy. Further more, the part-based learning features of ROIs also provides promising classification results. However, the average accuracy of malignancy classification of all methods are quite low, which implies that more features and more sophisticated methods should be considered.

Acknowledges. This work was partially supported by the MES Russian Federation grant 14.756.31.0001.

References

[1] C. DeSantis, J. Ma, L. Bryan, et al., "Breast cancer statistics, 2013", *CA: A Cancer Journal for Clinicians* 64 (1), 52–62 (2014).

[2] R.L. Siegel, K.D. Miller, and A. Jemal, "Cancer statistics, 2018", *CA: A Cancer Journal for Clinicians* 68 (1), 7–30 (2018).

[3] S. Yoon and S. Kim, "Adaboost-based multiple svm-rfe for classification of mammograms in ddsm", *BMC Medical Informatics and Decision Making* 9 (1), S1 (2009).

[4] S. Dhahbi, W. Barhoumi, and E. Zagrouba, "Breast cancer diagnosis in digitized mammograms using curvelet moments", *Computers in Biology and Medicine* 64, 79–90 (2015).

[5] A.M. Abdel-Zaher and A.M. Eldeib, "Breast cancer classification using deep belief networks", *Expert Systems with Applications* 46, 139–144 (2016).

[6] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, 2012, 1097–1105.

[7] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, et al., "Convo- lutional neural networks for speech recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (10), 1533–1545 (2014).

[8] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection", *Advances in Neural Information Process- ing Systems*, 2013, 2553–2561.

[9] F.A. Spanhol, L.S. Oliveira, C. Petitjean, et al., "Breast cancer histopathological image classification using convolutional neural networks", *Neural Networks (IJCNN), 2016 International Joint Conference on*, IEEE, 2016, 2560–2567.

[10] A. Dubrovina, P. Kisilev, B. Ginsburg, et al., "Computational mammography using deep neural networks", *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6 (3), 1–5 (2016).

[11] K. Sharma and B. Preet, "Classification of mammogram images by using cnn classifier", *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, IEEE, 2016, 2743–2749.

[12] J. Kurek, B. Swiderski, S. Osowski, et al., "Deep learning versus classical neural approach to mammogram recognition", *Bulletin of the Polish Academy of Sciences* (Accepted).

[13] J. Xu, L. Xiang, Q. Liu, et al., "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology im- ages", *IEEE Transactions on Medical Imaging* 35 (1), 119–130 (2016).

[14] M.A. Mazurowski, J.Y. Lo, B.P. Harrawood, et al., "Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis", *Journal of Biomedical Informatics* 44 (5), 815–823 (2011).

[15] M. Heath, K. Bowyer, D. Kopans, et al., "The digital database for screening mammography", *Digital Mammography* 431–434 (2000).

[16] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis", *Chemometrics and Intelligent Laboratory Systems* 2 (1–3), 37–52 (1987).

[17] G. Zhou, A. Cichocki, and D.P. Mandic, "Common compo- nents analysis via linked blind source separation", *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE Interna- tional Conference on*, IEEE, 2015, 2150–2154.

[18] G. Zhou, A. Cichocki, Y. Zhang, et al., "Group component analysis for multiblock data: Common and individual feature extraction", *IEEE Transactions on Neural Networks and Learning Systems* 27 (11), 2426–2439 (2016).

[19] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization", *Neural Computation* 19 (10), 2756–2779 (2007).

[20] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization", *Advances in Neural Information Processing Systems*, 2001, 556–562.

[21] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature* 401 (6755), 788 (1999).

[22] D. Guillamet, B. Schiele, and J. Vitria, "Analyzing non-negative matrix factorization for image classification", *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, IEEE, 2002, vol. 2, 116–119.

[23] Y. Zhang, G. Zhou, Q. Zhao, et al., "Spatial-temporal discriminant analysis for erp-based brain-computer interface", *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 21 (2), 233–243 (2013).

[24] Y. Zhang, G. Zhou, J. Jin, et al., "Sparse bayesian classification of eeg for brain–computer interface", *IEEE Transactions on Neural Tetworks and Learning Systems* 27 (11), 2256–2267 (2016).

[25] Y. Zhang, C.S. Nam, G. Zhou, et al., "Temporally constrained sparse group spatial patterns for motor imagery BCI", *IEEE Transactions on Cybernetics* (Accepted).

[26] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis", *Advances in Neural Information Processing Systems*, 2004, 1569–1576.

[27] L. Zhang and P.N. Suganthan, "Random forests with ensemble of feature spaces", *Pattern Recognition* 47 (10), 3429–3437 (2014).

[28] L.E. Peterson, "K-nearest neighbor", *Scholarpedia* 4 (2), 1883 (2009).

[29] Y. LeCun, B. Boser, J.S. Denker, et al., "Backpropagation applied to handwritten zip code recognition", *Neural Computation* 1 (4), 541–551 (1989).

[30] N. Srivastava, G. Hinton, A. Krizhevsky, et al., "Dropout: A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research* 15 (1), 1929–1958 (2014).

[31] Y. Bengio, P. Lamblin, D. Popovici, et al., "Greedy layer-wise training of deep networks", *Advances in Neural Information Processing Systems*, 2007, 153–160.

[32] C. Poultney, S. Chopra, Y. L. Cun, et al., "Efficient learning of sparse representations with an energy-based model", *Advances in Neural Information Processing Systems*, 2007, 1137–1144.