This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Deep Manifold Preserving Autoencoder for Classifying Breast Cancer Histopathological Images

Yangqin Feng, Lei Zhang, *Senior Member, IEEE*, Juan Mo

Abstract

Classifying breast cancer histopathological images automatically is an important task in computer assisted pathology analysis. However, extracting informative and non-redundant features for histopathological image classification is challenging due to the appearance variability caused by the heterogeneity of the disease, the tissue preparation, and staining processes. In this paper, we propose a new feature extractor, called deep manifold preserving autoencoder, to learn discriminative features from unlabeled data. Then, we integrate the proposed feature extractor with a softmax classifier to classify breast cancer histopathology images. Specifically, it learns hierarchal features from unlabeled image patches by minimizing the distance between its input and output, and simultaneously preserving the geometric structure of the whole input data set. After the unsupervised training, we connect the encoder layers of the trained deep manifold preserving autoencoder with a softmax classifier to construct a cascade model and fine-tune this deep neural network with labeled training data. The proposed method learns discriminative features by preserving the structure of the input datasets from the manifold learning view and minimizing reconstruction error from the deep learning view from a large amount of unlabeled data. Extensive experiments on the public breast cancer dataset (BreaKHis) demonstrate the effectiveness of the proposed method.

Index Terms—Histopathological image classification, breast cancer diagnose, manifold learning, autoencoder, deep neural networks.

1 Introduction

Breast cancer is the most common invasive cancer in females worldwide [1]. Finding breast cancer early and getting state-of-the-art cancer treatment are the key strategies to prevent deaths from breast cancer [2]. During past decades, it is a widely-used way to diagnose breast cancer by analyzing hematoxylin and eosin (H&E) stained histological slide preparations that are examined under a high powered microscope of the affected area of the breast. In clinical practice, classification of breast cancer biopsy result into different patterns (e.g. cancerous and non-cancerous) is manually assessed by experienced pathologists. Clearly, such an expert-involved way is exhaustive and time-consuming. Fortunately, emerging machine learning techniques and growing image volume make building automatic system for breast cancer classification possible and can help pathologists to obtain precise diagnosis more efficient [3].

Due to the large size of the histopathological images, the existing traditional machine learning methods and deep neural network models for directly analyzing those full-size histopathological images will result in very complex architectures which involve larger sets of parameters and are considered hard to train. Therefore, the strategies that are relying on the segmentation of cell nuclei, and then extracting features on the segmented cell nuclei is popular for breast cancer analysis over the past few decades. For instance, In [4], Kowal *et al.* applied four clustering algorithms in the color space along with adaptive thresholding in gray-scale for nuclei segmentation. Then, the morphological, topological and texture features are extracted for classification procedure with three different classifiers in the segmented nuclei. In [5], George *et al.* proposed an automated method for cell nuclei detection and segmentation for breast cytological images and developed a breast cancer classification system. The (12-dimensional) features about the shape and the texture of the detected nuclei are extracted for different supervised classification algorithms, e.g., learning vector quantization (LVQ), support vector machine (SVM), and multilayer perceptron (MLP). In [6] Filipczuk *et al.* presented a breast cancer diagnostic system for cytological images of fine needle biopsy. The breast cancer diagnostic system can classify an image as benign or malignant by extracting 25 features for each image and discriminating these features by four different classifiers, i.e., the k-nearest neighbor classifier (k-NN), naive Bayes classifier (NB), decision tree (DT), and SVM. However, breast cancer histopathology images are generally with high-resolution. The performance of the above breast cancer classification methods depends primarily on the appropriate representation of input data (morphological, topological and texture), such that many efforts are dedicated to feature engineering.

In recent years, deep learning (DL) architectures such

• *The authors are with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, 610065 Chengdu, P. R. China. Corresponding author: Lei Zhang (E-mail: leizhang@scu.edu.cn).*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

as deep neural networks [7], deep belief networks [8] and recurrent neural networks [9], [10] have been successfully applied to the fields of image classification, speech recognition, and natural language processing [11]. In the area of image classification, AlexNet [7], VGGNet [12] and GoogLeNet [13] are representative methods. AlexNet [7] was a large convolutional neural network which competed in the ImageNet Large-Scale Visual Recognition Challenge 2012. It trained to classify the 1.2 million high-resolution images into the 1000 different classes. VGGNet [12] also trained on the large-scale ImageNet dataset. It can be seen as a much deeper AlexNet. GoogLeNet [13] was the most effective method for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014. The above DL based model have shown great accomplishments in image classification tasks. However, they are all trained on large amounts of labeled data. Although the breast cancer database are much smaller than the database which mentioned above. Attempts have been made in the medical imaging domain after inspiration by the success of DL based methods in image classification. For example, [14] investigated the ability of CNN model to four distinct medical imaging applications in three specialties (radiology, cardiology, and gastroenterology) involving classification, detection, and segmentation by transfer learning.

In histopathological image classification, the number of the training samples is small and the size of the images is large. This makes the training of the deep model classifiers to be difficult or even impossible. The patch-based technique is a good alternative that can be used to solve this problem. It is robust to spatial transformation, scale variation, and cluttered background, and it has achieved a large number of successes in object classification and detection in medical image analysis field. For instance, to obtain a better representation of input data, patch-based deep learning methods have been proposed for breast cancer classification [15], [16]. Spanhol et al. used a convolutional neural network (CNN) which is a variant based on the AlexNet, to exact the features and classify the breast cancer histopathological images as benign or malignant based on the image patches sampled from the breast cancer images. In [17], a patch-based deep convolutional neural network (DCNN) approach is developed to classify epithelial and stromal compartments in histopathological images. In [18], Chang et al. proposed a method that automatically learns a series of dictionary elements for representing the underlying spatial distribution using stacked predictive sparse decomposition. It aims to classify components of each histology section in terms of distinct phenotypes (e.g., tumor, stroma, necrosis) on patch level. In [19], a patch-based CNN for whole slide tissue image classification has been proposed. It is a two-level model, the first-level (patch-level) model is an expectation maximization (EM) method combined with a patch-level CNN, and the second-level (image-level) is a multiclass logistic regression or SVM.

In this paper, we propose a new feature exactor, called deep manifold preserving autoencoder, and use it to construct a deep neural network for classifying breast cancer histopathology images. The proposed deep manifold preserving autoencoder learns hierarchal features from unlabeled image patches by minimizing the reconstruction error of input patches, and simultaneously minimizing the difference of the similarity between neighbor samples in the input space and the similarity between their corresponding representations in the hidden spaces. Then, the encoder layers of the trained deep manifold preserving autoencoder are further connected with an additional classifier to develop a cascade model. Next, the supervised end-to-end fine-tuning is conducted to optimize the model with the patches from the labeled samples. Finally, we can use the trained deep neural network to predict the class labels of input images. The proposed feature exactor synthesizes the autoencoder and the manifold learning techniques, which makes full use of their advantages as well as effectively avoiding their disadvantages. Also, it firstly learns the informative and non-redundant features with plenty of patches sampled form the unlabeled images, and then it is fine-tuned using a small number of patches sampled labeled samples, which is suitable and useful for breast cancer image classification tasks. In the end, it is trained end-to-end and hence requires significantly less supervision during training, makes it more generally applicable in realistic settings.

The remainder of this paper is organized as follows: Section 2 introduces the preliminaries. Our cell nuclei classification method based on the proposed DMAE is presented in Section 3. Section 4 provides experimental results to illustrate the effectiveness of the proposed method. Section 5 concludes this paper.

Notations: Unless specified otherwise, **lower-case bold letters** represent column vectors and **upper-case bold letters** represent matrices and collections respectively. $\mathbf{M}^T$ denotes the transpose of the matrix $\mathbf{M}$. Table 1 summarizes the notations used throughout the paper.

2 PRELIMINARIES

A brief review of the general manifold learning algorithms and a summary of autoencoder will be given in this section.

TABLE 1
Summary of notations used in this paper

Notation	Definition
d	dimension of input samples
k	dimension of representations in the hidden space
n	number of data points
m	number of the layers of neural networks
l	index of current layer
λ	weight decay parameter
β	sparsity parameter
ξ	parameter for the J_M
ρ	average target activation of hidden units
$\hat{\rho}_j$	j-th hidden unit average activation of the input data set
s_{ij}	similarity between the data points $\mathbf{x}(i)$ and $\mathbf{x}(j)$
α	learning rate
$\|\cdot\|$	result of ℓ_2-norm
$\mathbf{x} \in \mathbb{R}^d$	an input data point
$\mathbf{h}$	representation for $\mathbf{x}$ in the hidden layers
$\mathbf{y} \in \mathbb{R}^d$	output of the network for $\mathbf{x}$
$\mathbf{W}$	weight matrix
$\mathbf{P}$	training set for the DNN model
$\mathbf{S}$	similarity matrix of the input data set
$\mathbf{X}$	A set of input data points
$\mathbf{H}$	hidden representations of the input data set $\mathbf{X}$
$\mathbf{t}$	ground truth collection of training set $\mathbf{P}$
$\mathbf{Y}$	output collection from the neural network

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

2.1 Manifold learning

In many tasks, the measured data vectors are high-dimensional and they are multiple, indirect measurements of an underlying source, which typically cannot be directly measured. But we believe that the high-dimensional data vectors lie near a lower-dimensional manifold. Therefore, many manifold learning methods have been proposed to discover the underlying manifold structure of the input data set over the past decades. For example, principal component analysis (PCA) [20] is a very popular method for dimensionality reduction. It aims to find a linear approximations to a given high-dimensional observation. Linear discriminant analysis (LDA) [21] aims to find a linear combination of features that characterizes or separates two or more classes of data. Isometric feature mapping (ISOMAP) [22] is a nonlinear dimensionality reduction method. It is used for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. Locally linear embedding (LLE) [23] is a method to address the nonlinear dimensionality reduction by computing low-dimensional, neighbourhood preserving embedding of high-dimensional data. Laplacian eigenmaps (LE) [24] tries to build a graph incorporating neighborhood information of the data set, then compute a low dimensional representation using the notion of the Laplacian of the graph. They have been widely applied to the area of computer vision and image processing [25]–[27]. The goal of these methods is to obtain more useful data representations and these manifold methods can be summarized with a general manner [28]. For a given data set $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)\}$, manifold learning methods aim to find an embedding $g(\mathbf{x}(i))$ of each data point $\mathbf{x}(i)$ by minimizing the following objective function:

$$\Phi = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} s_{ij} L(g(\mathbf{x}(i)), g(\mathbf{x}(j))),$$

where $L(\cdot)$ is the loss function between pairs of data points, $g(\mathbf{x}(i))$ is the embedding representation of the input data point $\mathbf{x}(i)$, and s_{ij} is the similarity between the data points $\mathbf{x}(i)$ and $\mathbf{x}(j)$. The similarity s_{ij} is often computed with Gaussian kernel $s_{ij} = e^{-\frac{\|\mathbf{x}(i)-\mathbf{x}(j)\|^2}{\tau}}$, ($\tau \in \mathbb{R}$). For certain objective functions, balancing constraints may be required in practice.

2.2 Autoencoder

An autoencoder is a symmetrical neural network, which learns features in an unsupervised manner by enforcing the outputs to be close to the inputs. An example of the architecture of autoencoder is shown in Fig. 1. Let $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)\}, x(i) \in \mathbb{R}^{d \times n}$ be a given unlabeled training set. The autoencoder encodes the inputs $\mathbf{X}$ into hidden representations $\mathbf{H}$ in a latent subspace, and then decodes the hidden representations back into the original space, also called reconstruction space. Features can be obtained in the learning process by minimizing the loss function between the inputs and the outputs.

For each input data point $\mathbf{x}(i)$, autoencoder first learns its hidden representation $\mathbf{h}(i) \in \mathbb{R}^l$ by using a linear mapping and a nonlinear activation function:

$$\mathbf{h}(i) = f(\mathbf{W}^{(1)}\mathbf{x}(i) + \mathbf{b}^{(1)}), \quad (1)$$

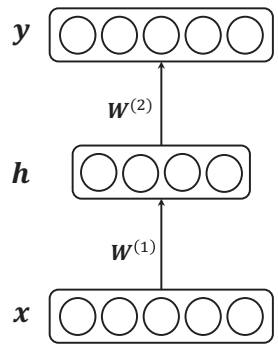


Fig. 1. An example of the architecture of a traditional autoencoder. It contains two parts: the encoder and decoder. The notation $\mathbf{x}$ is the input, $\mathbf{y}$ is the reconstruction of $\mathbf{x}$ by minimizing the distance between $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{h}$ is the representation of $\mathbf{x}$ in the hidden layer.

where $W^{(1)} \in \mathbb{R}^{l \times d}$ is the weight matrix between input layer and the hidden layer, $\mathbf{b}^{(1)} \in \mathbb{R}^l$ is the corresponding bias vector, and $f(\cdot)$ is the nonlinear activation function. After mapping the input $\mathbf{x}(i)$ to the hidden representation $\mathbf{h}(i)$, the output is calculated as:

$$\mathbf{y}(i) = f(\mathbf{W}^{(2)}\mathbf{h}(i) + \mathbf{b}^{(2)}), \quad (2)$$

where $\mathbf{W}^{(2)} \in \mathbb{R}^{d \times l}$ is the weight matrix between the hidden layer and the output layer, $\mathbf{b}^{(2)} \in \mathbb{R}^d$ is the corresponding bias vector, and $f(\cdot)$ is the nonlinear activation function. The features implied in the training data set $\mathbf{X}$ can be learned by minimizing the reconstruction error of the following cost function:

$$J_T(\mathbf{W}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{x}(i), \mathbf{y}(i)) + \frac{\lambda}{2} \|\mathbf{W}\|_2^2,$$

where the first term is the likelihood function, which can be a (one-half) squared error cost function $L(\mathbf{x}(i), \mathbf{y}(i)) = \frac{1}{2}\|\mathbf{x}(i) - \mathbf{y}(i)\|^2$, λ is a trade-off parameter, and the second term is a regularization term that tends to decrease the magnitude of the weights and helps prevent overfitting. It can discover more interesting structures by imposing other constraints on the above optimization problem. In particular, it often imposes a sparsity constraint on the hidden units by adding an additional penalty term to the optimization objective, and develops the sparse autoencoder [29], which is to minimize the reconstruction error with a sparsity constraint:

$$J_T(\mathbf{W}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{x}(i), \mathbf{y}(i)) + \frac{\lambda}{2} \|\mathbf{W}\|_2^2 + \beta \sum_{j=1}^{n} KL(\rho \| \hat{\rho}_j),$$

where β is the sparsity parameter, ρ is the target average activation of $\mathbf{h}$, and $\hat{\rho}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_j(i)$ is the average activation of the j-th hidden unit over the n training data. $KL(\rho \| \hat{\rho}_j)$ is the Kullback-Leibler divergence [30], which is define as $KL(\rho \| \hat{\rho}_j) = \rho log \frac{\rho}{\hat{\rho}_j} + (1-\rho) log \frac{1-\rho}{1-\hat{\rho}_j}$.

3 The Proposed Approach

In this section, we fist present our proposed deep manifold persevering autoencoder (DMAE), and then illustrate how to apply it to construct an end-to-end deep neural network for breast histopathology image classification.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

3.1 The proposed deep architecture for image classification

A patch-based deep learning method is proposed for breast histopathology image classification with the features learned from the DMAE. The features obtained from the breast histopathology images are identified in the deep model with the sampling patches. These sampled patches are labeled with the initial labels of the images. The procedure of the proposed method is shown in Fig. 2. The proposed framework contains two major stages. In the first stage, we construct an end-to-end deep neural network and train it using the extracted patches from the training histopathology images. In the second stage, we try to classify the new sampled patches of the test images and combine the classification results of these patches to predict the label of the test images.

3.2 DNN for breast histopathology image classification

3.2.1 Deep manifold persevering autoencoder

We first introduce our proposed manifold persevering autoencoder (MAE), which is the basic cell for the DMAE. The architecture of our proposed MAE is shown in Fig. 3. Given a dataset $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)\}$, $\mathbf{x}(i) \in \mathbb{R}^d$, each $\mathbf{x}(i)$ is represented as $f(\mathbf{x}(i)) = \mathbf{h}(i) \in \mathbb{R}^k$ in the hidden layer when it passed through the encoder. To model the relation between input data, we set $\mathbf{S} \in \mathbb{R}^{n \times n}$ as a similarity matrix whose element s_{ij} represents the similarity between $\mathbf{x}(i)$ and $\mathbf{x}(j)$, and is calculated with Gaussian kernel $s_{ij} = e^{-\frac{\|\mathbf{x}(i)-\mathbf{x}(j)\|^2}{\tau}}$, $(\tau \in \mathbb{R})$.

It is desirable to persevere the manifold structure in the input data set from our MAE model, which involves an additional constraint to the traditional autoencoder. To achieve this goal, we expect that the similarity between the representations of the input data can be preserved as the similarity between the corresponding input data. In other words, if $\mathbf{x}(i)$ and $\mathbf{x}(j)$ are neighbors in the input space, then the distance between their representations ($\mathbf{h}(i)$ and $\mathbf{h}(j)$) should be short, and their distance can be measured by the (one-half) squared Euclidean distance between the representation, and here we use the function $L(\mathbf{h}(i), \mathbf{h}(j))$ to denote it. The corresponding optimization problem is to minimize the following cost function:

$$J_M(\mathbf{W}, \mathbf{b}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} L(\mathbf{h}(i), \mathbf{h}(j)).$$

To obtain informative and non-redundant features, we formulate our MAE model as a minimization problem of the following cost function:

$$J(\mathbf{W}, \mathbf{b}) = J_T(\mathbf{W}, \mathbf{b}) + \xi J_M(\mathbf{W}, \mathbf{b}), \tag{3}$$

where ξ is a trade-off parameter.

To solve the optimization problem in Eq. (3), we use the gradient descent method to update the parameters $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$, where $m \in \{1, 2\}$ for the single hidden layer MAE. The gradient of the objective function $J(\mathbf{W}, \mathbf{b})$ with respect to the parameters $\mathbf{W}^{(m)}, \mathbf{b}^{(m)}$, can be computed as follows:

$$\frac{\partial J(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}^{(m)}} = \frac{\partial J_T(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}^{(m)}} + \xi \frac{\partial J_M(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}^{(m)}}$$

We use the method in [31] to compute the gradient of the objective function $J_T(\mathbf{W}, b)$ and the gradient of the objective function $J_M(\mathbf{W}, b)$ is computed as follows:

$$\frac{\partial J_M(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}^{(m)}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \left(\delta_{ij}^{(m)} \mathbf{h}(i)^T + \delta_{ji}^{(m)} \mathbf{h}(j)^T \right)$$

and

$$\frac{\partial J_M(\mathbf{W}, \mathbf{b})}{\partial \mathbf{b}^{(m)}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \left(\delta_{ij}^{(m)} + \delta_{ji}^{(m)} \right),$$

where $\mathbf{h}(i)$ and $\mathbf{h}(j)$ are the representations of the input data $\mathbf{x}(i)$ and $\mathbf{x}(j)$, respectively. For all layers of the MAE, we can compute the $\delta_{ij}^{(m)}$ with the following equations:

$$\delta_{ij}^{(2)} = (\mathbf{h}(i) - \mathbf{h}(j)) \cdot f' \left(\mathbf{z}^{(2)}(i) \right),$$

$$\delta_{ij}^{(1)} = \left({\mathbf{W}^{(2)}}^T \delta_{ij}^{(2)} \right) \cdot f' \left(\mathbf{z}^{(1)}(i) \right),$$

where

$$\mathbf{z}^{(2)}(i) = \mathbf{W}^{(2)} \mathbf{h}(i) + \mathbf{b}^{(2)},$$

$$\mathbf{z}^{(1)}(i) = \mathbf{W}^{(1)} \mathbf{x}(i) + \mathbf{b}^{(1)}.$$

After obtaining the gradient of the cost function, $\mathbf{W}^{(m)}$ and $\mathbf{b}^{(m)}$ can be updated by using the following gradient descent algorithm until an acceptable approximation to the minimum is obtained or the max iterative number is reached:

$$\mathbf{W}^{(m)} = \mathbf{W}^{(m)} - \alpha \left(\frac{\partial J_T(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}^{(m)}} + \xi \frac{\partial J_M(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}^{(m)}} \right),$$

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m)} - \alpha \left(\frac{\partial J_T(\mathbf{W}, \mathbf{b})}{\partial \mathbf{b}^{(m)}} + \xi \frac{\partial J_M(\mathbf{W}, \mathbf{b})}{\partial \mathbf{b}^{(m)}} \right),$$

where α is the learning rate.

We used the proposed MAE as a basic cell to built the DMAE. In other words, the DMAE is a neural network consisting of multiple hidden layers by stacking a number of MAE. The architecture of DMAE and the procedure for obtaining the hidden representations is shown in Fig. 4. After training the first single hidden layer MAE, the hidden representation $\mathbf{h}^{(1)}$ is treated as the input to the successive MAE. Then, the second level representation $\mathbf{h}^{(2)}$ is learned by reconstructing the $\mathbf{h}^{(1)}$. Repeating the procedure, more hidden layers can be added into the DMAE model. The greedy layer-wise approach is employed to pre-train our DMAE, and the details are summarized in Algorithm 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

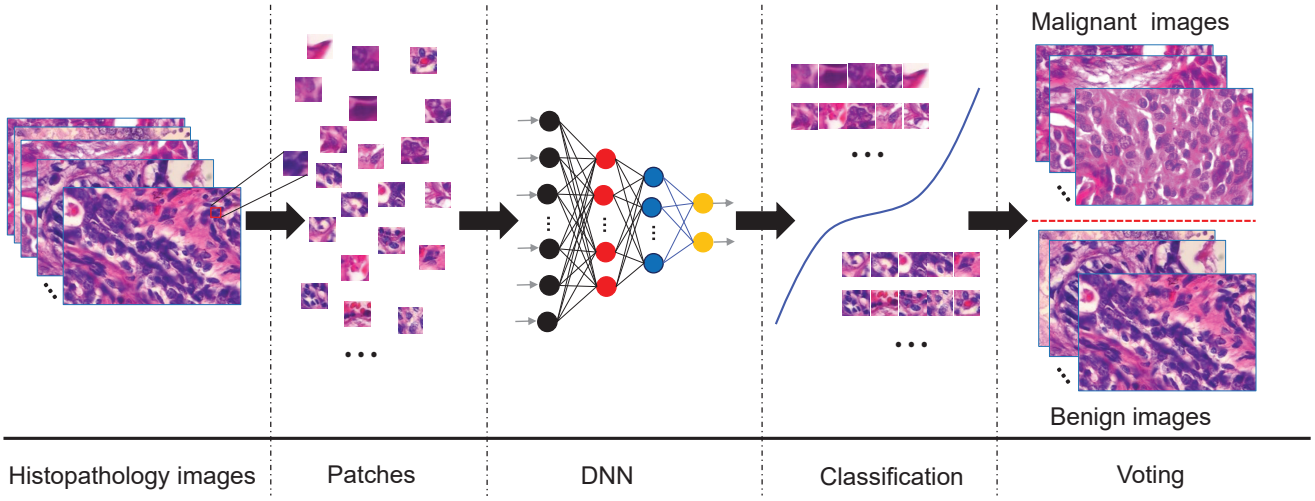


Fig. 2. The procedure of the proposed method for breast histopathology image classification. For the given breast cancer histopathology images, some image patches are sampled as the inputs to training the DNN model to classify the image patches. Then the model is applied to classify the new input histopathology image patches. Finally, combine all classification results for the whole images.

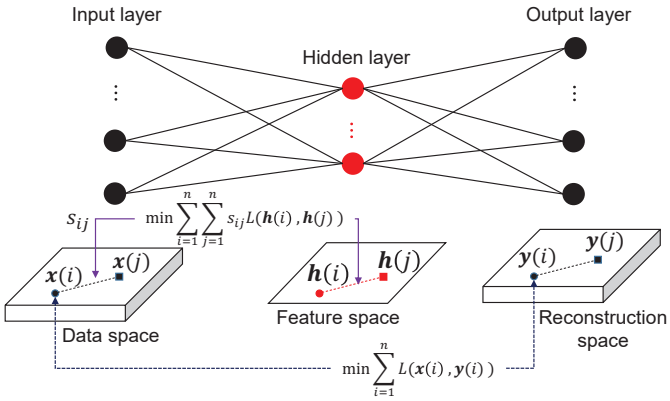


Fig. 3. Architecture of the MAE. For the breast cancer images patches, the MAE tries to learn the meaningful representations of the patches and preserve the manifold structure in the original space simultaneously.

3.2.2 *The architecture for the proposed DNN model*

After training the DMAE, encoder layers of the trained DMAE is employed to obtain representations for breast histopathology image patches. Then, an additional classifier layer capable of classifying the image patches as desired is employed in the last layer. The architecture of the final DNN model is shown in Fig. 5, where the encoder of the left DAME is used as the feature extractor in our DNN model. Finally, the parameters of the whole network are fine-tuned using a classical backpropagation (BP) algorithm with the labeled data. To increase the convergence rate, either the simple momentum method or more advanced optimization techniques, such as the L-BFGS or the conjugate gradient method, can be applied. Algorithm 2 summarizes the detailed procedure of the supervised end-to-end fine-tuning process for the DNN model.

3.3 Implementation details

We present the details of the nonlinear activation function and the initialization of the weight matrices and bias vectors in this subsection.

Algorithm 1 The training procedure for our DMAE

Input:

The training set: $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(n)\}$; learning rate: α; hyper parameter: λ; sparsity parameter: β; parameter for the J_M: ξ; and iterative number: I_t.

Output:

The weights and biases: $\left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\right\}_{l=1}^{L}$.

//Initialization

Initialize $\left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\right\}_{l=1}^{L}$ according to Equation (4).

Set $\mathbf{h}^{(0)}(i) = \mathbf{x}(i)$, $i \in [1, n]$.

//Greedy layer-wise approach pre-training DMAE

for each $l \in [1, L]$ **do**

  Set $\mathbf{x}^{(l)}(i) = \mathbf{h}^{(l-1)}(i)$.

  for each $t \in [1, I_t]$ **do**

    Do forward propagation to calculate $\mathbf{y}^{(l)}(i)$ according to Equation (1) and Equation (2).

    Solve the optimization problem in Equation (3) to compute $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$.

  end for

  Do forward propagation to obtain the representation $\mathbf{h}^{(l)}(i)$ for each $\mathbf{x}^{(l)}(i)$.

end for

return $\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L}$.

Activation Function: Many nonlinear activation functions are available to determine the node outputs of our deep neural network. In our experiments, the sigmoid function is used as the activation function. It is computed as follows:

$$sigm(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}}}.$$

Initialization: The initialization of all weight matrices and bias vectors are important to the gradient descent based method used in our deep neural networks. In our experiments, a simple normalized random initialization method was used [32], where the bias vectors $\mathbf{b}^{(l)}$ is initialized as $\mathbf{0}$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

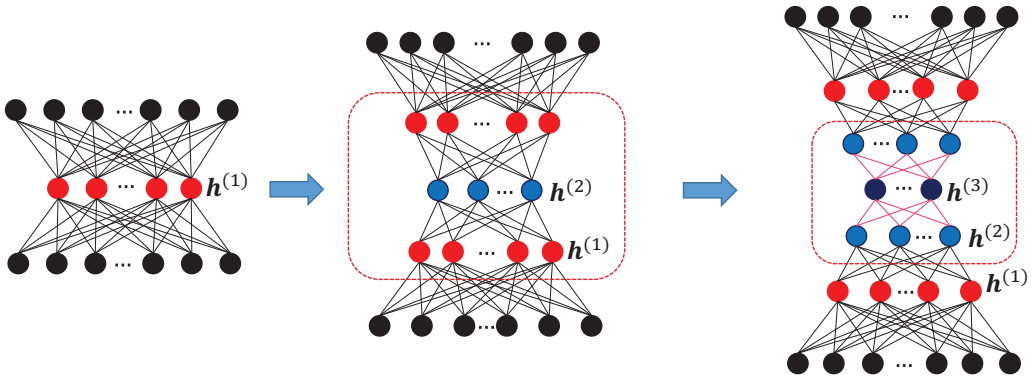


Fig. 4. The procedure of construct a DMAE. In the first step, there has a trained MAE. Then we add a new hidden layer between the encoder and decoder in the first MAE to construct a DMAE which the the three layers in the red box can be treated as another new MAE. In this manner, the DMAE which with more hidden layers can be built.

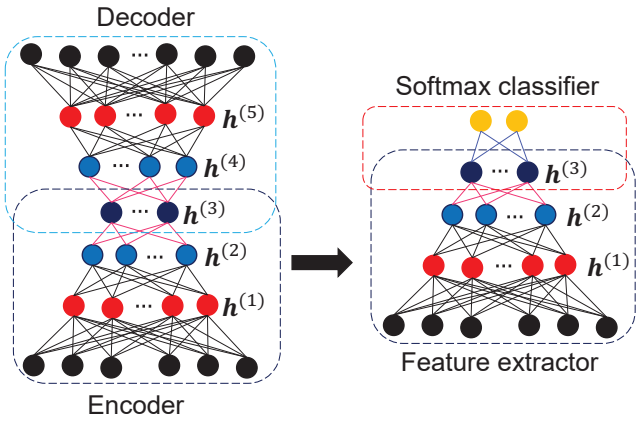


Fig. 5. The architecture of DMAE with five hidden layers at the left side and the architecture of the final end-to-end DNN model at the right side. The encoder of DMAE is used as the feature extractor in the DNN model.

and the weight matrices of each layer are initialized as the following uniform distribution:

$$\mathbf{W}^{(l)} \sim U\left[-\frac{\sqrt{6}}{\sqrt{d^{(l)}+d^{(l-1)}}}, \frac{\sqrt{6}}{\sqrt{d^{(l)}+d^{(l-1)}}}\right], \quad (4)$$

where $d^{(l)}$ is the dimension of the l-th layer, and $d^{(0)}$ is the dimension of input layer.

Algorithm 2 The training procedure for our DNN model

Input:

The training set: $\mathbf{P} = \{\mathbf{p}(1), \mathbf{p}(2), \ldots, \mathbf{p}(n)\}$; the training label: $\mathbf{t} = \{t(1), t(2), \ldots, t(n)\}$; learning rate: α; hyper parameter: λ; sparsity parameter: β; iterative number I_t.

Output:

The weights and biases: $\left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\right\}_{l=1}^{L+1}$.

//Initialization

Initialize $\left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\right\}_{l=1}^{L+1}$ according to the corresponding results gained from the trained DMAE.

Set $\mathbf{h}^{(L)}(i)$ as raw input to pre-train softmax classifier to obtain $\mathbf{W}^{(L+1)}$, $\mathbf{b}^{(L+1)}$.

Build a deep neural network with multi hidden layers form the trained DMAE and a final trained softmax classifier layer.

// Fine-tune all parameters

for each $t \in [1, I_t]$ **do**

  Set $\mathbf{x}^{(0)}(i) = \mathbf{p}(i), i \in [1, n]$.

  for each $l \in [1, L+1]$ **do**

    Do forward propagation.

  end for

  for each $l \in [1, L+1]$ **do**

    Fine-tune $\mathbf{W}^{(l)}$, $\mathbf{b}^{(l)}$ by minimizing cost function of softmax classifier.

  end for

end for

return $\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L+1}$.

4 Experiments

In this section, we conduct some experiments on a public breast cancer histopathological image database to evaluate the performance of our proposed method. Seven widely-used classification methods, i.e., 1-nearest neighbor classifier (1-NN) [33], support vector machine (SVM) [34], random forest classifier (RFC) [35], discriminant analysis classifier (DAC) [36], the convolutional neural networks (CNN) method employed in [15], multi-layer perceptron (MLP) [37] and stacked sparse autoencoder (SSAE) [38], [39], are compared in the experiments.

4.1 BreaKHis Database

The Breast Cancer Histopathological Image Classification database (BreakHis[1]) [40] contains images of two groups: benign breast tumors and malignant breast tumors. It was collected through a clinical study from January 2014 to December 2014. In BreakHis database, images were generated from breast tissue biopsy slides, stained with hematoxylin and eosin (H&E). The images were collected by surgical (open) biopsy (SOB), prepared for histological

1. The BreakHis database is downloaded from the website at http://web.inf.ufpr.br/vri/breast-cancer-database.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 2
Image distribution in the BreakHis database with different magnification factors and classes

Magnification	Benign	Malignant	Total
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total images	2480	5429	7909
Patient	24	58	82

TABLE 3
The sampling of patches from the imges

	Training set	Testing set
Benign image	50	30
Malignant image	30	30

study and labeled by pathologists of the P&D Lab, and they anonymized all the images.

The digitized images for the breast tissue biopsy slides were obtained by an Olympus BX-50 system microscope with a relay lens with a magnification of $3.3\times$ coupled to a Samsung digital color camera SCC-131AN. BreakHis database is composed of 7909 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors of $40\times$, $100\times$, $200\times$ and $400\times$, corresponding to objective lens of $4\times$, $10\times$, $20\times$ and $40\times$. It contains 2480 benign and 5429 malignant images (700×460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). Table 2 summarizes the image distribution in the BreakHis database. Fig. 6 shows some examples of the benign tumors and malignant tumors images with different magnification factors.

4.2 Experimental settings

The BreaKHis dataset has been divided into training (about 70%) and testing (about 30%) set according to the patient so that patients used to build the training set are not used for the testing set. This protocol was applied independently to each of the four magnifications that are available. The size of the histopathology images in the BreaKHis database is 700×460. The original images are reduced to 350×230 for $200\times$ and $400\times$ magnification factor. In the experiments, we sample patches in two different sizes (32×32 and 64×64). For each image, we obtain patches in a random manner. To overcome the imbalance of the benign and malignant images, we sample more patches for each benign image than malignant image. Table 3 summarizes the details of the strategy for sampling patches in each image. The parameters of $\alpha, \lambda, \beta, \xi$ and iterative number I_t for our method are equal to $1.0\times10^{-3}, 1.0\times10^{-3}, 1.0\times10^{-3}, 3.0\times10^{-3}$ and 3000, respectively. The parameter settings of the compared methods are selecting with the grid search strategy, and their best performance results are reported in this paper.

4.3 Performance Metrics

For each image, the voting strategy is applied for the final result. After training, the trained model has been used to classifying patches from the different images and get a result for each image patch, and then all the results of patches which from the same image vote and decide the final result for each image. The class receiving the highest vote is declared to be the predicted class for each image. After get the final results of each testing image, two popular performance metrics are used to report the results like that in [40] for the BreaKHis database. In the first case, the decision is patient-wise, therefore, the recognition rate is computed at the patient level. Let N_i be the number of cancer images of patient i. For each patient, if M_i cancer images are correctly classified, one can define a patient score (PS) as

$$PS(i) = \frac{M_i}{N_i},$$

and the global patient recognition rate (GPRR) is defined as

$$GPRR = \frac{\sum_{i=1}^{P_{all}} PS(i)}{P_{all}},$$

where P_{all} is the total number of patients.

In the second case, the recognition rate is computed at the image level (i.e. the patient information is not taken into account), thus providing a means to estimate solely the image classification accuracy of the DMAE models. The image recognition rate (IRR) is defined as

$$IRR = \frac{N_{rec}}{N_{all}},$$

where N_{all} is the number of cancer images of the test set and N_{rec} denotes number of images that are correctly classified.

4.4 Results and analysis

In the first experiment, we investigate the impact of a different number of hidden layers for feature extraction on the performance of the proposed method. To this end, we evaluate the performance by comparing our proposed method with a different number of hidden layers (i.e., 1-hidden layers, 2-hidden layers, 3-hidden layers, 4-hidden layers and 5-hidden layers). Once the feature representation was computed, the softmax classifier was employed to verify the performance of the different feature representations. We use two patch sizes for the experiment: 32×32 and 64×64. The dimensionality of the data was reduced by PCA with reserving 98% energy in the preprocessing. The aim of this experiment is to find out a good architecture for the breast cancer histopathology image classification. The result is shown in Fig. 7, from which we have following observations:

- The proposed method obtains a higher recognition rate for the images under $200\times$ magnifying factor;
- The proposed method with three hidden layers can perform well for the both image level, and the patient level indicated that the architecture with three hidden layers (totally with five layers) seems a suitable model for the four datasets. The recognition accuracy becomes higher with the increase of the number of hidden layer when it is less than three. However, when the number of the hidden layers is more than three, the recognition accuracy is not improved yet. This may be caused by the lack of sufficient image patches to fitting such a large number of parameters in the network;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

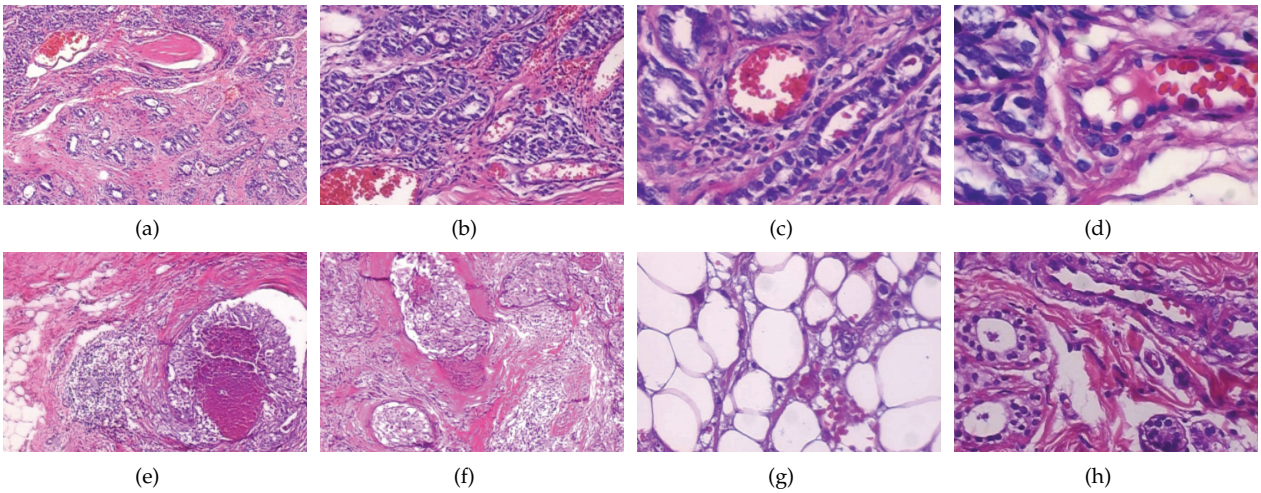


Fig. 6. Eight histopathology images of benign breast tumor and malignant breast tumor: (a) benign tumor with the magnification factors $40\times$. (b)benign tumor with the magnification factors $100\times$. (c) benign tumor with the magnification factors $200\times$. (d) benign tumor with the magnification factors $400\times$. (e) malignant tumor with the magnification factors $40\times$. (f) malignant tumor with the magnification factors $100\times$. (g) malignant tumor with the magnification factors $200\times$. (h) malignant tumor with the magnification factors $400\times$.

- The patch size 32×32 leads to higher recognition rate for $100\times$, $200\times$ and $400\times$ magnifying factor. Therefore, the patch size 32×32 is used in the following experiment.

In the second experiment, we compare the proposed DMAE with the SSAE [38], [39], which also can learn useful features from the unlabeled data. We use the same number of hidden layers, and the number of neurons in the hidden layers are set as same as our DMAE.

The results presented in this work are the mean accuracy and the standard deviation of 3-fold cross-validation. Table 5 shows the results of the network with SSAE and the softmax classifier (SSAE+SM), and the results of our method(DMAE+SM). From the results, we can find that:

- The SSAE+SM method also achieves promising performance both in image level and patient level. By comparing with the results of MLP in Tables 6 and 7, it illustrates that the unsupervised learning strategy can help the deep neural network to obtain better features for classification.
- The recognition rate of our method outperforms SSAE+SM in the image level and patient level on the four subsets, especially, the results on the magnification of $200\times$. It illustrates that the strategy of preserving the structure for the whole data set is significant for the classification of histopathological images.

In the third experiment, we compare the proposed method with six other methods, i.e., 1-NN [33], SVM [34], RFC [35], DAC [36], CNN [15], and MLP [37]. The first four methods are traditional machine learning methods, and the last two methods are neural networks. The employed CNN method in [15] was a variant based on the AlexNet. MLP is a feedforward neural network model that maps sets of input data onto a set of appropriate outputs, and here the used MLP has three hidden layers (five layers in total), and the each layer possess as many neurons as our method.

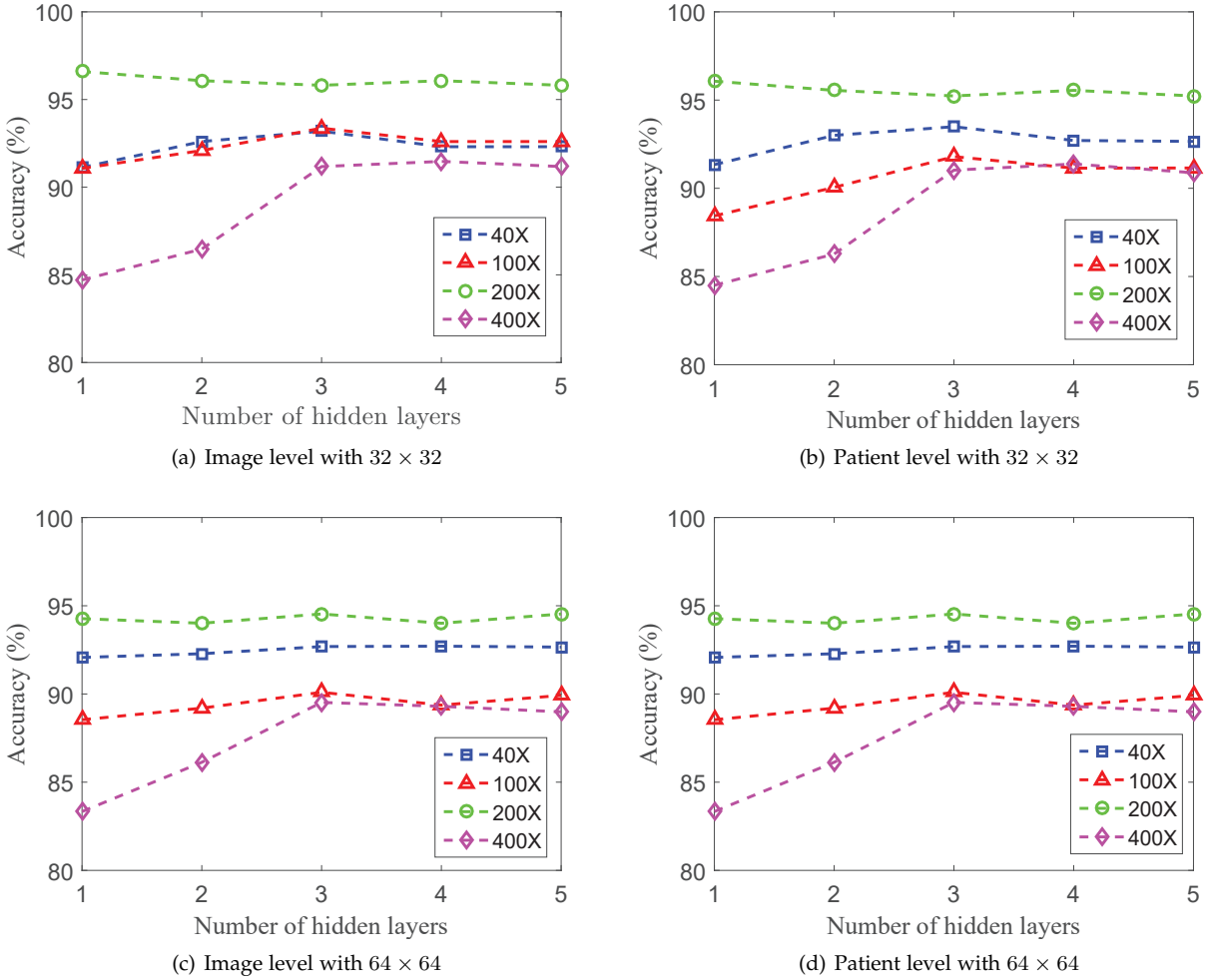
Tables 6 and 7 provide the classification results of the image level and the patient level on the four subsets, respectively. In order to get convincing results, the results presented in this work are the mean accuracy and the standard deviation of 3-fold cross-validation. For each trail, we picked up all images from 54 patients which randomly choose from the 82 patients as training set, and the all images for other 28 patients as testing set. From the results, we can obtain the following observations:

- The recognition rate of the images under $200\times$ magnifying factor performances best for the both image level and patient level;
- The proposed method outperforms all other examined classification algorithms on the four subsets. Specifically, the proposed method on the $200\times$ magnification factor is 3.01% higher than the best result reported by other methods that of MLP at the patient level, and 2.88% better than the best result reported by other methods at the image level;
- Our method outperforms the MLP which have the same architecture as our method means that our DMAE model can capture better feature representations than the traditional deep neural networks. It is mainly because the proposed method learns the informative and non-redundant features from plenty of the pathes from the unlabeled images.

5 Conclusion

In this paper, we propose a deep manifold preserving autoencoder, and then apply it to construct an end-to-end DNN model for breast cancer histopathology image classification. The proposed DMAE aims to to learn informative and non-redundant features from plenty of patches sampled from unlabeled histopathology images by minimizing the distance between the input patch and its output, and simultaneously preserving geometric structure of the whole input set of pathes. The proposed strategy makes DMAE benefits

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) Image level with 32×32 (b) Patient level with 32×32

(c) Image level with 64×64 (d) Patient level with 64×64

Fig. 7. Comparison of the classification accuracy of our proposed method with different number of hidden layers on four different databases.

TABLE 4
Comparison of the mean and standard deviation of classification accuracy (%) of our method (DMAE+SM) and SSAE+SM on four subsets in image level. The best mean results in different subsets are in bold.

Method	40×		100×		200×		400×	
	Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
SSAE	94.02 ± 1.28	82.34 ± 1.56	93.52 ± 1.01	81.45 ± 1.04	98.74 ± 0.35	82.38 ± 1.32	94.47 ± 1.21	82.32 ± 1.57
Ours	$\mathbf{94.43 \pm 1.02}$	$\mathbf{83.71 \pm 1.09}$	$\mathbf{93.15 \pm 0.89}$	$\mathbf{82.98 \pm 0.98}$	$\mathbf{99.36 \pm 0.20}$	$\mathbf{83.57 \pm 1.01}$	$\mathbf{94.86 \pm 1.03}$	$\mathbf{83.45 \pm 1.15}$

TABLE 5
Comparison of the mean and standard deviation of classification accuracy (%) of our method (DMAE+SM) and SSAE+SM on four subsets in patient level. The best mean results in different subsets are in bold.

Method	40×		100×		200×		400×	
	Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
SSAE	94.18 ± 1.13	83.09 ± 1.47	92.76 ± 1.54	82.32 ± 1.23	98.24 ± 0.53	82.79 ± 1.29	94.39 ± 1.29	82.58 ± 1.35
Ours	$\mathbf{95.03 \pm 0.99}$	$\mathbf{84.56 \pm 1.19}$	$\mathbf{92.53 \pm 1.10}$	$\mathbf{82.98 \pm 1.07}$	$\mathbf{99.30 \pm 0.39}$	$\mathbf{83.38 \pm 0.93}$	$\mathbf{95.27 \pm 1.07}$	$\mathbf{83.061.09}$

from the integration of the deep neural networks and manifold learning. Furthermore, the patch-based feature learning helps to overcome the challenge of big image size, and avoid to construct complex network architecture with larger sets of parameters. The results have demonstrated that our proposed DMAE can achieve promising performance compared with other seven popular classification methods on the BreaKHis dataset.

In the future, we would like to investigate how to adaptively determine the parameters of the proposed method, i.e., the parameters λ, β, ξ, patch size, the number of hidden units. Also, developing more efficient training method for our DMAE to handle huge volume of histopathology images is one of our research planes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 6
Comparison of the mean and standard deviation of classification accuracy (%) of the tested methods on four subsets in image level. The best mean results in different subsets are in bold.

Method	40×		100×		200×		400×	
	Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
1-NN	84.54 ± 3.93	53.78 ± 5.13	87.94 ± 3.52	56.18 ± 4.97	93.16 ± 2.23	61.89 ± 4.23	89.49 ± 2.15	61.54 ± 4.65
SVM	93.40 ± 3.72	65.63 ± 5.35	92.32 ± 3.67	62.75 ± 4.71	97.04 ± 2.12	77.78 ± 3.92	93.03 ± 1.87	73.64 ± 2.98
DAC	89.49 ± 4.15	61.54 ± 4.65	89.36 ± 3.25	67.28 ± 3.96	95.68 ± 2.39	79.65 ± 3.65	92.32 ± 2.67	62.75 ± 3.71
RFC	88.92 ± 5.85	67.32 ± 4.79	90.37 ± 4.20	66.53 ± 4.92	89.28 ± 3.24	65.91 ± 4.73	84.54 ± 3.93	53.78 ± 5.13
CNN	94.30 ± 1.67	79.86 ± 2.50	91.39 ± 1.78	71.51 ± 2.53	90.86 ± 3.12	76.59 ± 3.61	93.37 ± 4.20	70.53 ± 3.92
MLP	93.03 ± 1.87	73.64 ± 2.98	92.02 ± 1.69	68.76 ± 3.01	97.02 ± 0.31	75.86 ± 3.09	94.30 ± 1.67	79.86 ± 2.50
Ours	$\mathbf{94.43 \pm 1.02}$	$\mathbf{83.71 \pm 1.09}$	$\mathbf{93.15 \pm 0.89}$	$\mathbf{82.98 \pm 0.98}$	$\mathbf{99.36 \pm 0.20}$	$\mathbf{83.57 \pm 1.01}$	$\mathbf{94.86 \pm 1.03}$	$\mathbf{83.45 \pm 1.15}$

TABLE 7
Comparison of the mean and standard deviation of classification accuracy (%) of the tested methods on four subsets in patient level. The best mean results in different subsets are in bold.

Method	40×		100×		200×		400×	
	Malignant	Benign	Malignant	Benign	Malignant	Benign	Malignant	Benign
1-NN	84.69 ± 4.04	53.96 ± 5.97	87.21 ± 3.78	55.159 ± 5.72	93.23 ± 2.01	62.01 ± 4.41	90.24 ± 2.17	62.62 ± 4.07
SVM	94.49 ± 3.69	66.31 ± 4.19	90.13 ± 3.91	60.93 ± 5.19	97.31 ± 1.19	78.59 ± 3.64	93.81 ± 1.92	74.91 ± 3.01
DAC	89.73 ± 4.15	62.54 ± 4.65	87.09 ± 3.56	65.83 ± 3.99	96.39 ± 1.35	79.93 ± 3.29	93.14 ± 3.05	63.52 ± 3.34
RFC	87.92 ± 5.12	67.89 ± 6.01	88.33 ± 4.20	63.01 ± 5.37	90.01 ± 2.32	66.57 ± 4.28	85.93 ± 3.75	54.66 ± 5.72
CNN	94.98 ± 1.23	80.27 ± 2.01	92.18 ± 1.78	74.17 ± 3.19	91.43 ± 2.13	77.50 ± 3.13	94.17 ± 1.64	71.79 ± 3.67
MLP	93.87 ± 1.45	74.01 ± 2.62	91.24 ± 2.07	64.51 ± 4.17	97.69 ± 0.29	76.08 ± 2.87	94.99 ± 1.59	80.38 ± 2.23
Ours	$\mathbf{95.03 \pm 0.99}$	$\mathbf{84.56 \pm 1.19}$	$\mathbf{92.53 \pm 1.10}$	$\mathbf{82.98 \pm 1.07}$	$\mathbf{99.30 \pm 0.39}$	$\mathbf{83.38 \pm 0.93}$	$\mathbf{95.27 \pm 1.07}$	$\mathbf{83.06 \pm 1.09}$

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under grants 61772353 and 61332002, the Foundation for Youth Science and Technology Innovation Research Team of Sichuan Province under grant 2016TD0018, and the Fok Ying Tung Education Foundation under grant 151068.

REFERENCES

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: a cancer journal for clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[2] R. A. Smith, V. Cokkinides, A. C. von Eschenbach, B. Levin, C. Cohen, C. D. Runowicz, S. Sener, D. Saslow, and H. J. Eyre, "American cancer society guidelines for the early detection of cancer," *CA: a cancer journal for clinicians*, vol. 52, no. 1, pp. 8–22, 2002.

[3] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, I. Eric, and C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC bioinformatics*, vol. 18, no. 1, p. 281, 2017.

[4] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Computers in Biology and Medicine*, vol. 43, no. 10, p. 1563, 2013.

[5] Y. M. George, H. H. Zayed, M. I. Roushdy, and B. M. Elbagoury, "Remote computer-aided breast cancer detection and diagnosis system based on cytological images," *IEEE Systems Journal*, vol. 8, no. 3, pp. 949–964, Sept 2014.

[6] P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2169–2178, 2013.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, December 2012, pp. 1097–1105.

[8] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[9] L. Zhang and Z. Yi, "Dynamical properties of background neural networks with uniform firing rate and background input," *Chaos, Solitons & Fractals*, vol. 33, no. 3, pp. 979–985, Aug 2007.

[10] ——, "Selectable and unselectable sets of neurons in recurrent neural networks with saturated piecewise linear transfer function," *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1021–1031, 2011.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, June 2015, pp. 1–9.

[14] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[15] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *International Joint Conference on Neural Networks*, Vancouver, BC, Canada, July 2016, pp. 2560–2567.

[16] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *IEEE International Conference on Systems, Man, and Cybernetics*, Banff, AB, Canada, Oct 2017, pp. 1868–1873.

[17] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, 2016.

[18] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *International journal of computer vision*, vol. 113, no. 1, pp. 3–18, 2015.

[19] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, June 2016, pp. 2424–2433.

[20] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[21] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," *Chicago*, vol. 3, no. 6, pp. 27–33, 2013.

[22] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319, 2000.

[23] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2858763, IEEE/ACM Transactions on Computational Biology and Bioinformatics

by locally linear embedding." *Science*, vol. 290, no. 5500, p. 2323, 2000.
[24] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373—1396, 2003.
[25] L. Shang, J. C. Lv, and Z. Yi, "Rigid medical image registration using pca neural network," *Neurocomputing*, vol. 69, no. 13, pp. 1717–1722, 2006.
[26] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 796–809, 2008.
[27] Q. Guo, L. Zhang, S. Wang, and Z. Yi, "Rigid image registration via column sparse optimisation for seal registration," *Electronics Letters*, vol. 49, no. 17, pp. 1069–1071, August 2013.
[28] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *International Conference on Machine Learning*, Helsinki, Finland, July 2008, pp. 1168–1175.
[29] C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *International Conference on Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2006, Conference Proceedings, pp. 1137–1144.
[30] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
[31] B. Scholkopf, J. Platt, and T. Hofmann, "Efficient learning of sparse representations with an energy-based model," in *International Conference on Neural Information Processing Systems*, Vancouver, Canada, December 2006, pp. 1137 – 1144.
[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, Chia Laguna Resort, Sardinia, Italy, May 2010, pp. 249–256.
[33] K. Fukunaga and L. D. Hostetler, "K-nearest-neighbor bayes-risk estimation," *Information Theory, IEEE Transactions on*, vol. 21, no. 3, pp. 285–293, 1975.
[34] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.
[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[36] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley and Sons, 2004, vol. 544.
[37] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 454–459.
[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[39] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 119–130, 2016.
[40] F. Spanhol, L. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

Yangqin Feng received the B. Eng. and M. Sc. degrees from the Department of Computer Science, Southwest University of Science and Technology, Mianyang, China, in 2011 and 2014, respectively. Currently, she is a third-year Ph.D. candidate at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her research interests include machine intelligence and medical image processing.

Lei Zhang received the B. Sc. degree and M. Sc. degree in mathematics from the University of Electronic Science and Technology of China, Chengdu, China, in 2002 and 2005, respectively. She received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2008. Currently, she is a Professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. Her research interests include the theory and applications of Neural Networks.

Juan Mo received the B. Sc. and M. Sc. degrees in mathematics in 2006 and 2009, respectively, from the Inner Mongolia University, Hohhot, China. She is currently pursuing the Ph.D. degree at the Cognitive Computing Laboratory, School of Computer Science and Engineering, SiChuan University, Chengdu, China. Her research interests include deep learning and medical image analysis.