A Self-adaptive Discriminative Autoencoder for Medical Applications

Xiaolong Ge, Yanpeng Qu, Changjing Shang, Longzhi Yang and Qiang Shen

***Abstract*—Computer aided diagnosis (CAD) systems play an essential role in the early detection and diagnosis of developing disease for medical applications. In order to obtain the highly recognizable representation for the medical images, a self-adaptive discriminative autoencoder (SADAE) is proposed in this paper. The proposed SADAE system is implemented under a deep metric learning framework which consists of K local autoencoders, employed to learn the K subspaces that represent the diverse distribution of the underlying data, and a global autoencoder to restrict the spatial scale of the learned representation of images. Such community of autoencoders is aided by a self-adaptive metric learning method that extracts the discriminative features to recognize the different categories in the given images. The quality of the extracted features by SADAE is compared against that of those extracted by other state-of-the-art deep learning and metric learning methods on five popular medical image data sets. The experimental results demonstrate that the medical image recognition results gained by SADAE are much improved over those by the alternatives.**

***Index Terms*—Autoencoder network; Deep learning; Metric learning; Computer aided diagnosis.**

I. Introduction

As one of the essential topics for the medical application of computer aided diagnosis (CAD) systems, the research on medical image recognition (MIR) has made substantial progress during the past decades. In general, the performance of MIR can be improved from two perspectives. The first approach is to segment the local information of medical images and enhance those important for further feature extraction and MIR. In [1], each input image is decomposed into corresponding smooth layer, texture layer and edge layer by using the local extreme value defined in the spatial domain and low-pass filter. In [2], a fuzzy-rough refined image processing framework is proposed to segment the ROI region of each breast image and perform local enhancement in the region with the highest positive fuzzy region. The research in [3] proposes a new deep learning (DL) framework which uses local descriptor coding strategy and FV coding representation to classify melanoma images using support vector machines with Chi-square kernel. The second approach is to enhance the discernibility of features to represent samples. In [4], a two-channel convolutional neural network (CNN) model for medical hyperspectral image classification is proposed. In this network, an end-to-end network channel is designed to obtain the representative global fusion features through a pixel by pixel mapping between the original MHSI data and its principal components. In [5], a novel semantic similarity graph embedding (SSGE) framework is proposed to explicitly explore semantic similarity between images and optimize visual feature embedding to improve the performance of multi-label Chest X-ray image classification. And the research in [6] advances a deep convolution learning model for automatic multi-category classification of skin lesions. The network model adopts a multi-layer and multi-scale filter, which reduces the filter and parameter, and improves the efficiency and performance.

In addition to the DL algorithms, metric learning (ML) is an alternative way to produce the discriminative features. In general, the mainstream of ML can be conducted by unsupervised learning or supervised learning. When operate unsupervised learning, the ML methods reform the manifold structure in low dimensional subspace to obtain the discrimination information of samples [7], [8]. In the case of supervised learning, the ML methods learn the distance measures which can maximize the separability of the data in line with their category information [9]–[11]. Specifically, in [9], the KNN classification is implemented as a large marginal proximity classification method to achieve the largest branch among samples. In [10], a tensor local linear discriminant analysis is proposed for image representation. This method can preserve the local discriminant information of image data and the spatial localization of pixels in the image. The study in [11] relies on the convex optimization and proposes a learning algorithm for quadratic Gaussian metric for classification tasks. In order to obtain the discriminative features from the large-scale data, the collaboration of ML and DL has been paid much attention recently [12]–[17]. Generally, these outcomes implement the supervised ML methods based on the high-level features learned by a deep network model.

In order to simultaneously value the local information and extract the discriminative features that will constitute decent representations of the medical images, this paper co-opts the advantages of a self-adaptive local ML strategy [18] and proposes a self-adaptive discriminative autoencoder (SADAE). The framework of this approach consists of K local autoencoders (AEs) to learn K subspaces implying the diverse local distribution of data and a global autoencoder (AE) to restrict the spatial scale of the learned features. Aided

Corresponding author: Y. Qu.

X. Ge and Y. Qu are with the School of Artificial Intelligence, Dalian Maritime University, Dalian, 116026, China (e-mail: {xaolong_g, yanpengqu}@dlmu.edu.cn).

C. Shang and Q. Shen are with the Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB, U.K. (e-mail: {cns, qqs}@aber.ac.uk).

L. Yang is with the Department of Computer and Information Sciences, Northumbria University, London E1 7HT, U.K. (e-mail: longzhi.yang@northumbria.ac.uk).

by a self-adaptive ML strategy, the proposed SADAE model can automatically find a proper margin to split the different classes, so as to further improve the separability of diverse data. As a result, the image representation learned by SADAE will be nourished with the discriminative features in medical images for further recognition. The quality of the extracted features by SADAE is respectively compared against that of those extracted by Auto-sklearn [19], Auto-Keras [20], ResNet [21] and other five ML methods on certain medical image data sets, including three data sets in MedMNIST [22], [23]: PneumoniaMNIST [24], BreastMNIST [25] and DermaMNIST [26], two versions of MIAS with BI-RADS [27] and Tabár [28] labeling strategies. Moreover, a brief dip into biometric applications is also conducted over five face image data sets: AR face [29], LFW [30], CK+48 [31], ORL [32], Facescrub [33]. The experimental results demonstrate that, compared to the alternatives, SADAE can effectively improve the representation of medical and biometric images with discriminative features while leading to a better medical image recognition performance.

In general, the novelty and contributions of SADAE are summarized as follows.

1) SADAE proposes a novel self-adaptive deep ML strategy to simultaneously evaluate the local information extract by a community of AEs and optimize the margins between different classes.
2) The SADAE model can successfully function at the level of individual categories, thus SADAE enjoys a significant ability to generate the high-level discriminative representations of diverse data.
3) The outperformance produced by SADAE reveal the potential of deep metric learning for medical applications.

The remainder of the paper is organized as follows. Section II briefly reviews related background. The proposed SADAE algorithm is described in Section III. Results of comprehensive experiments are presented in Section IV, leading to conclusions in Section V.

II. Background

In this section, the literature is reviewed in two parts: ML and DL methods for image analysis task.

A. Metric Learning

In the field of image classification and image recognition, ML provides an effective proxy to manipulate complex objects. It is supposed to assign spatially large/small margin to the pairs of examples that are conceptually dissimilar/similar. Generally, ML can be implemented by either the global metric or the local metric.

The global metric only transforms the data scale on the feature space and ignores the local structure of the data space. It employs a global linear transformation to maximize the separability of different classes of data. For instance, in [11], a maximally collapsing ML method is constructed via a convex optimization whose solution aims to collapse the data in the same class to a single point and push these data in other classes infinitely far away. In [34], the information theory ML (ITML) method is proposed to learn a Mahalanobis distance matrix based on the multivariate Gaussian distribution. Relevant component analysis (RCA) [7] uses edge information in the form of equivalent constraints to solve the problem of metric learning. The authors demonstrate that this type of additional information can be obtained automatically without human intervention.This information is also presented to represent data to improve classification. Neighbourhood components analysis (NCA) [35] maximizes the random variable of leave-one-out kNN score on the training set, while learning a low dimensional linear embedding of labeled data. Least squared-residual ML [36] calculates least-squares (LS) residuals using previously estimated supports in place of observed compressed sensing (CS). The boundary of CS-residuals is determined, which is much smaller than the boundary of CS errors if the sparse mode changes slowly enough.

Local ML can discover the underlying local framework of certain classes in the complex data sets. For instance, via large margin nearest neighbor (LMNN) [9], the center of each class is surrounded by sample of its own class by pulling the data of the same class together and pushing the data of different classes away. Large margin nearest neighbor classification with privileged information (LMNN+) [37] is proposed based on the LMNN classification framework, which improves decision function learning by introducing visual features as well as depth features into training process. The supervised distance ML algorithm through maximization of the Jeffrey divergence (DMLMJ) [38] is an optimization model in which the Jeffrey divergence between two multivariate Gaussian distributions, respectively derived from the neighborhoods with the same label and different labels, is maximized. The parametric local ML method (PLML) [39] learns a smooth metric matrix function over the data manifold weighted by the neighborhood of each points. Sparse compositional metric learning (SCML) [40] produces a sparse combination of local discriminative metric. The framework can be naturally derived from the global, multi-task and local measure learning problems. Recently, a new supervised self-adaptive local ML method (SA-LM^2) is proposed in [18]. It adaptively adjust local neighborhood to construct the proper distance to separate the distinct categories.

B. Deep Learning and deep metric learning

To extract meaningful features that will constitute high-level representations of the image, many DL methods has been developed recently. For instance, to effectively use the information in the untagged medical images, the deep virtual confrontation self-training method [41] utilizes a virtual adversarial training strategy and a consistent regularization to exploit the latent knowledge between the labeled and unlabeled data. In [42], a framework that encodes deep vision and semantic embedding through a three-branch network in the coding stage is proposed. The output of these three branches is fused and input into a decoder to generate the report. Local deep-feature alignment (LDFA) [43] constructs a neighborhood for each data sample, from which a local stacked shrink AE is learned to extract local depth features. Affine transformation

is used to align the local deep features with the global feature in each neighborhood. In [44], a new attitude restoration framework is proposed, which adopts multi-manifold learning and the shared parameter is calculated. The model improves attitude recovery through global mapping and local refinement. In [45], a novel joint CNN architecture is designed as Face detection and attribute recognition networks (FDAR-Net). At the face detection stage, the face candidate extractor is used to get quick suggestions, and then a convolution neural network is used for further analysis.

In order to enhance the discernibility of the extracted high-level features, the collaboration of ML and DL has been investigated. In [12], a discriminative CNN is learned by explicitly imposing an ML regularization term on CNN features. Discriminative deep ML (DDML) [13] is a global deep metric method which learns a deep neural network to assign the distance between the samples in the same class less than a preset threshold. Deep localized ML (DLML) [14] learns multiple fine grained deep localized metric by multiple AEs. And the final sample pair distance is synthesized by calculating the weight proportion of each subnetwork. Deep clustering-based asymmetric metric learning (DECAMEL) [15] jointly learns the feature representation and the unsupervised asymmetric metric to ease the bias of different views and exploit the potential cross-view discrimination information of unsupervised person re-identification. Deep transfer metric learning (DTML) [16] transfers discriminative knowledge from labeled source domain to unlabeled target domain, and uses a set of deep ML networks for cross-domain visual recognition. Subtype clustering-based deep metric learning [17] defines a new clustering degree to mine classification-oriented subtype structure. Sample pairs for ML are selected according to the clustering results.

III. A Self-adaptive Discriminative Autoencoder

To further extend the usage of DL and ML in the field of medical image recognition, this paper presents a self-adaptive discriminative autoencoder (SADAE) model. As illustrated in Fig. 1, the overall network structure of SADAE mainly consists of one global AE and K local AEs. In the initialization stage of SADAE, the global AE and K local AEs are pre-trained via the clustering strategy of DLML [14] to capture the underlying spatial distinction of the training data. With the aid of a self-adaptive ML algorithm, these $K+1$ initialized AEs will be trained as an integral network to produce the representation of data with discriminative features. These resulting discriminative features can ensure that the spatial distance of two samples from different categories will be penalized to cross an adaptive threshold. In so doing, the intra-class diversity and the inter-class gap can be expected to degrade and increase, respectively, so that the classification performance relies on the resulting features will improve accordingly.

A. Initialization

The proposed SADAE method is due to adaptively produce discriminative features to enlarge the divergence between distinct classes. However, the resulting inter-class variance is not supposed to be irrationally overlarge. In view of such tradeoff, this paper adopts a community of AEs, which consists of K local AEs (AE_k, $k = 1, \ldots, K$) to learn K subspaces implying the diverse distribution of data and a global AE (AE_0) to restrict the spatial scale of the learned features. For each sample, its reconstruction loss in an AE can be used to gauge its significance in the corresponding subspace. A smaller reconstruction loss indicates a greater probability of belonging to this space. In this paper, the clustering strategy of DLML is employed to training these $K+1$ AEs to initialize SADAE.

In the initialization of SADAE, the updating process of AE_0 is independent from that of the local AEs. Specifically, given a data set X, for AE_0, the parameters are updated by using the gradient descent algorithm over the entire set of training samples. This iteration of the parameters in AE_0 only has one stop criterion that is the reconstruction loss over all samples is less than a pre-set threshold.

For AE_k, $k = 1, \ldots, K$, the training strategy is as follows.

1) Each sample $x \in X$ is clustered into the subspace induced by a local AE that enjoys minimum reconstruction loss of x among all local AEs. As a result, X will be partitioned into K subsets (subspaces), $X_1, \ldots, X_K$, subject to
$$X = X_1 \cup \ldots \cup X_K, \tag{1}$$
and
$$X_i \cap X_j = \varnothing, \ i \neq j. \tag{2}$$
2) In light of the clustering results, AE_k is updated only with the samples that belong to X_k. In so doing, the distance between different clusters is indirectly enlarged.
3) The above two steps are iterated until the clustering process converges. That is, every X_k, $k = 1, \ldots, K$ remains unaltered in two successive iterations. To restrict the running time, another stop criterion of the iterative process is that the reconstruction loss of each sample in its local AE is less than a pre-set threshold.

Assume that each AE_k, $k = 0, \ldots, K$, has M layers. Given a sample $x \in X_k$, let $O_k^{(0)}$ represent the input x. The output of the m-th layer of the k-th local AE is
$$O_k^{(m)} = \psi(W_k^{(m)} O_k^{(m-1)} + b_k^{(m)}), \ m = 1, \ldots, M. \tag{3}$$
Here, $W_k^{(m)}$ and $b_k^{(m)}$ are the weights and bias connect the $(m-1)$-th layer and m-th layer, respectively. $\psi(\cdot)$ is an element-wise nonlinear activation function such as ReLU, Tanh and Sigmoid. . The reconstruction loss of AE_k defined by the Euclidean distance is
$$\theta_k = ||O_k^{(M)} - x||_2^2. \tag{4}$$

Based on Eqs. (3) and (4), $W_k^{(m)}$ and $b_k^{(m)}$ can be updated as follows.
$$W_k^{(m)} = W_k^{(m)} - \eta \frac{\partial \theta_k}{\partial W_k^{(m)}}, \tag{5}$$
$$b_k^{(m)} = b_k^{(m)} - \eta \frac{\partial \theta_k}{\partial b_k^{(m)}}, \tag{6}$$

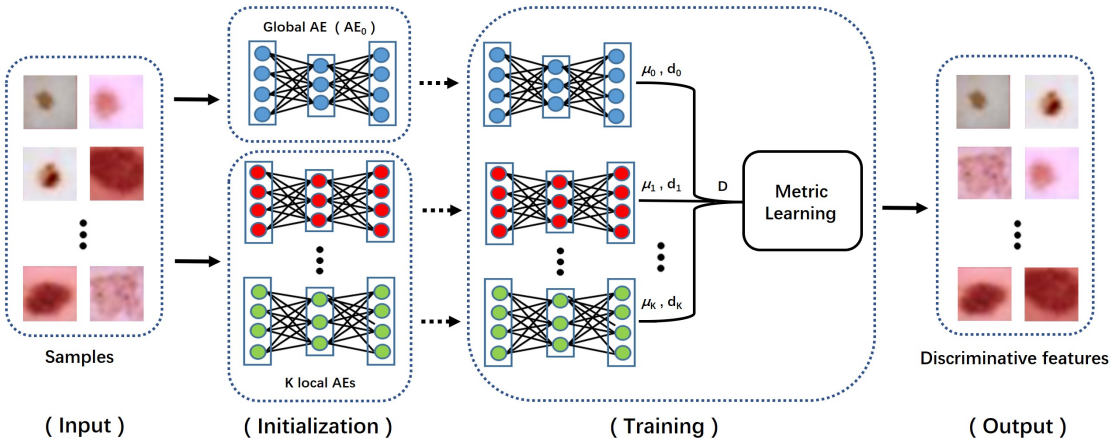


Fig. 1. Flowchart of SADAE

where,

$$\frac{\partial \theta_k}{\partial W_k^{(m)}} = ((O_k^{(m)} - x) \odot \psi'(O_k^{(m)}))(O_k^{(m-1)})^T, \quad (7)$$

$$\frac{\partial \theta_k}{\partial b_k^{(m)}} = (O_k^{(m)} - x) \odot \psi'(O_k^{(m)}). \quad (8)$$

Here, the operation $\odot$ represents the element-wise multiplication of two vectors. That is, for $\vec{a} = (a_1, a_2, \cdots, a_n)$ and $\vec{b} = (b_1, b_2, \cdots, b_n)$,

$$\vec{a} \odot \vec{b} = (a_1 b_1, a_2 b_2, \cdots, a_n b_n). \quad (9)$$

Based on above calculation, the initialization procedures of AE_0 and AE_k, $k = 1, \ldots, K$ are summarized in Algs. 1 and 2. It is noteworthy that the value of K determines the degree of refinement of the local relationship between samples. When $K = 0$, SADAE only has one global AE and is reduced to the normal deep ML hereafter. However, an excessively large value of K will consume much running time, and may result in certain null local AEs that can't be activated by any sample.

B. Training SADAE

Given a data set $X = \{x_i\}_{i=1}^n$, the set of corresponding outputs of the AE_k is $\{O_{k,i}^{(M)}\}_{i=1}^n$, $k = 0, \ldots, K$. Thus, the reconstruction loss of x_i, with regard to AE_k, is

$$\theta_{k,i} = \|O_{k,i}^{(M)} - x_i\|_2^2. \quad (10)$$

For each AE_k, $k = 0, \ldots, K$, the pair-wise feature distance with regard to (x_i,x_j) is measured by

$$d_k(x_i, x_j) = \|O_{k,i}^{(M)} - O_{k,j}^{(M)}\|_2^2. \quad (11)$$

For each $d_k(x_i, x_j)$, the associated weight is

$$\alpha_k(x_i, x_j) = \frac{\tau(\theta_{k,i}, \theta_{k,j})}{\sum_{l=0}^K \tau(\theta_{l,i}, \theta_{l,j})}, \quad (12)$$

where

$$\tau(\theta_{k,i}, \theta_{k,j}) = \frac{1}{\theta_{k,i} + \theta_{k,j}}. \quad (13)$$

Algorithm 1: Initialization of global AE

Input:
X, training set;
M, number of network layers;
S, maximum iterations;
η, learning rate;
ε, threshold of reconstruction loss.

;

Output:
$W_0^{(m)}$, $b_0^{(m)}$, parameters, $m = 1, \ldots, M$.

$\theta_0 \leftarrow 0$;
$s \leftarrow 1$;
Randomly initialize $W_0^{(m)}$, $b_0^{(m)} \in (-1, 1)$, $m = 1, \ldots, M$;
repeat
  $\widehat{\theta}_0 \leftarrow \theta_0$;
  $\theta_0 \leftarrow$ Eq. (4); // Calculate reconstruction loss
  if $|\ \theta_0 - \widehat{\theta}_0\ | < \varepsilon$ **then**
    break;
  end
  for $m = M, M-1, \ldots, 1$ **do**
    $\frac{\partial \theta_0}{\partial W_0^{(m)}} \leftarrow$ Eq. (7), $\frac{\partial \theta_0}{\partial b_0^{(m)}} \leftarrow$ Eq. (8);
    $W_0^{(m)} \leftarrow$ Eq. (5), $b_0^{(m)} \leftarrow$ Eq. (6);
  end
  $s \leftarrow s + 1$;
until $s == S$;
return $W_0^{(m)}$, $b_0^{(m)}$, $m = 1, \ldots, M$.

For the entire community of AE_k, the weighted sample pairwise distance of (x_i,x_j) is

$$D(x_i, x_j) = \sum_{k=0}^K \alpha_k(x_i, x_j) d_k(x_i, x_j). \quad (14)$$

With the use of the sample pairwise distance defined in Eq. (14), this paper co-opts the advantage of the ML method in [18] and further embeds a self-adaptive loss function into

Algorithm 2: Initialization of local AEs

Input:
X, training set;
K, number of local AEs;
M, number of network layers;
S, maximum iterations;
η, learning rate;
ε, threshold of reconstruction loss.
Output:
$W_k^{(m)}$, $b_k^{(m)}$, parameters, $k = 1, \ldots, K$, $m = 1, \ldots, M$.

$\theta_k \leftarrow 0$ on X, $X_k \leftarrow \varnothing$, $k = 1, \ldots K$;
$s \leftarrow 1$;
Randomly initialize $W_k^{(m)}$, $b_k^{(m)} \in (-1, 1)$, $k = 1, \ldots, K$, $m = 1, \ldots, M$;
repeat
  $\widehat{\theta}_k \leftarrow \theta_k$, $k = 1, \ldots K$;
  $\widehat{X}_k \leftarrow X_k$, $k = 1, \ldots K$;
  $\theta_k \leftarrow$ Eq. (4), $k = 1, \ldots K$;
  if $\theta_k == \min_l \theta_l$ **then**
    $X_k \leftarrow X_k \cup x$; // Update X_k
  end
  $\theta_k \leftarrow$ Eq. (4) on X_k;
  if $X_k == \widehat{X}_k \&\& \mid \theta_k - \widehat{\theta}_k \mid < \varepsilon$ **then**
    break;
  end
  for $m = M, M-1, \ldots, 1$ **do**
    $\frac{\partial \theta_k}{\partial W_k^{(m)}} \leftarrow$ Eq. (7), $\frac{\partial \theta_k}{\partial b_k^{(m)}} \leftarrow$ Eq. (8);
    $W_k^{(m)} \leftarrow$ Eq. (5), $b_k^{(m)} \leftarrow$ Eq. (6);
  end
  $s \leftarrow s + 1$;
until $s == S$;
return $W_k^{(m)}$, $b_k^{(m)}$, $k = 1, \ldots, K$, $m = 1, \ldots, M$.

these initialized $K+1$ AEs to produce the features with decent discernibility.

Let $U = [u_{ij}]_{n \times n}$, as follows, be an indicator matrix to identify the samples from different classes in $X = \{x_i\}_{i=1}^n$.

$$u_{ij} = \begin{cases} 0, & x_i \text{ and } x_j \text{ belong to the same class,} \\ 1, & x_i \text{ and } x_j \text{ belong to the different classes.} \end{cases} \tag{15}$$

Moreover, let ρ denote the uniform radius of the neighborhood of each sample in X. Intuitively, the neighborhood of a sample is not expected to contain any instance from other classes. That is, if the distance between two samples is less than ρ, they are supposed to belong to the same class. This intuition leads to the primary object rule of SADAE:

$$\forall x_i, x_j \in X : \text{if } D(x_i, x_j) < \rho, \text{ then } u_{ij} = 0, \tag{16}$$

where, $D(x_i, x_j)$ is the distance between x_i and x_j.

Above if-then rule can be inverted into the T-norm logical expression:

$$\prod_{x_i, x_j \in X} (D(x_i, x_j) < \rho \ \rightarrow \ \neg u_{ij}). \tag{17}$$

Since in the classical logic, $a \rightarrow b \Leftrightarrow \neg a \vee b$, the expression (17) can be further transformed into:

$$\prod_{x_i, x_j \in X} (D(x_i, x_j) \geq \rho \ \vee \ \neg u_{ij}). \tag{18}$$

Moreover, due to the property of the logical NOT operator, the following is further yielded

$$\sum_{i,j} (u_{ij} \wedge \nu(D(x_i, x_j) < \rho)), \tag{19}$$

where, ν is the indicator function

$$\nu(x) = \begin{cases} 1, & x = \text{true}, \\ 0, & x = \text{false}. \end{cases} \tag{20}$$

As a result, the primary objective rule (16) is equal to the following hinge loss function:

$$\begin{aligned} L_1(W, b, \rho) &= \sum_{i,j} u_{ij} [\rho - D(x_i, x_j)]_+ \\ &= \sum_{i,j} f(u_{ij}(\rho - D(x_i, x_j))). \end{aligned} \tag{21}$$

Here,

$$f(x) = \frac{1}{\beta} \log(1 + \exp(\beta x)) \tag{22}$$

is the generalized logistic approximation of the standard hinge loss $[x]_+ = \max(0, x)$, with a relatively large value of β. Geometrically, the intention of L_1 is to ensure that a sample won't lie in the neighborhood, where the radius is ρ, of those from different classes.

In light of the principle of structure risk minimization, two regularization terms will be added to L_1. First, in order to enlarge the margin between different classes, the value of ρ is designed to adaptively increase during the training procedure of SADAE. In this case, the first regularization term is

$$L_2(\rho) = -\rho. \tag{23}$$

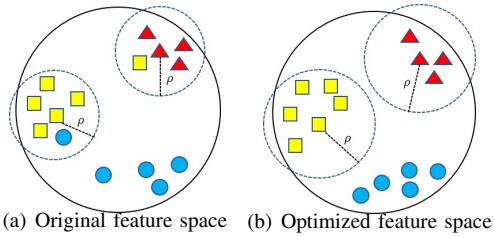
Another regularization term L_3 is to reduce the sophistication of W and b to prevent over fitting.

$$L_3(W, b) = \sum_{k=0}^{K} \sum_{m=1}^{M} (\|W_k^{(m)}\|_F^2 + \|b_k^{(m)}\|_2^2). \tag{24}$$

As the summation of L_1, L_2 and L_3, the general loss function of SADAE is

$$\begin{aligned} L(W, b, \rho) &= \lambda_1 L_1(W, b, \rho) + \lambda_2 L_2(\rho) + \frac{1}{2} \lambda_3 L_3(W, b) \\ &= \lambda_1 \sum_{i,j} f(u_{ij}(\rho - D(x_i, x_j))) - \lambda_2 \rho \\ &\quad + \frac{1}{2} \lambda_3 \sum_{k=0}^{K} \sum_{m=1}^{M} (\|W_k^{(m)}\|_F^2 + \|b_k^{(m)}\|_2^2). \end{aligned} \tag{25}$$

Fig. 2 illustrates the ML strategy used in SADAE. In the original feature space (as shown in Fig. 2(a)), some samples from different classes locate closer than ρ. Such abnormal distribution in the neighborhoods of samples may raise the risk of misrecognition. As shown in Fig. 2(b), by using the loss function (25), the value of ρ is penalized to increase and the



(a) Original feature space (b) Optimized feature space

Fig. 2. Illustration of ML strategy in SADAE

distance between the samples from different classes is enlarged to greater than ρ, simultaneously. In so doing, the discernibility of features is improved to distinguish different classes.

The stochastic gradient descent (SGD) algorithm is employed to update the parameters $\{W_k^{(m)}, b_k^{(m)}\}$, $k = 0, \ldots, K$, $m = 1, \ldots, M$, and the radius ρ in the loss function (25). The derivatives of $W_k^{(m)}$, $b_k^{(m)}$ and ρ are as follows.

$$\frac{\partial L}{\partial W_k^{(m)}} = \lambda_1 \sum_{i,j} (\Upsilon_{k,ij}^{(m)} (O_{k,i}^{(m-1)})^T + \Upsilon_{k,ji}^{(m)}, (O_{k,j}^{(m-1)})^T) + \lambda_3 W_k^{(m)} \tag{26}$$

$$\frac{\partial L}{\partial b_k^{(m)}} = \lambda_1 \sum_{i,j} (\Upsilon_{k,ij}^{(m)} + \Upsilon_{k,ji}^{(m)}) + \lambda_3 b_k^{(m)}, \tag{27}$$

$$\frac{\partial L}{\partial \rho} = \lambda_1 \sum_{i,j} u_{ij} f^{'}(c) - \lambda_2. \tag{28}$$

Here, $O_{k,i}^{(m)}$ represents the output of the m-th layer of x_i in AE_k. And c is defined as:

$$c \triangleq u_{ij}(\rho - D(x_i, x_j)). \tag{29}$$

Moreover, in AE_k, for $m = 1, \ldots, M-1$,

$$\Upsilon_{k,ij}^{(m)} = (W_k^{(m+1)})^T \Upsilon_{k,ij}^{(m+1)} \odot \psi^{'}(o_{k,i}^{(m)}), \tag{30}$$

$$\Upsilon_{k,ji}^{(m)} = (W_k^{(m+1)})^T \Upsilon_{k,ji}^{(m+1)} \odot \psi^{'}(o_{k,j}^{(m)}). \tag{31}$$

And for $m = M$,

$$\Upsilon_{k,ij}^{(M)} = -\alpha_k(x_i, x_j) u_{ij} f^{'}(c)(O_{k,i}^{(M)} - O_{k,j}^{(M)}) \odot \psi^{'}(o_{k,i}^{(M)}), \tag{32}$$

$$\Upsilon_{k,ji}^{(M)} = -\alpha_k(x_i, x_j) u_{ij} f^{'}(c)(O_{k,j}^{(M)} - O_{k,i}^{(M)}) \odot \psi^{'}(o_{k,j}^{(M)}), \tag{33}$$

where $o_{k,i}^{(m)}$ and $o_{k,j}^{(m)}$ are the intermediate functions as follows:

$$o_{k,i}^{(m)} \triangleq W_k^{(m)} O_{k,i}^{(m-1)} + b_k^{(m)}, \tag{34}$$

$$o_{k,j}^{(m)} \triangleq W_k^{(m)} O_{k,j}^{(m-1)} + b_k^{(m)}. \tag{35}$$

According to Eqs. (26), (27) and (28), $W_k^{(m)}$, $b_k^{(m)}$ and ρ are updated as follows.

$$W_k^{(m)} = W_k^{(m)} - \eta \frac{\partial L}{\partial W_k^{(m)}}, \tag{36}$$

$$b_k^{(m)} = b_k^{(m)} - \eta \frac{\partial L}{\partial b_k^{(m)}}, \tag{37}$$

$$\rho = \rho - \eta \frac{\partial L}{\partial \rho}, \tag{38}$$

where η represents the learning rate.

Algorithm 3 shows the details of the training process of SADAE. It is noteworthy that since the derivatives of μ_k is rather small in practice, the modification of μ_k is omitted to simplify the process of updating the SADAE model.

Algorithm 3: Training of SADAE model

Input:
X, training set;
AE_k, initialized AEs using Algorithms 1 and 2, $k = 0, \ldots, K$, $m = 1, \ldots, M$;
λ_1, λ_2, λ_3, trade-off parameters;
S, maximum iterations;
η, learning rate;
ε, threshold of reconstruction loss.
Output: O_i, $i = 1, \ldots, n$.

```
Initialize ρ;
s ← 1;
L ← 0;
repeat
    L̂ ← L;
    Calculate loss L ← Eq.(25);
    if |L − L̂| < ε then
        break;
    end
    for m = M, M − 1, ..., 1 do
        ∂L/∂W_k^(m) ← Eq. (26), ∂L/∂b_k^(m) ← Eq. (27);
        W_k^(m) ← Eq. (36), b_k^(m) ← Eq. (37);
    end
    ∂L/∂ρ ← Eq. (28);
    ρ ← Eq. (38);
    s ← s + 1;
until s == S;
O_i ← Eq. (39);
return O_i, i = 1, ..., n.
```

After training the SADAE model, the resulting parameters, i.e., $W_k^{(m)}$, $b_k^{(m)}$, $k = 0, \ldots, K$, $m = 1, \ldots, M$, are used to output the representation enjoys discriminative features for the input samples as follows.

$$O_i = \sum_{k=0}^{K} \frac{\frac{1}{\theta_{k,i}}}{\sum_{k=0}^{K} \frac{1}{\theta_{k,i}}} O_{k,i}^{(M)}, i = 1, \ldots, n. \tag{39}$$

Here, O_i is the improvement of x_i used in further medical image recognition.

IV. Experimental Evaluation

In this section, the medical image recognition tasks are conducted on three data sets in MedMNIST [22], [23]: BreastMNIST (small-size) [25], PneumoniaMNIST (medium-size) [24] and DermaMNIST (large-size) [26], and MIAS with two different labeling strategies: BI-RADS [27] and Tabár [28], respectively. Moreover, a brief dip into biometric applications is carried out on 5 data sets. The performance criteria used in this paper include recognition accuracy (ACC),

TABLE I
MEDMNIST DATA SETS USED FOR EVALUATION.

Data set	Samples	Training	Validation	Test	Classes	Channels
BreastMNIST	780	546	78	156	2	single-channel
PneumoniaMNIST	5856	4708	524	624	2	single-channel
DermaMNIST	10015	7007	1003	2005	7	triple-channel

area under curve (AUC), kappa statistic and confusion matrix. The performance of the SADAE model with different values of the local AEs is also discussed.

A. Experiments on MedMNIST data sets

MedMNIST [22] is a collection of 10 pre-processed medical image datasets with Creative Commons (CC) Licenses. It is designed to be used as a rapid-prototyping playground and a multi-modal machine learning/AutoML benchmark in medical image analysis. This experiment uses three data sets in MedMNIST: PneumoniaMNIST, BreastMNIST and DermaMNIST. For completeness, the official information of these three data sets is summarized in Table I.

BreastMNIST contains 780 breast ultrasound images [25], which are divided into three types: normal, benign and malignant. In MedMNIST, these images are simplified into binary classification by combing normal and benign as dissimilar, and classify them against malignant as similar. The source data set is divided into a training set, a verification set and a testing set in the ratio of 7:1:2, officially. The size of each image is adjusted to $1 \times 28 \times 28$.

PneumoniaMNIST is based on a prior data set [24] of 5856 pediatric chest X-ray images. The task is a binary classification of pneumonia and normal. In MedMNIST, the source training set is divided into a training set and a verification set in the ratio of 9:1 and takes the source verification set as the testing set, officially. Each image is a single channel with a size of $1 \times 28 \times 28$.

DermaMNIST is based on HAM10000 [26] which is a large multi-source dermatoscope image collection of common pigmented skin lesions. The data set consists of 10015 dermatological images, which are divided into seven different categories as a multi category classification task. These images are partitioned into a training set, a verification set and a testing set in the ratio of 7:1:2, officially. The source images are $3 \times 600 \times 450$ and resized into $3 \times 28 \times 28$ in DermaMNIST.

1) Experimental settings: The training samples and test samples in the three data sets used in the experiment split by official settings. Each image in BreastMNIST, PneumoniaMNIST and DermaMNIST is transformed into a 784-dimension vector and normalized via the max-min scaling. It is noteworthy that in DermaMNIST, each RGB triple-channel image will be converted into its grayscale equivalence before being flattened. This is done by setting the grayscale of each pixel as a weighted average of its RGB color components:

$$\text{grayscale} = 0.299R + 0.587G + 0.114B, \tag{40}$$

where, R, G and B represent the respective luminance values of a pixel in the RGB channels. And the weighting coefficients are set in proportion to the perceptual response of the human vision to each of the red, green and blue color channels [46]. The framework of SADAE consists of 5 local AEs and a global AE, each of which AEs has 7 layers. In Alg. 3, the parameters λ_1, λ_2, λ_3 are set to be 0.2, 0.8, 0.0005, respectively. The learning rate η is set to be 0.0005. The radius threshold ρ is a random number ranging from 2.4 to 3.0. Moreover, the classification tasks on the image representations coded by SADAE are performed in conjunction with the use of SVM using the linear kernel function, the cosine similarity (Cos) and the kNN method with 3 nearest neighbors (3NN).

2) Recognition accuracy: The performance of SADAE, in terms of recognition accuracy (ACC) and AUC, is compared to the benchmark results reported in [22], achieved by Auto-sklearn [19], Auto-Keras [20] and ResNet [21], respectively. Specifically, the model of ResNet is implemented with 18 layers (ResNet-18) and 50 layers (ResNet-50), respectively. Each implementation is trained with 100 epochs. Since the standard practice of training ResNet is at the resolution of 224×224, to make a comprehensive comparison, two input resolutions, 28 and 224 (resized from 28), are assigned in both ResNet-18 and ResNet-50.

In Table II, the respective results of ACC and AUC for the BreastMNIST, PneumoniaMNIST and DermaMNIST data sets are recorded with the best results for each data set marked in bold. For each method, the average results on ACC and AUC are summarized in the last two columns in Table II, respectively. On the grounds of the comparative results, it can be observed that SADAE achieves the highest ACC on the DermaMNIST (by Cos) data set and the largest values of AUC on both PneumoniaMNIST (by SVM) and DermaMNIST (by Cos) data sets. Although in other cases, SADAE takes the second best positions, its average results (by SVM) are the best in terms of both ACC and AUC. This demonstrates that SADAE can consistently provide decent representation for the medical image recognition tasks in different sizes. Such outperformance of SADAE may benefit from the use of local AEs, which can discover the underlying multi-modality concealed in data. Moreover, the utilization of the self-adaptive ML strategy can further improve the quality of discriminative features to recognize the difference between images.

3) Influence of number of local AEs: As mentioned previously, the framework of SADAE is composed of a global AE and K local AEs. When $K = 0$, SADAE is equivalent to a classical AE in structure. And with the growth of the value of K, SADAE can partition the data space into more subspaces. In this subsection, the impact of different values of K is investigated on the BreastMNIST, PneumoniaMNIST and DermaMNIST data sets.

With the use of a large number of local AEs, the initial-

TABLE II
COMPARISONS ON ACC AND AUC.

	BreastMNIST		PneumoniaMNIST		DermaMNIST		Average	
Methods	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Auto-sklearn	0.808	0.848	0.865	0.947	0.734	0.906	0.799	0.900
Auto-Keras	0.801	0.833	**0.918**	0.970	0.756	0.921	0.825	0.908
ResNet-18 (28)	0.859	0.897	0.843	0.957	0.750	0.911	0.817	0.922
ResNet-18 (224)	**0.878**	**0.915**	0.861	0.970	0.727	0.896	0.822	0.927
ResNet-50 (28)	0.853	0.879	0.857	0.949	0.727	0.899	0.812	0.909
ResNet-50 (224)	0.833	0.863	0.896	0.968	0.719	0.895	0.812	0.909
SADAE+SVM	0.859	0.897	0.901	**0.983**	0.756	0.925	**0.839**	**0.935**
SADAE+Cos	0.853	0.897	0.898	0.970	**0.759**	**0.927**	0.837	0.931
SADAE+3NN	0.846	0.877	0.876	0.965	0.753	0.923	0.825	0.922

ization and training process of SADAE will become rather time-consuming. Even worse, certain local AEs may suffer from useless empty subspaces, where the time is cost in vain. Thus, in this paper, the value of K is verified in a low value range.

Specifically, the 6 resulting frameworks of SADAE are implemented with the values of $K = 0, 1, 2, 3, 4, 5$, respectively. And the data representations coded by these SADAE models are classified by SVM, Cos and 3NN, again. The respective results are shown in Figs. 3, 4 and 5.

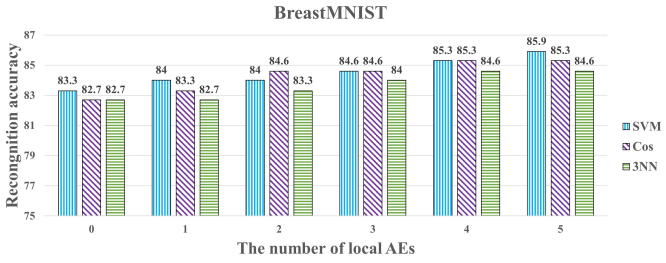


Fig. 3. Recognition accuracy of SADAE with regard to different values of K for BreastMNIST.

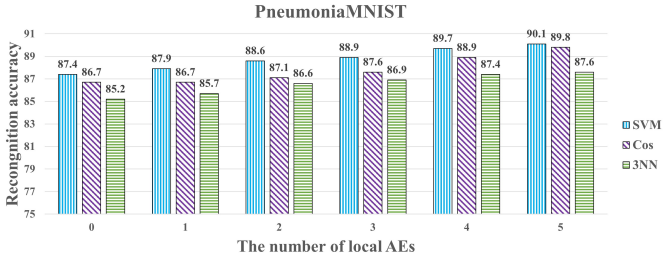


Fig. 4. Recognition accuracy of SADAE with regard to different values of K for PneumoniaMNIST.

It can be seen that, as the value of K grows, all of the recognition accuracies gained by SVM, Cos and 3NN exhibit improvement, consistently. However, the results in Figs. 3, 4 and 5 also demonstrate that, on the three used MedMINST data sets, the improvement between the recognition accuracies gained by SADAE with $K = 4$ and 5, is not significant already. Therefore, careful off-line selection of an appropriate K is necessary for the use of SADAE.

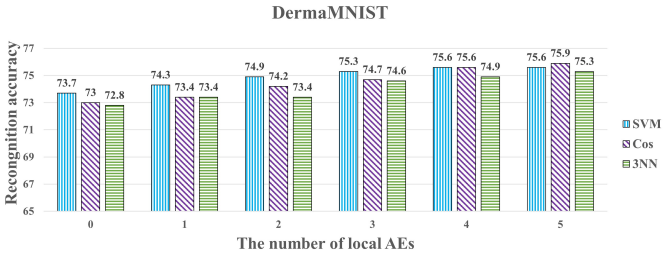
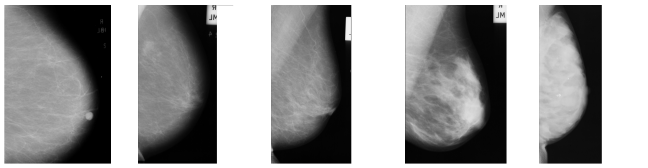


Fig. 5. Recognition accuracy of SADAE with regard to different values of K DermaMNIST.

B. Experiments on MIAS data set

The data employed in this experimental evaluation is derived from the mammographic image analysis society (MIAS) database [47]. It includes a set of Medio-Lateral-Oblique (MLO) left and right mammogram of 161 woman (322 samples). The spatial resolution of the image is $50\mu m \times 50\mu m$, quantized to 8 bits with a linear optical density in the range 0-3.2. In each image of the MIAS data set, an ROI of a 256×256 pixel size is extracted as the sample fibroglandular disk region [48]. Moreover, in this experiment, mammographic risk assessment are performed based on the BI-RADS [27], Tabár [28] labeling schemes (see Fig. 6 for examples).



(a) mdb098 (b) mdb304 (c) mdb252 (d) mdb116 (e) mdb318

Fig. 6. Example mammograms: (a) I, Pattern II; (b) II, Pattern III; (c) II, Pattern I; (d) III, Pattern IV; (e) IV, Pattern V.

In particular, BI-RADS is used to categorize a mammogram into one of four classes: BI-RADS I: Breast density is low; BI-RADS II: There exists some fibroglandular tissue; BI-RADS III: Breast density is high; BI-RADS IV: Breast is extremely dense. Numerically, the risk values for BI-RADS I–IV are 1, 1.6, 2.3 and 4.5, respectively.

Tabár describes breast composition of four building blocks: nodular density, linear density, homogeneous fibrous tissue and radiolucent adipose tissues. These blocks also define mammographic risk classification. In particular, the following

patterns are defined, with Patterns I-III corresponding to lower breast cancer risk and Patterns IV-V relating to higher risk.

1) Experimental settings: The framework of SADAE consists of 4 local AEs and a global AE, each of which has 7 layers. In Alg. 3, the parameters λ_1, λ_2, λ_3 are set to be 0.2, 0.8, 0.0005, respectively. The learning rate η is set to be 0.005. The radius threshold ρ is fixed to a random number ranging from 2.8 to 3.0. In addition, the MIAS data set with both the BI-RADS and Tabár labeling strategy is randomly split into a training set and a testing set in the ratio of 7:3.

2) Recognition accuracy: Since MIAS is not a very large medical image data set, the comparative studies on this data set is carried out between SADAE and other five ML algorithms, including ITML [34], LSML [36], NCA [35], LMNN [9] and SCML [40], which are implemented by using the metric-learn package [49] in Python. Since the performance of 3NN is rather poor on the data learned by these ML algorithms, in order to make a fair comparison, the following classification tasks are conducted by SVM, Cos and the kNN method with 5 nearest neighbors (5NN).

The resulting recognition accuracies are recorded in Table III. It can be seen that, via the tests by SVM, Cos and the 5NN, the proposed SADAE method consistently outperforms other ML methods on both the BI-RADS and Tabár data sets. This demonstrates that, even in the same sample space, SADAE can effectively adapt to the change of labels. In particular, the lowest improvement gained by SADAE is 0.0618, observed against ITML with Cos on the BI-RADS data set. Such superiority of SADAE is due to the use of the community of global and local AEs which can extract the information in the local subspaces for the subsequent self-adaptive ML strategy.

TABLE III
COMPARISON ON RECOGNITION ACCURACY.

Methods	Classifiers	BI-RADS	Tabár
ITML	SVM	0.6598	0.6495
	Cos	0.7835	0.7320
	5NN	0.7525	0.7113
LSML	SVM	0.6495	0.6289
	Cos	0.6598	0.6392
	5NN	0.7216	0.6907
NCA	SVM	0.6804	0.6082
	Cos	0.7320	0.6598
	5NN	0.7010	0.6701
LMNN	SVM	0.6186	0.5773
	Cos	0.5979	0.6186
	5NN	0.6186	0.5979
SCML	SVM	0.5773	0.5773
	Cos	0.6289	0.6082
	5NN	0.6082	0.5876
SADAE	SVM	**0.7629**	**0.7629**
	Cos	**0.8453**	**0.8041**
	5NN	**0.8556**	**0.8247**

3) Kappa statistic: To compare with the existing work, in this paper, the kappa statistic is employed to evaluate the experimental results also. The kappa statistic is generally thought to be a more robust measure than simple percent agreement calculation since it summarises the level of agreement between observers after agreement by chance has been removed. It tests how well observers agree with themselves (repeatability) and with each other (reproducibility).

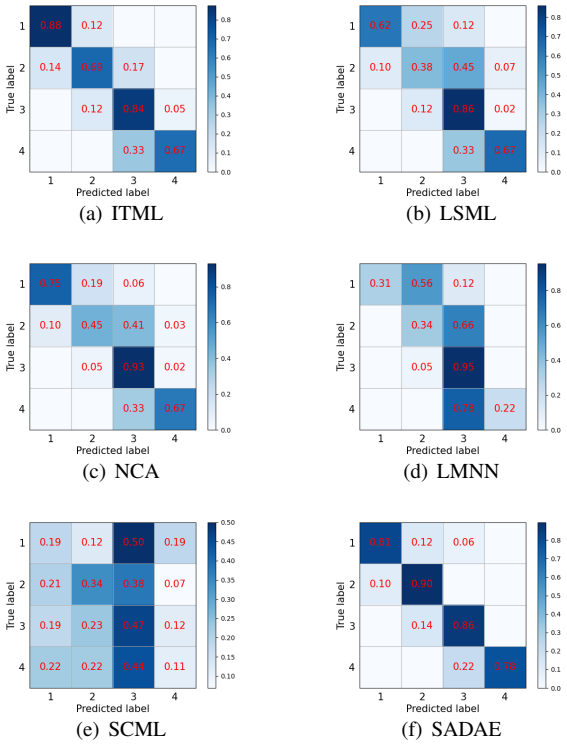
In Tables IV, the comparison is presented between SADAE and other ML methods. High values of kappa statistic are indicative of high agreement between the comparators. Thus, SADAE provides the best performance consistently in all cases. In particular, the values of kappa statistic of SADAE are all higher than 0.60, which means that the results gained by SADAE indicate highly moderate or substantial agreements between the comparators.

TABLE IV
COMPARISON ON KAPPA COEFFICIENT STATISTICS.

Methods	Classifiers	BI-RADS	Tabár
ITML	SVM	0.4607	0.4701
	Cos	0.6804	0.6180
	5NN	0.6279	0.4701
LSML	SVM	0.4451	0.4525
	Cos	0.4816	0.4890
	5NN	0.5781	0.5567
NCA	SVM	0.5011	0.3917
	Cos	0.5911	0.5140
	5NN	0.5381	0.5258
LMNN	SVM	0.3900	0.3529
	Cos	0.3404	0.4607
	5NN	0.3888	0.4044
SCML	SVM	0.2915	0.3928
	Cos	0.0542	0.4080
	5NN	0.3513	0.3744
SADAE	SVM	**0.6426**	**0.6796**
	Cos	**0.7873**	**0.7338**
	5NN	**0.8012**	**0.7622**

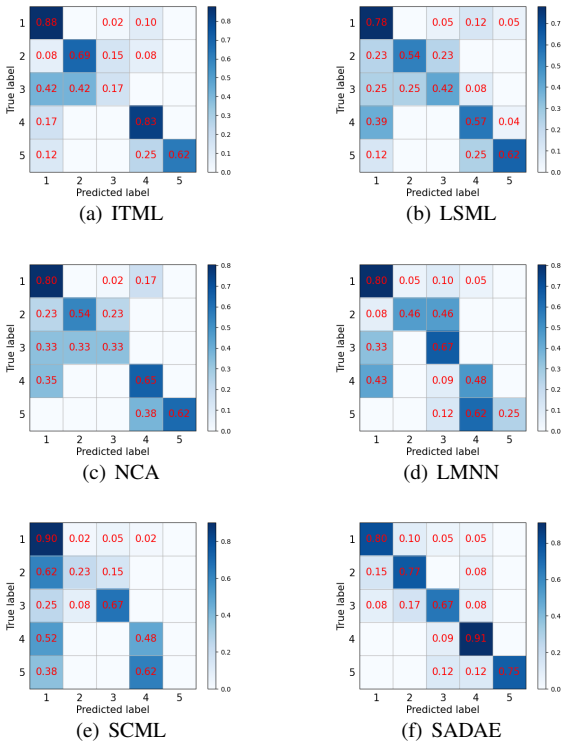
4) Confusion matrices using Cos classifier: In order to further demonstrate the advantage of SADAE, confusion matrix is employed to offer a standard means to support evaluation of recognition accuracy. In this work, confusion matrices are automatically plotted via SciKit-Learn. As shown in Figs. 7 and 8 in response to the use of the BI-RADS and Tabár criteria, the confusion matrices are yielded by using Cos on the data representations achieved by SADAE and other five ML methods, respectively. Specifically, in each confusion matrix, the entry in the i-th row and j-th column represent the ratio of the test image of the i-th class to be classified as the j-th class. And the null entry indicates the fact that no image in the i-th class is misclassified into the j-th class.

The experimental results in Fig. 7 demonstrate that SADAE perform consistently well on the identification of the 4 categories especially class II and class IV. Although ITML, LSML, NCA and LMNN misclassify less images than SADAE from class III to class II, SADAE assigns no image in class II to class III. In general, the SADAE model successfully reduce the class confusion, such as that between class II and class III in BI-RADS. This is of practical significance because these two classes constitute the majority of BI-RADS; it is therefore more useful, though more difficult, to identify class II and III separately. The experimental comparisons on the Tabár data set have also shown that the SADAE model can function better at the level of individual risk types. Considering the results shown in Figs. 7 and 8 jointly, it can be seen that SADAE enjoys a significantly better ability to distinguish distinct types of medical images than other ML methods.



(a) ITML (b) LSML (c) NCA (d) LMNN (e) SCML (f) SADAE

Fig. 7. Confusion matrices on BI-RADS data set



(a) ITML (b) LSML (c) NCA (d) LMNN (e) SCML (f) SADAE

Fig. 8. Confusion matrices on Tabár data set

C. Experiments on Biometrics

In order to further verify the performance of SADAE in other region, a comparative study on biometrics is conducted between SADAE and other five metric learning methods used in Section IV-B. The face image data sets used in this section are introduced as follows.

The AR [29] data set consists of 2600 images from 50 men and 50 women. Each person is described by 26 images. In this paper, 70% images of the AR data set are randomly selected as the training set, and the rest are used for testing.

The images in the LFW (Labled Face in Wild) [30] data set are all from natural scenes in life. This experiment uses a cropped grayscale version of LFW, which contains 500 images of 10 classes. In particular, 15 photos of each person are taken for testing and the rest are used for training.

CK+48 [31] is a data set for facial expression recognition. This paper applies a cropped version of CK+ 48 data set with top 5 emotions, anger, fear, happy, sadness and surprise. Such data set contains 750 emotional images which is divided into a training set and a test set in the ratio of 7:3.

The ORL [32] data set includes 400 face images of 40 different people. The 10 photos of each person represent different lighting, expressions and facial details. In particular, 7 photos from each person are taken as the training data set, and the remaining 3 photos are used for testing.

The FaceScrub [33] data set collects the face images celebrities. In this paper, the FaceScrub data set is cropped into two versions, FaceScrub10 and FaceScrub20, which contains 1041 photos of 10 celebrities and 2053 photos of 20 celebrities, respectively. For each category in these selected images, 20 photos were randomly selected for testing and the remaining photos were used for training.

Table V summarizes the information of above data sets.

TABLE V
BIOMETRIC DATA SETS USED FOR EVALUATION.

Data set	Samples	Training	Test	Classes
AR	2600	1820	780	100
LFW	500	350	150	10
CK+48	750	525	225	5
ORL	400	280	120	40
FaceScrub10	1041	841	200	10
FaceScrub20	2053	1653	400	20

1) Experimental details and parameter setting: In the experiments, let K be 5 for these data sets. Each of the global and local AEs adopts a 7-layer network structure. The parameters λ_1, λ_2, λ_3 are fixed as 0.2, 0.8, 0.0005, respectively. The learning rate is set to 0.0001. The radius threshold ρ is fixed to a random number ranging from 3 to 3.8.

2) Recognition results: Again, SVM, Cos and 5NN are used to perform classification in following experiments. As shown in Table VI, the features generated by SADAE consistently returns the best results for all data sets, even with different classifiers. Occasionally, ITML+SVM, NCA+SVM and LMNN+Cos perform competitively, compared against SADAE on the LFW and Facescrub data sets. These results further demonstrate the capacity of SADAE to extract discriminative features from diverse images and disclose the potential of SADAE for biometric application.

Together with the previous results, overall, it is clear that SADAE can effectively improve the representation of medical

TABLE VI
COMPARISONS ON RECOGNITION ACCURACY.

Methods	Classifiers	AR	LFW	CK+48	ORL	Facescrub10	Facescrub20
ITML	SVM	0.9833	0.6333	**0.9867**	0.9500	0.7000	0.6225
	Cos	0.9282	0.6400	0.9333	0.9500	0.7050	0.5550
	5NN	0.8346	0.6333	0.7911	0.9583	0.5050	0.3537
LSML	SVM	0.9782	0.6467	0.9822	0.9583	0.7100	0.6225
	Cos	0.6051	0.4867	0.9289	0.9333	0.5850	0.4775
	5NN	0.3897	0.5067	0.5511	0.9250	0.4800	0.3475
NCA	SVM	0.9731	0.6667	0.9778	**0.9667**	0.7200	0.6225
	Cos	0.9423	0.5533	0.9778	0.9417	0.6750	0.6025
	5NN	0.8923	0.6200	0.7867	0.9250	0.6200	0.5250
LMNN	SVM	0.9885	0.6333	0.9778	0.9417	0.7050	0.6200
	Cos	**0.9846**	0.6467	0.9773	0.9500	0.6850	0.6325
	5NN	0.9846	0.6867	0.9867	0.9750	0.6950	0.6700
SCML	SVM	0.9205	0.5933	0.9822	0.9500	0.6550	0.5500
	Cos	0.9282	0.5800	0.9778	0.9000	0.6600	0.5525
	5NN	0.9013	0.6000	0.9822	0.9500	0.6900	0.5225
SADAE	SVM	**0.9905**	**0.7133**	**0.9867**	**0.9667**	**0.7550**	**0.6525**
	Cos	**0.9846**	**0.7067**	**0.9867**	**0.9883**	**0.7500**	**0.6650**
	5NN	**0.9885**	**0.7400**	**0.9911**	**0.9883**	**0.7350**	**0.6875**

images and biometric images with discriminative features while leading to a better recognition performance. This outperformance of SADAE mainly thanks to the use of the global and local AEs to discover the local nonlinearity of data and the self-adaptive deep ML strategy to optimize the margins between different classes, so as to further improve the separability of data.

V. CONCLUSION

In this paper, a self-adaptive discriminative autoencoder (SADAE) algorithm is proposed for medical applications. The framework of SADAE consists of K local AEs to reveal the local nonlinearity of data and a global AE to restrict the spatial scale of the learned representation of images. By using a self-adaptive deep ML method, SADAE can automatically find a proper margin to distinguish different classes, so as to further improve the separability of data. The experimental results demonstrate that the recognition results gained by SADAE are much improved over those by other state-of-the-art DL and ML methods on the used image data sets.

Topics for further research include a more comprehensive investigation of an automatical selection of the optimal value of K for distinct medical image recognition data set. Moreover, when the medical images suffer from an irregular distribution of within-class multi-modality, the initialization process of SADAE may not successfully reach convergence within the preset maximum iteration. Thus, the method focused on sufficiently initializing SADAE for complex data is a worthwhile avenue of exploration. Last but not least, based on the framework of SADAE, potential alternative cooperations of DL and ML methods for image recognition tasks in diverse application domains remain active research.

ACKNOWLEDGMENTS

This work is jointly supported by Dalian High-Level Talent Innovation Program (No. 2018RQ70) and a Sêr Cymru II COFUND Fellowship, UK. The authors would like to thank the editor and anonymous referees for their constructive comments which have been very helpful in revising this work.

REFERENCES

[1] J. Du, W. Li, and H. Tan, "Three-layer image representation by an enhanced illumination-based image fusion method," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 1169–1179, 2020.

[2] Y. Qu, Q. Fu, C. Shang, A. Deng, Z. Reyer, G. Minu, and Q. Shen, "Fuzzy-rough assisted refinement of image processing procedure for mammographic risk assessment," *Applied Soft Computing*, vol. 91, p. 106230, 2020.

[3] Z. Yu, X. Jiang, F. Zhou, J. Qin, D. Ni, S. Chen, B. Lei, and T. Wang, "Melanoma recognition in dermoscopy images via aggregated deep convolutional features," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 4, pp. 1006–1016, 2019.

[4] X. Wei, W. Li, M. Zhang, and Q. Li, "Medical hyperspectral image classification based on end-to-end fusion deep neural network," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 11, pp. 4481–4492, 2019.

[5] B. Chen, Z. Zhang, Y. Li, G. Lu, and D. Zhang, "Multi-label chest x-ray image classification via semantic similarity graph embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2455–2468, 2022.

[6] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Computerized Medical Imaging and Graphics*, vol. 88, p. 101843, 2021.

[7] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, 2005.

[8] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *10th IEEE International Conference on Computer Vision (ICCV)*, pp. 1208–1213, 2005.

[9] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[10] Z. Zhang and W. S. Chow, "Tensor locally linear discriminative analysis," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 643–646, 2011.

[11] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in *Advances in Neural Information Processing Systems*, pp. 451–458, 2005.

[12] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[13] J. Lu, J. Hu, and Y. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4269–4282, 2017.

[14] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Deep localized metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2644–2656, 2018.

[15] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 956–973, 2020.

[16] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5576–5588, 2016.

[17] C. Ren, J. Li, P. Ge, and X. Xu, "Deep metric learning via subtype fuzzy clustering," *Pattern Recognition*, vol. 90, pp. 210–219, 2019.

[18] M. Taheri, Z. Moslehi, A. Mirzaei, and M. Safayani, "A self-adaptive local metric learning method for classification," *Pattern Recognition*, vol. 96, p. 106994, 2019.

[19] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Annual Conference on Neural Information Processing Systems*, pp. 2962–2970, 2015.

[20] H. Jin, Q. Song, and X. Hu, "Auto-keras: An efficient neural architecture search system," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1946–1956, 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision, Pattern Recognition*, pp. 770–778, 2016.

[22] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

[23] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification," *CoRR*, vol. abs/2110.14795, 2021.

[24] D. S. Kermany and M. Goldbaum, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[25] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

[26] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multisource dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[27] E. Mendelson, J. Baum, W. Berg, C. Merritt, and E. Rubin, *Breast imaging reporting and data system (BI-RADS).* American College of Radiology, 2003.

[28] L. Tabár, T. Tot, and P. B. Dean, "Breast cancer : the art and science of early detection with mammography : perception, interpretation, histopathologic correlation," *Jama the Journal of the American Medical Association*, vol. 300, no. 300, pp. 1822–1822, 2008.

[29] A. M. Martinez and R. Benavente, "The ar face database," *Tech. Rep. 24 CVC Technical Report*, vol. 01, 1998.

[30] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, *Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments*. Dept. Comput. Sci., University of Massachusetts, Amherst, Technical Report 07-49, 2007.

[31] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, 2010.

[32] C. Liu and H. Wechsler, "Independent component analysis of gabor features for face recognition," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, 2003.

[33] H. W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 343–347, 2014.

[34] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *ACM International Conference Proceeding Series*, vol. 227, pp. 209–216, 2007.

[35] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *NIPS*, pp. 513–520, 2005.

[36] N. Vaswani, "Ls-cs-residual (ls-cs): Compressive sensing on least squares residual," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4108–4120, 2010.

[37] J. He and D. Xu, "Large margin nearest neighbor classification with privileged information for biometric applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4567–4577, 2020.

[38] B. Nguyen, C. Morell, and B. De Baets, "Supervised distance metric learning through maximization of the jeffrey divergence," *Pattern Recognition*, vol. 64, pp. 215–225, 2017.

[39] J. Wang, A. Woznica, and A. Kalousis, "Parametric local metric learning for nearest neighbor classification," in *Advances in Neural Information Processing Systems*, vol. 2, pp. 1601–1609, 2012.

[40] Y. Shi, A. Bellet, and F. Sha, "Sparse compositional metric learning," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2078–2084, 2014.

[41] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Medical Image Analysis*, vol. 70, p. 102010, 2021.

[42] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint embedding of deep visual and semantic features for medical image report generation," *IEEE Transactions on Multimedia*, p. 1, 2021.

[43] J. Zhang, J. Yu, and D. Tao, "Local deep-feature alignment for unsupervised dimension reduction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2420–2432, 2018.

[44] J. Yu, C. Hong, Y. Rui, and D. Tao, "Multitask autoencoder model for recovering human poses," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 6, pp. 5060–5068, 2018.

[45] H. Liu, X. Shen, and H. Ren, "Fdar-net: Joint convolutional neural networks for face detection and attribute recognition," in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, pp. 184–187, 2016.

[46] C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab*. 01 2011.

[47] C. A. Ancy and L. S. Nair, "An efficient cad for detection of tumour in mammograms using svm," in *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 1431–1435, 2017.

[48] M. George, E. Denton, and R. Zwiggelaar, "Mammogram breast density classification using mean-elliptical local binary patterns," in *14th International Workshop on Breast Imaging (IWBI 2018)*, vol. 10718, p. 107180B, 2018.

[49] W. de Vazelhes, C. Carey, Y. Tang, N. Vauquier, and A. Bellet, "metric-learn: Metric Learning Algorithms in Python," *Journal of Machine Learning Research*, vol. 21, no. 138, pp. 1–6, 2020.