Computerized Medical Imaging and Graphics 84 (2020) 101765

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Automatic skin lesion classification based on mid-level feature learning

Lina Liu[a], Lichao Mou[b,c], Xiao Xiang Zhu[b,c], Mrinal Mandal[a,*]

[a] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G2V4, Canada
[b] Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany
[c] Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany

ARTICLE INFO

Article history:
Received 28 November 2019
Received in revised form 14 July 2020
Accepted 18 July 2020

Keywords:
Skin lesion analysis
Medical image analysis
Feature learning
Metric learning

ABSTRACT

Dermoscopic images are widely used for melanoma detection. Many existing works based on traditional classification methods and deep learning models have been proposed for automatic skin lesion analysis. The traditional classification methods use hand-crafted features as input. However, due to the strong visual similarity between different classes of skin lesions and complex skin conditions, the hand-crafted features are not discriminative enough and fail in many cases. Recently, deep convolutional neural networks (CNN) have gained popularity since they can automatically learn optimal features during the training phase. Different from existing works, a novel mid-level feature learning method for skin lesion classification task is proposed in this paper. In this method, skin lesion segmentation is first performed to detect the regions of interest (ROI) of skin lesion images. Next, pretrained neural networks including ResNet and DenseNet are used as the feature extractors for the ROI images. Instead of using the extracted features directly as input of classifiers, the proposed method obtains the mid-level feature representations by utilizing the relationships among different image samples based on distance metric learning. The learned feature representation is a soft discriminative descriptor, having more tolerance to the hard samples and hence is more robust to the large intra-class difference and inter-class similarity. Experimental results demonstrate advantages of the proposed mid-level features, and the proposed method obtains state-of-the-art performance compared with the existing CNN based methods.

1. Introduction

Melanoma is the most aggressive kind of skin cancer, whose incidence has risen rapidly over the last 30 years (Siegel et al., 2020). Early detection is the best way to treat melanoma since it is highly curable before it spreads into other body parts. To detect melanoma or suspected skin lesions, dermoscopy imaging is used as a primary step due to its non-invasive nature. Numerous clinical metrics based on the appearance of local color and texture patterns for the detection of melanoma have been proposed using dermoscopy images, such as ABCD rules (Stolz et al., 1994; Hazen et al., 1999), seven-point checklist (Argenziano et al., 1998) and classical pattern analysis (Pehamberger et al., 1987). However, due to the intrinsic visual similarity between different types of skin lesions, it is difficult to distinguish different types of skin lesions even for the dermatologists. Recent works have shown that the automatic learning method can obtain comparable performance with experienced dermatologists (Esteva et al., 2017), which demonstrates the appealing prospect of automatic skin lesion analysis.

Despite the current research achievement, skin lesion classification is still a challenging task due to the following reasons: (1) The pigment regions of skin lesion images may share strong visual similarity across different types of skin diseases. (2) Various visual patterns are observed within the same class of skin lesions. (3) Complex skin conditions, including color inconsistency and disturbing items, such as hairs, veins, color marks and other artifacts are also observed in the skin lesion images. Dermatologists have to focus on the subtlety of details in order to distinguish the malignant cases from benign ones, yet the large inter-class similarity and intra-class variations make it more formidable (Yu et al., 2018; Zhang et al., 2019). In addition, the existence of complex skin conditions may introduce noisy items which can affect the color and texture description of a given image and deteriorate the classification performance. Fig. 1 presents some example images from the ISIC 2017 dataset towards skin lesion analysis. Euclidean distances between inter-class and intra-class samples using features extracted via

[*] Corresponding author.
E-mail address: mmandal@ualberta.ca (M. Mandal).

https://doi.org/10.1016/j.compmedimag.2020.101765
0895-6111/© 2020 Elsevier Ltd. All rights reserved.

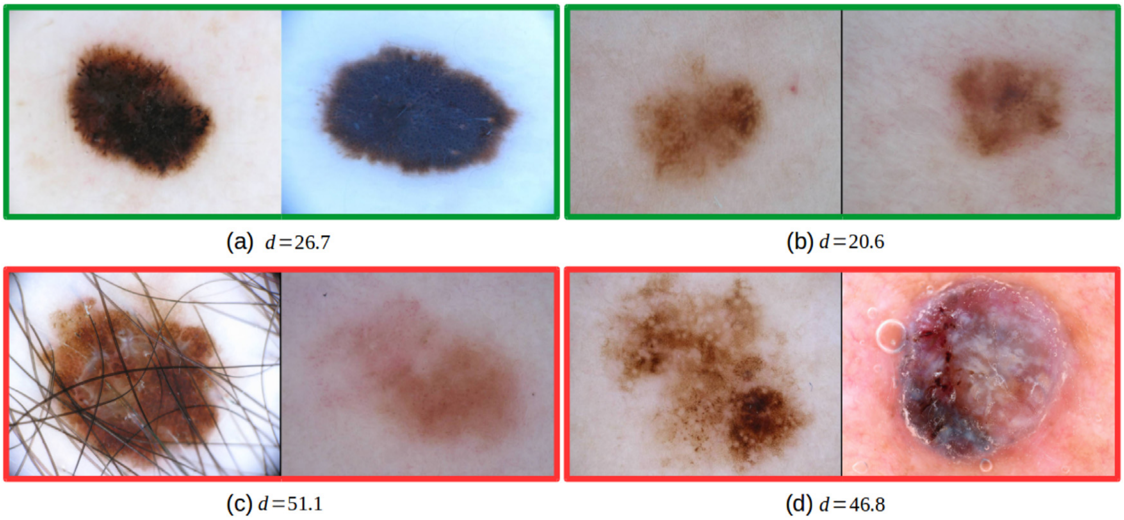
ResNet (see Section 2.2 for feature extraction) are given. Strong inter-class visual similarity and intra-class variations are observed across different types of skin lesions, which makes the diagnosis be difficult even for experienced dermatologists. More research can be done to improve the current machine learning methods so as to assist the diagnosis of skin lesions.

The general pipeline for existing automated methods follows three steps: preprocessing, feature extraction, and classification. As mentioned above, the original skin lesion images suffer from lighting condition change and interference from hairs and other artifacts. To address these problems, methods such as hair removal (Borys et al., 2015; Rebouças Filho et al., 2018) and color enhancement (Rebouças Filho et al., 2018) have been proposed for skin lesion analysis. After preprocessing, morphological or statistical features are extracted from dermoscopic images, and a classifier is then trained for melanoma detection. The image features play a key role in the skin lesion classification task, and many conventional methods with hand-crafted features (colors, textures, shapes, etc.) as inputs have been proposed. Previous works have shown that the combination of various types of feature representations, such as color, texture, and shape features is more beneficial for the skin lesion classification task than a single type of feature representation (Ma and Tavares, 2017; Oliveira et al., 2018). Unfortunately, hand-crafted features have limited discriminative power, and they perform poorly when dealing with complex problems.

Some methods perform skin lesion segmentation before the classification task, which aims at extracting the boundary information or detecting the ROI, to assist the subsequent classification task. Various skin lesion segmentation methods have been proposed in literature, including the thresholding-based methods (Humayun et al., 2011), region-merging based approaches (Wong et al., 2011), active contour models (Riaz et al., 2018; Abbas et al., 2014) and deep CNN models (Yuan et al., 2017).

Recently CNN (convolutional neural network) based methods have received much attention and many popular CNN architectures have been proposed for the image classification task with encouraging performance. A good initialization of neural networks with pretrained weights from a similar task is crucial for better performance and faster training. Consequently, a good starting point and the most intuitive way is to use these existing models and transform them to the task of skin lesion classification by finetuning parameters of the neural networks. Many methods based on this idea have been proposed for skin lesion classification, and techniques like ensemble and test augmentation are also used to boost the experimental results (Matsunaga et al., 2017; Menegola et al., 2017; Bi et al., 2017; Mahbod et al., 2019). The multi-task framework is also proposed for skin lesion analysis (Yang et al., 2017), which trained the segmentation and classification task simultaneously. Generally speaking, the multi-task framework can obtain better performance than the single task method since it can take advantage of the shared information between different tasks. González-Díaz (2019) proposed a method called DermakNet which used 50-layer ResNet (He et al., 2016) as a backbone. Dermatologists' knowledge (e.g. attributes, asymmetry information) modeled by different subsystems and meta-data are used to gain better performance and interpretability. The CNN models can extract global optimal features but miss the local information. To address this problem, Ge et al. (2017a) proposed to use both the global features and local features for melanoma detection. The global features are obtained using ResNet (He et al., 2016) and the local features are extracted using VGG-16 Network (Simonyan and Zisserman, 2015) with BP (*Bilinear Pooling*), which can differentiate skin conditions with subtle visual differences in local regions. To learn features with more discriminative power and take advantage of images from different sources, Ge et al. (2017b) proposed a siamese deep architecture with a pair of images from a single lesion as the input. Information of different modalities is shared in the middle layers of neural networks. The features are then spatially weighted using CAM (Class Activation Mapping), and BP is used to generate the feature representation.

In this paper, a novel mid-level feature learning method is proposed for skin lesion classification. Our motivation is that: the dermoscopic images suffer from strong visual similarity among different types of skin lesions and visual variations within the same class of samples (as shown in Fig. 1), and it is very difficult to learn an optimal feature representation that can well separate all the training images. Instead of using the original features as the input, relationships among different sample images are used as the feature representation. The relationships are modeled using the similarity measurement based on metric learning (learned using the training set) with a given reference set. An SVM classifier is finally used for the classification task by using the mid-level fea-


(a) $d = 26.7$

(b) $d = 20.6$

(c) $d = 51.1$

(d) $d = 46.8$

Fig. 1. Example images from ISIC 2017 dataset on skin lesion analysis towards melanoma detection. Each green box indicates a pair of hard inter-class samples (left: melanoma vs. right: benign), and each red box indicates a pair of intra-class samples (both images are melanoma). d is the Euclidean distance between two samples using features extracted via pretrained ResNet. The average distance for intra-class samples is 37.5. Strong inter-class visual similarity and intra-class variations are observed across different types of skin lesions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

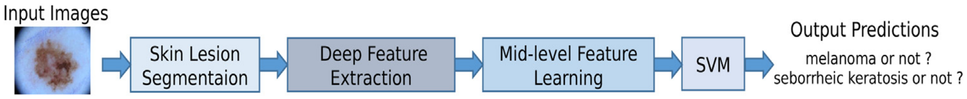


Fig. 2. Block diagram of the proposed method.

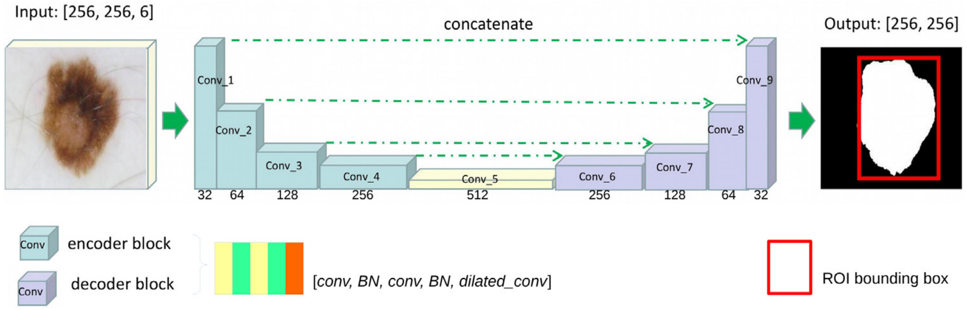


Fig. 3. Schematic of the proposed CNN-based model for skin lesion segmentation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tures as input. We name it mid-level feature representation, since it captures higher-level affinity information of the original features, and can be regarded as an intermediate semantic feature representation which bridges the raw features and the classification task. The distance metric learning can be regarded as a method for learning discriminative feature representation. Compared with the discriminative features learned by metric learning, the proposed mid-level feature is a soft discriminative feature representation, where the relationships of visual similarities and distinctions can be kept for some difficult cases (hard samples) as long as the remaining relationships are captured correctly. The learned image features are thus more robust to the large visual similarities between different classes of skin lesions and noisy items. Specifically, a CNN-based skin lesion segmentation model is used as a primary step for ROI detection. The skin lesion segmentation method is based on our previous work (Liu et al., 2019), which employs a U-Net architecture. We will not give detailed introduction about the skin lesion segmentation method, since our focus is the skin lesion classification task. Features are then extracted from the ROI via the pretrained neural networks (ResNet (He et al., 2016) and DenseNet (Huang et al., 2017).

The contributions of this paper can be summarized as follows: (1) A novel mid-level feature representation that utilizes the relationships among image samples (e.g. between an input image and the reference image set) is proposed for skin lesion classification. The new feature representation contains high-level affinity information between samples. It is a soft discriminative feature, having more tolerance to difficult cases, and is more robust to noise, large inter-class similarity and intra-class variations. (2) A novel framework for skin lesion classification is proposed. Pretrained CNN models are used as off-the-shelf feature extractors of ROI images, and metric learning is utilized to construct the mid-level features for classification. Extensive experiments have been conducted to show advantages of the proposed approach, and experimental results show that the proposed method outperforms state-of-the-art CNN based methods.

2. Proposed method

The proposed method contains four steps: skin lesion segmentation, feature extraction, mid-level feature learning and SVM classification. Block diagram of the proposed approach is shown in Fig. 2. Details about each step are presented in the following sections.

2.1. Skin lesion segmentation

It is difficult to process the original high-resolution dermoscopic images with machine learning algorithms, which contain much redundancy and require large computation memory. Many existing methods first downsample the dermoscopy images to the same scale for further analysis. However, the regions of interest (ROI) of skin lesions are generally of different scales, and some are so small that it is difficult to observe their patterns if we directly downsample the original images to the same size. Therefore, skin lesion segmentation is first performed to obtain the ROI.

Schematic plot of the proposed segmentation module is shown in Fig. 3. The inputs are images in RGB and HSV color spaces, and the output is a probability map of the foreground (i.e., the lesion). As observed from Fig. 3, the proposed method contains an encoder path and a decoder path, which is composed of a sequence of operation blocks. Both the encoder blocks and decoder blocks follow a structure of [*conv*, *BN*, *conv*, *BN*, *dilated_conv*], but with different pooling operations. Note that *conv* is a convolutional layer and *BN* is a batch normalization layer. Down-sampling is used at the end of each encoder block to reduce the resolution of feature maps by 2, while up-sampling is used at the beginning of each decoder block to increase the resolution of feature maps from the previous layer. Then the enlarged feature maps are concatenated with feature maps (of the same size) from the encoder path (as the dashed green arrows show in Fig. 3). The yellow block Conv_5 in Fig. 3 is the connection layer between the encoder and decoder paths.

For all convolution operations in this paper, 3 × 3 kernel with the "same" padding is used. The stride is set to be 1, and Rectified Linear Unit (ReLu) is used as the activation function. In addition, the number of output feature maps of each operation block is shown in Fig. 3. Note that the number of intermediate filters inside each encoder and decoder block is identical, which equals to the number of output feature maps. The proposed model is an improved version of the U-net (Ronneberger et al., 2015). Compared with the original U-net architecture, dilated convolution with a rate of 2 is used at the end of each operation block, which can increase the perceptive field of the output feature maps (at each depth of layer) without loss of resolution information. It is very suitable for the skin lesion segmentation task, since the ROI of dermoscopic images can be from different scales and with similar visual patterns. Note that the proposed skin lesion segmentation method is based on our previous work in (Liu et al., 2019). Different from the previous work, both

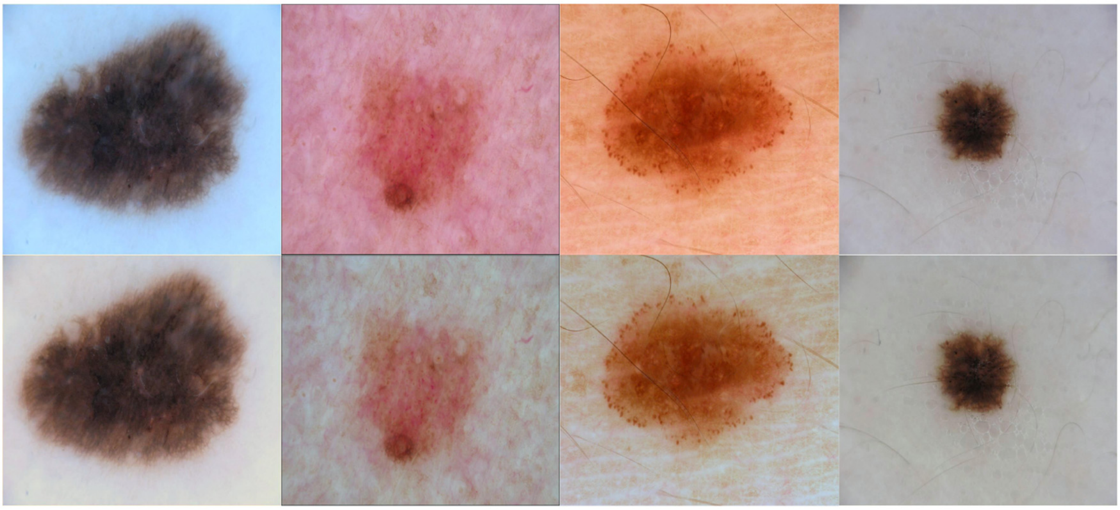


Fig. 4. The first row shows the original images, and the second row shows the images after pre-processing using Retinex method. The color distributions of different images are enhanced via the use of Retinex method. The resulted images are of similar lighting conditions.

binary focal loss and dice loss are used for CNN training, while the other implementations are the same.

2.2. *Deep feature extraction*

In this section, deep features are extracted from the ROI images using the pretrained neural networks. The pretrained neural networks can extract rich and meaningful texture information of images, and have been successfully used as offline feature extractors for medical image analysis (Gu et al., 2017; Mahbod et al., 2019). In this paper, pre-trained ResNet-50 (He et al., 2016) and DenseNet-201 (Huang et al., 2017) are used as off-the-shelf feature extractors. We denote them as ResNet and DenseNet for clarity in the following paper. Before feature extraction, Retinex algorithm (Ebner, 2007) is used to enhance the color consistency among different images. The resulting images are shown in Fig. 4. The colors of different images are more comprehensive and consistent after using the Retinex method. The ROI bounding boxes obtained by the segmentation module (Section 2.1) are superimposed on the pre-processed images to obtain the ROI images, which are then resized to 224 × 224 for feature extraction. For both CNNs, the output of the Global Average Pooling (GAP) layer is used as the feature. The output features corresponding to the two CNNs are of dimension 2048 (ResNet) and 1920 (DenseNet), respectively. Principal Component Analysis (PCA) is used to reduce the feature dimension by keeping 99% energy. The reduced feature dimensions for ResNet and DenseNet are 700 and 532, respectively.

2.3. *Mid-level feature learning*

Due to the complex skin conditions, noise, artifacts and severe visual similarities among different types of skin lesions, the extracted features may have limitations in describing characteristics of the original data and have poor discrimination power. Instead of using the original features as input, a novel mid-level feature representation is learnt, which describes relationships among image samples, and uses it as input feature of the classifier. The mid-level features of a sample are obtained by learning the similarities between a given sample and a reference image set. Since the discriminative power of the original features on the Euclidean space is poor due to the strong visual similarities among different classes of skin lesions, metric learning is used to address this problem. The metric learning method can learn a similarity measure to separate samples of different classes. Here, we present a brief introduction of the metric learning method. The squared Euclidean distance between two features x_i and x_j can be calculated by:

$$d(x_i, x_j) = (x_i - x_j)^T(x_i - x_j) = (x_i - x_j)^T I(x_i - x_j)$$

where I is an identity matrix. Similar to the formulation of Euclidean distance, instead of using an identity matrix, the Mahalanobis distance between features is defined as:

$$d_M(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j), \quad s.t. \quad M \geq 0 \tag{1}$$

where M is a positive semidefinite matrix to be learned during the training procedure. Since M is a positive semidefinite matrix, it can be represented as $L^T L$, and the above function can be reformulated as:

$$\begin{aligned} d_M(x_i, x_j) &= (x_i - x_j)^T M(x_i - x_j) \\ &= (x_i - x_j)^T L^T L(x_i - x_j) \\ &= \left\|L(x_i - x_j)\right\|^2 \\ s.t. \quad M \geq 0 \end{aligned} \tag{2}$$

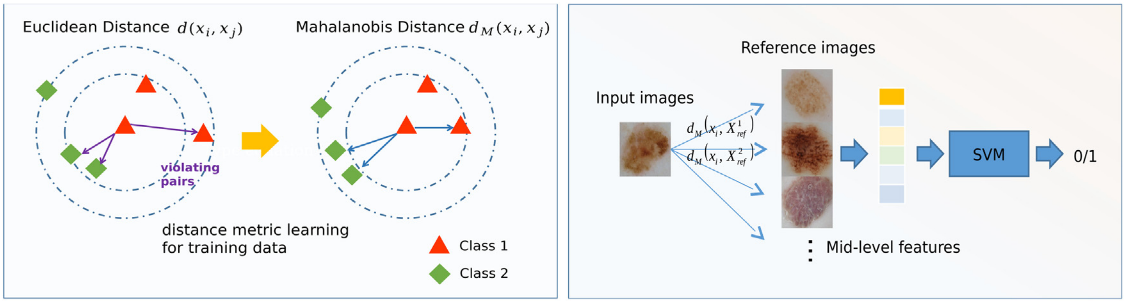
By observing Eq. (2), the distance metric learning method can also be treated as a discriminative subspace learning problem that aims at learning L, and the new discriminative feature of x_i is denoted as Lx_i. The metric learning method expects the distance between within-class samples to be small, and the distance between inter-class samples to be large. Compared with the discriminative feature Lx_i, the learned feature in this paper is a soft discriminative feature. Experiments in Section 3.9.3 show the advantage of the soft discriminative features over the discriminative features.

The optimal distance metric can be learned by separating samples of the same class and different classes by a margin of μ. The objective function can then be formulated as follows by using the logistic loss function:

$$f_M(x_i, x_j) = \log(1 + e^{y_{ij}(d_M(x_i, x_j) - \mu)}), \tag{3}$$

$$y_{ij} = \begin{cases} 1 & \text{if} \quad y(x_i) = y(x_j) \\ -1 & \text{if} \quad y(x_i) \neq y(x_j) \end{cases}$$

where $y(x_i)$ is the label of input feature x_i. The above function can drive distances between intra-class samples to become smaller than μ, and distances between inter-class samples to become larger



(a)

(b)

Fig. 5. (a) Schematic plot of the distance metric learning algorithm; (b) Schematic plot for the skin lesion classification method.

than μ. In this paper, we set μ as the average Euclidean distance between samples of the same class. The optimal solution is learned by minimizing the following function:

$$P(M) = \sum_{x_i, x_j \in V} w_{ij} f_M(x_i, x_j), \tag{4}$$

where w_{ij} is a weighting factor for each training pair. Instead of using a fixed weight for each pair of inputs, we update w_{ij} according to its difficulty of training the input pair (x_i, x_j). Especially, we only focus on the violating pairs and give higher weights to those who violate the rules more. A violating pair is defined as a pair of samples that violates the learning rule. For instance, if the distance between two samples of the same class is larger than μ, the two samples are regarded as a violating pair. Similarly, two samples are also regarded as a violating pair if their distance is smaller than μ and they are from two different classes. Examples of violating pairs are shown in Fig. 5. In this paper, we use V to denote the collection of violating pairs. The value of w_{ij} is initialized as 1 and it is updated in each training iteration according to:

$${w_{ij}}^{\tau} = \mathcal{N}(|d_{M_{\tau-1}}(x_i, x_j) - \mu|), \quad (x_i, x_j) \in V \tag{5}$$

in which τ is the number of iteration. $\mathcal{N}(.)$ is the normalization process, which is min–max normalization in this paper.

The learned distance values among different samples vary dramatically, some violating pairs' distance values are an order of magnitude different compared with the others. These violating input pairs are known as the hard samples (difficult cases), which is mainly caused by the appearance variation within the same class. Directly normalizing the distance values using standard min-max normalization will make the majority of weights w_{ij} be closer to 0, which indicates the algorithm will only use these hard samples. Therefore, min-max normalization at a cutoff distance value is performed to normalize the distance difference values $|d_{M_{\tau-1}}(x_i, x_j) - \mu|$ to a fixed range [0, 1]. In this paper, the cutoff value is decided automatically by calculating the cumulative histogram of the distance differences. For the cumulative histogram, the y-axis of a bin represents the percentage of observations that are smaller than a specific value (x-axis of the bin). We use the bin value that accounts for 97% as the cut-off value, and distance values that are larger than the cutoff value are set to be 1. In this case, the weights are updated dynamically in each iteration and hard violating samples are given more importance during training.

The objective function in Eq. (4) can be solved using the APG (Accelerated Proximal Gradient) algorithm (Liao and Li, 2015). Liao and Li (2015) used fixed weights for input training pairs. Different from (Liao and Li, 2015), only violating pairs are used in this paper and weight w_{ij} is updated in each iteration to give different pairs of samples different importance. After learning the metric M, a new feature representation is obtained by using the similarity information among a reference set. In this paper, the validation set is used as the reference set. Let $X_r \in \mathbb{R}^{p \times N_r}$ denote feature representations of the reference set. N_r is the number of images in the reference set, and ${X_r}^j$ is the jth column of X_r, which represents the feature vector of jth image in the reference set. For a feature vector x_i coming from training or testing set, its corresponding new feature representation v_i is obtained by calculating its distance with all samples in the reference set.

$$v_i = \left\{ d_M(x_i, {X_r}^1), \quad d_M(x_i, {X_r}^2), \quad \ldots, \quad d_M(x_i, {X_r}^{N_r}) \right\}, \tag{6}$$

After v_i is calculated, "L2" normalization is performed. The dimension of the new feature space equals to the number of samples in the referent set N_r, which is 150 in this paper.

2.4. Classification using SVM

In this paper, SVM with the radial basis function (RBF) kernel is used for classification, which is a common choice due to its good generalization ability and competing performance (Oliveira et al., 2018). There are two parameters that need to be tuned for the RBF kernel SVM, the parameter C, which is known as the capacity constant, the parameter g for the RBF kernel, which is a multiplier for the squared Euclidean distance between the two feature vectors. Details about the parameter selection can be found in Section 3.5.

3. Experiments and results

3.1. Dataset

For performance evaluation of the proposed method, we have used the dataset from ISIC 2017 for skin lesion detection (Codella et al., 2018), which is a very challenging dataset for skin lesion classification. There are 2000 images in the training set, including 374 melanoma, 254 seborrheic keratosis, and 1372 benign nevi. The validation dataset contains 150 images and the final testing dataset contains 600 images. All the images are of various resolutions, ranging from 767 × 1022 to 4499 × 6748 pixels. Severe illumination variation, noise and various artifacts are also witnessed in this dataset.

3.2. Evaluation metrics

To evaluate the classification results, Accuracy (ACC) and Area Under Curve (AUC) are used as the evaluation metrics. The criteria are defined as:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Table 1
Parameters used for performance evaluation.

	Proposed		Without MFL	
	C	*g*	*C*	*g*
SK	1	0.001	0.5	0.0156
MEL	2	0.002	0.5	0.0313

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{AUC} = \int_0^1 T(F_0)\mathrm{d}F_0$$

where TP is the True Positive number, TN is the True Negative number, FP is the False Positive number and FN is the False Negative number. $T(F_0)$ is the corresponding true positive rate (TPR) when the false positive rate (FPR) is F_0.

3.3. Platform information

The skin lesion segmentation method is implemented using Keras while the classification method is implemented using Matlab. All the experiments were conducted on a desktop with Intel(R) i7-7700 4.2 GHz CPU and a GPU of Nvidia GeForce GTX 1080Ti with 11GB memory.

3.4. Parameter selection

There are two parameters of SVM that need to be tuned for the proposed skin lesion classification method. The best parameters *C* and *g* of the proposed method are selected by conducting 5-fold cross-validation on the training dataset using the deep features extracted by ResNet. Note that 2 binary classifiers, SK for seborrheic keratosis and MEL for melanoma, are trained in the skin lesion classification task, therefore, we select different sets of parameters for different tasks. The best parameters used in this paper are shown in Table 1. In addition, the best parameters for the proposed method without MFL are also given.

3.5. Comparison with features extracted via pretrained CNN

To show advantages of the learned mid-level feature representation, we first compare the learned mid-level features with the raw features obtained by the pretrained CNN models. This is done by comparing the proposed method with and without the MFL on the same test data. Best parameters of the proposed method with and without the MFL module are used to make a fair comparison. Test data augmentation is used to increase performance as previous work (Mahbod et al., 2019; González-Díaz, 2019). Similarly, to show influence of the input features, we provide experimental results with input features extracted via the pretrained ResNet and DenseNet. Dimensions of the extracted features for ResNet and DenseNet are 2048 and 1920, respectively. After applying PCA (with 99% energy preserved), the reduced dimensions are 700 and 532 correspondingly. Performance of this part is shown in Table. 2 .

As observed in Table 2, the learned mid-level features consistently outperform the raw features extracted by pretrained CNN. Features extracted via ResNet achieve comparable performance with the features extracted via DenseNet. Especially, for the ResNet features, the proposed method with MFL module achieves 4.7% higher for the average AUC score and 3.2% higher for the average ACC score compared with the proposed method without MFL module. For the features extracted by DenseNet, the proposed method with mid-level features as input outperforms the one with original features as input by 4.0% and 2.5% for the average AUC and ACC scores. Experimental results show that the mid-level features can significantly improve the performance, this is because additional discriminative power is gained by using metric learning.

To visualize the distribution of raw features obtained using pretrained ResNet and corresponding mid-level feature representation, t-distributed stochastic neighbor embedding (t-SNE) is used to visualize the high dimensional data following Mahbod et al. (2019). The t-SNE first reduces the dimension of original features to 50 by PCA (for speed up), and then to 2 by using the t-SNE Barnes-Hut algorithm (Maaten and Hinton, 2008). It allows us to visualize the cluster of high dimensional data to some degree. The Visualization plot is shown in Fig. 6. From Fig. 6, we can see that the raw features are more likely to mix together (for both the training and testing data), especially for the samples of melanoma and nevus. This means the raw features have limitations in dealing with those hard samples. The mid-level features learned from pretrained ResNet show apparent grouping behavior for the training data. Samples are more likely to cluster together if they are of the same class, and hence the three-class skin lesions become more discriminative after using MFL. For the mid-level features of testing data, the melanoma and nevus become more distinguishable, although not completely separable. Some hard samples are identified along with the learning phase.

3.6. Comparison with features extracted via finetuned CNN

In this section, we first finetune the pretrained ResNet and DenseNet for the classification tasks by changing the output dimensions of the last fully connected layers to be 2 (i.e., the number of classes). For each neural network, we set the batch size to be 24 and train it for 100 epochs. Adam optimization algorithm with a learning rate of 0.0001 is used. The best models are selected by the validation performance. Data augmentation techniques including random resize cropping (70% to 100% of the original size), random horizontal and vertical flipping, random rotation (-20° to 20°) and normalization are used. Afterward, finetuned features are extracted and SVM classifiers are trained in order to compare the mid-level features with the finetuned CNN features. The same steps described in Section 3.5 are used but with finetuned CNN features as inputs. It is worth noting that parameter selection and test augmentation are also performed so as to make a fair comparison.

Experimental results are shown in Table 3. Method “ResNet” is the finetuned ResNet for classification. “ResNet + SVM” method

Table 2
Comparison of the proposed method with and without MFL module.

Networks	ResNet				DenseNet			
Metric	AUC (in %)		ACC (in %)		AUC (in %)		ACC (in %)	
Task	Without MFL	With MFL	Without MFL	With MFL	Without MFL	With MFL	Without MFL	With MFL
Mel	77.67	**84.29**	81.17	**84.33**	76.34	**84.29**	81.17	**84.33**
SK	91.00	**93.71**	87.00	**90.17**	93.32	**93.42**	87.83	**89.67**
Average	84.34	**89.00**	84.09	**87.25**	84.83	**88.85**	84.50	**87.00**

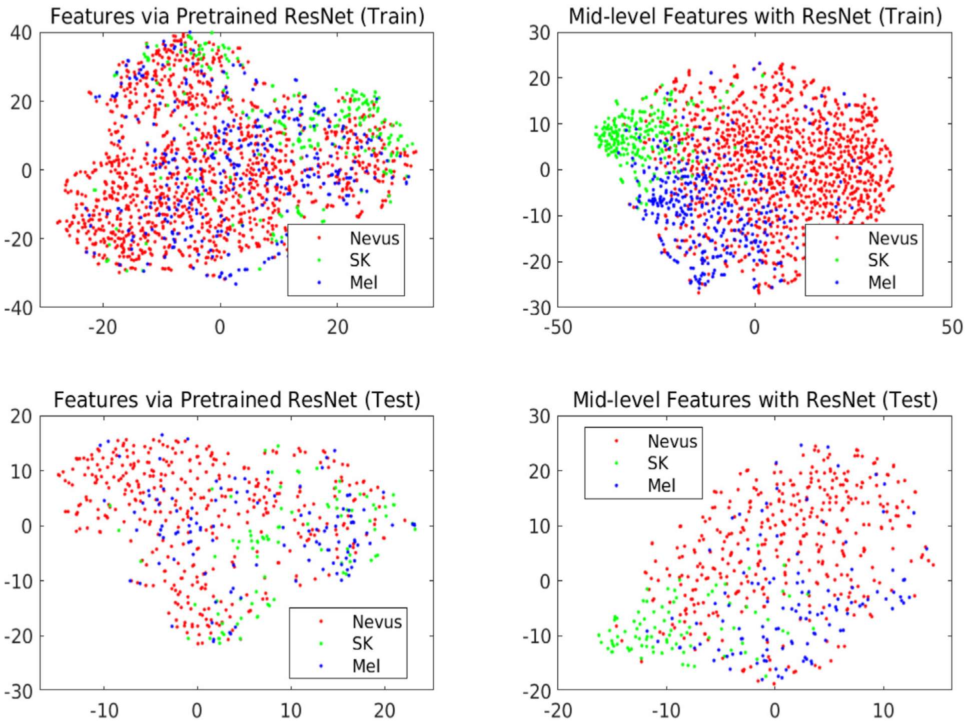


Fig. 6. t-SNE visualization of the raw features extracted via pretrained ResNet and the learned mid-level features given raw features obtained from pretrained ResNet. The first and second row show scatter plots of the training data and testing data, respectively.

Table 3
Comparison of the proposed method with the finetuned CNN and finetuned features using AUC scores (in %).

Method	Mel	SK	Avg
ResNet	77.30	**94.19**	85.75
ResNet + SVM	80.92	93.94	87.43
Proposed (ResNet)	**84.29**	93.71	**89.00**
DenseNet	84.52	92.51	88.52
DenseNet + SVM	**84.66**	90.62	87.64
Proposed (DenseNet)	84.29	**93.42**	**88.85**

Table 4
Average AUC of the proposed method with different input sizes.

Scale	ResNet	DenseNet	Fusion
$I_{[224,224]}$	89.00	**88.85**	**90.67**
$I_{[448,448]}$	88.22	88.13	89.35
$I_{[672,672]}$	**89.04**	88.28	89.40

uses the SVM to classify the features extracted via finetuned ResNet. “Proposed (ResNet)” is the proposed method which uses features extracted via pretrained ResNet. The same definitions are used for DenseNet. As shown in Table 3, the proposed method achieves the best average performance for both features extracted by pretrained ResNet and DenseNet. The proposed method (ResNet) outperforms the finetuned ResNet by 3.25%, and ResNet + SVM by 1.57%. The three methods obtain similar performance when using the DenseNet models. The proposed method outperforms the finetuned DenseNet by 0.33%, and DenseNet + SVM by 1.21%. The best performance for seborrheic keratosis classification is obtained when using the finetuned ResNet (94.19%). For melanoma classification, the best performance is obtained by using SVM over finetuned DenseNet (84.66%).

3.7. Comparison with state-of-the-art methods

In this section, we compare the proposed method with state-of-the-art methods. Due to the fact that ISIC 2017 dataset is a challenge dataset, tricks such as ensemble are widely used among the existing methods (see Table 6). Prevalent methods get the final performance by fusing outputs from different trained neural networks. Here we give a brief introduction about the compared methods in Table 6: Matsunaga et al. (2017) trained ResNet-50 with different optimization methods, and selected the best combination of fine-tuned CNNs through cross validation. Besides, a manual decision rule with metadata (age, sex information) is also adopted. Menegola et al. (2017) used ResNet-101 and Inception-v4 models. The final results were obtained by ensembling 7 trained neural networks with a meta learning model to assemble these models. Bi et al. (2017) fused outputs of the binary ResNet and 3-class ResNet to get the final results. Mahbod et al. (2019) used AlexNet, VGG16, ResNet-18 and ResNet-101 models. Extensive models are used to boost performance. The final results of a single architecture (e.g. ResNet-18) were acquired from 18 different models (obtained by different training settings). Yang et al. (2017) used multi-task framework (GoogleNet and U-net) for learning skin lesion segmentation and classification jointly. González-Díaz (2019) trained a Fully Convolutional Network (FCN) for detecting ROI. In addition, González-Díaz (2019) also incorporates the meta-data information and attribute information to improve performance.

In the proposed method, ensemble method is used during the testing time to improve the performance. Our final model is obtained by fusing outputs given input images from multiple scales (based on performance on the validation set), which does not require extra training process. Performances of inputs with different scales are shown in Table 4, and experimental results regarding the ensemble of scales on the validation set are shown in Table 5. As shown in Table 4, for input images with different scales, the best performance is obtained at scale 224. Scale 672 comes the second best, and scale 448 gives the least satisfacoty performance. From Table 5, we can see that, in general, adding more scale information can improve the performance, but the best per-

Table 5
Ensemble performance with input of different scales on the validation set.

224	224 _*entire*	448	672	AUC
√	√			94.0
√		√		93.6
√			√	94.1
√	√	√		94.5
√	√		√	**95.2**
√		√	√	94.3
√	√	√	√	94.9

Table 6
Performance comparison with state-of-the-art methods on ISIC 2017 dataset (AUC score).

Method	Ensemble	External data	Mel	SK	Avg
Matsunaga et al. (2017)	Y	1444	86.8	95.3	91.1
Menegola et al. (2017)	Y	7544	**87.4**	94.3	90.8
Bi et al. (2017)	Y	1600	87.0	92.1	89.6
Yang et al. (2017)	N	0	83.0	94.2	88.6
González-Díaz (2019)	N	2828	87.3	96.2	91.7
Mahbod et al. (2019)	Y	187	87.3	95.5	91.4
Proposed	Y	0	87.0	**97.1**	**92.1**

formance (AUC of 95.2%) is obtained by fusing results with input scale 224, 224 _*e ntire* (i.e., the whole image without ROI segmentation), and 672. The fusion with scale 448 gives less satisfactory performance compared with the other scales. This is consistent with the results reported in Table 4 that scale 448 performs the least satisfactory and fusing a less satisfactory output (obtained with input scale 224) would not increase the performance. This may due to the fact that, the parameters of SVM are not selected for these inputs with different scales. When the input size is larger than [224 × 224], the output of pretrained neural networks will be multi-channel features instead of one feature vector. We reshape the multi-channel features into one feature vector, and PCA is then used to reduce the feature dimension, which is described in Section 2.2. This will result in raw features with different input dimensions, given the inputs of different scales. The finetuned parameters of scale 224 are used for the proposed method for simplicity. The use of entire images is to get information about the lesion size and skin regions, which can benefit the proposed model (Bissoto et al., 2019). Therefore, our final performance is obtained by fusing the outputs with input scale [224 × 224], [672 × 672] and the entire image with scale [224 × 224]. It is worth noting that consistent ensemble trend regarding the fusing of different input scales has also been found on the testing set.

Final results compared with state-of-the-art methods are shown in Table 6. The column "ensemble" indicates whether the compared methods use ensemble technique or not, and column "external data" shows the number of external data used for training neural networks. The external data plays an important role in the training of CNN models for the skin lesion classification task. For instance, (González-Díaz, 2019) got the best performance of 90.8% (vs. 91.7%) for the models with less external training sets, even when it incorporated the meta-data and attribute data. From Table 6, we can see that the proposed method provides a superior performance compared with state-of-the-art methods without using external data. The proposed method achieves the best AUC of 97.1% for seborrheic keratosis, which verifies the effectiveness of the proposed method.

Table 7
Average AUC of the proposed method with and without skin lesion segmentation.

Input	ResNet	DenseNet	Fusion
I_{whole}	84.11	84.62	86.10
I_{ROI}	**89.00**	**88.85**	**90.67**

We also display some challenging images that have been correctly classified by the proposed method in Fig. 7. The left images are skin lesions of melanoma, while the right images are skin lesions of seborrheic keratosis. Strong visual similarity and artifacts are observed in these two types of images, yet the proposed method successfully classifies these hard samples, which implies that the proposed method can tackle difficult samples.

3.8. Time complexity

There are two binary classification tasks in this paper. For simplicity, the following time complexity is calculated on the classification task of seborrheic keratosis using features extracted via ResNet. The platform information has been described in Section 3.3. The training time of the proposed method is a total of 96 min. Out of 96 min, 92.4 min are spent to train the segmentation network, and about 3.6 min are used to train the classification model. The testing time is 0.39 s on average for one given image. Typically, the extra time induced by the proposed method (mid-level feature processing) is 0.13 s, which is relatively fast. The fast inference time of the proposed method indicates the potential in clinical application.

3.9. Discussions

3.9.1. Effect of lesion segmentation

To determine the influence of the skin lesion segmentation, we conduct experiments with and without skin lesion segmentation as a primary step, and experiment results are shown in Table 7. I_{whole} means using the whole images as input, and I_{ROI} means using

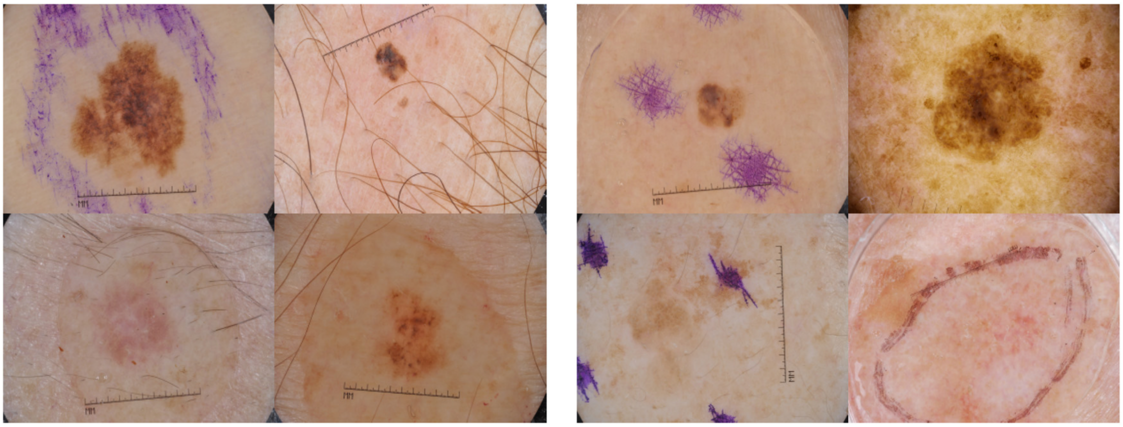


Fig. 7. Examples of correctly classified images: left: melanoma; right: seborrheic keratosis.

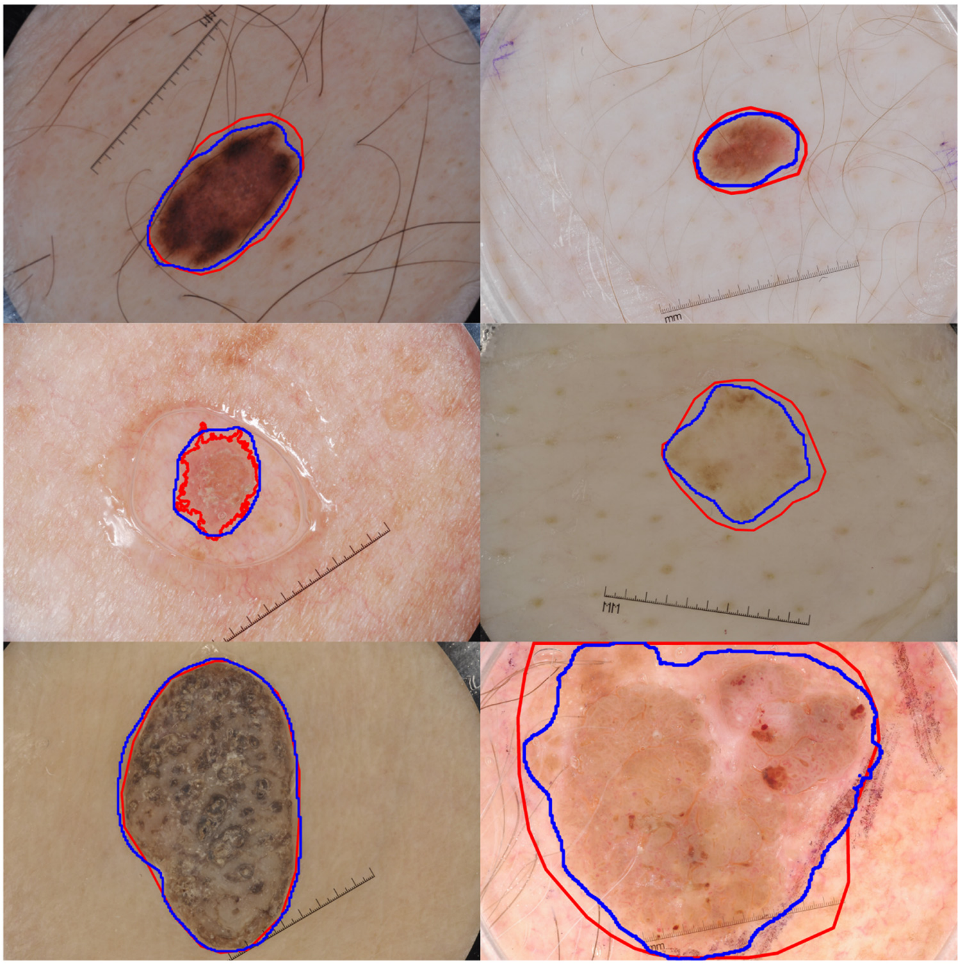


Fig. 8. Segmentation results of the proposed method. The red contours are the ground truths, and the blue contours are the segmentation results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Effect of weighting factor for APG algorithm.

Weighting	ResNet	DenseNet
Uniform	87.88	88.60
Proposed	**89.00**	**88.85**

Table 9
Comparison of AUC scores of the proposed method using discriminative features and mid-level features.

Input	ResNet	DenseNet	Fusion
Discrim Fea	86.53	87.68	88.71
Mid-level Fea	**89.00**	**88.85**	**90.67**

the ROI images obtained with skin lesion segmentation method as input. As shown in Table 7, the proposed method obtains significant improvement if we use the ROI images as input. This is mainly because the interest regions of skin lesions are of various scales, and some targets are very small and only occupy a small region of the whole image. Directly downsampling all images to the same size ([224, 224]) will lose detail information about skin lesions, and make it even difficult to observe patterns of skin lesions. Some examples of the predicted binary masks are shown in Fig. 8.

3.9.2. *Effect of weighting factor for APG algorithm*

In this section, we conduct experiments to show the benefits of the weighting scheme introduced in Section 2.3. We compare the proposed method with uniform weight and experiment results are shown in Table 8. The weighting scheme used in this paper can get an improvement of 1.12% and 0.15% with features extracted via ResNet and DenseNet, respectively. Though minor improvement is observed for the features extracted via DenseNet with online weighting, it can still be regarded as useful overall. One reason for this may be that focusing more on hard violating pairs can benefit the training phase of metric learning.

3.9.3. *Advantage of soft discriminative feature*

As shown in Eq. (2), the metric learning problem can also be regarded as a discriminative subspace learning problem. The new feature representation can be represented as Lx_i given the input feature x_i, which is a discriminative feature representation. Compared with the discriminative feature Lx_i learned based on metric learning, the proposed mid-level feature representation v_i in Eq. (6) is a soft descriptor which uses affinity information as the new feature representation. In this section, we also implement experiments to compare the proposed mid-level features v_i with the discriminative feature Lx_i. To make a fair comparison, best parameters of the discriminative features based on metric learning are also selected as in Section 3.5. Experimental results are shown in Table 9.

As shown in Table 9, the mid-level features outperform the discriminative features for both the features extracted via ResNet and DenseNet. This is because learning an optimal feature representation that can well separate all the samples (especially the hard

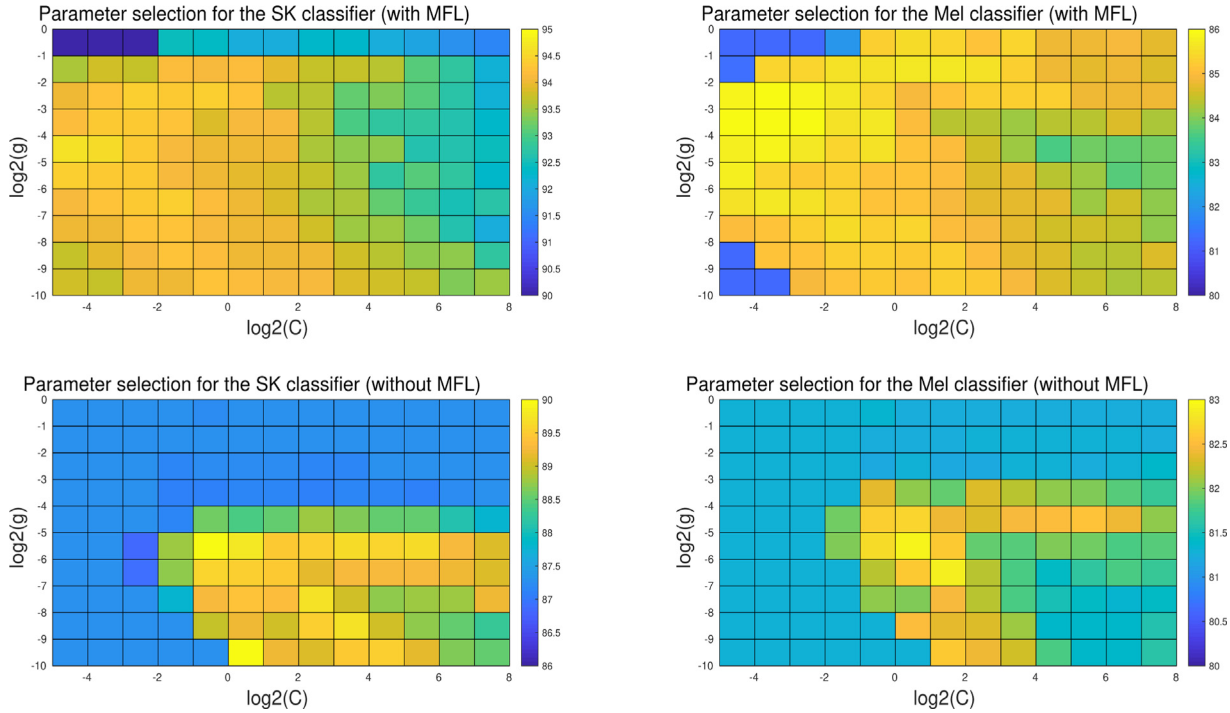


Fig. 9. Performance of the proposed method (with and without the MFL module) on the parameter space. The learned mid-level features are more robust and discriminative compared with the original features. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

samples) is very difficult. In contrast, the proposed mid-level feature representation is a soft discriminative descriptor, where the relationships of visual similarities and distinctions can be kept for some difficult cases (hard samples) as long as the remaining relationships are captured correctly. Also note that, compare with Table 2, the discriminative features obtain better performance than the original features, which demonstrates that the original features have poor discriminative power, and using the discriminative features can promote the classification task.

3.9.4. Robustness of the proposed mid-level features against parameters

In this section, to show the robustness of the proposed mid-level feature representation against parameters, performance (5-fold cross-validation on the training set) of the proposed method on the parameter space is given in Fig. 9. As shown in Fig. 9, the mid-level features outperform the original features by a large margin for both binary classifiers. Especially, the AUC of proposed method for the SK classifier ranges from 90% to 95% across the parameter space, while the AUC of proposed method without MFL ranges from 86% to 90%. The best performance of the proposed method is about 5% higher than the proposed method without MFL, which proved that the mid-level features contain more discriminative power. A similar trend is also observed for the MEL classification. Also note that the ranges of color bars are similar for the method with and without MFL, and as the score increases, the color of the parameter space changes from blue to yellow. A large area of the proposed method's parameter space is yellow while only a minor part is yellow for the proposed method without MFL, which demonstrates the robustness of the proposed mid-level features.

4. Conclusions

Automatic melanoma detection is a challenging task due to the large inter-class similarity and intra-class variation, and complex skin conditions among different skin lesions. In this paper, a novel framework for skin lesion classification is proposed. Skin lesion segmentation is first performed to get the ROI images for the later classification task. A novel mid-level feature representation is obtained by using metric learning and a reference set. The learned mid-level feature representation contains affinity information among image samples, which is a soft discriminative feature, having more tolerance to the hard samples thus being more robust. Experimental results show that skin lesion segmentation can benefit the subsequent classification task. Meanwhile, the learned mid-level features obtain much better performance compared with the original features. Experimental results show that the proposed method outperforms state-of-the-art CNN based methods, which verify the effectiveness of the proposed method.

Authors' contributions

Lina Liu: Methodology, Software, Investigation, Validation, Writing – Original Draft

Lichao Mou: Methodology

Xiao Xiang Zhu: Methodology

Mrinal Mandal: Conceptualization, Supervision, Writing – Review & Editing, Funding acquisition, Project administration

Conflict of interest

The authors declare no conflict of interest.

References

Abbas, Q., Fondón, I., Sarmiento, A., Celebi, M.E., 2014. An improved segmentation method for non-melanoma skin lesions using active contour model. International Conference Image Analysis and Recognition, 193–200.

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M., 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch. Dermatol. 134, 1563–1570.

Bi, L., Kim, J., Ahn, E., Feng, D., 2017. Automatic Skin Lesion Analysis Using Large-Scale Dermoscopy Images and Deep Residual Networks. arXiv preprint arXiv:1703.04197.

Bissoto, A., Fornaciali, M., Valle, E., Avila, S., 2019. de) constructing bias on skin lesion datasets. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Borys, D., Kowalska, P., Frackiewicz, M., Ostrowski, Z., 2015. A simple hair removal algorithm from dermoscopic images. International Conference on Bioinformatics and Biomedical Engineering, 262–273.

Codella, N.C., Gutman, D., Celebi, M.E., et al., 2018. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging. International Symposium on Biomedical Imaging, 168–172.

Ebner, M., 2007. Color Constancy, vol. 7. John Wiley & Sons.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115.

Ge, Z., Demyanov, S., Bozorgtabar, B., et al., 2017. Exploiting local and generic features for accurate skin lesions classification using clinical and dermoscopy imaging. International Symposium on Biomedical Imaging, 986–990.

Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., Garnavi, R., 2017. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. International Conference on Medical Image Computing and Computer-Assisted Intervention, 250–258.

González-Díaz, I., 2019. Dermaknet: incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. IEEE J. Biomed. Health Inform. 23, 547–559.

Gu, Y., Zhou, J., Qian, B., 2017. Melanoma detection based on mahalanobis distance learning and constrained graph regularized nonnegative matrix factorization. IEEE Winter Conference on Applications of Computer Vision (WACV), 797–805.

Hazen, B.P., Bhatia, A.C., Zaim, T., Brodell, R.T., 1999. The clinical diagnosis of early malignant melanoma: expansion of the abcd criteria to improve diagnostic sensitivity. Dermatol. Online J. 5.

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

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708.

Humayun, J., Malik, A.S., Kamel, N., 2011. Multilevel thresholding for segmentation of pigmented skin lesions. In: 2011 IEEE International Conference on Imaging Systems and Techniques, IEEE, pp. 310–314.

Liao, S., Li, S.Z., 2015. Efficient psd constrained asymmetric metric learning for person re-identification. IEEE International Conference on Computer Vision, 3685–3693.

Liu, L., Mou, L., Zhu, X.X., Mandal, M., 2019. Skin lesion segmentation based on improved u-net. In: 2019 IEEE Canadian Conference of Electrical and Computer Engineering, Edmonton, AB, Canada.

Ma, Z., Tavares, J.M.R., 2017. Effective features to classify skin lesions in dermoscopic images. Expert Syst. Appl. 84, 92–101.

Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C., 2019. Fusing fine-tuned deep features for skin lesion classification. Comput. Med. Imaging Graphics 71, 19–29.

Matsunaga, K., Hamada, A., Minagawa, A., Koga, H., 2017. Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. arXiv preprint arXiv:1703.03108.

Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E., 2017. Recod Titans at ISIC Challenge 2017. arXiv preprint arXiv:1703.04819.

Oliveira, R.B., Papa, J.P., Pereira, A.S., Tavares, J.M.R., 2018. Computational methods for pigmented skin lesion classification in images: review and future trends. Neural Comput. Appl. 29, 613–636.

Pehamberger, H., Steiner, A., Wolff, K., 1987. In vivo epiluminescence microscopy of pigmented skin lesions. i. pattern analysis of pigmented skin lesions. J. Am. Acad. Dermatol. 17, 571–583.

Rebouças Filho, P.P., Peixoto, S.A., da Nóbrega, R.V.M., Hemanth, D.J., Medeiros, A.G., Sangaiah, A.K., de Albuquerque, V.H.C., 2018. Automatic histologically-closer classification of skin lesions. Comput. Med. Imaging Graphics 68, 40–54.

Riaz, F., Naeem, S., Nawaz, R., Coimbra, M., 2018. Active contours based segmentation and lesion periphery analysis for characterization of skin lesions in dermoscopy images. IEEE J. Biomed. Health Inform. 23, 489–500.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241.

Siegel, R.L., Miller, K.D., Goding Sauer, A., Fedewa, S.A., Butterly, L.F., Anderson, J.C., Cercek, A., Smith, R.A., Jemal, A., 2020. Colorectal cancer statistics, 2020. CA: A Cancer J. Clin. 70, 145–164.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Stolz, W., Riemann, A., Cognetta, A., Pillet, L., Abmayr, W., Holzel, D., Bilek, P., Nachbar, F., Landthaler, M., 1994. Abcd rule of dermatoscopy – a new practical method for early recognition of malignant-melanoma. Eur. J. Dermatol. 4, 521–527.

Wong, A., Scharcanski, J., Fieguth, P., 2011. Automatic skin lesion segmentation via iterative stochastic region merging. IEEE Trans. Inf. Technol. Biomed. 15, 929–936.

Yang, X., Zeng, Z., Yeo, S.Y., Tan, C., Tey, H.L., Su, Y., 2017. A Novel Multi-Task Deep Learning Model for Skin Lesion Segmentation and Classification. arXiv preprint arXiv:1703.01025.

Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., Wang, T., 2018. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Trans. Biomed. Eng. 66, 1006–1016.

Yuan, Y., Chao, M., Lo, Y.C., 2017. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. IEEE Trans. Med. Imaging 36, 1876–1886.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging 38, 2092–2103.