IET Journals
The Institution of Engineering and Technology

Deep Residual Network with Regularized Fisher Framework for Detection of Melanoma

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Nazneen N Sultana[1], *Bappaditya Mandal*[2] *and N. B. Puhan*[1]

[1] *School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar, Odisha 752050, India*
[2] *School of Computing and Mathematics, Keele University, Newcastle ST5 5BG, United Kingdom.*
** E-mail: {nns11, nbpuhan}@iitbbs.ac.in, b.mandal@keele.ac.uk.*

Abstract: Of all the skin cancer that is prevalent, melanoma has the highest mortality rates. Melanoma becomes life threatening when it penetrates deep into the dermis layer unless detected at an early stage, it becomes fatal since it has a tendency to migrate to other parts of our body. This paper presents an automated non-invasive methodology to assist the clinicians and dermatologists for detection of melanoma. Unlike conventional computational methods which require (expensive) domain expertise for segmentation and hand crafted feature computation and/or selection, a deep convolutional neural network based regularized discriminant learning framework which extracts low dimensional discriminative features for melanoma detection is proposed. Our approach minimizes the whole of within-class variance information and maximizes the total class variance information. The importance of various subspaces arising in the within-class scatter matrix followed by dimensionality reduction using total class variance information are analyzed for melanoma detection. Experimental results on ISBI 2016, MED-NODE, PH2 and the recent ISBI 2017 databases show the efficacy of our proposed approach as compared to other state-of-the-art methodologies.

1 Introduction

According to the American Cancer Society, there will be an estimated 1,735,350 new diagnosed cancer cases and 609,640 cancer deaths in the United States out of which 87,290 new cases of melanoma will be diagnosed in 2018 [1]. Although melanoma accounts for around 1% of all the skin cancer, it has the highest mortality rate. The rate of occurrence of melanoma from 2004-2014 has increased by 2-3% per year. In 2018, an estimated 9,320 death will occur due to melanoma [1]. Early diagnosis is quite important because melanoma if detected at an early stage, it can be curable. The five year survival rate is 95% when detected early and this reduces to around 13% if detected at an advance stage of melanoma and the cost of treatment is also quite high [2]. With the advent of dermoscopy (also referred as epiluminescence microscopy), it has been possible to assist clinicians through computer aided diagnosis efficiently since dermatoscope captures the dermal features and eliminates the surface glare. Dermatoscope is a non-invasive technique which magnifies the structures otherwise invisible to naked eyes; thus helps in detection of melanoma from other types of skin cancer. There is an improvement of 5-30% in the detection while using dermoscopy and clinical images as compared to the naked eye examination [3]. Some examples of the benign and melanoma cases from the International Skin Imaging Collaboration (ISIC) database [4] are shown in Fig. 1.

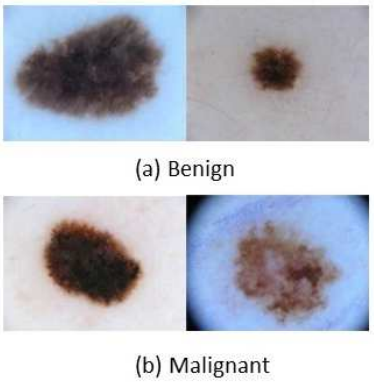

Fig. 1: Example of dermoscopy images from ISIC database. Using naked eye examination it is difficult to differentiate melanoma (malignant) from non-melanoma (benign) due to the low inter-class variation and high intra-class variation.

Clinically, there are numerous empirical approaches like, ABCD rule (A - stands for Asymmetrical shape, B - border of the lesion, melanomas generally have irregular borders, C - Color, uneven distribution of color can sometimes be a warning sign of melanoma, D - Diameter melanoma lesions are often greater than 6mm in diameter) [5], Menzies method [6] and CASH (color, architecture, symmetry and homogeneity) [7], which have been developed to enhance the ability of clinicians and dermatologists to distinguish between melanoma from benign nevi. However, even for expert clinicians to have correct diagnosis is not trivial and very often the decision taken by human visual inspection is subjective in nature. Hence automated methods have been developed to assist clinicians, primary care physicians and staffs as the screening tool for referrals. However, unsatisfactory accuracy and results are still an issue for diagnosing this particular disease. Immense challenges are involved when the acquired images have low contrast and obscure boundaries. The presence of artifacts both natural (hairs, veins) and man-made (presence of small air bubbles, ruler marking) aggravates this situation. The other significant property of melanoma skin lesion is that it has huge intra-class variation in terms of color, texture, shape, size and location in the dermoscopy images and high degree of visual similarity between melanoma and non-melanoma lesions making it difficult to discriminate them.

In 2017 Nature article, Esteva *et al.* [8] described a computer vision based artificial intelligence (AI) system trained on a data set of 129,450 clinical images of 2,032 different diseases and compared its diagnostic performance against 21 board-certified dermatologists. They found that the AI system is quite capable of classifying skin cancer at a level of competence comparable to the dermatologists. Earlier approaches used low-level visual feature representations such

© The Institution of Engineering and Technology 2015

as color and texture features for melanoma detection. Color features include mean and standard deviation, color variance, color histogram, color asymmetry using different color spaces. Texture features include gradient histogram [9, 10], gray level co-occurrence matrix [11], which computes the dissimilarity index, mean and standard deviation. The color constancy method [12] uses a multisource image database (such as interactive atlas of dermoscopy from Environmental Design Research Association (EDRA)) acquired under different setups. They have shown that changes in the illumination and acquisition devices could alter the color of images and often reduce the performance of such systems. Using experiments on single source dataset, they showed that there was no significant improvement in the accuracy level using color constancy. The MLR_JRC [13] is a detection methodology for dermoscopy images using multi-scale lesion-biased representation (MLR) and joint reverse classification (JRC). This method alleviates the problems arising from the various lesion sizes and shapes, fuzzy lesion boundaries, different skin color images and presence of hair. The fusion of structural and textural features in [14] involves features from wavelet and curvelet transforms for capturing structural and different variants of local binary patterns as textual features for melanoma detection. Abuzaghleh *et al.* [15] extracted the color and texture geometry features using which they integrated a two-level classifier: the first classifier divides into normal and abnormal skin lesion and the second classifier further divides the abnormal into atypical and melanoma.

To extract the lesion features, one needs to acquire significant domain knowledge in this particular field, to recognize the importance of different features in the detection of melanoma. One of the important steps in low-level classical approach is segmentation and/or pre-processing. Traditionally, many researchers apply segmentation before feature extraction, which is cumbersome, labor intensive and prone to errors. Such segmentation separates the lesion from the surrounding normal skin in order to perform accurate lesion analysis and feature extraction. Thus, the conventional methods are time consuming and certainly require (expensive) domain expertise. Removal of artifacts is done in the pre-processing step using DullRazor hair removal algorithm [16], median filtering [17], directional filters and illumination enhancement [18]. Lesion segmentation and complex pre-processing are non-trivial steps where errors propagate and may require human intervention. For example, the skin lesion segmentation approach of Li *et al.* [19] require manual initialization, post-processing and depth information. Further, they report large variations in the manual lesion segmentations done by dermatologists, which may indicate segmentations are subjective. In addition, these low-level hand-engineered features have limited discrimination capability to differentiate melanoma from non-melanoma skin lesion. Thus, in this work a state-of-the-art image feature extractor that does not require lesions' segmentation nor any complex preprocessing, in the form of a pretrained fully-convolutional neural network is developed. Overviews of different conventional methods can be found in [11, 20].

Recently, deep learning techniques have led important breakthroughs in many image processing tasks including medical image analysis [21]. Researchers have applied convolutional neural networks for the detection of melanoma, taking the advantage of its discrimination capability. Lopez *et al.* [22] compares three different methods of using VGGNet architecture, i.e. training from scratch, transfer learning and fine-tuning. It has been observed that small sized dataset fine-tuning is the most preferred approach for classification. Fine-tuning updates the weights of the pretrained network by continuing the backpropagation on the present data, since the later layers of the ConvNet are more specific to the details of the classes contained in the original dataset. Zhen *et al.* [23] combines deep convolutional neural network with fisher vector encoding and support vector machine (SVM) classifier. Fisher vector encoding improves the invariant and more discriminative properties of the deep features extracted from a pretrained model, however they are of very high dimensional ($> 12,000$).

Yu *et al.* [24] integrate the fully convolutional residual network for segmentation (FCRN) and very deep residual networks for classification to form two-stage framework. Deep residual network ResNet solves the vanishing gradient and over-fitting problem as the network goes deeper. Their integrated network was tested on different models (VGG-16, GoogleNet and DRN-50); DRN-50 gave the best results among all. Their paper was placed as 1^{st} place in the ISBI 2016 challenge with average precision score of 0.637. Codella *et al.* [25] applies fully convolutional network similar to that used in U-Net architecture for segmentation and non-linear SVM classifier is used to classify the individual features. For classification, they have an ensemble of features like color histogram, edge histogram, multi-scale color LBP, sparse coding, convolutional neural network (CNN) (FC6, 4096 dimension feature vector), DRN-101 (1000 dimension), fully convolutional U-Net used as a shape descriptor. Ensemble of different features increases the dimension and hence increased their computational complexity.

Residual CNN with hierarchical feature learning capability have lead breakthroughs in many medical image analysis tasks. The main concern compared to natural image processing problem is that the datasets used for training in medical imaging are quite small. This makes the deep networks difficult to train effectively with large amount of parameters (in millions [26]). Another challenge is that the interclass variation in medical image analysis tasks are usually much smaller than that in natural image processing tasks (the interclass variation between non-melanoma from the melanoma is very small as compared to interclass variation between a man and a horse, for example). Thus in addition to the deep network, a technique is required which would minimize this intraclass variation and maximize the interclass variation.

Motivated by these works, in this paper, a new efficient deep convnet framework based on statistical discriminant analysis for distinguishing melanoma from non-melanoma cases is proposed. A very deep residual neural network (*i.e.* 50 layers) pre-trained on ImageNet is first applied to each input image. In our proposed method, fine-tuning approach is used which includes replacing of the last few fully connected layers and retraining the whole network and then the local deep descriptors are extracted from the dense activation maps of the last convolutional layer. Discriminative features are then extracted from these local deep descriptors using the variance information from intra-class and inter-class among the training samples. They are regularized and finally low dimensional features are extracted to distinguish between melanoma and non-melanoma images. Experimental results on four benchmarking databases demonstrate the effectiveness of our proposed approach as compared to other state-of-the-art methods.

2 Proposed Approach

2.1 Problem Formulation

Recent works on large scale image recognition tasks have demonstrated that increasing the number of layers in convolutional neural network can improve the efficiency rate. But simply stacking the layers to increase the network depth leads to saturation of accuracy and then further degrades rapidly [27]. This degradation is not as a result of over-fitting but because of adding more layers. He *et al.* [26] gave a perfect solution to the above problem by introducing deep residual network (ResNet) and reduced the error rate to 3.57% on the ImageNet test set. Their basic building block is shown in Fig. 2.

As shown in Fig. 2, deep residual network is composed of a set of residual blocks, each of which consists of convolutional layers, batch normalization layers and rectified linear unit (RELU) as an activation function. RELU due to its linear and non-saturating form, greatly accelerate the convergence of stochastic gradient descent compared to *tanh* and *sigmoid* functions. Batch normalization properly initializes neural networks by explicitly forcing the activation's throughout a network to take on an unit Gaussian distribution at the beginning of the training.

CNN has the ability to effectively learn the features automatically useful for recognition according to the given training dataset. The model contains multiple processing layers to learn different level features. The lower level layers of the network contain more generic features such as to detect simple edges or colors of the image. As

© The Institution of Engineering and Technology 2015

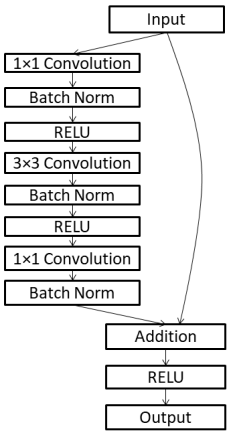

Fig. 2: Illustration of a typical residual block of ResNet (50 layers), where each layer consists of Convolutional (conv), Rectified Linear Unit (RELU) and Batch Normalization (Batch Norm) layers. The shortcut connections perform identity mapping and their outputs are added to the outputs of the stacked layers. The three convolutional layers are 1×1, 3×3 and 1×1. The 1×1 layers are responsible for reducing the dimensions and then increasing (restoring) the dimensions. After constructing the residual block, very deep networks are built by stacking residual blocks [26, 27].

the layers are increased, more and more specific features are recognized according to the training dataset. Thus the higher level features which are more specific to the problem are extracted for further processing. The only difference between ResNets and normal ConvNets is that ResNet provides a clear path for gradients to back propagate to early layers of the network thus making the learning process faster. ResNets can be seen as multiple basic blocks that are serially connected to each other and these skip connections parallel to each basic block gets added to its output. As shown in [27], it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping.

In our framework, we first apply the input image to a very deep (50 layers) residual neural network pre-trained on ImageNet. A fine-tuning approach is used which includes replacing of the last few fully connected layers and retraining the whole network. Local deep descriptors are extracted from the dense activation maps of the last convolutional layer. Discriminative features are then extracted from these local deep descriptors using the variance information from intra-class and inter-class among the training samples. They are regularized and finally low dimensional features are extracted to distinguish between melanoma and non-melanoma images. Since our problem contains very small inter-class variation and high intra-class variation, we propose to extract the deep features and perform discriminant analysis to maximize the total class variation and minimize the within-class variation among the training samples.

2.2 Discriminant Analysis

Linear discriminant analysis projects the high dimensional features onto a lower-dimensional space while taking the class information to have good class-separability in order to avoid over-fitting and also to reduce computational costs. It tries to find out the axes which maximizes the between-class scatter matrix $\sum_b$, while minimizes the within-class scatter matrix $\sum_w$ in the projective subspace. To find the fisher discriminants for a set of images, at first, the within-class scatter matrix is computed as

$$\sum_w = \sum_{i=1}^{p} \sum_{j=1}^{q_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T, \tag{1}$$

where $\sum_w$ is the within-class scatter matrix which computes the amount of scatter between images of the same class. Here x_{ij} represents the j^{th} normalized image features belonging to class i, $\bar{x}_i$ indicates the mean of the training samples in class i. p is the number of classes and q_i is the number of samples in i^{th} class. After computation of within-class scatter matrix $\sum_w$, the between-class covariance matrix $\sum_b$ is computed using the following equation:

$$\sum_b = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T, \tag{2}$$

where n_i the number of images in the class i, $\bar{x}_i$ is the mean of the images in the class and $\bar{x}$ is the mean of all the images. Fisher criterion is defined as the ratio of between-class and within-class variances, given by:

$$J(W) = \frac{|W^T \sum_b W|}{|W^T \sum_w W|}. \tag{3}$$

Here W is the weight vector. The goal is to maximize $J(W)$ by minimizing $\sum_w$ and maximizing $\sum_b$. To extract discriminative features, at first eigen decomposition of the within-class scatter matrix $\sum_w$ is performed, given by:

$$\sum_w = \Phi \Lambda \Phi^T. \tag{4}$$

Here, Φ are the eigenvectors and Λ are the eigenvalues. We regularize this subspace by performing experiments on different selected subspaces to find the best subspace for projection. This selection of subspace is defined in the Section 2.3.

Let the projection matrix be Ψ, which contains the eigenvectors obtained from (4), the output y after projecting the training features x onto this matrix, is given by:

$$y = \Psi^T x. \tag{5}$$

Since our problem is a binary classification problem, original Fisher criteria [28] cannot be directly applied as the numerator ($\sum_b$) in (3) would provide us only $p - 1$ number of features, which is just 1 in our case. There are some prior works [29–31], reporting that when the training data are small, in place between-class scatter matrix, total scatter matrix performs better with Fisher criteria (3) and is less sensitive to noises and different training databases. Hence for our binary classification, between-class scatter matrix is modified to total scatter matrix using the data after projecting the training features onto the selected eigenvectors of the within-class scatter matrix. This would help us to extract final number of features, which is $\sum_{i=1}^{p} q_i - 1$ as described in [28].

To extract the discriminative projection vectors, using the data matrix y, total scatter matrix is computed as:

$$\sum_t = \sum_{i=1}^{p} \sum_{j=1}^{q_i} (y_{ij} - \bar{y})(y_{ij} - \bar{y})^T \tag{6}$$

After computing the covariance matrix the projection matrix Ω is selected by eigen decomposition of $\sum_t$ and selecting the eigenvectors in Φ_{wy} according to the most significant eigenvalues Λ_{wy}. Eigen decomposition of $\sum_t$ is given by:

$$\sum_t = \Phi_{wy} \Lambda_{wy} \Phi_{wy}^T. \tag{7}$$

Thus, Ψ and Ω, are the two different projection matrices obtained using two covariance matrices $\sum_w$ and $\sum_t$, respectively. The final set vectors that are used for classification is computed as:

$$z = \Psi^T \Omega^T x, \tag{8}$$

where x represents either the training or testing features. The block diagram of the proposed method is summarized in Fig. 3.

© The Institution of Engineering and Technology 2015

2.3 Regularization in the Subspace

In the conventional linear discriminant analysis (LDA) following the Fisher criteria (3) and (4), many researchers have been selecting only the eigenvectors corresponding to the principal eigenvalues (non-zero eigenvalues) [30, 32, 33] from the within-class scatter matrix eigen analysis, thereby, losing crucial within-class discriminatory information. This has been pointed out in many applications such as face recognition [34] and person re-identification [35]. Since the intra-class (within-class) variation is high in melanoma, this large variation is modeled using our regularized discriminant analysis by using the LDA concept of minimizing the within class variation and maximizing the total variation of the lesion images. But since ours is a binary classification problem it is not possible to apply Fisher criteria directly and compute between-class scatter matrix for final feature selection, we modify this to compute the total scatter matrix, which will still minimize the within-class variability and maximize some portion of the between-class variability [36]. After decomposition of the within-class scatter matrix using (4), the projection matrix Ψ is divided into three subspaces to study the importance of different eigenspaces. Below three cases are created which show the differences of each of these subspaces and their formulations:

Case 1: selects the eigenvectors corresponding to the non-zero eigenvalues in Λ and dividing each eigenvector vector with its corresponding square root of eigenvalues, given by:

$$\Psi = \Phi_i(\Lambda_i)^{-\frac{1}{2}}, \tag{9}$$

where Φ_i is the eigenvector corresponding to Λ_i eigenvalue.

Case 2: selects the null space eigenvectors (corresponding to zero eigenvalues only) and divide each with the square root of the smallest eigenvalue,

$$\Psi = \Phi_i(\Lambda_{smallest})^{-\frac{1}{2}}, \tag{10}$$

where $\Lambda_{smallest}$ is the smallest non-zero eigenvalue from the non-zero eigenspace.

Case 3: is the concatenation / merging of the described two cases, containing regularized eigenvectors corresponding to both zero and non-zero eigenvalues,

$$\Psi = [\Phi_i(\Lambda_i)^{-\frac{1}{2}} \qquad \Phi_i(\Lambda_{smallest})^{-\frac{1}{2}}]. \tag{11}$$

The input training features are then projected onto this projection matrix Ψ, given by (5). Dividing each eigenvectors with square root of the eigenvalues improves the performance metrics quite significantly. It is expected that selecting the eigenvectors of null space would give better results since the null space eigenvalue means that there is zero variance between the within-class properties and also eigenfeatures are regularized using our proposed model, *i.e.* our approach is able to extract low dimensional discriminatory information arising from the within-class (scatter) variance information for detection of melanoma.

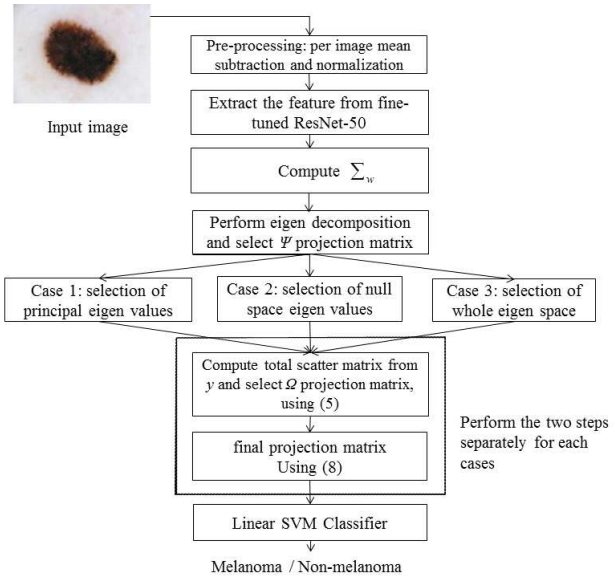

Fig. 3: Block diagram of the proposed method.

2.4 Training Procedure

Since we are using pre-trained ResNet-50 model, while training the network, each image is re-sized into a fixed size of 224 $\times$ 224. The images are then normalized by subtracting channel-wise mean intensity values from corresponding individual image channels in the color image. Each channel of the color image is also made to unit standard deviation. ResNet-50 pretrained network is fine-tuned using our dataset by retraining the whole network. Training only the higher-level layers helps in extraction of features specific to the present dataset. It also avoids overfitting, since the number of images are few as compared to ImageNet which contains millions of images. Additionally it avoids overfitting and increases the robustness. Data augmentation is applied in our framework to increase the performance of the network via a set of random transformation to the images. The augmentation operators include rotation (90, 180 and 270 degrees), translation, horizontal and vertical flipping of the original images.

Let $I(width, height, channels)$ represent the resized and normalized input image of size $width \times height$ and number of channels as depth. Since RGB color images are used, the depth dimension is 3. $C(m, n, f)$ represents the convolutional layer, where m is the filter size, n is the strides and f is the number of filter banks. $P(s, r)$ represents the pooling layer, where s is the number of strides and r is the size of window for subsampling. Each convolutional layers is followed by a batch normalization layer and RELU as a non-linearity function. The summations at the end of each residual unit are followed by a ReLU unit. Each repetitive residual unit is presented inside R. $F(e)$ denotes the fully connected layer where e is the number of neurons. Thus, the fine-tuned ResNet-50 can be represented as:

$$\begin{aligned}\Theta_R = {} & I(224, 224, 3) \rightarrow C(7, 2, 64) \rightarrow P(2, 3) \rightarrow \\ & 3 \times R(C(1, 1, 64) \rightarrow C(3, 1, 64) \rightarrow C(1, 1, 256)) \rightarrow \\ & R(C(1, 2, 128) \rightarrow C(3, 2, 128) \rightarrow C(1, 2, 512)) \rightarrow \\ & 3 \times R(C(1, 1, 128) \rightarrow C(3, 1, 128) \rightarrow C(1, 1, 512)) \rightarrow \\ & R(C(1, 2, 256) \rightarrow C(3, 2, 256) \rightarrow C(1, 2, 1024)) \rightarrow \\ & 5 \times R(C(1, 1, 256) \rightarrow C(3, 1, 256) \rightarrow C(1, 1, 1024)) \rightarrow \\ & R(C(1, 2, 512) \rightarrow C(3, 2, 512) \rightarrow C(1, 2, 2048)) \rightarrow \\ & 2 \times R(C(1, 1, 512) \rightarrow C(3, 1, 512) \rightarrow C(1, 1, 2048)) \rightarrow \\ & P^*(1, 7) \rightarrow F(e) \rightarrow Softmax\end{aligned} \tag{12}$$

The length of $F(e)$ depends on the number of categories to classify, e is the number of classes. P^* refers to average pooling rather than max pooling as used in literature. The softmax function or the normalized exponential function used during the training process is described as:

$$S(F)_j = \frac{exp^{F_j}}{\sum_{k=1}^{e} exp^{F_k}}, for\ j = 1, 2, ...e. \tag{13}$$

Our final classification during the test is done using SVM classifier [37]. In all our experiments, our proposed method is validated with the existing ones using the same corresponding datasets and protocols. They are implemented on a system with Intel Core i7 processor, 16GB RAM, and NVIDIA GeForce GTX-1050Ti GPU card.

Our proposed framework is summarized in Algorithm 1.

© The Institution of Engineering and Technology 2015

Algorithm 1 : Proposed Approach

Input: Normalized images $I(224 \times 224 \times 3)$
Output: Melanoma / Non-melanoma

procedure
  Fine-tune ResNet-50 model
  Extract training and testing features
Training data:
  Compute $\sum_w$ and perform its eigen decomposition using (4)
  Enter Case $\leftarrow$ goto line 11
  Compute y using (5) to calculate $\sum_t$.
  Perform eigen decomposition of $\sum_t$ using (7)
  Selection of Ω matrix with significant eigenvalues
  Compute final set of vectors z using (8)
Switch (Case):
  if Case = 1 : **then**
    the eigen matrix , $[\Psi = \Phi_i(\Lambda_i)^{-\frac{1}{2}}]$ (9).
    return;
  if Case = 2 : **then**
    the eigen matrix , $[\Psi = \Phi_i(\Lambda_{smallest})^{-\frac{1}{2}}]$ (10).
    return;
  if Case = 3 : **then**
    the eigen matrix ,
    $[\Psi = \Phi_i(\Lambda_i)^{-\frac{1}{2}} \quad \Phi_i(\Lambda_{smallest})^{-\frac{1}{2}}]$ (11).
    return;
Classification:
  Linear SVM classifier

3 Experiments and Results

3.1 Datasets

Our proposed method is experimented and validated on the publicly challenging datasets of skin lesion analysis towards melanoma detection released by ISBI (International Symposium on Biomedical Imaging), PH2 and MED-NODE. The datasets released by ISBI are based on the International Skin Imaging Collaboration (ISIC) archive which is an international effort to improve melanoma diagnosis, sponsored by the International Society for Digital Imaging of the Skin. The ISIC Archive [38] contains the largest publicly available collection of quality controlled dermoscopic images of skin lesions. They are briefly described below:

• ISBI 2016 [4]: This challenge employs a subset from the ISIC archive, containing 900 dermoscopic lesion images in JPEG format. The testing dataset contains 379 images in the same format as the training data.
• PH2 dataset [39]: This is a dermoscopic image database acquired at the Dermatology Service of Hospital Pedro Hispano, Portugal, under the same conditions through Tuebinger Mole Analyzer system using a magnification of 20x. They are 8-bit RGB color images with a resolution of 768 $\times$ 560 pixels. It contains a total number of 200 melanocytic lesions, including 80 common nevi, 80 atypical nevi, and 40 melanomas. Following the validation method in [12], we perform $5-$folds of cross validation for our experiments on this database.
• MED-NODE dataset [40]: This dataset consists of 70 melanoma and 100 naevus images from the digital image archive of the department of dermatology, university medical center Groningen used for the development and testing of the MED-NODE system for skin cancer detection from macroscopic images. Following [12], we perform $5-$folds of cross validation for our experiments on this database.
• ISBI 2017 [41]: 2017 challenge consists of more number of images and also included Seborrheic keratosis which is a benign skin tumor, derived from keratinocytes (non-melanocytic) along with benign nevus (melanocytic) and melanoma (melanocytic). The training data consists of 2000 images of which 374 are melanoma (malignant skin tumour), 254 seborrheic keratosis and 1626 benign nevus. The testing data consists of total 600 images of which 117 are melanoma images. This is the largest among all current state-of-the-art datasets.

3.2 Evaluation Metrics

For comparison of the proposed methodology with the existing methods, the following criteria are used. They are defined as

• Accuracy: The number of correct predictions divided by the total number of predictions. It is also defined as ratio of true detected cases to all cases, given by:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{14}$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = Flase Negative.
• Sensitivity: It is the ability of the diagnosis to correctly identify the diseased cases (*i.e.* malignant), given by:

$$SE = \frac{TP}{TP + FN} \tag{15}$$

• Specificity: It is the ability of the diagnosis to correctly identify the non-diseased cases (*i.e.* benign), given by:

$$SP = \frac{TN}{TN + FP} \tag{16}$$

• AUC: Area under receiver operating characteristic. It is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. It is the graph between true positive rate vs. false positive rate.
• Average Precision: Average precision (AP) is the integral under the precision-recall curve within the most confident image and the maximum rank that contains all positively labeled instances. The detailed explanation can be found in [4].
• Positive Predictive Value (PPV): It is the probability whether the subject with a positive test, truly have the disease, given by:

$$PPV = \frac{TP}{TP + FP} \tag{17}$$

• Negative Predictive Value (NPV): It is the probability whether the subjects with a negative screening test, truly don't have the disease, given by:

$$NPV = \frac{TN}{TN + FN}. \tag{18}$$

3.3 Experiments on ISBI 2016

We compare our proposed approaches with the state-of-the-art methodologies on ISBI 2016 dataset, the results are shown in Table 1. The accuracy of the deep features with SVM as the classifier (ResNet features+SVM) is much less as compared to our proposed method of regularized discriminant analysis for all the three cases. Features obtained from projection matrices from the null space (case 2) and whole space (case 3) give better results as compared to selecting only the principal eigenvalues (case 1). The results of with fine-tuning and without fine-tuning the network are also shown in Table 1, as evident, there is an obvious increase in performance when we further retrain the higher level layers as compared to simply extracting the features from the pre-trained model.

Lequan Yu *et al.* [24] and Codella *et al.* [25] have performed prudent segmentation followed by classification. However, our method outperforms them even without performing segmentation. Fisher vector based methods [42] and [23] use 32,768 (even after dimensionality reduction using principal component analysis) and 12,800 feature dimensions, respectively, for final feature matching, which are very high as compared to ours that uses only 2048 feature dimensions of deep residual network and 899 (number of samples -1) for

© The Institution of Engineering and Technology 2015

Table 1 Comparison of the proposed method with the existing state-of-the-art on ISBI 2016 dataset.

Methods	Accuracy	AUC	AP	SE	SP
LDF-FV (fusion) [23]	**0.868**	**0.852**	**0.684**	0.426	**0.977**
CNN-FV (fusion) [42]	0.831	0.796	0.535	-	-
FCRN+deep ResNet [24]	0.855	0.804	0.637	0.507	0.941
Ensemble model [25]	0.805	0.838	0.645	**0.693**	0.832
Deep Bayesian Active Learning [43]	-	0.750	-	-	-
ResNet features+SVM	0.738	0.620	0.313	0.347	0.835
Proposed method (without fine-tuning)					
Case 1	0.768	0.664	0.509	0.493	0.835
Case 2	0.775	0.679	0.529	0.520	0.842
Case 3	0.802	0.721	0.584	0.586	0.855
Proposed method (with fine-tuning)					
Case 1	0.841	0.803	0.650	0.480	0.930
Case 2	0.854	0.827	0.676	0.573	0.924
Case 3	0.861	0.835	**0.684**	0.560	0.924

final classification purpose. Ensemble method [25] combines many different features like sparse coding, low-level color and texture features resulting in large number of features, higher complexity and computing time. Our proposed approach of case 3 is comparable to LDF-FV [23] with similar accuracy. There exists a tradeoff between sensitivity and specificity. Higher sensitivity means that the test is able to correctly identify those with the disease. Higher specificity means that the test correctly identifies patients without the disease. It is generally suggested that patients who are initially positive to a test with high sensitivity/low specificity test, must undergo to a second test with low sensitivity/high specificity. Thus to compare the performance measures, AUC and average precision are generally taken into account and in these we have achieved the third best AUC 0.835 as compared to 0.852 [23] and 0.838 [25], while maintaining the same average precision of 0.684. Our method is simple, efficient, requires less computing time and complexity, yet achieves state-of-the-art performances without any segmentation procedure.

The receiver operating characteristics curves (ROCs) of the three different cases on ISBI 2016 dataset are shown in Fig. 4(a). ROC is a function of true positive rate (sensitivity) against the false positive rate (100-specificity) for different cut-off points. A test with perfect discrimination has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test. The best result is with Case 3 having ROC curve area as 0.83.

3.4 Experiments on PH2

The performance of the proposed method using PH2 dataset is shown in Table 2. Barata *et al.* [9] uses the color and texture features for classification using SVM and Adaboost classifier. To deal with class imbalance they repeated the melanoma features belonging to each training set until the number of examples on both the classes is same. To prevent having the same number of examples they added Gaussian noise to each repeated feature vector. In our method data augmentation is applied to deal with class imbalance while training the model, thus our method is more deterministic. The performance is superior with sensitivity 90.9% and specificity 100%. The other performance measures are unavailable, but due to importance of these measures we have computed and have achieved 97.5% accuracy with AUC 99.7% and average precision 99.4%. The color constancy method [12] improves the classification of multi-source images like EDRA database [44]. They implemented their method of color constancy using Shades of Gray [45] and found that this method does not have any improvement using PH2 data since the images are from same source. In [14], [13] and [15], the features used for classification is conventional hand-crafted features. Our approach of using deep features with regularized discriminant analysis gives better result as compared to theirs using hand-crafted

Table 2 Performance comparison of the proposed method with the existing state-of-art on PH2 dataset.

Methods	Accuracy	AUC	AP	SE	SP
Global features [9]	-	-	-	0.960	0.800
Local features [9]	-	-	-	**1.000**	0.750
Color constancy [12]	0.843	-	-	0.925	0.763
Fusion of Structural and textural features[14]	0.860	-	-	0.789	0.932
MLR_JRC [13]	0.920	0.937	-	0.875	0.931
ResNet features+SVM	0.875	0.963	0.924	0.750	0.916
Proposed Method (with fine-tuning)					
Case 1	**0.985**	**0.999**	**0.998**	0.930	**1.000**
Case 2	0.975	0.997	0.994	0.909	**1.000**
Case 3	0.980	0.996	0.985	0.909	**0.993**

features. In addition all the three cases of our proposed approach outperforms the baseline methodology of ResNet features+SVM. The Fig. 4(b) shows the ROC curves of our proposed methods for the PH2 dataset. Similar to the previous experiments, cases 2 and 3 have higher performance as compared to case 1.

3.5 Experiments on MED-NODE

Table 3 compares our method with the MED-NODE color and texture descriptors as reported in [40] and ResNet features+SVM. For cases 2 and 3, which keeps the null space information outperform ResNet features+SVM. The dataset has 170 clinical images from which 25 / 75 % ratio split is done for training and testing respectively, which is the same protocol as used in the original paper [40]. We can see that our performance of 77.1%, is better than existing methods for accuracy. Due to class imbalance accuracy cannot be considered as the sole estimator of performance. We have also computed AUC which is 90.3%. A direct comparison is not possible because of their unavailability of the evaluation metrics. Fig. 4(c) shows the ROC curve for the MED-NODE dataset using our methodology on all the three cases. It can be clearly seen that both case 2 and case 3, that keeps the null space information, has much higher performance as compared to case 1.

Table 3 Performance comparison of the proposed method with the existing state-of-art on MED-NODE dataset.

Methods	Accuracy	AUC	AP	SE	SP	PPV	NPV
MED-NODE - texture descriptor [40]	0.760	-	-	0.620	**0.850**	0.738	0.771
MED-NODE - color descriptor [40]	0.730	-	-	0.740	0.720	0.638	0.806
ResNet features+SVM	0.688	0.744	0.611	0.620	0.733	0.607	0.743
Proposed Method (with fine-tuning)							
Case 1	0.658	0.883	0.871	0.617	0.659	**0.745**	**0.872**
Case 2	**0.771**	**0.903**	**0.882**	0.807	0.737	**0.745**	**0.872**
Case 3	0.717	0.881	0.876	**0.848**	0.668	0.680	0.868

3.6 Experiments on ISBI 2017

We further experimented our method on the recently available ISBI 2017 dataset which has 2000 dermoscopic image as training and 600 as testing, currently largest in the research community. Table 4 shows the different performance measures with our best accuracy as 83.2%. Menegola *et al.* [46] achieved the 1^{st} place in ISBI 2017 challenge with accuracy of 87.2 %. Since deep models crave for data, their training data not only comprised of 2000 dermoscopic images, but also they increased their data by including all the publicly available datasets in skin lesion, *e.g.* ISBI 2017, ISIC Archive, Interactive Atlas of Dermoscopy, Dermofit, IRMA and PH2, we anticipate that their final feature dimensions could be very high. They experimented

© The Institution of Engineering and Technology 2015

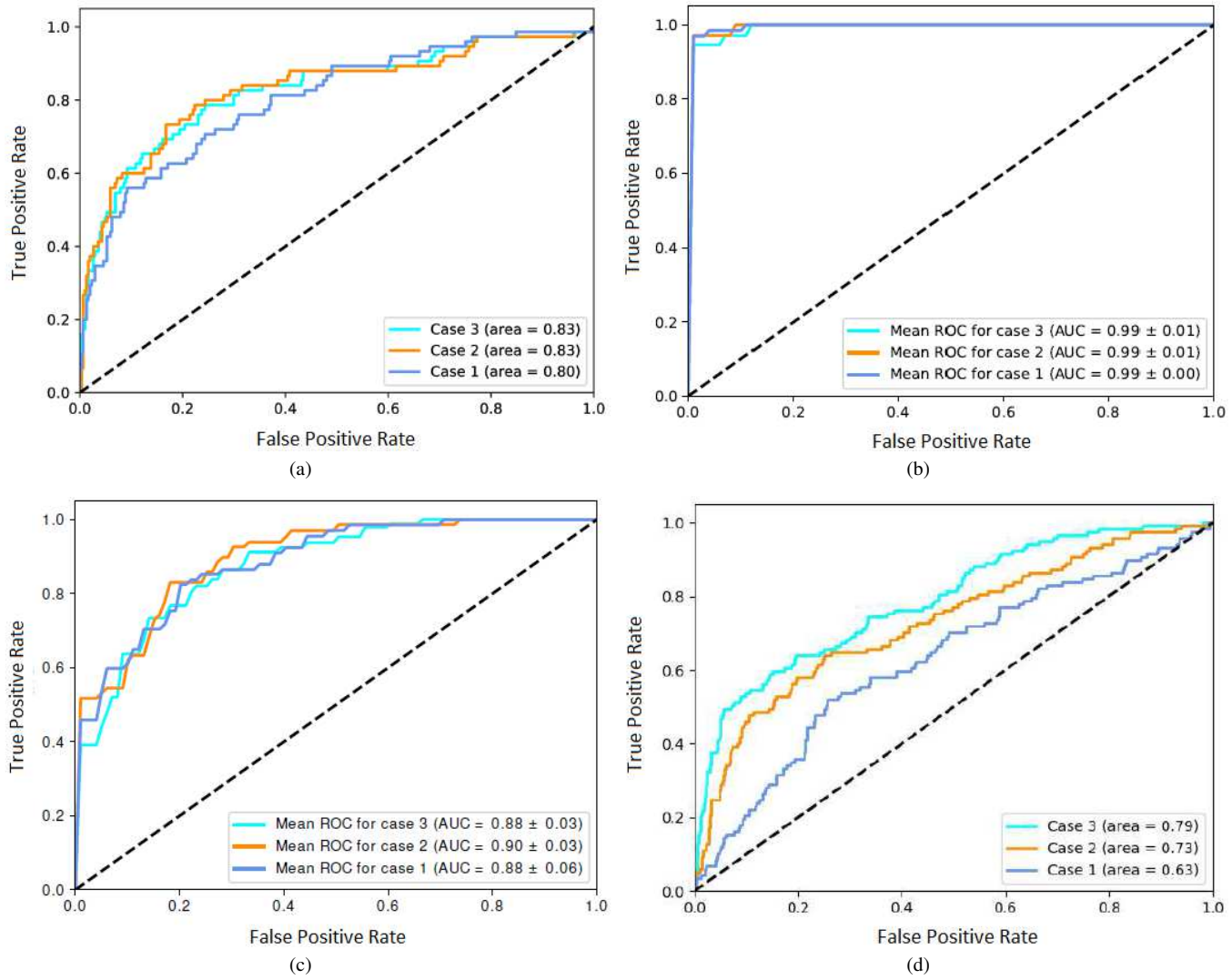

Fig. 4: Receiver operating characteristics for (a) ISBI 2016 dataset, (b) PH2 dataset, (c) MED-NODE dataset and (d) ISBI 2017 dataset. Area represents the AUC for each of the methods (best viewed in color).

Table 4 Results on ISBI 2017 dataset

Methods	Accuracy	AUC	AP	SE	SP
RECOD-TITANS [46]	**0.872**	**0.874**	**0.715**	0.547	**0.950**
ResNet features+SVM	0.783	0.689	0.379	0.350	0.888
Proposed Method (with fine-tuning)					
Case 1	0.630	0.631	0.297	**0.581**	0.642
Case 2	0.808	0.731	0.441	0.487	0.886
Case 3	0.832	0.789	0.549	0.529	0.905

using two pretrained CNN models ResNet-101 and Inception-v4. Experimentation using large number of data requires huge computational horsepower such as large memory CUDA-compatible GPUs. The training time and complexity are huge as compared to our approach, which uses only 2048 features and still achieve competitive accuracy performances. Fig. 4(d) shows the ROC curve for the ISBI 2017 dataset.

Similar to the previous experiments, on this large dataset, case 3 with AUC 0.79, which keeps both principal and null space information performs better than both case 2 with AUC 0.73 (keeping only the null space information) and case 1 with AUC 0.63 (that keeps only principal space information). The difference is much more significant and can be seen clearly for all the three cases on this large, most challenging state-of-the-art dataset in Fig. 4(d). From all the experimental results and ROC curves on four databases, it can be concluded that the null space eigenvectors (corresponding to zero eigenvalues) is important in differentiating melanoma from the non-melanoma images. Selecting and regularizing the features from all the three subspaces (case 3) of the within-class scatter matrix has given best performance results.

4 Conclusions and Future Work

In this paper, we have proposed a deep convolutional neural network based regularized discriminant learning framework which extracts low dimensional discriminative features for melanoma detection. Our work presents an automated non-invasive methodology to assist the clinicians and dermatologists for detection of melanoma. Unlike traditional methods, it does not require domain expertise for segmentation and hand crafted feature computation and/or selection. In our approach, we propose to fine-tuning the deep convolutional neural network based residual network and extract rich representations of the melanoma and non-melanoma images. On these features regularized discriminant analysis is performed which minimizes the whole of within-class variances and maximizes the total class variance information. Three cases involving the subspaces arising in the within-class scatter information followed by dimensionality reduction using total class variance information are analyzed and their

© The Institution of Engineering and Technology 2015

importance are shown for melanoma detection. The main advantages of our approach are that, (a) it extracts low dimensional discriminative features, (b) it eliminates the segmentation procedure and (c) retraining the whole deep residual model from scratch is not required. Experiment results on 4 benchmark publicly available challenging skin lesion datasets show the superiority of our proposed approach as compared to other state-of-the-art methodologies. Future work will include experimentation over different architectures and taking different number of layers into account which might further improve the accuracy and also train our regularized discriminant analysis in an end-to-end fashion which would increase the discriminative power of the neural network.

5 References

1 Siegel, R.L., Miller, K.D., Jemal, A.: 'Cancer statistics, 2018,' *CA: a cancer journal for clinicians*, 2018, **68(1)**, pp. 7–30.

2 Satheesha, T., Satyanarayana, D., et al.: 'Melanoma is Skin Deep: A 3D reconstruction technique for computerized dermoscopic skin lesion classification,' *IEEE Journal of Translational Engineering in Health and Medicine*, 2017, **5**, pp. 1–17.

3 Garnavi, R., Aldeen, M., Bailey, J.: 'Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis,' *IEEE Transactions on Information Technology in Biomedicine*, 2012, **16(6)**, pp. 1239–1252.

4 Gutman, D., Codella, N.C., et al.: 'Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC),' *arXiv preprint arXiv:1605.01397*, 2016.

5 Stolz, W.: 'ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma,' *Eur. J. Dermatol.*, 1994, **4**, pp. 521–527.

6 Menzies, S.W., Ingvar, C., et al.: 'Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features,' *Archives of Dermatology*, 1996, **132(10)**, pp. 1178–1182.

7 Henning, J.S., Dusza, S.W., et al.: 'The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy,' *Journal of the American Academy of Dermatology*, 2007, **56(1)**, pp. 45–52.

8 Esteva, A., Kuprel, B., et al.: 'Dermatologist-level classification of skin cancer with deep neural networks,' *Nature*, 2017, **542(7639)**, pp. 115–118.

9 Barata, C., Ruela, M., et al.: 'Two systems for the detection of melanomas in dermoscopy images using texture and color features,' *IEEE Systems Journal*, 2014, **8(3)**, pp. 965–979.

10 Marques, J.S., Barata, C., Mendonça, T.: 'On the role of texture and color in the classification of dermoscopy images,' in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012 (pp. 4402–4405).

11 Maglogiannis, I., Doukas, C.N.: 'Overview of advanced computer vision systems for skin lesions characterization,' *IEEE transactions on information technology in biomedicine*, 2009, **13(5)**, pp. 721–733.

12 Barata, C., Celebi, M.E., Marques, J.S.: 'Improving dermoscopy image classification using color constancy,' *IEEE journal of biomedical and health informatics*, 2015, **19(3)**, pp. 1146–1152.

13 Bi, L., Kim, J., et al.: 'Automatic melanoma detection via multi-scale lesion-biased representation and joint reverse classification,' in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, IEEE, 2016 (pp. 1055–1058).

14 Adjed, F., Gardezi, S.J.S., et al.: 'Fusion of structural and textural features for melanoma recognition,' *IET Computer Vision*, 2017.

15 Abuzaghleh, O., Barkana, B.D., Faezipour, M.: 'Automated skin lesion analysis based on color and shape geometry feature set for melanoma early detection and prevention,' in *Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island*, IEEE, 2014 (pp. 1–6).

16 Lee, T., Ng, V., et al.: 'DullrazorⓇ: A software approach to hair removal from images,' *Computers in biology and medicine*, 1997, **27(6)**, pp. 533–543.

17 Celebi, M.E., Kingravi, e.: 'A methodological approach to the classification of dermoscopy images,' *Computerized Medical Imaging and Graphics*, 2007, **31(6)**, pp. 362–373.

18 Barata, C., Marques, J.S., Rozeira, J.: 'A system for the detection of pigment network in dermoscopy images using directional filters,' *IEEE transactions on biomedical engineering*, 2012, **59(10)**, pp. 2744–2754.

19 Li, X., Aldridge, B., et al.: 'Depth data improves skin lesion segmentation,' in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2009 (pp. 1100–1107).

20 Korotkov, K., Garcia, R.: 'Computerized analysis of pigmented skin lesions: A review,' *Artificial intelligence in medicine*, 2012, **56(2)**, pp. 69–90.

21 Shen, D., Wu, G., Suk, H.I.: 'Deep learning in medical image analysis,' *Annual review of biomedical engineering*, 2017, **19**, pp. 221–248.

22 Lopez, A.R., Giro-i Nieto, X., et al.: 'Skin lesion classification from dermoscopic images using deep learning techniques,' in *Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on*, IEEE, 2017 (pp. 49–54).

23 Yu, Z., Jiang, X., et al.: 'Aggregating Deep Convolutional Features for Melanoma Recognition in Dermoscopy Images,' in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2017 (pp. 238–246).

24 Yu, L., Chen, H., et al.: 'Automated melanoma recognition in dermoscopy images via very deep residual networks,' *IEEE transactions on medical imaging*, 2017, **36(4)**, pp. 994–1004.

25 Codella, N.C., Nguyen, Q.B., et al.: 'Deep learning ensembles for melanoma recognition in dermoscopy images,' *IBM Journal of Research and Development*, 2017, **61(4)**, pp. 5–1.

26 He, K., Zhang, X., Ren, S.e.: 'Deep residual learning for image recognition,' in *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016 (pp. 770–778).

27 Ebrahimi, M.S., Abadi, H.K.: 'Study of Residual Networks for Image Recognition, 2017,' .

28 Fukunaga, K.: *Introduction to statistical pattern recognition*, Academic Press, second edition, USA, 1991.

29 Martinez, A.M., Kak, A.C.: 'PCA versus LDA,' *IEEE PAMI*, 2001, **23(2)**, pp. 228–233.

30 He, X., Yan, S., et al.: 'Face recognition using Laplacianfaces,' *IEEE PAMI*, 2005, **27(3)**, pp. 328–340.

31 Jiang, X.D., Mandal, B., Kot, A.: 'Eigenfeature Regularization and Extraction in Face Recognition,' *IEEE PAMI*, 2008, **30(3)**, pp. 383–394.

32 Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: 'Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,' *IEEE PAMI*, 1997, **19(7)**, pp. 711–720.

33 Liang, Z., Li, Y.: 'Multiple kernels for generalised discriminant analysis,' *IET Computer Vision*, 2010, **4(2)**, pp. 117–128.

34 Liu, W., Wang, Y., et al.: 'Null space approach of Fisher discriminant analysis for face recognition,' in *ECCV*, 2004 (pp. 32–44).

35 Zhang, L., Xiang, T., Gong, S.: 'Learning a Discriminative Null Space for Person Re-identification,' in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016 (pp. 1239–1248).

36 Mandal, B., Jiang, X.D., Kot, A.: 'Dimensionality Reduction in Subspace Face Recognition,' in *IEEE ICICS*, 2007 (pp. 1–5).

37 Cortes, C., Vapnik, V.: 'Support-Vector Networks,' *Machine Learning*, 1995, **20(3)**, pp. 273–297.

38 'International Skin Imaging Collaboration Website.' URL `http://www.isdis.net/index.php/isic-project`.

39 Mendonça, T., Ferreira, P.M., et al.: 'PH 2-A dermoscopic image database for research and benchmarking,' in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, IEEE, 2013 (pp. 5437–5440).

40 Giotis, I., Molders, N., et al.: 'MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images,' *Expert systems with applications*, 2015, **42(19)**, pp. 6578–6585.

41 Codella, N.C., Gutman, D., et al.: 'Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),' *arXiv preprint arXiv:1710.05006*, 2017.

42 Yu, Z., Ni, D., et al.: 'Hybrid dermoscopy image classification framework based on deep convolutional neural network and Fisher vector,' in *14th IEEE International Symposium on Biomedical Imaging, ISBI 2017, Melbourne, Australia, April 18-21, 2017*, 2017 (pp. 301–304).

43 Gal, Y., Islam, R., Ghahramani, Z.: 'Deep Bayesian Active Learning with Image Data,' *arXiv preprint arXiv:1703.02910*, 2017.

44 Argenziano, G., Soyer, H., et al.: 'Interactive atlas of dermoscopy (Book and CD-ROM),' 2000.

45 Finlayson, G.D., Trezzi, E.: 'Shades of gray and colour constancy,' in *Color and Imaging Conference*, volume 2004, Society for Imaging Science and Technology, 2004 (pp. 37–41).

46 Menegola, A., Tavares, J., et al.: 'RECOD Titans at ISIC Challenge 2017,' *arXiv preprint arXiv:1703.04819*, 2017.

© The Institution of Engineering and Technology 2015