Computerized Medical Imaging and Graphics 71 (2019) 19–29

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

# Fusing fine-tuned deep features for skin lesion classification

Amirreza Mahbod[a,b,*], Gerald Schaefer[c], Isabella Ellinger[a], Rupert Ecker[b], Alain Pitiot[d], Chunliang Wang[e]

[a] Institute of Pathophysiology and Allergy Research, Medical University of Vienna, Vienna, Austria
[b] Research and Development Department of TissueGnostics GmbH, Vienna, Austria
[c] Department of Computer Science, Loughborough University, Loughborough, United Kingdom
[d] Laboratory of Image and Data Analysis, Ilixa Limited, Nottingham, United Kingdom
[e] School of Technology and Health, KTH Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

*Article history:*
Received 31 December 2017
Received in revised form
30 September 2018
Accepted 30 October 2018

*Keywords:*
Skin cancer
Melanoma
Dermoscopy
Medical image analysis
Deep learning

ABSTRACT

Malignant melanoma is one of the most aggressive forms of skin cancer. Early detection is important as it significantly improves survival rates. Consequently, accurate discrimination of malignant skin lesions from benign lesions such as seborrheic keratoses or benign nevi is crucial, while accurate computerised classification of skin lesion images is of great interest to support diagnosis. In this paper, we propose a fully automatic computerised method to classify skin lesions from dermoscopic images. Our approach is based on a novel ensemble scheme for convolutional neural networks (CNNs) that combines intra-architecture and inter-architecture network fusion. The proposed method consists of multiple sets of CNNs of different architecture that represent different feature abstraction levels. Each set of CNNs consists of a number of pre-trained networks that have identical architecture but are fine-tuned on dermoscopic skin lesion images with different settings. The deep features of each network were used to train different support vector machine classifiers. Finally, the average prediction probability classification vectors from different sets are fused to provide the final prediction. Evaluated on the 600 test images of the ISIC 2017 skin lesion classification challenge, the proposed algorithm yields an area under receiver operating characteristic curve of 87.3% for melanoma classification and an area under receiver operating characteristic curve of 95.5% for seborrheic keratosis classification, outperforming the top-ranked methods of the challenge while being simpler compared to them. The obtained results convincingly demonstrate our proposed approach to represent a reliable and robust method for feature extraction, model fusion and classification of dermoscopic skin lesion images.



## 1. Introduction

Malignant melanoma (MM) is a very aggressive form of skin cancer. Although occurrences of non-melanoma skin cancer are far more common (MM represents less than 5% of all skin cancers), 70% of skin cancer deaths are due to MM. 132,000 melanoma skin cancers occur globally each year (WHO, 2018), and both incidence and mortality rates have increased throughout most of the developed world over the past 30 years (Apalla et al., 2017). Prevention as well as early detection are crucial to reverse this trend (Shellenberger et al., 2016). If identified early enough, skin cancer can be cured through a simple excision, while diagnosis at later stages is associated with a greater risk of death – the estimated 5-year survival rate is over 95% for early stage diagnosis, but below 20% for late stage detection (Esteva et al., 2017; Balch et al., 2009).

Seborrheic keratosis (SK) is one of the most common benign skin lesions. SKs can exhibit wide variations in its clinical features, and some types of SK resemble melanomas or other skin tumors. Moreover, melanomas may appear adjacent to or within SKs. Likewise, benign nevi (BN), which are pigmented skin growths with no current signs of pathology, can appear similar to melanomas, while patients with numerous nevi have a significantly higher risk of developing skin cancer (Thomas and Puig, 2017).

Pathological analysis of a biopsy specimen enables differentiation between different types of skin lesions with certainty, but this type of analysis is both time and labour intensive and not always possible. Dermoscopy, in contrast, is a non-invasive, microscopy-based diagnostic method, which allows for enhanced visualisation of the internal structures of lesions (Argenziano et al., 2002). When

* Corresponding author at: Amirreza Mahbod at Institute of Pathophysiology and Allergy Research, Medical University of Vienna, Vienna, Austria.
*E-mail address:* amirreza.mahbod@tissuegnostics.com (A. Mahbod).

https://doi.org/10.1016/j.compmedimag.2018.10.007
0895-6111/

performed by well-trained and experienced dermatologists, dermoscopy supports a diagnostic accuracy of about 80% (Carli et al., 2003; Vestergaard et al., 2008) and leads to a reduced number of unnecessary excisions (Thomas and Puig, 2017). However, visual inspection of dermoscopic images by dermatologists requires training and experience, since the diagnostic accuracy achieved by non-experts using dermoscopy is no better than with the unaided eye (Kittler et al., 2002).

Despite the definition of commonly employed diagnostic schemes such as the ABCD rule (Stolz et al., 1994) or the 7-point checklist (Argenziano et al., 1998), due to the difficulty and subjectivity of human interpretation as well as the variety of lesions and confounding factors encountered in practice, computerised analysis of dermoscopic images has become an important research area to support diagnosis (Fleming et al., 1998; Carrera et al., 2016). Conventional computer-aided methods for dermoscopic lesion classification typically involve three main stages: segmenting the lesion, extracting hand-crafted image features from the lesion area and its border, and classification (Celebi et al., 2007; Oliveira et al., 2016). In addition, often extensive pre-processing is involved to improve image contrast (Oliveira et al., 2016; Jaisakthi et al., 2017), perform white balancing based on colour constancy algorithms (Barata et al., 2015), apply colour normalisation (Schaefer et al., 2011) or calibration (Iyatomi et al., 2011), colour space transformation (Oliveira et al., 2016), illumination correction (Oliveira et al., 2016), or remove image artefacts such as hairs (Fleming et al., 1998; Abbas et al., 2011) or bubbles (Fleming et al., 1998).

Accurate segmentation of the lesion area is considered important, since the shape of the lesion gives crucial clues for diagnosis, while the subsequent processing steps rely on a precise division between lesion and skin areas. A variety of segmentation algorithms have been developed for border detection (Celebi et al., 2009; Oliveira et al., 2016) including thresholding-based methods (Celebi et al., 2013), region merging approaches (Celebi et al., 2008), clustering techniques (Zhou et al., 2009), active contours (Zhou et al., 2011) and machine learning techniques such as artificial neural networks (Jaisakthi et al., 2017). Based on the segmented lesion area, domain specific features are then extracted. These features can relate to lesion type (primarily morphological features), lesion configuration (secondary morphological features), colour, shape, texture and lesion border (Celebi et al., 2007; Lopez et al., 2017). In order to select the most relevant features and to reduce the dimensionality of the feature space, a number of feature selection methods can be utilised, which in turn can lead to improved classification performance and lower training and testing time (Oliveira et al., 2018).

In supervised approaches, where the ground truth of a subset of data is available, the selected features together with the corresponding labels are used to train a classifier (such as support vector machines (SVMs), random forest classifiers or multi-layer perceptrons (MLP) (Oliveira et al., 2018; Esteva et al., 2017; Codella et al., 2015)). The trained model can then be employed for classifying new skin lesion images. An overview of classifiers that have been used for skin lesion classification can be found in Oliveira et al. (2018) and shows that a SVM is a common choice due to its relatively good generalisation properties (Oliveira et al., 2018), the possibility to incorporate kernel functions to simplify and enhance the classification of non-linear feature distributions in high-dimensional spaces, and competitive classification performance compared to the more complex classifiers (Oliveira et al., 2018; Gessert et al., 2018).

The main drawback of conventional approaches is a lack of generalisation capability due to high variations in dermoscopic images, different artefacts and insufficient training data. Variations in dermoscopic images are due to different zooming configurations, lighting conditions, instruments or operators, while common artifacts in dermoscopic images include not just skin hair and bubbles but also, among others, dark corners/borders, light reflections or shadows, skin lines, ruler or calibration chart artefacts or ink markings, which can lead to failures of segmentation algorithms, changes in extracted image features and consequently a negative effect on classification accuracy (Mahbod et al., 2017; Oliveira et al., 2016).

Deep neural networks (DNNs), in particular convolutional neural networks (CNNs), are superior to other methods for tasks such as object detection and natural image classification (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). To achieve high accuracy, well-established CNN architectures such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016) are typically trained on large image databases such as ImageNet (Deng et al., 2009) which comprise millions of heterogenous images. However in medicine, access to validated data is heavily restricted and expensive to obtain, which makes training such networks from scratch problematic (Vasconcelos and Vasconcelos, 2017). One way to address this problem is to use transfer learning, which employs a pre-trained network (i.e., one trained on other tasks such as generic image classification) and adapt it to the problem at hand. This pre-training allows the network to identify useful features even when training samples are limited (Lopez et al., 2017).

In medical image analysis, transfer learning has been used for a variety of applications including radiology, cardiology, ultrasound imaging, gastroenterology, retinopathy, microscopic imaging as well as dermoscopy (Tajbakhsh et al., 2016; Lopez et al., 2017). So far, mainly two different approaches of transfer learning were used for medical image analysis and in particular for skin lesion classification (Tajbakhsh et al., 2016). On the one hand, pre-trained CNNs were used as feature generators. In this setting, images are fed to pre-trained models and deep features extracted from a certain fully connected (FC) layer or convolution layer. The generated extracted features are then used to train a classical classifier such as an SVM (Kawahara et al., 2016; Lopez et al., 2017). In some extended studies, these features were encoded to more invariant and discriminative representations (Yu et al., 2017) or combined with other hand-crafted feature descriptors (Codella et al., 2016, 2015) to enhance classification performance. On the other hand, trained models can be adapted to the problem at hand by fine-tuning. To fine-tune deep models, FC layers of the pre-trained networks were typically replaced by one or more new logistic layers and then the networks re-trained to adapt the weights of the newly added layers for classifying skin lesions (DeVries, 2017). The pre-trained models used in both approaches for skin lesion classification varied in different studies and include AlexNet (Kawahara et al., 2016; Mahbod et al., 2017; Codella et al., 2016), VGG16 (Mahbod et al., 2017; Lopez et al., 2017; Harangi, 2015), VGG19 (Mahbod et al., 2017), GoogleNet (Yang et al., 2017; Nader and Nader, 2018; Harangi, 2015), ResNet-50 (Harangi, 2015; Matsunaga et al., 2017; Gonzalez-Díaz, 2017), ResNet-101 (Menegola et al., 2017), ResNet-152 (Li and Li, 2018; Li and Shen, 2018), Inception-v3 (DeVries, 2017; Mirunalini et al., 2017), Inception-v4 (Menegola et al., 2017; Li and Li, 2018), variations of DenseNets (Gessert et al., 2018; Li and Li, 2018), SeNets (Gessert et al., 2018; Li and Shen, 2018) and PolyNets (Gessert et al., 2018). Moreover, ensembles of fine-tuned deep networks (Menegola et al., 2017; Matsunaga et al., 2017) and fusing outputs of classical and deep models (Codella et al., 2016; Yu et al., 2017) were utilised to boost classification performance.

In this paper, in contrast to former studies, we utilise both schemes of transfer learning in one single approach. We exploit several well-known CNNs pre-trained on ImageNet and fine-tune them on a limited dataset of dermoscopic lesion images. We ensemble deep features, that is the outputs of the last few fully-connected layers, in an SVM classifier that then gives the classification of the lesion type. Unlike previous works using deep features for skin lesion classification (Lopez et al., 2017; Codella et al., 2015;

Kawahara et al., 2016), which were limited to exploit specific network architectures or using specific layers for extracting features, in our approach, we hypothesise that extracting features from different layers of different abstraction levels and from different deep models can improve the classification results. More importantly, we fine-tune the pre-trained networks multiple times with different settings to achieve more stable classification performance for skin lesion categorisation. Compared to conventional methods where each network architecture is used once, the diverse fine-tuning mechanism boosted the performance of a single architecture and the final fused results. The proposed method avoided using extensive pre-processing steps, lesion segmentation masks or engineered hand-crafted feature descriptors, which can potentially increase the generalisation ability and at the same time its adaptability to be extended for other classification tasks. Finally, we perform a thorough investigation of the performance of each component of our proposed method to justify our approach and to provide a useful guideline for further developments of CNN-based algorithms for skin lesion classification.

## 2. Materials and methods

Our proposed skin lesion classification method consists of the following major steps: image pre-processing, deep neural network fine-tuning and feature extraction to train a SVM classifier, and ensembling the model outputs. In the following, we describe the utilised datasets, and cover in detail each of the stages of our approach.

### 2.1. Dataset

We used the training, validation and test images of the ISIC 2016 challenge (Gutman et al., 2016) as well as the training and validation images of the ISIC 2017 challenge (Codella et al., 2017). These probably represent the most challenging skin lesions datasets that are publicly available to date for ternary skin lesion classification. From these two datasets, 2187 images were extracted for training which included 441 MMs, 296 SKs and 1450 BN images. We tested our trained model on the 600 images that comprise the test set of the ISIC 2017 challenge and which were not used in the training phase. All training and test images were 24-bit RGB images of various sizes (ranging from $1022 \times 767$ to $6748 \times 4499$ pixels), perspectives, and lighting conditions, while a significant number of images contained various artefacts.

### 2.2. Pre-processing

In our proposed pipeline, we aimed to keep the pre-processing steps to a minimum to support better generalisation ability when tested on other datasets. Three pre-processing steps were applied in our approach where only one was task specific (related to skin lesion classification) while the other two were standard pre-processing steps to prepare the images before feeding them to deep networks.

#### 2.2.1. Colour standardisation

As the images were acquired under different lighting conditions and with different devices, we performed colour normalisation using the gray world colour constancy algorithm, which has been reported to support improved skin lesion classification (Barata et al., 2015; Matsunaga et al., 2017).

#### 2.2.2. Normalisation

In order to utilise pre-trained deep networks, a common normalisation technique is to subtract the mean RGB value of the ImageNet dataset from all training and test images (Krizhevsky et al., 2012). Other approaches were also tested, including subtracting mean RGB values computed over each individual image and subtracting mean RGB values computed over the whole training dataset from all training and test images as suggested in Kawahara et al. (2016) and Yu et al. (2017).

#### 2.2.3. Resizing

Since all pre-trained networks used in our implementation expect the input images to be of the same size defined during training, we resized all images to the appropriate size ($227 \times 227$ and $224 \times 224$ pixels) using bicubic interpolation. For non-square images, the aspect ratio was changed during this resizing step.

### 2.3. Pre-trained deep learning models and fine-tuning

In order to extract optimised features from the images, we used well-established CNN architectures, namely AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014) and two variations of ResNet (He et al., 2016) which have shown excellent performance in previous classification tasks such as the Image Large Scale Visual Recognition Challenge (VGGNet was the runner-up of the challenge in 2014, while AlexNet and ResNet were the winners of the challenge in 2012 and 2015, respectively (Russakovsky et al., 2015)). While AlexNet has a well-established architecture with 5 convolutional layers and 3 FC layers, the original implementations of VGGNet and ResNet come with several variations. In our work, we used VGG16, which has 16 weight layers, 13 convolutional and 3 FC layers as well as ResNet-18 and ResNet-101 which exhibit different depths. In general, ResNet's architecture consists of special building blocks called residual blocks and one FC layer on top which performs the classification.

In order to extract features from these CNNs, one approach is to simply run the images through the pre-trained networks and take the output of the FC layers as was done in some previous works (Lopez et al., 2017; Kawahara et al., 2016). However, we hypothesise that fine-tuning of pre-trained networks using skin lesion images should lead to higher quality features from the images.

Fine-tuning of the selected networks was performed as follows. First, the last FC layer and the output layer of all pre-trained networks were removed and replaced by two new FC layers with 64 and 3 nodes to solve the ternary (MM/SK/BN) classification problem, as shown in Fig. 1 with ResNet as an example. The weights of the added fully connected layers were randomly generated from a Gaussian distribution with zero mean and standard deviation of 0.01. In order to prevent overfitting and to speed up the training, we froze the weights of the initial layers of the deep models. For AlexNet and VGG16, we froze the initial layers up to the 4-th and 10-th convolutional layers, while we froze the layers up to the 4-th and 30-th residual blocks for ResNet-18 and ResNet-101, respectively.

We tested different optimisers with regularisation terms for the loss function in order to perform fine-tuning. In particular, we utilised stochastic gradient descent with momentum (SGDM) (Bishop, 2006; Murphy, 2012), root mean square propagation (RMSProp) (Tieleman and Hinton, 2012) and adaptive moment estimation (Adam) (Kingma and Ba, 2014) optimisers in our experiments.

The SGDM optimiser updates the weights and biases of the network in each iteration in order to minimise the error (i.e., minimise the loss function output) by taking small steps in the negative direction of the gradient. We used the momentum term in order to prevent oscillations along the steepest descent path. The general SGDM term employed in our approach is

$$\theta_{i+1} = \theta_i - \alpha \nabla E_R(\theta_i) + \gamma(\theta_i - \theta_{i-1}), \quad (1)$$

where $\theta$ is the parameter vector of the network, $i$ represents the iteration number, $\alpha$ is the learning rate, $E_R$ indicates the loss

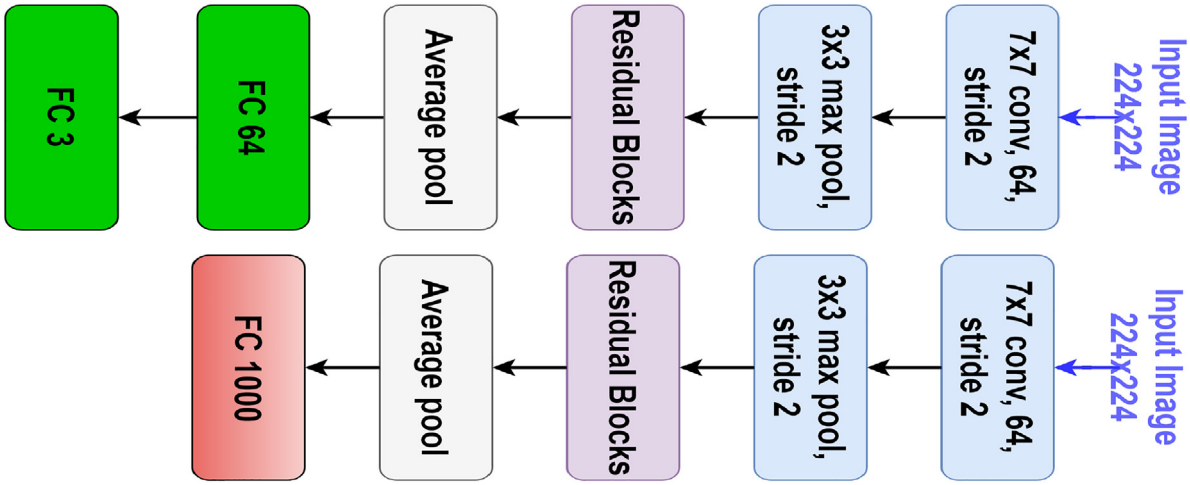


**Fig. 1.** Generic structure of the original ResNet (top) and the modified architecture adapted for fine-tuning in our proposed approach (bottom). The final FC layer of the original architecture (red block) is replaced by two FC layers (green blocks). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

function, and $\gamma$ is the momentum term which determines the contribution of the previous gradient step to the current iteration. We used the cross-entropy loss function in the optimisation process as

$$E(\theta) = -\sum_{l}\sum_{m=1}^{k} t_{lm} \ln(y_m(x_l, \theta)) \tag{2}$$

and

$$E_R(\theta) = E(\theta) + \lambda\Omega(w), \tag{3}$$

where $k$ is the number of classes, $t_{lm}$ indicates that the $l$-th sample belongs to the $m$-th class and $y_m(x_l, \theta)$ is the network output of the $l$-th sample. The added term in Eq. (3) is the regularisation term, where $w$ is the weight factor, $\lambda$ is the regularisation factor coefficient and $\Omega(w)$ is the regularisation function defined as

$$\Omega(w) = \frac{1}{2}w^T w. \tag{4}$$

RMSProp minimises the loss function based on

$$\theta_{i+1} = \theta_i - \frac{\alpha\nabla E(\theta_i)}{\sqrt{v_i} + \epsilon}, \tag{5}$$

where $v_i$ is

$$v_i = \beta_2 v_{i-1} + (1 - \beta_2)[\nabla E(\theta_i)]^2, \tag{6}$$

and $\beta_2$ is the decay rate which needs to be set as an hyperparameter ($\epsilon$ is a very small number and prevents division by zero).

While SGDM uses a single learning rate for updating the parameters, RMSProp tries to adapt the learning rate for different parameters based on the loss function being optimised. In the RMSProp optimisation approach, the learning rate of the parameters with large gradients will be reduced and the learning rate of the parameters with relatively small gradients will be increased.

The Adam optimiser, similar to RMSProp, adapts the learning rate for optimisation but with a momentum term as

$$\theta_{i+1} = \theta_i - \frac{\alpha m_i}{\sqrt{v_i} + \epsilon}, \tag{7}$$

where $m$ is

$$m_i = \beta_1 m_{i-1} + (1 - \beta_1)\nabla E(\theta_i), \tag{8}$$

and $v$ is as in Eq. (6). $\beta_1$ is the gradient decay factor, another hyperparameter. The added momentum term in Adam controls the parameter updates. If the gradients over many iterations are similar, the updates will be larger and if the gradient varies a lot (e.g. through noise) then the updates will be small.

In our experiments, we set the initial learning rate to 0.001 for the SGDM optimiser and to 0.0001 for RMSProp and Adam, but we kept the learning rate of the new FC layers 10 times bigger compared to all other learnable layers. Weight decay was set to 0.0001 and the momentum term for SGDM was set to 0.9. $\beta_1$ and $\beta_2$ in Eq. (6) and Eq. (8) were set to 0.9 and 0.999, respectively. For AlexNet, the batch size was set to 128, for VGG16 to 32, and for the ResNets to 16 in order to fit into the GPU memory. The learning rate was dropped by a factor of 10 after 6 epochs and we retrained all models for 12 epochs.

In order to prevent overfitting of the networks to our limited training dataset, we artificially increased the training size by data augmentation. For this, we used rotation (90, 180 and 270°) and horizontal flipping as main data augmentation techniques. Moreover, the images randomly underwent small changes in each iteration in the training process. These changes included random rotations (−5 to 5°), random scaling (0.9 to 1.1) and random shearing (−2 to 2°). From the derived modified training data, we randomly split the dataset to 90% for training and 10% for validation.

### 2.4. Ensembling deep features and fusion of networks

The deep features are the outputs of the FC layers from the pre-trained or fine-tuned DNNs. We tested two strategies to extract deep features from DNNs. The first was to use the output of only the first FC layer following the convolutional layers. The second was to concatenate the outputs of all FC layers. For the fine-tuned networks, we also included the outputs of the two added/replaced layers in the modified networks, i.e. the FC64 and FC3 outputs in Fig. 1.

The extracted deep features along with the corresponding labels identifying the lesion types were used to train a ternary SVM classifier. We tested both linear and radial basis function (RBF) kernels and observed slightly better performance with the RBF kernel, similar to others (Oliveira et al., 2016; Gessert et al., 2018). We therefore utilised one-versus-all multi-class SVM classifier with RBF kernels in our final models. The SVM scores were mapped to probabilities using logistic regression (Platt, 1999), and the classification results were the probabilistic prediction vectors derived from the trained SVMs for the three different classes, which can also be used to identify the predicted lesion type. Data augmentation, similar to that

**Table 1**
Effects of gray world colour constancy (using fine-tuned ResNet-18).

| | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| No standardisation | 80.23 ± 1.77 | 89.64 ± 0.99 | 84.93 ± 0.56 |
| Colour constancy | 83.48 ± 0.74 | 91.39 ± 1.33 | 87.44 ± 0.57 |

employed during the CNN fine-tuning step, was also performed. During the inference stage on testing data, 8 copies of a single test image (0, 90, 180 and 270° rotation, with and without horizontal flipping) were fed to the pipeline. The final classification for each individual test image was based on the average probabilities of the 8 results for each model.

Finally, we employed an extensive yet straight-forward ensembling approach to boost our classification performance and to improve the robustness of our approach. For each architecture, we took the average over different prediction vectors which were acquired from the same model architecture, but with different training settings. The varying settings in the fine-tuning step were the normalisation technique (ImageNet mean subtraction or training mean subtraction) and the optimizer (SGDM, RMSProp or Adam). Moreover, we trained each model 3 times and took the average over the results. Hence, the final results of a single architecture (e.g. ResNet-18) were acquired from 18 different models.

### 2.5. Evaluation

Evaluation of the proposed method was performed by calculating the area under the receiver operating characteristics curve (AUC) which was the main evaluation metric in the ISIC 2017 challenge (Codella et al., 2017).

Since the ISIC 2017 challenge evaluation was based on two binary classification tasks (MM vs. all and SK vs. all), we converted our three elemental prediction vectors to two elemental binary vectors by a one-versus-all approach. For these binary tasks, we also evaluated the results based on the accuracy at the threshold of 50%. Moreover, optimal sensitivity and specificity of our best performing approach were calculated using Youden index method (Youden, 1950).

## 3. Results

The obtained results are derived from the 600 test images of the ISIC 2017 challenge. These are comprised of 117 MMs, 90 SKs, and 393 BN images not used in the training phase. All test images underwent the same pre-processing steps that were applied to the training images.

For most of the hyperparameter searches and to show the effect of the individual components of the proposed methods on the classification results, we utilise the ResNet-18 model since its single model performance is very competitive (see Table 7) and as due to its shallower depth compared to ResNet-101, its training is faster. In all experiments, we use the RMSprob optimiser, ImageNet mean subtraction, gray world normalisation and feature extraction from all FC layers, unless stated otherwise in the text.

We started our experiments by examining the effect of colour standardisation and normalising the images prior to feature extraction as described in Section 2.2. The obtained results are given in Table 1 and Table 2 where the average and standard deviation were calculated by running each setting 3 times. Since we observed better performance using ImageNet normalisation and training mean subtraction normalisation, we did not use the other settings in subsequent experiments. Similarly, as colour constancy was found to be beneficial, subsequent experiments always incorporated the colour standardisation step.

**Table 2**
Effects of various normalisation techniques (using fine-tuned ResNet-18).

| | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| No normalisation | 74.38 ± 0.19 | 86.00 ± 1.59 | 80.19 ± 0.89 |
| ImageNet mean | 83.48 ± 0.74 | 91.39 ± 1.33 | 87.44 ± 0.57 |
| Image mean | 75.89 ± 0.51 | 83.53 ± 1.44 | 79.70 ± 0.97 |
| Training mean | 84.36 ± 0.45 | 91.88 ± 0.85 | 88.12 ± 0.61 |

**Table 3**
Effects of various optimisers (using fine-tune ResNet-18).

| | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| SGDM | 83.30 ± 0.64 | 91.64 ± 0.99 | 87.47 ± 0.81 |
| RMSProp | 83.48 ± 0.74 | 91.39 ± 1.33 | 87.44 ± 0.57 |
| Adam | 84.38 ± 0.41 | 91.81 ± 0.64 | 88.10 ± 0.50 |

**Table 4**
Effect of weight initialisation on performance of ResNet-18 model.

| | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| ImageNet | 83.48 | 91.39 | 87.44 |
| Random | 64.07 | 84.25 | 74.16 |

**Table 5**
Classification results from the fine-tuned networks from different abstraction levels for ResNet-18.

| | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| Single FC | 82.17 | 90.97 | 86.57 |
| All FCs | 83.48 | 91.39 | 87.44 |

In the next experiment, we investigated the effect of optimiser on the classification performance. Table 3 shows the results of this comparison, i.e. the results of using the SGDM, RMSProp and Adam optimiser.

In order to investigate the generalisability of the employed transfer learning approach (i.e., extracting features from the pre-trained and the fine-tuned CNNs), we performed dimensionality reduction to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) (Hinton, 2008). This method allows to visualise the natural clusters of the high-dimensional features which we use. We used the extracted features from the first FC layer of the pre-trained network and first FC layer of the modified fine-tuned network, and utilised the Barnes-Hut Variation of t-SNE (Van Der Maaten, 2013) to speed up the algorithm while setting the dimensionality of the principal component analysis to 50. The obtained results for pre-trained and fine-tuned ResNet-18 architectures for the both training and test dataset are shown in Fig. 2.

Moreover, we performed experiments to fine-tune ResNet-18 with the same model architecture but with random weight initialisation in order to compare the obtained performance with ImageNet weight initialisation. The same initialisation method as described in Section 2.3 was used for random weight initialisation. The results of this experiments are shown in Table 4.

In the next experiment, we evaluated the effect of feature extraction from different abstraction levels of the fine-tuned ResNet-18 model. Table 5 shows the obtained results and allows to compare the performance of using features from a single FC and from all FCs.

As the results confirm, there is a level of variation in all results when running the experiments multiple times. Moreover, the models with different parameters (e.g., different optimisers) lead to slightly different but yet comparable classification results. As explained in Section 2.4, to achieve more robust and improved classification performance, we took the average over 18 models of a single architecture. The results of this fusion scheme are given in

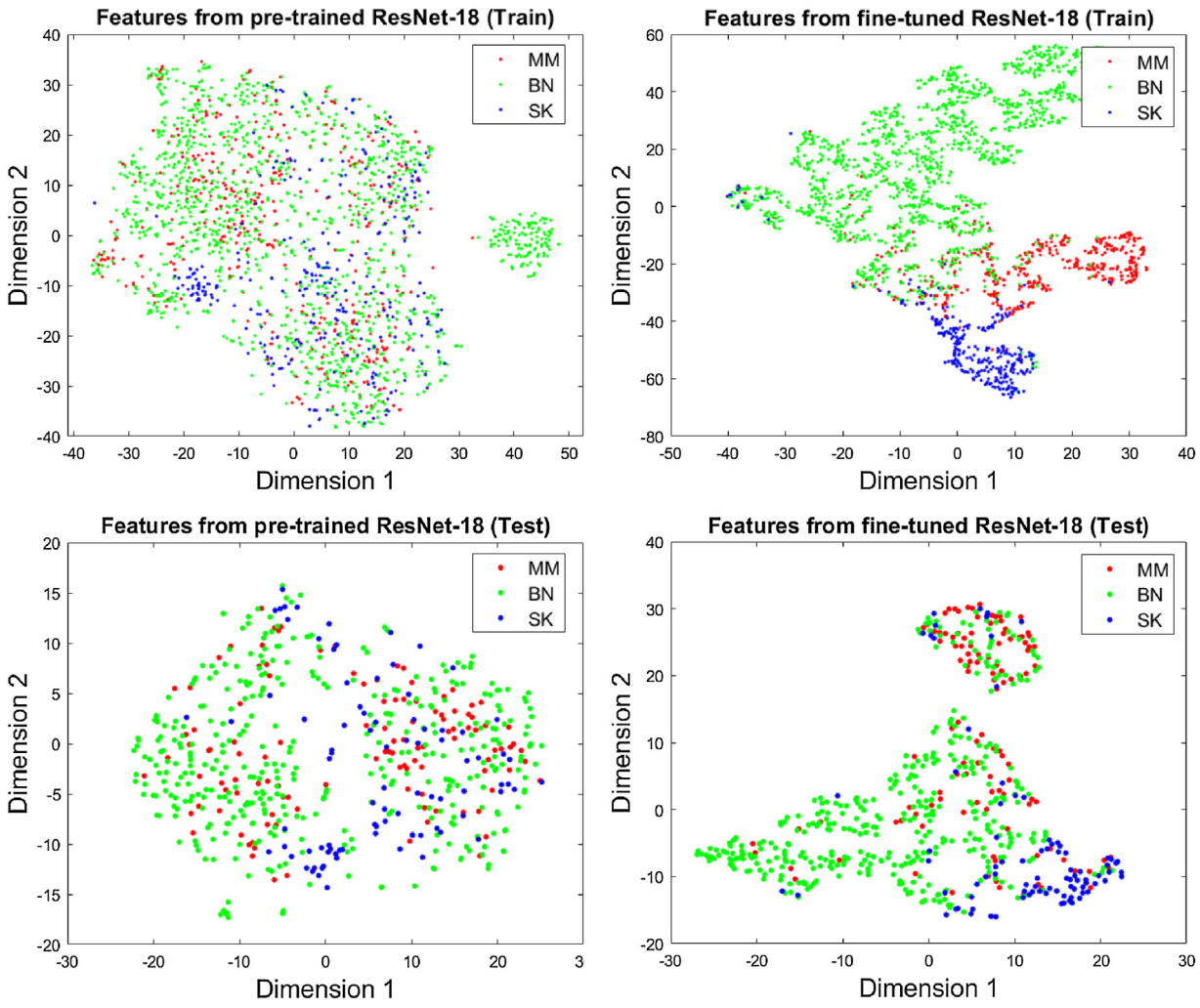


**Fig. 2.** t-SNE visualisation of the extracted features for the train (first row) and test (second row) dataset from pre-trained and fine-tuned ResNet-18 model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 6**
Fusion scheme over 18 ResNet-18 models.

| Optimiser | Normalisation | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|---|
| | | (average over 3 runs) | | |
| Adam | ImageNet mean | 85.24 | 93.20 | 89.22 |
| RMSProp | ImageNet mean | 84.70 | 93.18 | 88.94 |
| SGDM | ImageNet mean | 84.18 | 92.85 | 88.52 |
| Adam | training mean | 84.28 | 93.23 | 88.76 |
| RMSProp | training mean | 85.02 | 93.09 | 89.05 |
| SGDM | training mean | 85.54 | 92.93 | 89.23 |
| Average over above models | | 85.65 | 94.04 | 89.85 |

**Table 7**
Classification results from plain pre-trained networks, fine-tuned networks, and fusion of networks.

| | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| Plain AlexNet | 72.04 | 91.43 | 81.73 |
| Plain VGG16 | 69.85 | 89.71 | 79.78 |
| Plain ResNet-18 | 72.51 | 89.72 | 81.11 |
| Plain ResNet-101 | 74.31 | 91.90 | 83.10 |
| Fine-tuned AlexNet | 80.31 | 88.49 | 84.40 |
| Fine-tuned VGG16 | 84.16 | 93.51 | 88.83 |
| Fine-tuned ResNet-18 | 85.65 | 94.04 | 89.85 |
| Fine-tuned ResNet-101 | 85.54 | 92.24 | 88.89 |
| Fusion of all pre-trained networks | 73.19 | 93.02 | 83.10 |
| Fusion of all fine-tuned networks | 87.26 | 95.52 | 91.39 |

Table 6 for ResNet-18 networks. We performed the same fusion approach for the other deep models (i.e., AlexNet, VGG16 and ResNet-101).

Table 7 compares the performance of different deep feature extraction strategies and fusion schemes, showing the results obtained based on deep features from pre-trained single networks (plain AlexNet, plain VGG16, plain ResNet-18, and plain ResNet-101), from fine-tuned single networks (fine-tuned AlexNet, fine-tuned VGG16, fine-tuned ResNet-18 and fine-tuned ResNet-101) as well as the results obtained based on the fusion scheme of the networks. Receiver operating characteristics (ROC) curves of the fusion models (fusion of plain pre-trained networks and fusion of the fine-tuned networks) are shown in Fig. 3 and Fig. 4 for the MM and SK classification problems, respectively.

We also investigated the contribution of each single model to the final classification results. To do so, we removed one of the model at a time in the fusion scheme, calculated the resulting AUC, and report the results in Table 8.

Table 9 summarises the performance of the best performing approach of our proposed method (i.e. fusion of all fine-tuned network, the last row in Table 7) and compares it to the top three teams that participated in the ISIC 2017 challenge (ranked based

**Table 8**
Effects of removing a model in the fusion scheme.

| Fused networks (dropped model) | AUC MM (%) | AUC SK (%) | Average AUC (%) |
|---|---|---|---|
| ResNet-18+ResNet-101+VGG16 (AlexNet) | 87.01 | 95.36 | 91.18 |
| ResNet-18+ResNet-101+AlexNet (VGG16) | 87.14 | 94.84 | 90.99 |
| ResNet-101+AlexNet+VGG16 (ResNet-18) | 86.59 | 94.13 | 90.36 |
| ResNet-18+AlexNet+VGG16 (ResNet-101) | 86.76 | 95.43 | 91.09 |
| All (none) | 87.26 | 95.52 | 91.39 |

**Table 9**
Comparison of selected algorithms on the ISIC 2017 challenge.

| Authors | Approach | AUC MM (%) | AUC SK (%) | Average AUC (%) | Average accuracy (%) |
|---|---|---|---|---|---|
| Matsunaga et al. (2017) | ResNet-50 Ensemble | 86.8 | 95.3 | 91.1 | 81.6 |
| Gonzalez-Díaz (2017) | ResNet-50+Segmentation | 85.6 | 96.5 | 91.0 | 84.9 |
| Menegola et al. (2017) | ResNet-101+Inception-v4 | 87.4 | 94.3 | 90.8 | 88.3 |
| Mahbod et al. (2017) | pre-trained AlexNet+VGG | 71.5 | 90.8 | 81.1 | 81.1 |
| Proposed approach | See Table 7 | 87.3 | 95.5 | 91.4 | 87.7 |

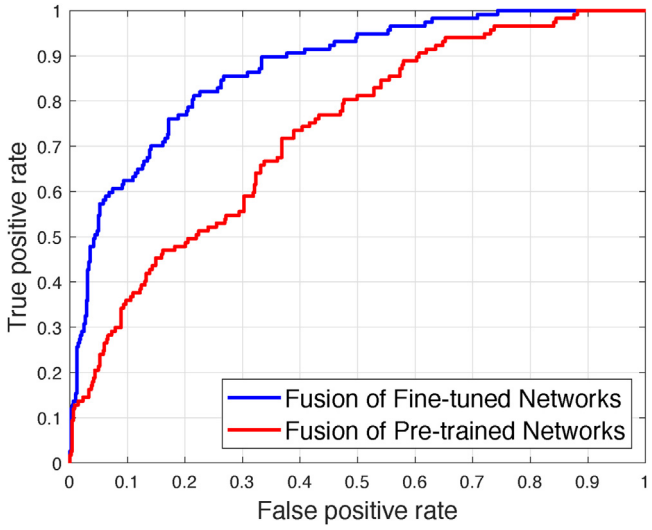


**Fig. 3.** ROC curve of MM vs. all classification. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

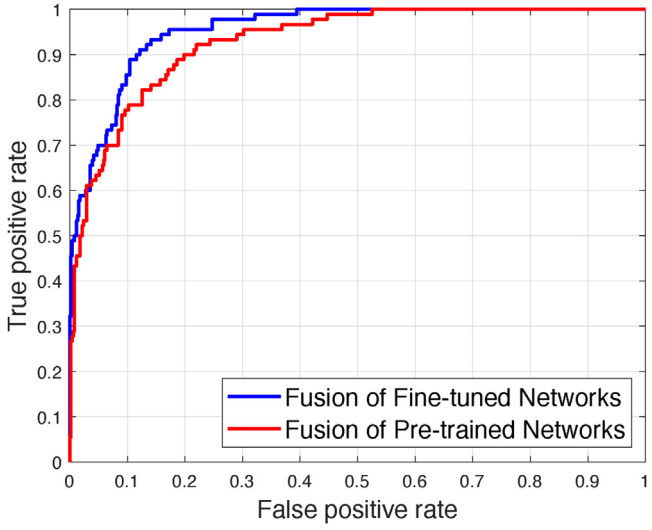


**Fig. 4.** ROC curve of SK vs. all classification. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on average AUC), as well as an earlier approach of our work that was submitted to the final classification phase of the ISIC 2017 challenge and that was obtained by feature extraction and combination of VGG16 and AlexNet pre-trained models.

The top-ranked approach by Matsunaga et al. (2017) used colour constancy (Barata et al., 2015) as a main pre-processing step and a variation of fine-tuned ResNet-50 networks to obtain the final classification. Gonzalez-Díaz (2017), the runner-up, performed lesion segmentation using a fully convolutional network (Long et al., 2015) and trained a structure segmentation network to produce a set of eight global and local structures which were assumed to be beneficial for dermatologists in their routine diagnosis procedure. In a final step, the produced set of structures along with augmented data were fed to a modified ResNet-50 network for classification. Menegola et al. (2017), whose approach was ranked third, utilised extensive data sources for fine-tuning an ensemble of seven models, six based on Inception-v4 (Szegedy et al., 2017) and one based on ResNet101 (He et al., 2016). As the comparison shows, our proposed approach outperforms all other algorithms submitted on average AUC, while it would be ranked 2nd both for the MM vs. all and for the SK vs. all classification tasks among 23 participating teams in the final test phase of the ISIC 2017 challenge (Codella et al., 2017).

Figs. 5 and 6 show examples of skin lesion images correctly and incorrectly classified by our best performing approach. Moreover, in Fig. 7, the effect of fine-tuning and model fusion in terms of accuracy for MM classification is illustrated. Here, the fusion approaches from Table 7 are selected for comparison.

The algorithm was implemented in MatLab (versions 2017b and 2018a) using the MatConvNet framework (Vedaldi and Lenc, 2015) and the MatLab Neural Network Toolbox. All experiments were performed on a single desktop computer. For the pre-processing steps an Intel Corei5-6600k 3.50 GHz CPU was utilised. The model training was performed on a single NVIDIA GTX 1070 with 8 GB of installed memory. The training of the models took around 25 min, 90 min, 70 min, and 230 min for the AlexNet, VGG16, ResNet-18 and ResNet-101, respectively.

## 4. Discussion

The main contribution of our approach is proposing a hybrid CNN ensemble scheme for skin lesion classification that combines intra-architecture and inter-architecture network fusion. Through fine-tuning the networks of different architecture multiple times with different settings and combine the results from multiple sets of fine-tuned networks the proposed method yields very accurate results without requiring extensive pre-processing, or segmenta-

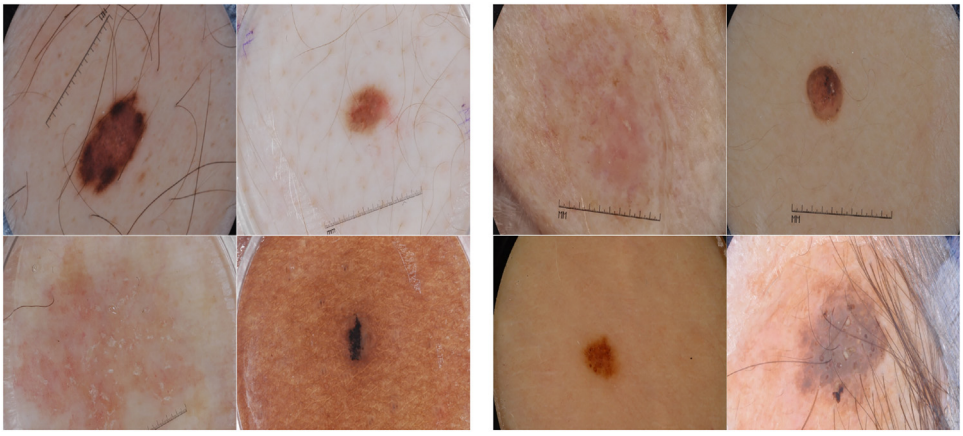

**Fig. 5.** Examples of correctly classified images for MM vs. all (left) and SK vs. all (right) tasks.

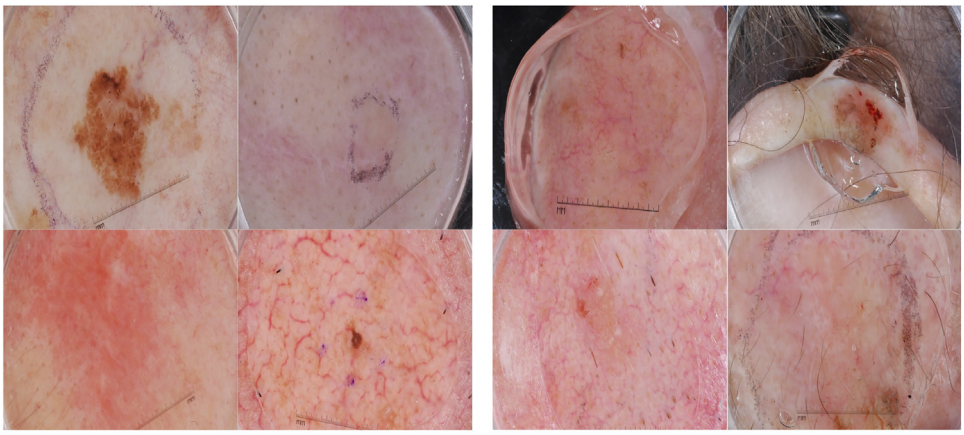

**Fig. 6.** Examples of incorrectly classified images for MM vs. all (left) and SK vs. all (right) tasks.

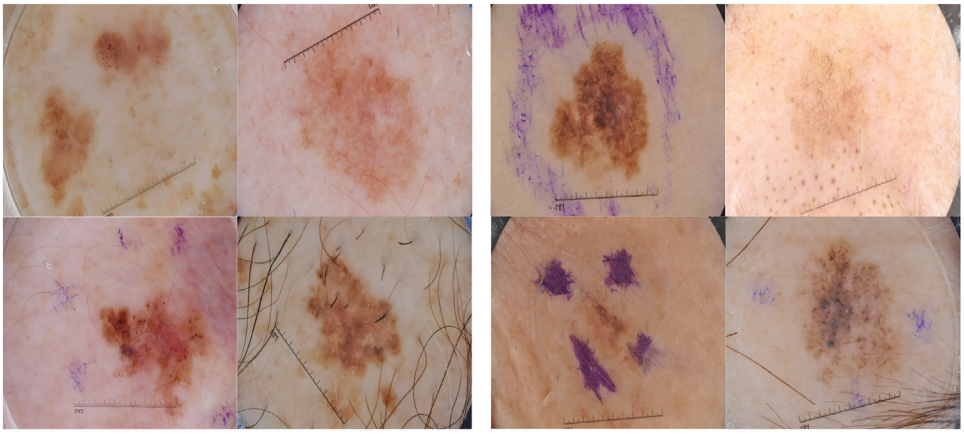

**Fig. 7.** Comparison of different fusion approaches – fusion of plain pre-trained networks and fusion of fine-tuned networks – for MM classification: MM examples that are correctly classified by both fusion approaches (left). Challenging MM examples that are only correctly classified by fusion of fine-tuned networks but not by fusion of pre-trained networks (right).

tion of the lesion area, or addition training data, which the other winning algorithms required.

To justify the intra-architecture fusion, we compared the fused probabilities of 18 different models from a single architecture as shown in Table 6. As the results clearly show, averaging over the models' outputs yields better performance compared to individual models. Moreover, it reduces the chances of degradation in results which can be caused by random weight initialisation or other factors.

Inter-architecture CNN fusion is demonstrated to further boost the prediction accuracy. Since the depths of the networks are different for AlexNet, VGG16, ResNet18 and ResNet-101, we can anticipate that deep features from different networks may provide information complementary to each other. Moreover, from Table 8 we can see that dropping each model from the fusion scheme results in a slight degradation in classification performance. However, one can use fewer networks in order to reduce the com-

putational complexity with only a relatively small performance drop.

As shown in Table 9, in comparison to other methods evaluated on the same dataset, our best performing approach delivers better performance compared to the ISIC 2017 competition winner and clearly outperforms the results of our earlier submission to the contest. However, our proposed method has lower complexity compared with those of the top 3 teams in Table 9, and was not trained on extensive external data sources. Direct comparison of the methodologies of the algorithms is challenging since different teams implemented different pre-processing steps and used various training datasets. Moreover, our algorithm can be easily used for other classification task with minimal changes in the models.

The results in Table 1 and Table 2 show the effects of pre-processing schemes on the classification results. From Table 1, it can be seen that a colour constancy algorithm can improve the performance and we hence used the colour corrected images in the remainder of the experiments. Table 2 shows the effects of different normalisation approaches to prepare the images before feeding them to the selected deep models. Among these normalisation techniques, ImageNet mean subtraction and training mean subtraction delivered better results compared to no normalisation and per image mean normalisation. Instead of choosing one of them (which delivers slightly better performance), we used both of them in our ensembling approach which leads to an improvement in the classification results.

In order to validate the generalisability of the extracted features from the pre-trained and fine-tuned networks visually, we mapped the high-dimensional feature maps to two dimensions as shown in Fig. 2 for both the training and test dataset. As can be seen, the extracted features from the plain pre-trained ResNet-18 model is distinguishable between the different classes to some extent even without any training. Hence, it can be inferred that ImageNet features are indeed well-generalised to our ternary classification task for skin lesions. It can further be observed that, although not completely separable, by fine-tuning the network using only a limited dataset, the three skin lesion classes become more distinguishable. As we fine-tuned our models using training images, the separation between classes in the training dataset is much clearer compared to the test dataset. However, if we compare the test data feature distribution between the pre-trained and fine-tuned models for the test images, it can be seen visually that fine-tuning helped a lot for separation between skin lesion classes. While Fig. 2 illustrates the applicability of the extracted features initiated by ImageNet weights visually, Table 4 confirms this quantitatively. As the results demonstrate, ImageNet weight initialisation clearly yields better performance compared to random weight initialisation which is in agreement with former studies (Tajbakhsh et al., 2016).

The results in Table 5 suggest that ensembling deep features from all FC layers in an SVM classifier delivers better performance compared to extracting features only from the first FC layer. Different FC layers are often thought to represent different levels of abstraction. Hence, our data suggest that combining features of different abstraction levels leads to improved classification accuracy.

From Table 7, we can observe that the performance of the SVM classifier when trained on features from fine-tuned networks is better compared to pre-trained networks for skin lesion classification, which is in agreement with our hypothesis. Comparing the results from different architectures shows that, although all single models delivers quite impressive classification performance, the results of the ResNet-18 model are slightly better compared to the other models. This is probably because of the shallower depth of ResNet-18 compared to the depth of ResNet-101. While generally ResNet-101 should deliver better performance (He et al., 2016), our training data size is relatively small, and the deeper model thus likely overfits to our limited dataset while the shallower network shows a better generalisation ability under these circumstances. Compared to AlexNet and VGG16, although ResNet-18 is still deeper, it consists of residual blocks which in general deliver better performance compared to regular convolutional blocks.

From Fig. 5 it can be seen that even challenging lesion images are correctly classified, while instances where an incorrect classification is obtained often include samples where the lesion is difficult to be marked out as illustrated in Fig. 6. It can also be observed from Fig. 7 that more challenging examples with vague lesion borders, low contrast and more severe artefacts can be correctly classified when fusion of fine-tuned networks is employed.

The last column of Table 9 shows the average accuracy of our best performing approach and other state-of-the-art algorithms. It should be noted that the accuracy numbers in this column are derived with mapping the score vectors to binary numbers using a probability of 50% which may not be the optimal thresholding. Optimal thresholding can be derived from the ROC curve of our best performing approach. Using Youden index method, our best performing approach yields a sensitivity of 81.20% and a specificity of 78.47% for MM vs. all classification. Likewise, a sensitivity of 93.33% and a specificity of 85.88% can be driven for SK vs. all classification from the ROC curve. However, by considering the clinical importance of not missing any MM lesions, it is possible to choose a threshold from ROC curve that improves the sensitivity of the MM vs. all classification at the expense of reduced specificity. From the ROC curve of MM vs. all, a sensitivity of 85%, 90% and 95% can be reached with corresponding specificity of 73.29%, 62.32% and 44.72%, respectively.

As stated in Section 2.1, the training images in the three classes are not well-balanced as there are relatively few MM and SK images in comparison to BN lesions. While it is common practice to balance a dataset in such cases using e.g. boot strapping, class-balanced cost functions or through resampling, we have not gained any improvement in performance by balancing the dataset (we performed resampling of the minority classes to deal with class imbalance), while the training time drastically increased. This appears to confirm experiments on weighting strategies reported in Menegola et al. (2017). We therefore do not explicitly address class imbalance in our approach, nor do (Matsunaga et al., 2017; Barata et al., 2015; Menegola et al., 2017) i.e. the three top teams of the ISIC 2017 contest.

There are some limitations of our current approach that can be explored in future work. First, even though we show that fusing deep features from different CNNs can improve the classification accuracy, the number of networks investigated is limited. Extending this study by incorporating other CNN architectures, such as GoogleNet (Szegedy et al., 2015) or DensNet (Huang et al., 2017), may result in further improvements. Moreover, ensembling hand-crafted feature descriptors as used in conventional methods alongside proposed fused deep features could lead to better classification performance (Oliveira et al., 2017; Codella et al., 2016), but also increases the complexity of the method. Second, the employed training data are limited. The amount of training data is important for appropriately training or fine-tuning DNNs. Hence, having access to additional reliable skin lesion data sources can lead to better results. Third, using pre-trained networks for skin lesion classification requires the images to be resized to a certain dimension that is pre-defined for other image classification tasks. Some valuable information may be lost during the downsampling step. Although in some works, images were resized to higher resolutions (e.g. $339 \times 339$ pixels in Kawahara et al. (2016), $448 \times 448$ pixels in DeVries (2017) and up to $512 \times 512$ in Yu et al. (2017)), they are still significantly smaller compared to their original sizes. However, these input sizes are still significantly bigger compared to our approach and they may hence capture more useful information. Finally, using more extensive pre-processing steps or data

augmentation techniques might further improve the classification performance (Oliveira et al., 2016; Schaefer et al., 2011; Iyatomi et al., 2011).

## 5. Conclusions

In this paper, a fully automatic computerised method with minimal pre- and post-processing operations is proposed for accurate skin lesion classification. The proposed algorithm ensembles deep features from multiple pre-trained and fine-tuned DNNs at multiple abstraction levels and fuses the prediction probability vectors of different models. The obtained results show that such fusion of features provides better discrimination ability and is complementary to the individual networks. The general performance of the proposed method is competitive with other state-of-the-art algorithms, while the generalisation ability of the proposed approach for other medical imaging classification tasks is subject for future work.

## Acknowledgments

This work was supported by the European Union Horizon 2020 Research and Innovation Program ("CaSR Biomedicine", 675228). The authors appreciate the help of the TissueGnostics support team http://tissuegnostics.com/en/ for providing valuable comments and feedback. Moreover, we would like to thank Prof. Örjan Smedby since part of this study was conducted in his research group.

## References

Abbas, Q., Celebi, M.E., Garcia, I.F., 2011. Hair removal methods: a comparative study for dermoscopy images. Biomed. Signal Process. Control 6 (4), 395–404.
Apalla, Z., Nashan, D., Weller, R.B., Castellsagué, X., 2017. Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. Dermatol. Ther. 7 (1), 5–19.
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 (12), 1536–1570.
Argenziano, G., Soyer, H.P., De Giorgi, V., Piccolo, D., Carli, P., Delfino, M., 2002. Dermoscopy: a tutorial. EDRA, Medical Publishing & New Media.
Balch, C.M., Gershenwald, J.E., Soong, S.-j., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., 2009. Final version of 2009 AJCC melanoma staging and classification. J. Clin. Oncol. 27 (36), 6199–6206.
Barata, C., Celebi, M.E., Marques, J.S., 2015. Improving dermoscopy image classification using color constancy. IEEE J. Biomed. Health Inform. 19 (3), 1146–1152.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.
Carli, P., Quercioli, E., Sestini, S., Stante, M., Ricci, L., Brunasso, G., De Giorgi, V., 2003. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. Br. J. Dermatol. 148 (5), 981–984.
Carrera, C., Marchetti, M.A., Dusza, S.W., Argenziano, G., Braun, R.P., Halpern, A.C., Jaimes, N., Kittler, H.J., Malvehy, J., Menzies, S.W., 2016. Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: a web-based International Dermoscopy Society study. JAMA Dermatol. 152 (7), 798–806.
Celebi, M.E., Kingravi, H., Uddin, B., Iyatomi, H., Aslandogan, A., Stoecker, W.V., Moss, R.H., 2007. A methodological approach to the classification of dermoscopy images. Comput. Med. Imag. Grap. 31 (6), 362–373.
Celebi, M.E., Kingravi, H.A., Iyatomi, H., Aslandogan, Y.A., Stoecker, W.V., Moss, R.H., Malters, J.M., Grichnik, J.M., Marghoob, A.A., Rabinovitz, H.S., Menzies, S.W., 2008. Border detection in dermoscopy images using statistical region merging. Skin Res. Technol. 14 (3), 347–353.
Celebi, M., Iyatomi, H., Schaefer, G., Stoecker, W., 2009. Lesion border detection in dermoscopy images. Comput. Med. Imag. Grap. 33 (2), 148–153.
Celebi, M.E., Wen, Q., Hwang, S., Iyatomi, H., Schaefer, G., 2013. Lesion border detection in dermoscopy images using ensembles of thresholding methods. Skin Res. Technol. 19 (1), e252–e258.
Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R., 2015. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: International Workshop on Machine Learning in Medical Imaging, Springer, pp. 118–126.
Codella, N., Nguyen, Q.-B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., Smith, J.R. Deep learning ensembles for melanoma recognition in dermoscopy images. arXiv preprint 1610.04662.
Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. 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 1710.05006.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 248–255.
DeVries, T., Ramachandram, D. Skin lesion classification using deep multi-scale convolutional neural networks. arXiv preprint 1703.01402.
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 (7639), 115–118.
Fleming, M.G., Steger, C., Zhang, J., Gao, J., Cognetta, A.B., Dyer, C.R., 1998. Techniques for a structural analysis of dermatoscopic imagery. Comput. Med. Imag. Grap. 22 (5), 375–389.
Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A. Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting. arXiv preprint 1808.01694.
Gonzalez-Díaz, I. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. arXiv preprint 1703.01976.
Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A. 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 1605.01397.
Harangi, B. Skin lesion detection based on an ensemble of deep convolutional neural networks. arXiv preprint 1705.03360 (2015) 1–4.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 770–778.
Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. CVPR, vol. 1, 3.
Iyatomi, H., Celebi, M.E., Schaefer, G., Tanaka, M., 2011. Automated color calibration method for dermoscopy images. Comput. Med. Imag. Grap. 35 (2), 89–98.
Jaisakthi, S.M., Chandrabose, A., Mirunalini, P. Automatic skin lesion segmentation using semi-supervised learning technique. arXiv preprint 1703.04301.
Kawahara, J., BenTaieb, A., Hamarneh, G., 2016. Deep features to classify skin lesions. In: 13th International Symposium on Biomedical Imaging, IEEE, pp. 1397–1400.
Kingma, D.P., Ba, J. Adam: a method for stochastic optimization. arXiv preprint 1412.6980.
Kittler, H., Pehamberger, H., Wolff, K., Binder, M., 2002. Diagnostic accuracy of dermoscopy. Lancet Oncol. 3 (3), 159–165.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.
Li, K.M., Li, E.C. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. arXiv preprint 1807.08332.
Li, Y., Shen, L., 2018. Skin lesion analysis towards melanoma detection using deep learning network. Sensors 18 (2), 556.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.
Lopez, A.R., Giro-i Nieto, X., Burdick, J., Marques, O., 2017. Skin lesion classification from dermoscopic images using deep learning techniques. In: 13th IASTED International Conference on Biomedical Engineering (BioMed), IEEE, pp. 49–54.
Mahbod, A., Ecker, R., Ellinger, I. Skin lesion classification using hybrid deep neural networks. arXiv preprint 1702.08434.
Matsunaga, K., Hamada, A., Minagawa, A., Koga, H. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv preprint 1703.03108.
Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E. RECOD titans at ISIC challenge 2017. arXiv preprint 1703.04819.
Mirunalini, P., Chandrabose, A., Gokul, V., Jaisakthi, S.M. Deep learning for skin lesion classification. arXiv preprint 1703.04364.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT press.
Nader, C., Nader, B., 2018. Experiments using deep learning for dermoscopy image analysis. Pattern Recogn. Lett., 1–9.
Oliveira, R.B., Mercedes Filho, E., Ma, Z., Papa, J.P., Pereira, A.S., Tavares, J.M.R.S., 2016. Computational methods for the image segmentation of pigmented skin lesions: a review. Comput. Methods Programs Biomed. 131, 127–141.
Oliveira, R.B., Pereira, A.S., Tavares, J.M.R.S., 2017. Skin lesion computational diagnosis of dermoscopic images: ensemble models based on input feature manipulation. Comput. Methods Programs Biomed. 149, 43–53.
Oliveira, R.B., Papa, J.P., Pereira, A.S., Tavares, J.M.R.S., 2018. Computational methods for pigmented skin lesion classification in images: review and future trends. Neural Comput. Appl. 29 (3), 613–636.
Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10 (3), 61–74.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Schaefer, G., Rajab, M.I., Celebi, M.E., Iyatomi, H., 2011. Colour and contrast enhancement for improved skin lesion segmentation. Comput. Med. Imag. Grap. 35 (2), 99–104.
Shellenberger, R., Nabhan, M., Kakaraparthi, S., 2016. Melanoma screening: a plan for improving early detection. Ann. Med. 48 (3), 142–148.
Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint 1409.1556.
Stolz, W., Riemann, A., Cognetta, A.B., Pillet, L., Abmayr, W., Holzel, D., Bilek, P., Nachbar, F., Landthaler, M., Braun-Falco, O., 1994. ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. 4 (7), 521–527.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 1–9.
Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. Proceedings of the Thirty-First Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, 4278–4284.
Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imag. 35 (5), 1299–1312.
Thomas, L., Puig, S., 2017. Dermoscopy, digital dermoscopy and other diagnostic tools in the early detection of melanoma and follow-up of high-risk skin cancer patients. Acta Derm.-Venereol. 97, 14–21.
Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning 4 (2), 26–31.
Van Der Maaten, L. Barnes-hut-sne. arXiv preprint 1301.3342.
Vasconcelos, C.N., Vasconcelos, B.N. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. arXiv preprint 1702.07025.
Vedaldi, A., Lenc, K., 2015. MatConvNet: convolutional neural networks for MATLAB. In: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, pp. 689–692.
Vestergaard, M.E., Macaskill, P., Holt, P.E., Menzies, S.W., 2008. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br. J. Dermatol. 159 (3), 669–676.
WHO. Ultraviolet radiation and the INTERSUN Programme (Data Accessed May 11, 2018). http://www.who.int/uv/faq/skincancer/en/index1.html.
Yang, X., Zeng, Z., Yeo, S.Y., Tan, C., Tey, H.L., Su, Y. A novel multitask deep learning model for skin lesion segmentation and classification. arXiv preprint 1703.01025.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3 (1), 32–35.
Yu, Z., Jiang, X., Wang, T., Lei, B., 2017. Aggregating deep convolutional features for melanoma recognition in dermoscopy images. In: International Workshop on Machine Learning in Medical Imaging, Springer, pp. 238–246.
Zhou, H., Schaefer, G., Sadka, A., Celebi, M.E., 2009. Anisotropic mean shift based fuzzy c-means segmentation of dermoscopy images. IEEE J. Sel. Topics Signal Process. 3 (1), 26–34.
Zhou, H., Schaefer, G., Celebi, M.E., Lin, F., Liu, T., 2011. Gradient vector flow with mean shift for skin lesion segmentation. Comput. Med. Imag. Grap. 35 (2), 121–127.