SKIN LESION CLASSIFICATION USING HYBRID DEEP NEURAL NETWORKS

Amirreza Mahbod[⋆†], *Gerald Schaefer*[‡], *Chunliang Wang*[§], *Rupert Ecker*[†], *Isabella Ellinger*[⋆]

[⋆]Institute for Pathophysiology and Allergy Research, Medical University of Vienna, Austria
[†]Department of Research and Development, TissueGnostics GmbH, Austria
[‡]Department of Computer Science, Loughborough University, U.K.
[§]Department of Biomedical Engineering & Health Systems, KTH Royal Institute of Technology, Sweden

ABSTRACT

Skin cancer is one of the major types of cancers with an increasing incidence over the past decades. Accurately diagnosing skin lesions to discriminate between benign and malignant skin lesions is crucial to ensure appropriate patient treatment. While there are many computerised methods for skin lesion classification, convolutional neural networks (CNNs) have been shown to be superior over classical methods. In this work, we propose a fully automatic computerised method for skin lesion classification which employs optimised deep features from a number of well-established CNNs and from different abstraction levels. We use three pre-trained deep models, namely AlexNet, VGG16 and ResNet-18, as deep feature generators. The extracted features then are used to train support vector machine classifiers. In a final stage, the classifier outputs are fused to obtain a classification. Evaluated on the 150 validation images from the ISIC 2017 classification challenge, the proposed method is shown to achieve very good classification performance, yielding an area under receiver operating characteristic curve of 83.83% for melanoma classification and of 97.55% for seborrheic keratosis classification.

Index Terms— Medical imaging, skin cancer, melanoma classification, dermoscopy, deep learning, network fusion.

1. INTRODUCTION

Skin cancer is one of the most common cancer types worldwide [1]. As an example, skin cancer is the most common cancer type in the United States and it is estimated that one in five Americans will develop skin cancer in their lifetime. Among different types of skin cancers, malignant melanoma (the deadliest type) is responsible for 10,000 deaths annually just in the United States [2]. However, if detected early it 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 [3].

There are a number of non-invasive tools that can assist dermatologists in diagnosis such as macroscopic images which are acquired by standard cameras or mobile phones [1]. However, these images usually suffer from poor quality and resolution. Significantly better image quality is provided by dermoscopic devices which have become an important non-invasive tool for detection of melanoma and other pigmented skin lesions. Dermoscopy supports better differentiation between different lesion types based on their appearance and morphological features [4].

Visual inspection of dermoscopic images is a challenging task that relies on a dermatologist's experience. Despite the definition of commonly employed diagnostic schemes such as the ABCD rule [5] or the 7-point checklist [6], due to the difficulty and subjectivity of human interpretation as well as the variety of lesions and confounding factors encountered in practice (see Fig. 1 for some examples of common artefacts encountered in dermoscopic images), computerised analysis of dermoscopic images has become an important research area to support diagnosis [7]. Conventional computer-aided methods for dermoscopic lesion classification typically involve three main stages: segmenting the lesion area, extracting hand-crafted image features from the lesion and its border, and classification [8]. In addition, often extensive preprocessing is involved to improve image contrast, perform white balancing, apply colour normalisation or calibration, or remove image artefacts such as hairs or bubbles [1, 9].

With the advent of deep convolutional neural networks (CNNs) and considering their excellent performance for natural image classification, there is a growing trend to utilise them for medical image analysis including skin lesion classification [10]. Likewise, in this paper, we exploit the power of deep neural networks for skin lesion classification. Using CNNs, which are pre-trained on a large dataset of natural images, as optimised feature extractors for skin lesion images can potentially overcome the drawbacks of conventional approaches and can also deal with small task-specific training datasets. A number of works [11, 12, 13, 10] have

This research has received funding from the Marie Sklodowska-Curie Actions of the European Union's Horizon 2020 programme under REA grant agreement no. 675228.

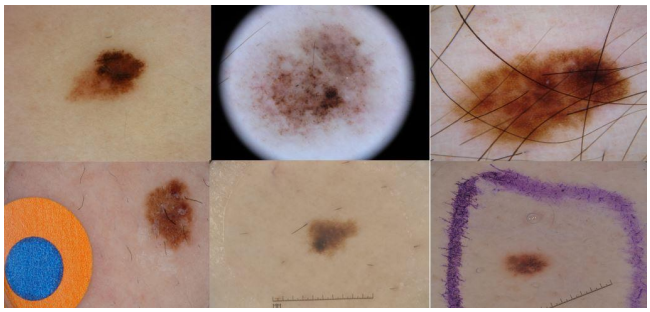

Fig. 1. Common artefacts in dermoscopic images from the ISIC challenge. Normal image, dark corner artefact, skin hair artefacts, colour chart artefact, ruler marker artefact, ink marker artefact (left to right, top to bottom).

tried to extract deep features from skin lesion images and then train a classical classifier. However, these studies are limited by exploiting specific pre-trained network architectures or using specific layers for extracting deep features. Also, the utilised pre-trained networks were limited to a single network. In [11], a single pre-trained AlexNet was used while [10] employed a single pre-trained VGG16, and [13] utilised a single pre-trained Inception-v3 [14] network.

In this work, we hypothesise that using different pre-trained models, extracting features from different layers and ensemble learning can lead to classification performance competitive with specialised state-of-the art algorithms. In our approach, we utilise three deep models, namely AlexNet [15], VGG16 [16] and ResNet-18 [17], which are pre-trained on ImageNet [18], as optimised feature extractors and support vector machines, trained using a subset of images from the ISIC archive[1], as classifiers. In the final stage, we fuse the SVM outputs to achieve optimal discrimination between the three lesion classes (malignant melanoma, seborrheic keratosis and benign nevi).

2. MATERIALS AND METHODS

2.1. Dataset

We use the training, validation and test images of the ISIC 2016 competition [19] as well as the training set of the ISIC 2017 competition[2] for training the classifiers. In total, 2037 colour dermoscopic skin images are used which include 411 malignant melanoma (MM), 254 seborrheic keratosis (SK) and 1372 benign nevi (BN). The images are of various sizes (from 1022×767 to 6748×4499 pixels), photographic angles and lighting conditions and different artefacts such as the ones shown in Fig. 1. A separate set of 150 skin images is provided as a validation set. It is these validation images that we use to evaluate the results of our proposed method.

[1]https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main

[2]https://challenge.kitware.com/#phase/5840f53ccad3a51cc66c8dab

2.2. Pre-processing

A generic flowchart of our proposed approach is shown in Fig. 2.

In our approach, we try to keep the pre-processing steps minimal to ensure better generalisation ability when tested on other dermoscopic skin lesion datasets. We thus only apply three standard pre-processing steps which are generally used for transfer learning. First, we normalise the images by subtracting the mean RGB value of the ImageNet dataset as suggested in [15] since the pre-trained networks were originally trained on those images. Next, the images are resized using bicubic interpolation to be fed to the networks (227×227 and 224×224). Finally, we augment the training set by rotating the images by 0, 90, 180 and 270 degree and then further applying horizontal flipping. This augmentation leads to an increase of training data by a factor of eight.

2.3. Deep Learning Models

Our deep feature extractor uses three pre-trained networks. In particular, we use AlexNet [15], a variation of VGGNet named VGG16 [16], and a variation of ResNet named ResNet-18 [17] as optimised feature extractors. These models have shown excellent classification performance for natural image classification in the Image Large Scale Visual Recognition Challenges [20] and various other tasks. We choose the shallowest variations of VGGNet and ResNet to prevent overfitting since the number of training images in our study is limited. We explore extracting features from different layers of the pre-trained models to see how they can affect classification results. The features are mainly extracted from the last fully connected (FC) layers of the pre-trained AlexNet and pre-trained VGG16. We use the first and second fully connected layers (referred to as FC6 and FC7 with dimensionality 4096) and the concept detector layer (referred to as FC8 with dimensionality 1000). For ResNet-18, since it has only one FC layer, we also extract features from the last convolutional layer of the pre-trained model.

2.4. Classification and Fusion

The above features along with the corresponding labels (i.e., skin lesion type) are then used to train multi-class non-linear support vector machine (SVM) classifiers. We train different classifiers for each network and then, to fuse the results, average the class scores to obtain the final classification result. To evaluate the classification results, we map SVM scores to probabilities using logistic regression [21]. Since the classifiers are trained for a multi-class problem with three classes, we combine the scores to yield results for the two binary classification problems defined in the ISIC 2017 challenge, which are malignant melanoma vs. all and seborrheic keratosis vs. all classifications.

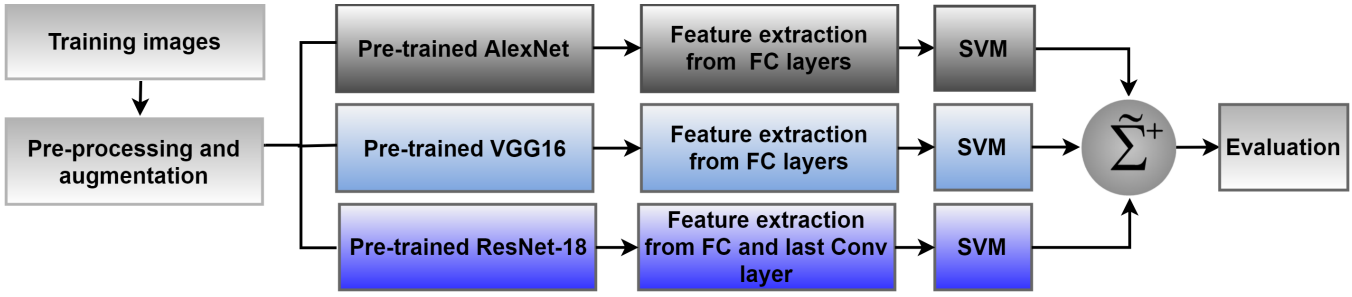

Fig. 2. Flowchart of the proposed method.

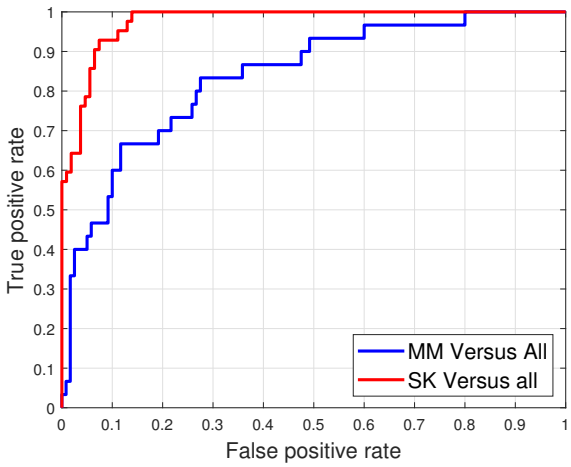

Fig. 3. ROC curve of the best performing approach.

3. RESULTS

As mentioned above, evaluation is performed based on the 150 validation images provided by the ISIC 2017 challenge. The validation set comprises 30 malignant melanoma, 42 seborrheic keratosis and 78 benign nevus images. For evaluation, we employ the suggested performance measure of area under the receiver operating characteristics curve (AUC). The raw images are resized to 227×227 pixels for AlexNet and to 224×224 pixels for VGG16 and ResNet-18. For each individual network and also for each fusion scheme, the results are derived by taking the average of the outputs over 5 iterations.

The obtained classification results are shown in Table 1 for all single networks and for all fused models.

Fig. 3 shows the receiver operating characteristic (ROC) curve of our best performing approach (i.e., fusion of all networks) while Fig. 4 show examples of skin lesion images that are incorrectly classified by this approach.

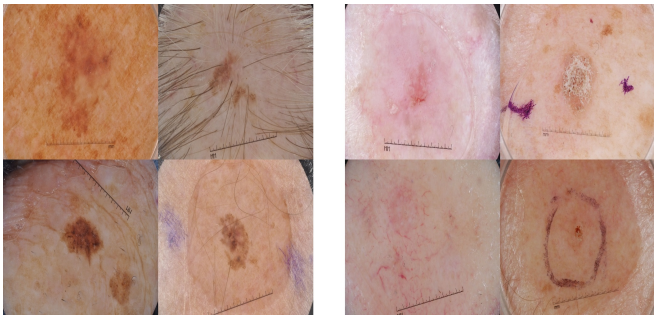

Fig. 4. Examples of incorrectly classified images for malignant melanoma vs. all (left) and seborrheic keratosis vs. all (right) tasks.

4. DISCUSSION

The main contribution of this study is proposing a hybrid approach for skin lesion classification based on deep feature fusion, training multiple SVM classifiers and combining the probabilities for fusion in order to achieve high classification performance.

From the classification results in Table 1, we can infer a number of observations. First of all, for all approaches, even for the worst performing approach, the classification results are far better than pure chance (i.e. AUC of 50%) which confirms that the concept of transfer learning can be successfully applied to skin lesion classification. Besides this, for all single networks, fusing the features from different abstraction levels leads to better classification performance compared to extracting features from a single FC layer.

Features extracted from AlexNet lead to the best performance of a single network approach. This could be potentially related to the network depth. Since our training dataset is not very big, using a shallower network may lead to better results.

The single network approaches are however outclassed by our proposed method of employing multiple CNNs and fusing their SVM classification outputs. The obtained results demonstrate that significantly better classification performance can be achieved.

Table 1. Experimental results on ISIC 2017 validation dataset.

network	feature layers	MM AUC	SK AUC	avg. AUC
AlexNet	FC8	80.67	94.95	87.81
AlexNet	all FC	82.81	96.65	89.73
VGG16	FC8	82.61	90.94	86.78
VGG16	all FC	82.06	95.46	88.76
ResNet-18	FC	81.00	91.93	86.47
ResNet-18	FC + last convol. layer	82.81	94.22	88.51
AlexNet + VGG16 fusion	all FC	83.56	97.05	90.30
AlexNet + ResNet-18 fusion	all FC	83.53	97.05	90.29
VGG16 + ResNet-18 fusion	all FC	83.69	95.97	89.83
fusion of all networks	all FC	**83.83**	**97.55**	**90.69**

While our proposed method is shown to give very good performance on what is one of the most challenging public skin lesion dataset, there are some limitations that can be addressed in future work. First, the number of pre-trained networks that we have studied so far is limited. Extending the model to incorporate more advanced pre-trained models such as DenseNets [22] could lead to further improved classification performance. Second, extending the training data is expected to lead to better results for each individual network as well as their combinations. Finally, resizing the images to very small patches might removing some useful information from the lesions. Although in a number of studies bigger training patches were used (e.g. 339×339 in [11] or 448×448 in [23]), these are still significantly smaller compared to the captured image sizes. Cropping the images or using segmentation masks to guide the resizing could be a potential solution for dealing with this.

5. CONCLUSIONS

In this paper, we have proposed a fully automatic method for skin lesion classification. In particular, we have demonstrated that pre-trained deep learning models, trained for natural image classification, can also be exploited for dermoscopic image classification. Moreover, fusing the deep features from various layers of a single network or from various pre-trained CNNs is shown to lead to better classification performance. Overall, very good classification results have been demonstrated on the challenging images of the ISIC 2017 competition, while in future work fusing more deep features also from further CNNs can potentially lead to even better predictive models.

6. REFERENCES

[1] R. B. Oliveira, J. P. Papa, A. S. Pereira, and J. M. R. S. Tavares, "Computational methods for pigmented skin lesion classification in images: review and future trends," *Neural Computing and Applications*, vol. 29, no. 3, pp. 613–636, 2018.

[2] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the U.S. population, 2012," *JAMA Dermatology*, vol. 151, no. 10, pp. 1081–1086, 2015.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] K. Steiner, M. Binder, M. Schemper, K. Wolff, and H. Pehamberger, "Statistical evaluation of epiluminescence dermoscopy criteria for melanocytic pigmented lesions," *Journal of the American Academy of Dermatology*, vol. 29, no. 4, pp. 581–588, 1993.

[5] W. Stolz, A. Riemann, A. B. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, M. Landthaler, and O. Braun-Falco, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," *European Journal of Dermatology*, vol. 4, no. 7, pp. 521–527, 1994.

[6] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "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.," *Archives of Dermatology*, vol. 134, no. 12, pp. 1536–1570, 1998.

[7] M. G. Fleming, C. Steger, J. Zhang, J. Gao, A. B. Cognetta, I. Pollak, and C. R. Dyer, "Techniques for a structural analysis of dermatoscopic imagery," *Computerized Medical Imaging and Graphics*, vol. 22, no. 5, pp. 375–389, 1998.

[8] M. E. Celebi, H. Kingravi, B. Uddin, H. Iyatomi, A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A

methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007.

[9] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[10] A. R. Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *13th IASTED International Conference on Biomedical Engineering*, 2017, pp. 49–54.

[11] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *13th International Symposium on Biomedical Imaging*, 2016, pp. 1397–1400.

[12] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *International Workshop on Machine Learning in Medical Imaging*, 2015, pp. 118–126.

[13] P. Mirunalini, A. Chandrabose, V. Gokul, and S. M. Jaisakthi, "Deep learning for skin lesion classification," *arXiv preprint arXiv:1703.04364*, 2017.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[19] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "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.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[21] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] T. DeVries and D. Ramachandram, "Skin lesion classification using deep multi-scale convolutional neural networks," *arXiv preprint arXiv:1703.01402*, 2017.