Received: 6 June 2018 | Revised: 20 July 2018 | Accepted: 30 August 2018

DOI: 10.1111/exd.13777

ORIGINAL ARTICLE

WILEY Experimental Dermatology

Multimodal skin lesion classification using deep learning

Jordan Yap[1] | William Yolland[1] | Philipp Tschandl[2,3]

[1]MetaOptima Technology Inc., Vancouver, British Columbia, Canada

[2]Department of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada

[3]Department of Dermatology, Medical University of Vienna, Vienna, Austria

Correspondence
Philipp Tschandl, Department of Dermatology, Medical University of Vienna, Vienna, Austria.
Email: philipp.tschandl@meduniwien.ac.at

Abstract

While convolutional neural networks (CNNs) have successfully been applied for skin lesion classification, previous studies have generally considered only a single clinical/macroscopic image and output a binary decision. In this work, we have presented a method which combines multiple imaging modalities together with patient metadata to improve the performance of automated skin lesion diagnosis. We evaluated our method on a binary classification task for comparison with previous studies as well as a five class classification task representative of a real-world clinical scenario. We showed that our multimodal classifier outperforms a baseline classifier that only uses a single macroscopic image in both binary melanoma detection (*AUC* 0.866 vs 0.784) and in multiclass classification (*mAP* 0.729 vs 0.598). In addition, we have quantitatively showed the automated diagnosis of skin lesions using dermatoscopic images obtains a higher performance when compared to using macroscopic images. We performed experiments on a new data set of 2917 cases where each case contains a dermatoscopic image, macroscopic image and patient metadata.

KEYWORDS

deep learning, dermatology, dermatoscopy, feature fusion, multimodal

1 | INTRODUCTION

Dermatoscopy is regarded as the state of the art technique in skin cancer screening which provides a higher diagnostic accuracy than the unaided eye.[1,2] Increasing the sensitivity for diagnosing melanoma is key as detecting melanoma in an early stage can decrease the mortality rate.[3] Although the incidence rate of melanoma is increasing,[4] keratinocyte cancer such as squamous cell carcinomas (including actinic keratoses and Bowen's disease) and basal cell carcinomas are far more common.[5] While those diseases rarely result in fatal outcomes compared to melanoma, the economic burden has been shown to be one of the highest for Medicare patients.[6] Especially for basal cell carcinomas, costs rise significantly if they have to be treated in an advanced stage due to delayed diagnosis.[7]

In previous studies, teledermatology guided referrals using dermatoscopy have been shown to be accurate,[8] reduce burden on healthcare systems, and reduce waiting times for necessary skin cancer surgery.[9] Automated classification systems can be one tool to help quickly screen a large number of patients and identify those most at risk. This may help to reduce unnecessary visits to the clinic and allow skin cancer to be detected while it is still at an early stage.

Automated analysis of dermatoscopic images, specifically with neural networks, has been studied for many years[10] but recently gained traction with promising results when compared against physicians.[11] Clinical close-up (*macroscopic*) images can also be harnessed for evaluation by a neural network for diagnosing skin cancer; however, this technique has been demonstrated to provide lower accuracy when predicting multiple disease classes.[12] In a clinical practice, dermatologists very rarely evaluate only one image

modality but rather see patients in person across one or more visits. Thus, physicians are able to combine a dermatoscopic view with a clinical view and patient information (eg, time of onset, change of lesion, approximate age, gender and location of the disease) in their analysis of each lesion. The availability of multiple feature sources is equally true for most teledermatoscopy evaluations as well.[9]

The focus of this work is to explore the importance of the dermatoscopic imaging modality specifically in conjunction with its macroscopic counterpart for the task of automated lesion diagnosis. We also include comparisons with previous studies that leverage patient-level metadata as this has been shown to improve diagnostic accuracy.[13] Our network architecture, shown in Figure 1, is chosen using a grid-search technique while trying to maintain overall simplicity where possible. We employ two ResNet-50[14] convolutional neural network (CNN) architectures followed by a late fusion technique to combine features. We show through our experiments that, just as physicians are able to integrate an abundance of data when making a diagnosis, it is beneficial for our network to integrate data from multiple modalities.

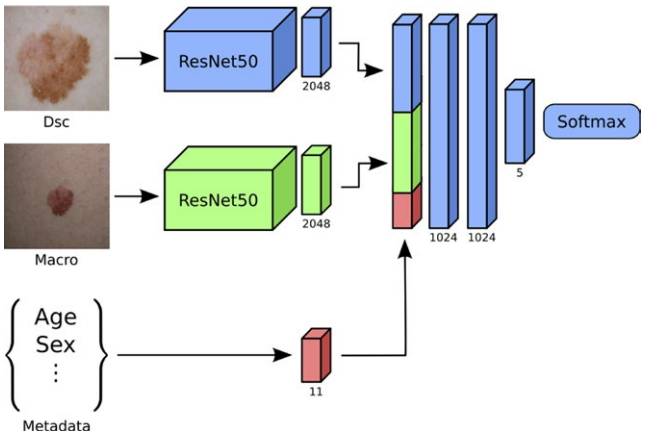

FIGURE 1 Diagram of network architecture for multimodal classification

2 | RELATED WORK

2.1 | Macroscopic image analysis

Breakthroughs in classifying macroscopic images have recently been made by two studies. The first, by Esteva et al[12] collected over 100 000 macroscopic images from undisclosed online databases and the Stanford University Medical Center. From this, they fine-tuned an Inception-V3 network to distinguish between a variety of skin conditions. Instead of using a flat class-partitioning scheme, they employed a hierarchical partitioning algorithm using a taxonomy tree to balance an otherwise imbalanced data set. The algorithm selects a class label from nodes in the tree whose aggregated descendants have sufficient images on which to train. Since their taxonomy tree was not made public we were unable to compare against their work. Once trained, their network performed on par with board-certified physicians at detecting keratinocyte cancer or detecting melanoma in a binary classification setting.

The second, and more recent study is by Han et al[15] who reported on a collection of over 20 000 macroscopic images covering 12 disease classes[1] from their own proprietary data set as well as publicly available data sets. They employed fine-tuning from a pretrained model (while freezing early layers) using a deep ResNet-152[14] architecture trained on manually cropped images. Similar to Esteva et al., they achieved a classification accuracy which was competitive with trained physicians; the Top-1 accuracy ranged between 55% and 57.3% across different subsets.

Regarding human performance, a recent study from Sinz et al[2] reported increased accuracy of physicians who view both dermatoscopic and macroscopic images of the same lesion in a challenging data set of nonpigmented cases. This supports previous findings which report higher diagnostic accuracy of physicians when using dermatoscopy as opposed to only using the unaided eye.[16]

2.2 | Dermatoscopic image analysis

There were early efforts to apply neural networks to dermatoscopic skin lesion classification[10]; however, the subsequent years focused mainly on image processing techniques[17–19] and techniques for feature extraction.[20–22] In recent years, there has been a shift back towards the end-to-end application of neural networks.[11–13,15,23] This is largely thanks to an exponential increase in GPU computing capability as well as an overall improvement in the effectiveness of convolutional neural networks (both through significant research in network design[14,24,25] and the curation of large data sets such as ImageNet[26]). This interest has been further fuelled by the efforts of the International Skin Imaging Collaboration (ISIC) who have successfully released thousands of high-quality images to the public. Through the *ISIC-archive*[2] *images are publicly available along with a diagnosis, metadata and segmentation masks. In the most recent ISIC challenge*[11] an ensemble model combining several of the most accurate neural networks was able to outperform dermatologists on binary classification tasks.[23]

2.3 | Modality fusion

Previous works demonstrate the ability of neural networks to leverage additional data by integrating multiple modalities into a general framework.[11,27,28] Furthermore, each modality does not need to belong to the same domain; fusion has been explored across different image domains[28] as well as across textual domains representing semantic information[29] and metadata,[11,13] The integration of these modalities can occur at different stages in a pipeline depending on what information is present. For example, pixel-level image fusion

[1]Actinic Keratosis, Basal Cell Carcinoma, Dermatofibroma, Hemangioma, Intraepithelial Carcinoma (ie, Bowen's disease), Lentigo, Melanoma, Nevus, Pyogenic Granuloma, Squamous Cell Carcinoma, Seborrhoeic Keratosis and Wart.

[2]https://isic-archive.com.

has been widely explored in radiology whereby multiple registered images containing different signals are combined prior to being introduced to the model.[30,31] Late stage feature fusion techniques such as bilinear-gating[32] show slight improvements over more basic techniques such as max-pooling.

3 | METHODS

3.1 | Data set

The original histopathologic diagnoses of images found in our data set spanned many fine-grained classes and were aggregated into a higher level disease class through manual inspection by a dermatologist; we only used disease classes with more than 100 cases. Only cases that contained metadata, a macroscopic image, a dermatoscopic image and a histopathological diagnosis were retained. Notably, the cases found in our data set are inherently challenging; all cases include a histopathological diagnosis which indicates that, after a physical examination by an expert dermatologist using dermatoscopy, excision was believed to be necessary in order to confirm a diagnosis. Through repeated manual screening of all images, we only selected cases where images are of sufficient quality and free of any identifiable features (ie, eyes, multiple facial landmarks, jewellery or parts of a garment). We did this in an attempt to remove any possible biases from the data set; as for example, basal cell carcinomas (BCCs) may be more commonly found on the nose and therefore the network may learn to predict BCC if a nose is visible. The final data set is composed of 2917 cases, from five classes (naevus, melanoma, basal cell carcinoma, squamous cell carcinoma, pigmented benign keratoses (bkl)), see the Appendix S1 for more information on the diagnosis included in these classes.

3.2 | Network architecture

To obtain image features we used a modified ResNet-50 architecture.[14] The softmax and 1000 dimensional fully connected layer were removed from the end of the network, and the flattened output from the average pooling layer was used as our 2048-dimensional image feature vector. We refer to this as our *image feature extraction network*. Transfer learning was used to combat common problems, such as over-fitting, which come along with having a relatively small data set with which to train a neural network. We initialized the weights of the image feature extraction network from a model that had been pretrained for the task of 1000-way classification on ILSVRC 2015.[33] In order to leverage data from different modalities we chose to perform late fusion using an *embedding network* composed of two 1024-dimensional fully connected layers with a ReLU activation function[27] and a 5-way softmax layer. See Appendix S1 for more details on deciding the depth and width of the layers in the embedding network and parameters used to train the network.

The network architecture changed slightly depending on the modalities being used. Parts of the network were omitted if a given modality was not being used in an experiment. Figure 1 shows a diagram of the complete network architecture used when all modalities were present.

3.3 | Full multimodality classification

When all three modalities (macroscopic image, dermatoscopic image and metadata) were present we created a network that is composed of two towers of the image feature extraction network, one for dermatoscopic images and one for macroscopic images. In our experiments, we observed that letting each tower learn its own set of parameters as opposed to sharing weights between them led to better performance. To perform multimodality classification, we use a late fusion technique[34] in our network: after each image was sent through its respective feature extraction tower the image feature vectors were concatenated together with the metadata feature vector and sent through the embedding network.

3.4 | Partial multimodality classification

In the case where only one image modality (either dermatoscopic or macroscopic) and metadata was present for classification we omitted the other tower from the full network. We therefore calculated only a single image feature vector and concatenated it with the metadata feature vector before sending it through the embedding network.

3.5 | Single image classification

When only one image modality was present for classification without metadata, the image was simply passed through our image feature extraction network and the image feature vector was then sent through the embedding network. In our experiments, we found that the addition of the embedding network for single image classification achieved similar results compared to a standard ResNet-50 network.

3.6 | Evaluation metrics

Achieving a high classification performance on all classes is desirable, but correctly predicting all skin malignancies, especially tumors with a high mortality rate (ie, melanoma), is much more important than incorrectly predicting a benign lesion. In an effort to address both paradigms we report mean average precision (*mAP*), Top-1 Accuracy (*Top-1 Acc*) as well as the area under the ROC curve for detecting melanoma ($AUC_{Melanoma}$) or any kind of skin cancer (AUC_{Cancer}).

4 | RESULTS

4.1 | Metadata only classification

To evaluate the performance of metadata without any image information we trained a random forest classifier to predict the diagnosis of a single lesion based on age, gender and body location. We chose random forests for their desirable property of calculating

TABLE 1 Performance of different modality combinations on a held-out test set. Values depict mean (standard deviation) from fivefold cross-validation. Values in bold show the best modality combination for every network and metric.

	Top-1 Acc	*mAP*	$AUC_{Melanoma}$	AUC_{Cancer}
Random forest classifier				
meta	0.544 (0.006)	0.402 (0.005)	0.634 (0.010)	0.810 (0.004)
CNN – no embedding network				
macro	0.647 (0.016)	0.598 (0.009)	0.784 (0.005)	0.858 (0.007)
macro + meta	0.645 (0.009)	0.603 (0.012)	0.794 (0.011)	0.862 (0.004)
dsc	0.705 (0.013)	0.682 (0.015)	0.830 (0.010)	0.870 (0.007)
dsc + meta	0.700 (0.008)	0.672 (0.008)	0.832 (0.005)	0.871 (0.004)
dsc + macro	0.716 (0.012)	**0.720** (0.011)	0.846 (0.007)	**0.888** (0.005)
dsc + macro + meta	**0.719** (0.011)	0.714 (0.007)	**0.849** (0.010)	0.881 (0.004)
CNN – with embedding network				
macro	0.647 (0.010)	0.598 (0.005)	0.791 (0.009)	0.854 (0.004)
macro + meta	0.652 (0.005)	0.604 (0.009)	0.787 (0.007)	0.859 (0.004)
dsc	0.707 (0.010)	0.669 (0.010)	0.831 (0.004)	0.871 (0.004)
dsc + meta	0.701 (0.011)	0.691 (0.014)	0.840 (0.008)	0.872 (0.005)
dsc + macro	**0.721** (0.007)	0.726 (0.012)	**0.866** (0.006)	**0.888** (0.005)
dsc + macro + meta	0.720 (0.007)	**0.729** (0.009)	0.861 (0.006)	**0.888** (0.002)

feature importance. After searching for optimal parameters through an extensive fivefold cross-validation grid search, the random forest model resulted in a *mAP* on the test set of 0.402. Feature importance inspection revealed that *age* and the *head/neck/face* location were the most influential features in metadata-only prediction. We also tried using only the embedding network for metadata classification however it resulted in a slightly lower *mAP* of 0.391.

4.2 | Single image classification

To ensure that our single image tower model has a comparable performance to existing state-of-the-art models we fine-tuned it on the ISIC 2017 classification challenge training data. We compare results to the "ISIC 2017 Part 3: Lesion Classification - Final test submission" leader board.[3] After training, we reached an average AUC of 0.858, placing the base model among the top 30% of the ISIC 2017 ranked submissions.[11] From this, we infer that the single-image performance of the network trained on our data set in the following experiments reflects a competitive performance in automated dermatoscopic skin lesion classification. However, we note that the focus of this paper was not to achieve the best possible single-image classification performance.

In order to ensure that the addition of the embedding layers have no detrimental effect on the overall performance, we repeated experiments for all modality combinations by directly classifying the images without using an embedding network. This would be equivalent to a standard ResNet-50 architecture modified for 5-way classification on a single image. Results in Table 1 show consistently higher performance across almost all metrics when an embedding network is used.

4.3 | Multimodal network performance

Results show that combining dermatoscopic with macroscopic images can increase the accuracy of skin lesion classification with a summary of experimental results shown in Table 1. Looking at the embedding network results, distinguishing melanoma from non-melanoma ($AUC_{Melanoma}$) improves from 0.831 (*dsc*) to 0.866 (*dsc + macro*), Figure 2). There is, however, a slight decrease in performance once metadata was added to 0.861 (*dsc + macro + meta*). Combining dermatoscopic and macroscopic images also improves performance for multiclass classification with an increase in *mAP* from 0.669 (*dsc*) to 0.726 (*dsc + macro*, see Figure 3). The addition of metadata slightly boosted the *mAP* to 0.729 (*dsc + macro + meta*).

Confusion matrices shown in Figure 4 reveal that the addition of the macroscopic modality (*dsc + macro*) primarily helps to classify squamous cell carcinomas. In both cases where metadata is added (*dsc + meta*, *dsc + macro + meta*) we can see that it helps to classify basal cell carcinomas but also begins to misclassify more squamous cell carcinoma as basal cell carcinoma.

5 | DISCUSSION

Multimodal image analysis is a common technique employed across the domain of radiology images where it can often be translated into a channel-wise fusion technique thanks to the registered nature of the images. In contrast, our modalities are distinct to the extent that no image registration readily exists; therefore we opt to combine modalities in some common latent space.

Previously, Binder et al[35] combined age, body site, naevus count, proportion of dysplastic nevi, personal history and family history of melanoma with a neural network-based classifier. By using this metadata they increased their AUC for distinguishing nevi from melanoma from 0.942 to 0.968. In our experiments, the inclusion of the metadata fields age, location and sex did not significantly improve accuracy for pigmented skin lesions (Figure 4). Thus, we infer that the other clinical attributes which are known to indicate higher melanoma risk (naevus count, proportion of

[3]https://challenge.kitware.com/#phase/584b0afccad3a51cc66c8e38.

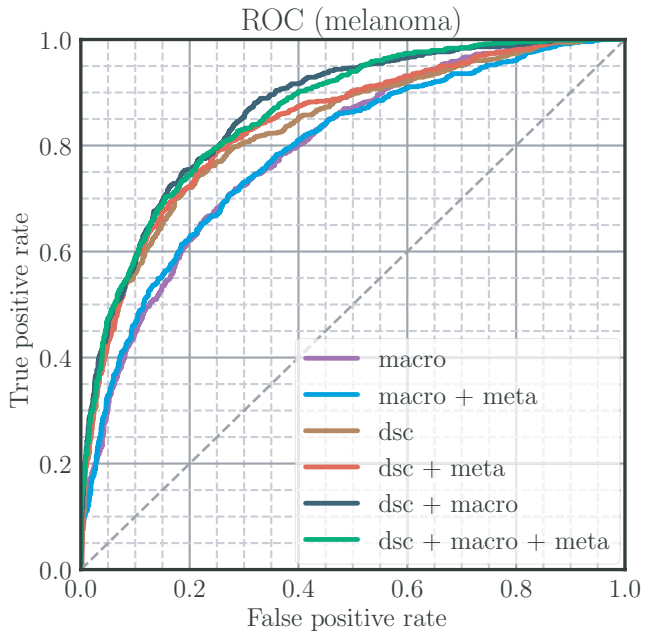

FIGURE 2 ROC curves for all image modality combinations for distinguishing melanoma from non-melanoma on the test set using embedding network

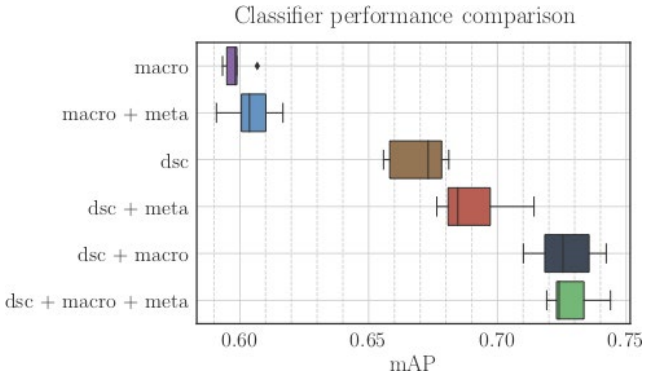

FIGURE 3 Mean average precision (mAP) of different modality combinations for all five disease classes. Box plots include results of fivefold cross-validation with each trained model evaluated on the same held-out test set

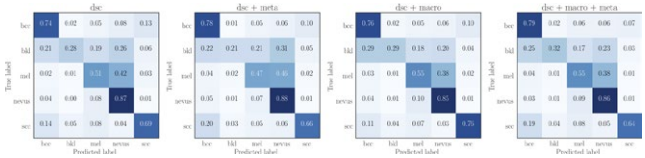

FIGURE 4 Class confusion matrix normalized by the ground truth label from models trained on various combinations of modalities

atypical nevi and history of melanoma) are more informative to help distinguish nevi from melanoma. As this clinical information was not found in the data set available to us, we could not estimate whether it has value in helping to distinguish between diagnoses in a multi-class problem.

Kharazmi et al[13] added age, gender, location as well as lesion size and elevation to features extracted from a sparse autoencoder in a binary bcc vs non-bcc classification task. The accuracy of detecting bcc was improved from 0.847 to 0.911, clinical metadata alone achieved an accuracy of 0.756. The location metadata values "head/neck/face" and "age", being markers for sun- damaged skin, seem to be important factors when trying to solve a binary decision of "bcc versus benign non-bcc diagnoses", for our more challenging 5-way classification we suspect it becomes less informative. The authors in Ref.[13] describe their fusion step as integration of clinical information to a "feature set" before the softmax layer. If we infer this to be a simple concatenation of feature vectors, this method is similar to our setup without the added fully connected embedding layers (corresponding to "No Embedding" in Figure S1). While we cannot compare our system directly, Table S2 shows results of our proposed architecture when restricting our test data set and predictions to the diagnoses that were mentioned by Ref.[13] Interestingly, diagnosing using only clinical metadata results in very similar performance. However, since our dermatoscopic single-modality network performs better than their sparse autoencoder, we suspect there is little value adding clinical metadata to increase accuracy further.

Finally, Ge et al[36] described the automated analysis of clinical and dermatoscopic images by using the average output of two separate networks, siamese networks, or training of a third network of fused feature maps. They achieved up to 8% increase in accuracy at a multi-task problem, with their data set incorporating expert-labelled data without pathologically verified cases. Lack of description where feature maps originate from, which network architecture was used, how fusion was performed exactly and use of proprietary data impedes making direct comparisons to their result.

For single-image classification the test set predictions showed that the automated classification of dermatoscopic images yielded consistently higher performance when compared against macroscopic images; similar to what is known from previous work with human participants.[2,16] Even more important for clinical application, dermatoscopic images may incorporate less variability than macroscopic images since size, lighting and distance are generally restricted at the hardware level.

It is worth noting, for comparison with other studies using clinical images, that we did not manually crop any of our images while creating the data set. The images used in our data set were captured centred on the lesion so instead we chose to automatically crop the largest square from the centre of our images. Previous work from Han et al[15] also report that they manually cropped every image in their data set to ensure that the lesion in question is centred in the image. Their public model (https://modelderm.com/) also gives strict definitions on the amount of covered area that the lesion must occupy on the image, and gives a suggested imaging distance to the lesion. We hypothesize that when using clinical images, especially within the context of a small data set, image composition must fit to strict requirements in order for neural networks to be applied for analysis. We therefore conclude that current state of the art models are heavily dependent on restrictive image specifications and may

be unstable when used in clinical practice. Even more so, selective data sets may include unwanted biases for certain patterns (eg, images from BCCs are commonly from the nose, therefore a network may learn to predict BCC if a nose is visible). Verification of these problems and solving them falls under the scope of future work.

From the confusion matrices shown in Figure 4, we hypothesize that the increase in performance on classifying squamous cell carcinomas when macroscopic images are combined with dermatoscopic could be explained by these tumors showing scale or keratin plugs on the surface—features which may be rendered less visible by using dermatoscopy with immersion fluid (a common technique used throughout our data set). The positive effect of adding macroscopic images was less evident for basal cell carcinomas and melanomas, possibly because these tumors already show a distinctive dermatoscopic appearance to which other imaging modalities are unable to add significant information. The addition of macroscopic images slightly increased the accuracy of seborrhoeic keratoses (*bkl*); however, the sample size of this group is very small rendering a final conclusion for this group impossible.

6 | LIMITATIONS

Since we restrict cases to those which are histopathologically verified, results for both benign groups (nevi and bkl) have to be interpreted with caution. About 15% of nevi were falsely labelled as a malignant class by our dsc + macro model, outwardly suggesting a myriad of unnecessary excisions. In reality though, 100% of those benign cases were excised in clinical practice due to sufficient suspicion from an expert physician.

By this, our study suffers also from the common verification bias in dermatoscopic studies with only pathologically diagnosed cases included. Having these reliable labels is in part necessary for machine learning, but makes all resulting models biased to this case distribution. Thus, in daily practice such models may not only fail for benign lesions that are not present in current training sets but we also do not know when it is misinterpreting a case. We probably also cannot preselect cases that are more representative of the training data, because if we already knew what would be a lesion that needs histopathologic verification, we would not need a decision support after all.

We suggest that future studies need to integrate more benign non-excised skin lesions including seborrhoeic keratoses, nevi, angiomas and dermatofibromas, and further stratify them based on suspicion; however, available data for this is scarce due possibly to a relatively low incentive of physicians to document this information.

CONFLICT OF INTEREST

Jordan Yap and William Yolland are employees of MetaOptima Technology Inc. MetaOptima provided and unrestricted research grant to Philipp Tschandl to for conducting a one-year post-doc fellowship at Simon Fraser University.

AUTHOR CONTRIBUTIONS

JY, WY and PT each contributed substantially to experimental design, result analysis and manuscript preparation.

ORCID

Jordan Yap http://orcid.org/0000-0001-6391-5302

REFERENCES

[1] H. Kittler, *Arch. Dermatol.* **2008**, *144*, 533.
[2] C. Sinz, P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, J. Lapins, A. A. Marghoob, S. W. Menzies, J. Paoli, H. S. Rabinovitz, C. Rinner, A. Scope, H. P. Soyer, L. Thomas, L. Zalaudek, H. Kittler, *J. Am. Acad. Dermatol.* **2017**, *77*, 1100.
[3] M. Claeson, M. Gillstedt, D. C. Whiteman, J. Paoli, *Acta Derm. Venereol.* **2017**, *97*, 1206.
[4] C. A. Clarke, M. McKinley, S. Hurley, R. W. Haile, S. L. Glaser, T. H. M. Keegan, S. M. Swetter, *J. Invest. Dermatol.* **2017**, *137*, 2282.
[5] J. H. Lee, Y. H. Kim, K. D. Han, Y. M. Park, J. Y. Lee, Y. G. Park, Y. B. Lee, *Acta Derm. Venereol.* **2018**, *98*, 382.
[6] T. S. Housman, S. R. Feldman, P. M. Williford, A. B. Fleischer, N. D. Goldman, J. M. Acostamadiedo, G. J. Chen, *J. Am. Acad. Dermatol.* **2003**, *48*, 425.
[7] M. Migden, J. Xie, J. Wei, W. Tang, V. Herrera, J. B. Palmer, *J. Am. Acad. Dermatol.* **2017**, *77*(1), 55.
[8] E. Tan, A. Yung, M. Jameson, A. Oakley, M. Rademaker, *Br. J. Dermatol.* **2010**, *162*, 803.
[9] A. Borve, J. DahlenGyllencreutz, K. Terstappen, E. JohanssonBackman, A. Aldenbratt, M. Danielsson, M. Gillstedt, C. Sandberg, J. Paoli, *Acta Derm. Venereol.* **2015**, *95*(2), 186-190.
[10] M. Binder, A. Steiner, M. Schwarz, S. Knollmayer, K. Wolff, H. Pehamberger, *Br. J. Dermatol.* **1994**, *130*, 460.
[11] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection: A challenge at the, international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC), **2017**, CoRR, vol. abs/1710.05006.
[12] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, *Nature* **2017**, *542*, 115.
[13] P. Kharazmi, S. Kalia, H. Lui, Z. J. Wang, T. K. Lee, *Skin Res. Technol.* **2017**, *24*, 256.
[14] K. He , X. Zhang , S. Ren , J. Sun , in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June **2016**, pp. 770-778.
[15] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, S. E. Chang, *J. Invest. Dermatol.* **2018**, *138*, 1529.
[16] H. Kittler, H. Pehamberger, K. Wolff, M. Binder, *Lancet Oncol.* **2002**, *3*, 159.
[17] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, R. H. Moss, *Comput. Med. Imaging Graph.* **2007**, *31*, 362.
[18] B. Shrestha, J. Bishop, K. Kam, X. Chen, R. H. Moss, W. V. Stoecker, S. Umbaugh, R. J. Stanley, M. E. Celebi, A. A. Marghoob, G. Argenziano, H. P. Soyer, *Skin Res. Technol.* **2010**, *16*(1), 60.
[19] Q. Abbas, M. Emre Celebi, I. F. Garcia, W. Ahmad, *Skin Res. Technol.* **2013**, *19*(1), 93.
[20] M. Sadeghi, M. Razmara, M. Ester, T. K. Lee, M. S. Atkins, "Graph-based Pigment Network Detection in Skin Images", SPIE Medical Imaging Conference, Town and Country Resort and Convention Center, San Diego, California, USA, 13 - 18 February 2010.

[21] M. Sadeghi , M. Razmara , P. Wighton , T. K. Lee , M. S. Atkins , in Medical Imaging and Augmented Reality. Springer, Berlin Heidelberg, **2010**, pp. 467-474.
[22] A. Dalal, R. H. Moss, R. J. Stanley, W. V. Stoecker, K. Gupta, D. A. Calcara, J. Xu, B. Shrestha, R. Drugge, J. M. Malters, L. A. Perry, *Comput. Med. Imaging Graph.* **2011**, *35*, 148.
[23] A. Menegola, J. Tavares, M. Fornaciali, L. Tzy Li, S. Avila, E. Valle, Imagenet classification with deep convolutional neural networks, RECOD Titans at ISIC Challenge 2017, ArXiv e-prints. **2017**.
[24] A. Krizhevsky , I. Sutskever , G. E. Hinton , in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, ser. NIPS'12, **2012**, pp. 1097-1105.
[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR, vol. abs/1409.1556, **2014**.
[26] J. Deng , W. Dong , R. Socher , L.-J. Li , K. Li , L. Fei-Fei , in CVPR09, **2009**.
[27] T. Xu, H. Zhang, X. Huang, S. Zhang, D. N. Metaxas, Multimodal Deep Learning for Cervical Dysplasia Diagnosis, Springer International Publishing, Cham **2016**, Ch.115-123.
[28] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, D. Shen, *NeuroImage* **2015**, *108*, 214.
[29] J. C. Caicedo, J. A. Vanegas, F. Pez, F. A. Gonzlez, *J. Biomed. Inform.* **2014**, *51*, 114.
[30] J. Du, W. Li, K. Lu, B. Xiao, *Neurocomputing* **2016**, *215*, 3.
[31] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, *Inform. Fusion* **2017**, *33*, 100.
[32] D. Kiela, E. Grave, A. Joulin, T. Mikolov, Efficient Large-Scale Multi-Modal Classification, ArXiv e-prints, **2018**.
[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, *Int. J. Comput. Vis.* **2015**, *115*, 211.
[34] D. Ramachandram, G. W. Taylor, *IEEE Signal Process. Magaz.* **2017**, *34*(6), 96.
[35] M. Binder, H. Kittler, S. Dreiseitl, H. Ganster, K. Wolff, H. Pehamberger, *Melanoma Res.* **2000**, *10*, 556.
[36] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, R. Garnavi, Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. in: Medical Image Computing and Computer-Assisted Intervention MICCAI 2017, (Eds: M Descoteaux, L Maier-Hein, A Franz, P Jannin, DL Collins, S Duchesne), Springer International Publishing, Cham **2017**, Ch.250-258.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Grid search of embedding dimensions with final validation loss and mAP of the resulting model shown. Points and whiskers depict mean and 95%-confidence intervals of all fivefold cross-validation results (fivefold cross-validation was performed on the following modality combinations then averaged: dsc + meta, dsc + macro and dsc + macro + meta).

Table S1 Dataset diagnosis distribution. SCC – Squamous Cell Carcinoma, KA – Keratoacanthoma, LPLK – Lichen planus-like keratosis (a seborrheic keratosis or solar lentigo in regression). **Table S2** Diagnostic values for BCC-detection of our method in comparison to Kharazmi et al.[13]

Appendix S1 Supplementary information including network implementation and dataset details.

How to cite this article: Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol.* 2018;27:1261–1267. https://doi.org/10.1111/exd.13777