FUSING METADATA AND DERMOSCOPY IMAGES FOR SKIN DISEASE DIAGNOSIS

Weipeng Li[1,2,*] *Jiaxin Zhuang*[1,2,*] *Ruixuan Wang*[1,2] *Jianguo Zhang*[3] *Wei-Shi Zheng*[1,2]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
[3] Department of Computer Science and Engineering,
Southern University of Science and Technology, China

ABSTRACT

To date, it is still difficult and challenging to automatically classify dermoscopy images. Although the state-of-the-art convolutional networks were applied to solve the classification problem and achieved overall decent prediction results, there is still room for performance improvement, especially for rare disease categories. Considering that human dermatologists often make use of other information (e.g., body locations of skin lesions) to help diagnose, we propose using both dermoscopy images and non-image metadata for intelligent diagnosis of skin diseases. Specifically, the metadata information is innovatively applied to control the importance of different types of visual information during diagnosis. Comprehensive experiments with various deep learning model architectures demonstrated the superior performance of the proposed fusion approach especially for relatively rare diseases. All our codes will be made publicly available[1].

***Index Terms*—** Skin disease classification, metadata, data fusion.

1. INTRODUCTION

Nowadays, skin cancer is one of the most common cancers, with over 5,000,000 patients in the United States [1]. Among skin cancers, melanoma is the most serious form and causes most of skin cancer deaths. While early detection and diagnosis of melanoma can largely increase the survival of patients, the diagnosis accuracy of melanoma from expert's visual inspection is only about 60% [2]. In this case, the state-of-the-art Artificial Intelligence(AI) techniques could potentially help clinicians more accurately diagnose skin cancers.

Deep learning, one stream of the most successful AI techniques, has been rapidly developed and applied to various scenarios, such as face identification [3], surveillance [4], and healthcare [5], etc. In particular, the convolutional neural networks (CNNs) have shown excellent performance in classifying images and achieved human-level performance in various medical image diagnosis tasks [6, 7, 8]. However, these deep models (such as AlextNet [9] and ResNet [10]) often requires large training data in order to perform well for relevant tasks. This has impeded the application of deep learning to small-sample or data-imbalance tasks where training data at least for part of classes are difficult to collect. In this paper, we are interested in the intelligent diagnosis of skin diseases where the number of available images for model training are very different between classes. One way to alleviate this data-imbalance issue is to transfer the pretrained deep learning models based on a large dataset and finetune it based on task-specific small dataset. Such transfer learning technique recently has shown to perform well in many domains, including medical image analysis tasks [11]. Another popular way is to set higher penalty coefficients for small-sample classes in the loss function, which can improve the cost of mis-classifying each training example coming from small-sample classes and therefore force the classifier pay enough attention to the classification of images from small-sample classes during learning [12].

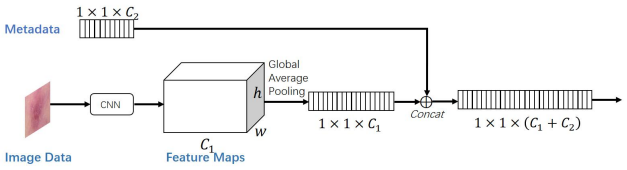
Different from these existing approaches, this paper proposes using metadata to help improve the performance of intelligent skin disease diagnosis. This is inspired partially by the finding that different skin diseases may often appear at different body parts [13], such that such kind of non-image metadata could help differentiate one disease from another. While existing multi-modality medical image classification tasks have been explored recently [14], there exists few work in fusing image and non-image metadata for the medical image classifications, although the general data-fusion ideas have been applied to natural image classifications, e.g., by concatenating metadata information and the extracted visual features from a CNN [15]. In this paper, we applied a novel data-fusion framework to the data-imbalance skin image classification task, and experiments showed that the metadata can effectively improve the recall rate of the skin diseases particularly with smaller training data.

*These authors contributed equally to this work.
[1] https://github.com/fatetail/MetaNet

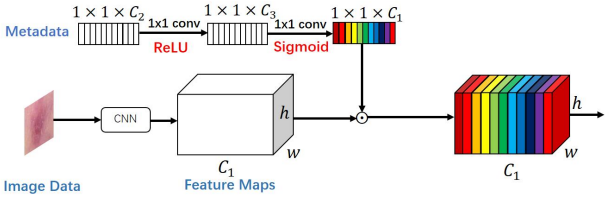
2. MULTIPLICATION-BASED DATA FUSION

In this section, we introduce a new framework to fuse metadata and image data for skin image classification. The metadata used here includes the body location of each image, the gender and the age of the patient. Suppose the metadata information can be represented by a one-dimensional vector, one general data-fusion method is to directly concatenate the feature vector of the metadata and the visual feature vector extracted from the last layer of a CNN (Figure 1(a)), followed by one or more fully connected layers. Such concatenation does not directly consider potential effect of the metadata information on the visual feature extraction process. For example, if one specific body location is related to just one or some of skin diseases, such metadata information could be used to directly suppress the prediction probability of all the other irrelevant diseases. Simply direct concatenation of metadata and visual features would not directly capture such kind of metadata effect.

Different from such concatenation-based data fusion, we propose a multiplication-based data fusion to make the metadata directly interact with the visual features. Inspired by the Squeeze-and-Excitation operation in the SENet network [16], we propose using the metadata to control the importance of each feature channel at the last convolutional layer, expecting that the network would be able to focus on specific part of the feature channels based on specific metadata information. Figure 1(b) demonstrates the data fusion process. Specifically, the metadata feature vector is fed into a two-layer fully connected sub-network, with first layer (i.e., a 1×1 convolution) followed by a rectified linear unit (ReLU) and second layer followed by a Sigmoid activation function, and the output size of the sub-network is the same as that of the feature channels at the last convolutional layer. Then, each feature map in the output of the last convolutional layer is weighted (i.e., multiplicated) by one corresponding vector element in the sub-network output, resulting in the re-weighted feature maps. Obviously, such data-fusion operation can be embedded into any CNN model, such as AlexNet [9], VGGNet [17], ResNet [10], DenseNet [18] and even the SENet [16].



(a) Conventional concatenation-based data fusion



(b) Proposed multiplication-based data fusion

Fig. 1. The proposed multiplication-based data fusion can make the metadata directly control the importance of each feature channel, helping the network focus on more discriminative channels, while the conventional concatenation-based method may not.

3. EXPERIMENT

3.1. Experimental settings

The dataset was from the ISIC'2018 challenge on skin lesion diagnosis[2], containing totally 10015 images for seven skin diseases, namely Melanocytic Nevi (NV, 6705 images), Melanoma (MEL, 1113 images), Benign Keratosis (BKL, 1098 images), Basal Cell Carcinoma (BCC, 514 images), Actinic Keratosis / Bowens disease (AKIEC, 327 images), Vascular Lesion (VASC, 143 images), and Dermatofibroma (DF, 115 images). Along with each image, three types of metadata information were also provided, including the patient's age, gender, and body part location of the image. Considering that skin images were captured with different background illumination and normal skin image regions are often in various colors, we applied the color constancy algorithm [19] to normalize original skin images. Then, the images (originally around 600×450 pixels) are resized to 300×300 pixels when the backbone CNN architecture is AlexNet [9], VGG [17], ResNet [10], DenseNet [20], or SENet [16], and to 441×441 pixels for PNASNet [18]. During model training, the image dataset was augmented by multiple transformations, including randomly cropping (patch size 331×331 pixels for PNASNet, and 224×224 pixels for all other CNN backbones), randomly flipping horizontally and vertically, randomly changing image brightness, contrast, saturation, and random rotation within certain ranges. All backbone CNN models are pretrained on ImageNet Dataset.

During the training of each model, a class-weighted loss function [21] was used to (partially) alleviate the data imbalance issue. Stochastic gradient descent (SGD) with a mini-batch size of 32 was used, with the learning rate starting from 0.001 and then divided by 10 at the 50^{th}, 120^{th}, and 200^{th} epoch respectively. Weight decay (coefficient 0.0001) and a momentum of 0.9 was also applied. Each model was trained for up to 250 epochs, with the consistent observation of training convergence within 250 epochs. Considering the imbalanced sample size across classes, we used 5-fold cross-validation to evaluate the performance of each model. More specifically, the average recall over the seven categories

[2]https://challenge2018.isic-archive.com/task3

(i.e., *mean class recall*, MCR) and its standard deviation over 5-fold cross-validation sets were reported for each model.

3.2. Metadata preprocessing

To quantitatively represent metadata, the gender of each patient was encoded as a two-dimensional one-hot vector, and the body location (totally 14 possible options) of each image was encoded by a 14-dimensional one-hot vector. The age of each patient was normalized to the range $[0, 1]$. The three types of metadata information were concatenated to form a 17-dimensional feature vector for each corresponding image. Because one or two types of metadata information were originally missing for a small number (around 50-100 per type of metadata) of images, the average of the provided age values was used to fill the missing age information, while the most frequent gender and body location from the provided metadata were used to respectively fill the missing gender and location information.

3.3. Comparison with baseline approaches

To evaluate the effectiveness of the proposed multiplication-based fusion approach, we compared our approach with two baseline approaches on multiple CNN backbone architectures. One baseline approach is just based on the image data without using the metadata, and the other baseline is the concatenation-based fusion approach. Table 1 shows that the traditional concatenation-based fusion approach is generally better than the basic approach without using metadata (except on the PNASNet backbone), and the proposed multiplication-based fusion approach (last column in table) significantly outperforms both baseline approaches (p-values are smaller than 0.05 with Wilcoxon's Sign Rank Test), whatever the backbone CNN architecture is. This strongly supports that the proposed multiplication-based fusion approach is more effective than the traditional concatenation-based fusion approach in improving the classification performance.

Table 1. Comparison of the proposed multiplication-based fusion approach with two baseline approaches on multiple backbone CNN architectures.

Backbones	No metadata	Concatenation-based	Ours
AlexNet	74.68 ± 0.92	76.55 ± 1.25	$\mathbf{78.26}\pm1.55$
VGG19	81.60 ± 1.67	82.35 ± 1.68	$\mathbf{84.06}\pm1.16$
ResNet50	82.50 ± 1.31	82.98 ± 1.35	$\mathbf{84.02}\pm1.50$
DenseNet161	84.59 ± 1.42	85.85 ± 0.92	$\mathbf{87.03}\pm1.40$
SENet154	85.44 ± 1.09	86.46 ± 0.69	$\mathbf{87.64}\pm0.52$
PNASNet-5	87.90 ± 1.32	87.25 ± 0.73	$\mathbf{89.09}\pm0.67$

To further inspect the detailed effect of the proposed fusion approach, the classification performance for each individual disease class was also reported with the SENet154 backbone in Table 2. Interestingly, compared to the basic approach without using metadata, the performance of the proposed approach increases on the two smallest disease classes (DF and VASC), with the largest performance boost from the smallest (rare) disease (DF), while the performance of the proposed approach decreases on the two largest disease classes (NV and MEL). This suggests that the appropriate use of metadata as in the proposed approach may help partially alleviate the data imbalance issue. One possible reason is that the rare disease is more location-specific such that the location metadata can largely help the classifier differentiate the rare disease from others.

Table 2. Classification performance of each approach on each disease class. The backbone CNN architecture is SENet154.

Diseases	Baseline	Fusion-network	Meta-network
NV (6705)	95.31 ± 1.35	$\mathbf{95.38}\pm1.68$	93.42 ± 1.19
MEL (1113)	$\mathbf{84.24}\pm0.79$	76.54 ± 1.25	78.26 ± 0.73
BKL (1098)	81.31 ± 0.45	84.47 ± 0.69	$\mathbf{85.64}\pm1.03$
BCC (514)	90.58 ± 1.25	91.99 ± 1.35	$\mathbf{92.02}\pm1.38$
AKIEC (327)	$\mathbf{83.45} \pm1.23$	80.84 ± 0.92	80.23 ± 1.20
VASC (143)	99.23 ± 0.86	99.22 ± 0.73	$\mathbf{99.36}\pm0.67$
DF (115)	63.56 ± 1.32	76.78 ± 0.73	$\mathbf{84.55}\pm0.47$

3.4. Effect of metadata elements

To explore the effect of each metadata information and their combinations on the classification performance, an ablation study of the metadata is performed by using each individual metadata and each possible combination of two metadata information respectively for the proposed approach. Interestingly, Table 3 shows that while the individual or combined *age* and *location* metadata can help improve the classification performance, the *gender* metadata actually degraded the classification performance when used individually. The combination of *gender* with any other metadata also slightly degraded the classification performance compared to that without using the *gender*. Actually, the best performance of the proposed approach was obtained when just using *age* and *location* metadata together. Such observation strongly suggests that *age* and *location* are related to skin diseases, while *gender* is not, i.e., these seven skin diseases are neither male- nor female-preferred.

3.5. Effect of fully connected layers

The above tests are based on two fully connected layers for the metadata transformation in the proposed fusion approach. Here we also evaluate the effect of the number of fully connected layers on the classification performance. As Table 4 (first row) shows, too few (i.e., one) or too many (i.e., four) fully connected layers would make the classifier perform relatively worse than that with two or three fully connected lay-

Table 3. Ablation study of metadata effect with the SENet154 as backbone for the proposed fusion approach. A tick means the corresponding metadata is used during model training.

Age	-	✓	-	-	✓	-	✓	✓
Gender	-	-	✓	-	✓	✓	-	✓
Localization	-	-	-	✓	-	✓	✓	✓
MCR	85.44 (1.09)	85.84 (0.85)	84.04 (0.05)	87.06 (0.20)	84.16 (0.71)	85.91 (0.58)	87.70 (0.99)	87.64 (0.52)

ers. Since the performance of the proposed approach with one fully connected layer for metadata transformation performs actually similar to the traditional fusion approach in which the metadata is directly concatenated with the visual features without any fully connected layer, here we also performed a comparative study by respectively inserting one to four fully connected layers to transform the metadata information before concatenation in the traditional approach (Table 4, last row). In this comparative study, the output size (i.e., 2048) of the last fully connected layer is the same as that of the feature vector from the last convolutional layer, while the output size of any other possible hidden fully connected layer(s) was set the same size as that for the corresponding proposed approach (e.g., for two fully connected layers, the output size of the hidden layer is 1024, while for three fully connected layers, the output sizes of two hidden layers are 512 and 1024 respectively). Table 4 (last row) shows that the performance of the traditional approach performs slightly better with two or three fully connected layer than with one or four layers. However, the best performance with two or three layers are from the proposed approach, supporting that the reported superior performance of the proposed approach above is not due to the two-layer fully connected layers for metadata transformation.

Table 4. Effect of number of fully connected layers for metadata transformation, with SENet154 as the backbone.

Approaches	One layer	Two layers	Three layers	Four layers
Ours	86.47 (0.53)	87.64 (0.52)	**87.71** (0.87)	86.21 (0.70)
Concatenation-based	86.54 (0.74)	86.81 (0.14)	**86.88** (0.54)	85.53 (0.67)

3.6. Qualitative analysis

Here we also provide representative qualitative analysis to further demonstrate the effect of metadata in the proposed approach. The classification activation map (CAM) method was used to show the attended image region based on the last set of feature maps when predicting the disease for any specific image [22]. Figure 2 shows that when both the baseline approach without using metadata and the proposed fusion approach predicted correctly (first row), the attended regions (red in the heatmaps, superimposed on the original image) are also similar between the two approaches. In comparison, when the baseline approach predicted incorrectly (second row, middle), the approach seems not focusing on the possibly key region (i.e., the bright region around the image center), while the proposed approach did with correct prediction (second row, right). These results indicate that the proposed approach may more appropriately use metadata to help the classifier focus on relevant regions during diagnosis.

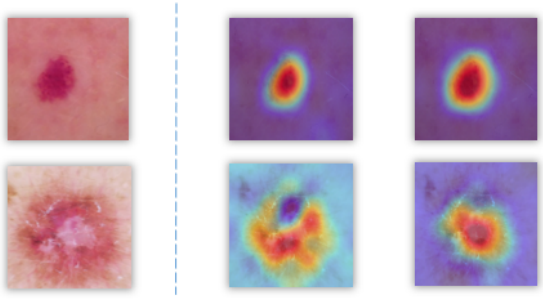


Fig. 2. Representative activation maps by the proposed approach (right) and the approach without using metadata (middle) for two input images (left), respectively.

4. CONCLUSIONS

This paper applied a novel multiplication-based fusion strategy to the intelligent diagnosis of skin disease. Experiments showed that the proposed approach can effectively improve the diagnosis performance particularly for small-sample classes. It was also observed that not all metadata helped improve diagnosis performance, indicating that effective metadata selection before data fusion may be necessary if the number of metadata types becomes large. Qualitative analysis also supported that the proposed fusion approach can help classifiers more accurately focus on lesion regions during diagnosis. Future work includes exploring effects of more types of metadata for more skin diseases.

Acknowledgement. This work is supported in part by the National Key Research and Development Plan (grant No. 2018YFC1315402), the Guangdong Key Research and Development Plan (grant No. 2019B020228001), the National Natural Science Foundation of China (grant No. U1811461) and the Guangzhou Science and Technology Program (grant No. 201904010260).

5. REFERENCES

[1] Howard W Rogers, Martin A Weinstock, Steven R Feldman, and Brett M Coldiron, "Incidence estimate of non-

melanoma skin cancer (keratinocyte carcinomas) in the us population, 2012," *JAMA Dermatology*, vol. 151, no. 10, pp. 1081–1086, 2015.

[2] Harold Kittler, H Pehamberger, K Wolff, and M Binder, "Diagnostic accuracy of dermoscopy," *The Lancet Oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[4] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011.

[5] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[6] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017.

[8] Lihteh Wu, Priscilla Fernandez-Loaiza, Johanna Sauma, Erick Hernández-Bogantes, and Marissé Masis, "Classification of diabetic retinopathy and diabetic macular edema.," *World Journal of Diabetes*, vol. 4 6, pp. 290–4, 2013.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[11] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[12] Jiaxin Zhuang, Jiabin Cai, Ruixuan Wang, Jianguo Zhang, and Weishi Zheng, "Care: Class attention to regions of lesion for classification on imbalanced data," in *MIDL*, 2019.

[13] Alireza Firooz, Bardia Sadr, Shahab Babakoohi, Maryam Sarraf-Yazdy, Ferial Fanian, Ali Kazerouni-Timsar, Mansour Nassiri-Kashani, Mohammad Mehdi Naghizadeh, and Yahya Dowlati, "Variation of biophysical parameters of the skin with age, gender, and body region," *The Scientific World Journal*, vol. 2012, 2012.

[14] Zhiqin Zhu, Yi Chai, Hongpeng Yin, Yanxia Li, and Zhaodong Liu, "A novel dictionary learning approach for multi-modality medical image fusion," *Neurocomputing*, vol. 214, pp. 471–482, 2016.

[15] Tom Zahavy, Abhinandan Krishnan, Alessandro Magnani, and Shie Mannor, "Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce," in *AAAI*, 2018.

[16] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[18] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy, "Progressive neural architecture search," in *ECCV*, 2018.

[19] Catarina Barata, M Emre Celebi, and Jorge S Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1146–1152, 2014.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[21] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.