CLASSIFICATION OF NON-TUMOROUS FACIAL PIGMENTATION DISORDERS USING IMPROVED SMOTE AND TRANSFER LEARNING

Jiawei Peng*, Ruihan Gao*, Long Nguyen*, Yunfeng Liang§, Steven Thng¶, Zhiping Lin*
* School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
§ Interdisciplinary Graduate School, Nanyang Technological University, Singapore
¶ National Skin Center, Singapore

***Abstract*—Classification of non-tumorous facial pigmentation disorders is an important but overlooked problem. Recently, a voting-based probabilistic linear discriminant analysis (V-PLDA) method was developed to address this problem by extracting hand-craft features from a given image set of rather small size, with limited classification accuracy. In this paper, we propose an improved Synthetic Minority Over-sampling Technique (improved SMOTE) with several parameters tuned to fully utilize the available images. Moreover, transfer learning is applied to reduce the data size requirement of the deep learning model. By combining the improved SMOTE and transfer learning, a classification accuracy gain (10%) is attained compared to the state-of-the-art V-PLDA method.**

***Keywords*—improved SMOTE, facial pigmentation disorders, biomedical images analysis and classification, transfer learning**

I. INTRODUCTION

Biomedical image analysis allows rapid automatic classification and diagnosis which can accelerate medical research and clinical practices [1, 2]. In dermatology, though significant researches have been achieved in automatic classification of tumorous facial pigmentation disorders using image processing and analysis [3-5], much remains to be explored in dealing with non-tumorous facial pigmentation disorders. Unlike those of its tumorous counterpart, non-tumorous pigmentation disorders have features varying significantly in shapes, sizes and colors even within the same class [6]. Hence, methods developed for tumorous pigmentation disorders may not be directly applied to solve the classification of non-tumorous pigmentations. Moreover, it is important to develop an automatic classification process for non-tumorous pigmentation disorders as they can reveal mild health conditions which are otherwise unnoticeable [7].

Recently, a voting-based probabilistic linear discriminant analysis (V-PLDA) method was developed to address the classification of non-tumorous pigmentation disorders [6]. Proper features reflecting the color and texture are extracted and fed into the V-PLDA model. However, exhaustively attempting all combinations of features would require great human effort. The classification accuracy is also limited by the small number of accessible pigmentation images provided by the dermatologists.

In this paper, we address the classification of non-tumorous pigmentation disorders using an improved Synthetic Minority Over-Sampling Technique (improved SMOTE) in conjunction with transfer learning. SMOTE [8] is an over-sampling method which originally aims to deal with imbalanced datasets. It is improved and utilized to augment the small image set in our experiments on the classification of non-tumorous pigmentation disorders. Transfer learning can be employed to utilize deep learning for a small-size dataset as it uses pre-trained models that have been previously trained over a large dataset [9, 10]. The deep neural network can extract generic features instead of hand-craft features with the connection of numerous layers of neurons. With the combination of the improved SMOTE and transfer learning, a significant improvement in the overall classification result (10% increase in accuracy) for non-tumorous facial pigmentation is achieved for the same image set used in [6].

The contributions made in this paper are mainly in the improved SMOTE which are summarized as followed: i) Instead of interpolating two samples using a random weight between 0 and 1, the range of the weight is optimized to obtain synthesized image sets with more diversity; ii)The effect on accuracy by synthesized image sets of different sizes are compared, and the size of the augmented input image set is adjusted to improve the classification results; iii) A redundancy reduction scheme is developed such that the similarity between the generated images and the original images is assessed using Structural Similarity index (SSIM index) [11, 12], and images having higher similarity with the original dataset are deleted to avoid redundancy.

II. METHODOLOGY

A workflow of the proposed method combining the improved SMOTE and transfer learning for the classification of non-tumorous pigmentation disorders is shown in Fig.1. As described in Fig.1, the improved SMOTE is implemented to produce more suitable and effective training images. Transfer learning with the base model of Inception-ResNet-v2 [13] is applied to exploit deep learning on small dataset to make full use of the limited image set. The combination of two techniques turns out to be effective.

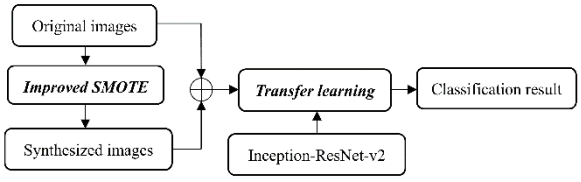

Fig. 1. Workflow of the proposed mrthod

A. SMOTE

1) Original SMOTE

SMOTE is an oversampling approach initially invented to process imbalanced datasets by enlarging the minority class to balance the training samples in each class. In this research, it is employed to augment images in every class to fulfill the requirement of deep neural networks for large dataset size.

SMOTE generates a new sample by taking a random point along the line joining two adjacent samples [8]. For every sample *s* in a class, *N* out of its *k* nearest neighbor (KNN) within the same class are randomly chosen with KNN determined using Euclidean distance. One new sample is generated by interpolating the original sample *s* and one of the chosen neighbors pixel by pixel as expressed as (1):

$$s' = s + w \cdot (s_{nn} - s) \quad (1)$$

which is equivalent to

$$s' = (1 - w) \cdot s + w \cdot s_{nn} \quad (2)$$

It is seen from (2) that a new image s' is generated from the original sample s and the chosen nearest neighbor s_{nn} based on a weight coefficient *w*. s and s_{nn} can be referred as the parent images of s'. Every original sample produces *N* new samples with its *N* nearest neighbors. With the process repeated using every sample in the class, a set of new images with *N* times the size of original image sets in each class is synthesized and *N* is referred to as the data augmentation multiplier in the SMOTE process.

In the original settings [8], *k* is set to be 5 with *N* set to be 2. The weight coefficient *w* is a random number drawn from a uniform distribution between 0 and 1. The improved SMOTE is to explore different parameter settings as follows.

2) Improved SMOTE

The improved SMOTE enhances the original SMOTE by fine-tuning the augmentation process. This includes setting the range for the randomly selected weight coefficient, optimizing the data augmentation multiplier and removing redundant images in the enlarged training set.

a) Weight Coefficient w

In this method, the effect of the range of the weight coefficient on the training process is explored. When the weight coefficient is extremely close to 0 or 1, the synthesized images will be highly similar to one of their parent images. This will lead to overfitting and hence is not beneficial for extracting generic features from the augmented image set. Due to the large intra-class variance of the dataset, redundant images may cause the classification machine to be biased towards certain images. However, if the weight coefficients are restrained to a small range, the variations in the generated images will be limited, and the randomness of the SMOTE algorithm will be undermined. The aim is to find a satisfying range for the weight coefficient which can produce the best result in terms of classification accuracy.

b) Data augmentation multiplier N

In the original SMOTE [8], *N=2* was used to generate synthesized data. In this paper, experiments with datasets generated using different data augmentation multipliers *N* are performed. With a larger dataset, the classifier is able to extract more information to accomplish a higher accuracy. However, due to the small size of the original dataset, as the data augmentation multiplier *N* increases, some of the newly generated images are redundant and cannot offer additional information. With the above consideration, experiments with *N*=2, *N*=3 and *N*=4 are carried out respectively when k is fixed at 5.

c) SSIM index-assisted redundancy reduction

To ensure the variety in the training images, the similarities between newly generated images and original images are assessed to prevent new images from being too similar to the original images. Those extremely similar images will cause overfitting in training and hence synthesized images with very high similarity to the original images will be removed.

The measurement criteria adopted is Structural Similarity index (SSIM index) [11, 12]. It is originally used to assess the quality of digital videos and images projected onto a screen by comparing the luminance(l)， contrast(c), and structure(s) of the projected image and that of the original image. The adopted SSIM index for two images represented by *x* and *y*, is expressed in (3) [11, 12].

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x{}^2 + \mu_y{}^2 + c_1)(\sigma_x{}^2 + \sigma_y{}^2 + c_2)} \quad (3)$$

with μ_x: average value of *x*;
μ_y: average value of *y*;
σ_x: variance of *x*;
σ_y: variance of *y*;
σ_{xy}: covariance of *x* and *y*;
$c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$,
where $L = 2^{\#bits\ per\ pixel} - 1$, $k_1 = 0.01$ and $k_2 = 0.03$.

The SSIM indexes of R, G, B components of the images are calculated separately, and the average value of them is used as the SSIM index between two images. For each newly-generated image, its SSIM indexes against all original images are calculated, out of which the highest SSIM index (HSI) is taken. The HSI of every synthesized image is sorted and the 10 images with the highest HSI in each class are then excluded from the training set to reduce the redundancy during each training stage.

B. Transfer learning with the pre-trained model Inception-Resnet-v2

Transfer learning is a method to reuse a pre-trained model on a new task by applying previously acquired knowledge to learn new knowledge. Since the cost of learning directly from the target dataset at the beginning is too high, it is better to utilize relevant generic features to assist in extracting new and more specific features as quickly as possible. In this paper, transfer learning is employed to apply a deep neural network that has already been trained on ImageNet to a relatively small dataset available to us. By taking advantages of the pre-trained model as a feature extractor, the amount of images required by the deep neural network is significantly reduced.

Inception-Resnet-v2, which is a hybrid inception model with residual connections [13], is chosen as it is a state-of-art pre-trained model. Inception network has been demonstrated to achieve superb performance at a relatively low computational cost. With the introduction of residual connections, Inception-ResNet-v2 with deeper and denser layers has higher training speed than Inception architecture and degradation problem is also avoided [9]. The network has 164 layers in depth and can classify images into 1000 object categories [13].

III. EXPERIMENT

A. *The original dataset*

This experiment concentrates on improving the accuracy of classification of five common types of non-tumorous facial pigmentation disorders based on a real-world image set consisting of 30 clinical images per class: freckles, lentigines, melasma, Hori's nevus, and nevus of Ota [14]. Images of the whole facial region of patients are taken to crop out the areas containing fully or partially the region of interest (ROI). The cropped images are pre-processed to a consistent dimension of 100×100×3 for subsequent classifications. A representative image in each class is displayed in Fig.2.

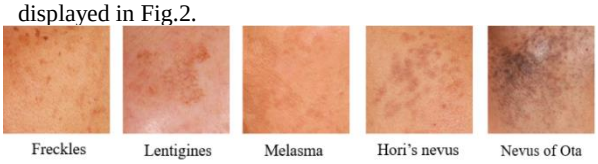

Fig. 2. Sample images from each class

B. *Enlarged dataset using improved SMOTE*

To assess the classification performance with different SMOTE parameter settings, several input datasets are prepared. During the generation process, the images are normalized to the same size (200×200×3) and stored in RGB image format for Euclidean distance measurement and the interpolation of the pixels storing information of the images.

Six different ranges of the weight coefficient are chosen and experimented for comparison: 0-1(original), 0.1-0.9, 0.2-0.8, 0.3-0.7, 0.4-0.6. 0.5. For the case of using a fixed number 0.5, new images generated with the same two parent images are exactly the same. The repeated images are deleted from the image set. N is chosen to be 2 at all time in the weight coefficient comparison test. Some representative images synthesized using weight coefficients in large difference and their parent images from Hori's nevus class are demonstrated below.

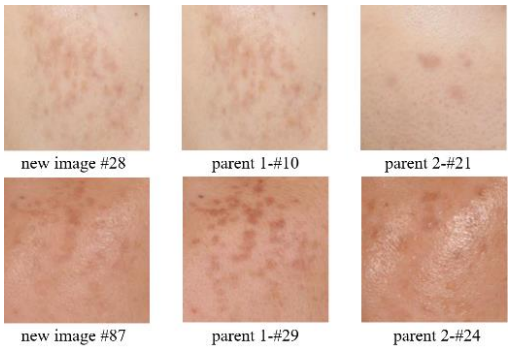

Fig. 3. Sample images generated using different weight coefficients

As shown in Fig.3, new image 28 is derived from original image 10 and 21 using weight coefficient 0.075, which is extremely close to 0. The HSI evaluated is 0.999. On the other hand, new image 87 is generated from original image 29 and 24 using weight coefficient 0.449 and the HSI is 0.921. To further illustrate the similarity between new images and the original images when different weight coefficients are used, the HSI ranges of images generated with different weight coefficient ranges are shown in Table I. Five classes are simplified by their capital letters: Freckles: F; Lentigines: L; Melasma: M; Hori's nevus: H; Nevus of Ota: O.

TABLE I. HSI OF IMAGES USING DIFFERENT WEIGHT RANGE

Weight coefficient range	F	L	M	H	O
0-1	0.876-1.000	0.794-0.999	0.863-1.000	0.908-0.999	0.883-1.000
0.1-0.9	0.900-0.996	0.918-0.998	0.840-0.997	0.889-0.996	0.849-0.997
0.2-0.8	0.898-0.990	0.897-0.993	0.864-0.991	0.822-0.990	0.859-0.992
0.3-0.7	0.891-0.989	0.826-0.981	0.860-0.991	0.875-0.987	0.850-0.988
0.4-0.6	0.853-0.977	0.834-0.968	0.872-0.973	0.816-0.973	0.859-0.977
0.5	0.856-0.974	0.826-0.962	0.849-0.972	0.821-0.966	0.840-0.967

It is clearly shown that when the weight coefficients range is amended to 0.1-0.9, there are no images which closely resemble the original images in the enlarged image set. Hence, 0.1-0.9 is chosen as the weight coefficient range in the subsequent experiments.

Following this, *N* is varied for data augmentation multiplier comparison test. Image set with 90 new images in each class (N=3) and that with 120 images in each class (N=4) are obtained for training. Table II lists the HSI ranges of the image set generated using N=4. The 20 highest HSI and 10 lowest HSI are tabulated. High degree of similarity is observed for the 10 highest HSI images. Hence, removing highly redundant images is necessary.

TABLE II. SELECTED HSI RANGE IN DATASET GENERATED USING N=4

HSI	F	L	M	H	O
1^{st}-10^{th}	0.996-0.998	0.995-0.997	0.995-0.997	0.995-0.998	0.995-0.999
11^{th}-20^{th}	0.994-0.996	0.991-0.995	0.992-0.994	0.992-0.995	0.991-0.995
…	…				
111^{th}-120^{th}	0.878-0.935	0.877-0.915	0.882-0.905	0.908-0.924	0.882-0.899

C. *Experiment set-up*

As stated in section IIB, the pre-trained model used in this paper is Inception-Resnet-v2 [13]. During the model set-up, the original images in each class are randomly allocated to ten subsamples for a ten-fold cross-validation test such that each subsample contains three images per class. In every validation process, one subsample is selected as the testing set and the images in the remaining nine subsamples are included in the training set. Subsequently, newly generated images except for those whose parent images are selected as testing images are added into the training set. In the redundancy reduction experiment, 10 images having the highest HSI among the remaining generated training images in each class are removed from the training set. Subsequently, a grid search using four-fold cross-validation within the training set is conducted for parameter selection and early stopping is triggered when constant validation performance is observed in 500 continuous iterations to prevent overfitting. Weights of the processing units in the network are randomly initialized with a Gaussian distribution of zero mean and a standard deviation of 0.001. The learning rate is refreshed with an exponential decay factor shown in (4), with

α denoting the learning rate. The decay step is selected to be 1000 [9].

$$\alpha_{adaptive} = \alpha \times decay\ rate^{(\frac{step}{decay\ step})} \quad (4)$$

In the ten-fold cross-validation, the validation process is repeated 10 times until all the subsamples are taken as the testing subsample once. The ten-fold cross-validation is performed independently 10 times with the images being randomly shuffled into different subsamples to produce statistically dependable results and the average result is computed.

D. Experiment results and discussion

To pinpoint the optimal parameter settings for the improved SMOTE, experiment results using different SMOTE parameters with transfer learning are presented. In Table III, overall accuracy and standard deviation obtained with different weight coefficient ranges are tabulated.

TABLE III. RESULTS USING DIFFERENT WEIGHT COEFFICIENT RANGE

Weight coefficient range	Accuracy %	Standard deviation
0.5	81.07	0.0941
0.4 - 0.6	85.53	0.0929
0.3 - 0.7	85.47	0.0705
0.2 - 0.8	85.33	0.0842
0.1 - 0.9	86.13	0.0681
0 - 1	84.67	0.0825

As shown in Table III, the range 0.1-0.9 yields the highest accuracy, at an improvement of 1.5% compared to applying original SMOTE (corresponding to the weight coefficient range 0-1). This is because with weight coefficient close to 0 or 1, the generated images are extremely similar to one of the parent images and provide no further useful information to the feature extractor. Conversely, this may even cause the classifier to overfit those images. Moreover, it is discovered that with the weight coefficient fixed at 0.5, the accuracy is much lower than the other cases as images generated are not diversified. To further improve the classification result, we set the weight coefficient range to be 0.1-0.9 and augment more images for training. The results with different data augmentation multiplier are shown in Table IV.

TABLE IV. RESULTS USING DIFFERENT DATA AUGMENTATION MULTIPLIER

Multiplier	Accuracy %	Standard deviation
2	86.13	0.0681
3	86.67	0.0874
4	86.67	0.0772

It is observed from Table IV that when images with 3 times of the number of original images are generated, the accuracy increases by 0.5% as larger image set contains more general features. At higher N, the accuracy approaches to a constant value. This may be constrained by small size of the original dataset. Despite the effort to generate more images, there is an intrinsic limit to the possible improvement as images synthesized become redundant and provide no additional information to the classifier. Based on the above observation, the redundancy reduction experiment is carried out with the weight coefficient range set to be 0.1 to 0.9 and the multiplier at 4. The results are tabulated in Table V with comparison with the V-PLDA method [6] and Inception-ResNet-v2 without data augmentation using SMOTE.

TABLE V. RESULTS OF DIFFERENT CLASSIFIER

Method	Accuracy %	Standard Deviation
V-PLDA [6]	77.33	0.0982
Inception-ResNet-v2	81.87	0.0889
Inception-ResNet-v2 with Improved SMOTE	87.33	0.0767

From Table V, several points can be noticed: i) With redundancy reduction, the accuracy is further improved by 0.66%. This proves that duplicate images will cause the model to be over-fitting, resulting in poor generalization to images which are new to the training set. ii) With the employment of transfer learning using pre-trained model Inception-ResNet-v2, the accuracy achieves a gain of more than 4% compared to the V-PLDA method [6]. This is because the pre-trained model has been trained on a large image set and thus it is able to learn high-level features. iii) With the additional application of the improved SMOTE, the accuracy is further improved by 5.5%. Overall, our proposed method achieves a significant accuracy improvement of 10% as compared to V-PLDA method. This improvement is significant, and our proposed method is also practical in developing other classification models for similar applications, as for most biomedical image analysis practices, the small and domain-specific dataset is a common problem.

IV. CONCLUSION

In this paper, to address the problem of limited real-world training images in the classification of five common non-tumorous facial pigmentation disorders, we have proposed to combine an improved SMOTE with transfer learning. In the improved SMOTE, the image generation process is adjusted based on the similarity measurement. Transfer learning with the pre-trained network architecture Inception-ResNet-v2 is employed to apply deep learning on small datasets. With the combination of the improved SMOTE and transfer learning, a significant improvement is achieved reflected by the overall accuracy increase (10%) compared to the V-PLDA method [6]. Specifically, by fine-tuning the range of weight coefficient, and adjusting the data augmentation multiplier of SMOTE and by removing those highly redundant images from the newly generated training set, the classification accuracy has improved by about 5.5% compared to the method using transfer learning only. This is promising for tackling other similar medical image applications limited by a small dataset in the future.

ACKNOWLEDGEMENTS

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on Campus (URECA) program.

References

[1] G. Dougherty, "Image analysis in medical imaging: recent advances in selected examples," Biomedical Imaging and Intervention Journal, vol. 6, no. 3, p. e32, Jul-Sep 2010.

[2] M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, "Medical image processing, analysis and visualization in clinical research," in Proceedings 15th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001, Bethesda, MD, USA, 2001.

[3] N. J. Dhinagar and M. Celenk, "Analysis of regularity in skin pigmentation and vascularity by an optimized feature space for early cancer classification," in 2014 7th International Conference on Biomedical Engineering and Informatics Biomedical Engineering and Informatics (BMEI), Dalian, China, 2014.

[4] G. Zouridakis, M. Doshi, and N. Mullani, "Early diagnosis of skin cancer based on segmentation and measurement of vascularization and pigmentation in Nevoscope images," in 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Francisco, CA, USA, 2004.

[5] Z. Ma and J. M. R. Tavares, "A review of the quantification and classification of pigmented skin lesions: from dedicated to hand-held devices," Journal of medical systems, vol. 39, no. 11, pp. 177, 2015.

[6] Y. Liang, L. Sun, W. Ser, F. Lin, S T. G. Thng, Q. Chen and Z. Lin, "Classification of non-tumorous skin pigmentation disorders using voting T based probabilistic linear discriminant analysis," Computers in Biology and Medicine, vol. 99, pp. 123-132, 2018.

[7] E. J. Parra, "Human pigmentation variation: evolution, genetic basis, and implications for public health," American Journal of Physical Anthropology, vol. 50, pp. 85-105, 2007.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, no. 16, pp. 321-357, 2002.

[9] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, "Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation," in 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 2018

[10] X. Liu, C. Wang, Y. Hu, Z. Zeng, J. Bai, and G. Liao, "Transfer Learning with Convolutional Neural Network for Early Gastric Cancer Classification on Magnifiying Narrow-Band Imaging Images," in 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, Oct 2018.

[11] Z. Wang, B. A. Conrad, S. H. Rahim, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[12] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in The 37th Asilomar Conference on Signals, Systems & Computers, vol. 2, pp. 1398-1402, 2003.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2016.

[14] S. G. Y. Ho and H. H. L. Chan, "The Asian Dermatologic Patient," American Journal of Clinical Dermatology, vol. 10, no. 3, pp. 153-168, June 2009.