Classification of Skin Lesions Based on Data Collaboration Under Imbalance Dataset

Weijia Ji[1,2], Lizhi Cai[1,2], Mingang Chen[2](✉), and Naiqi Wang[1,2]

[1] School of Information Science and Engineer,
East China University of Science and Technology, Shanghai, China
asdfvl929@gmail.com, slytherinwnq@163.com
[2] Laboratory of Computer Software Testing and Evaluating,
Shanghai Development Center of Computer Software Technology,
Shanghai, China
{clz,cmg}@ssc.stn.sh.cn

Abstract. Imbalance data is a common problem in machine learning task, which often impacts the accuracy of models. An effective way to solve it is to increase the number of minority class samples in the dataset. Many methods are put forward to solve the problem of imbalance data in machine learning. But these are all for low-dimensional data. For high-dimensional data, such as images, these methods are not well applicable. In this paper, an image generation method based on generative adversarial network is introduced to do pattern learning for samples of minority class in the dataset, so as to realize the expansion of data for minority class. And finally the classification networks for skin lesions are trained by data collaboration which consist of real images and generated images. The experimental results indicate that the accuracy of networks are further improved by the addition of generated images while alleviating the imbalance problem to some extent.

Keywords: Imbalance data · Generative adversarial network · Data collaboration · Skin image classification

1 Introduction

Classification task is always being a research hotspot in machine learning field. The existing classification methods have been able to achieve good performance on the classification task of conventional dataset. However, the research of these classifiers are based on an assumption that the distribution of sample categories is roughly the same. That means the dataset used for training is balanced. In simple, the number of data samples contained in each category is basically equal. But this assumption does not exist in many practical problems. The truth is the number of samples in one or even several categories in dataset is much smaller than that of other categories. For example, imbalance data exists in these application scenarios, such as information retrieval, credit card illegal transaction [1], medical diagnosis [2, 3], etc. And the recognition rate of minority class in these tasks appears to be more important. As far as medical diagnosis task is concerned, if a normal person is misdiagnosed as a patient, it will

X. Wang et al. (Eds.): CollaborateCom 2019, LNICST 292, pp. 295–306, 2019.
https://doi.org/10.1007/978-3-030-30146-0_20

bring him some mental burdens, but the fact is that the person is healthy. However, if a patient is misdiagnosed as a normal person, it may bring about the patient to miss the best treatment period and some serious consequences. So it often fails to achieve satisfied results when applying the model based on imbalance data to the above scenarios.

The reason for this problem is that the classification model with the overall classification accuracy as the learning target pays too much attention to the majority class samples. That means the model has a high recognition rate of the majority class samples, resulting in the degradation of performance about minority class samples. So it is particularly necessary to solve the problem of imbalance data in some specific fields.

When performing a lesion recognition study on the skin images dataset, we found that the number of image samples for skin cancer diseases was far less than that of benign lesions. This leads to the model's accuracy in predicting malignant lesions is much lower than that of benign lesions. Thus in order to improve the recognition accuracy of skin cancer diseases, it is first necessary to solve the problem that the sample numbers differ greatly, that is, sample imbalance problem.

We firstly adopt the methods [4] commonly used in machine learning to solve the imbalance problem in skin lesion images, such as random sampling, SMOTE, Borderline-SMOTE. But most of these methods are aimed at low-dimensional sample data. For high-dimensional data such as skin images, these solutions cannot play their due role well, and cannot improve the classification accuracy of malignant diseases.

Therefore, this paper introduces an image generation method based on generative adversarial network (GAN) [5]. The images of the same distribution as the corresponding class are generated by learning the pattern features of minority class samples in the dataset. The model is then trained in a way of data collaboration with real images and generated images, thereby further improving the accuracy of lesion classification, especially the prediction of skin cancer.

2 Related Methods

Many methods have been proposed to solve the problem of imbalance data in machine learning classification tasks. We list several commonly used methods here.

2.1 Random Sampling

The sampling algorithm uses some strategies to change the class distribution in the dataset to convert the imbalanced sample into a relatively balanced sample. The random sampling method is the simplest and most intuitive one of the sampling algorithms. There are two types of random sampling: RandomUnderSampling and RandomOverSampling. RandomUnderSampling refers to the random deletion of some data from majority class, so that the data amount of majority class and minority class is basically the same. Another random sampling method, RandomOverSampling, is to achieve the relative balance of classes by adding data to minority class. The method of adding is generally to randomly extract data from and put them back into the minority class, and finally make the number of minority and majority class equal. The essence of RandomOverSampling is to copy some samples from minority class to achieve the effect of increasing the dataset size.

2.2 SMOTE

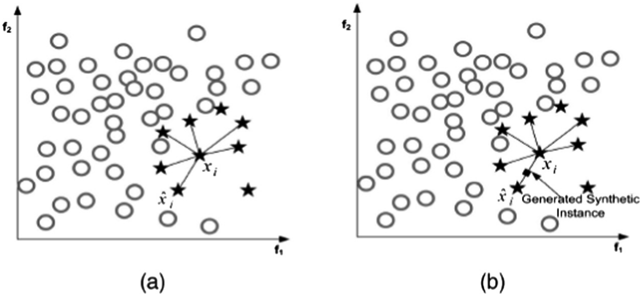
SMOTE (Synthetic Minority Oversampling Technique) [6] is an improved method based on random oversampling algorithm. Due to random oversampling method uses a strategy of simply copying data form the minority class to increase the sample size of the minority class. It is easy to cause over-fitting problem in the training process of network. So the SMOTE algorithm has made an effective improvement to this problem. Its basic idea is to analyze minority class samples and synthesize new samples, then add them to the dataset.

The operation flow of SMOTE is as follows:

(1) For each sample x in minority class, calculating the distance to all other samples in minority class $S_{minority}$ by the Euclidean distance, and obtain k neighbor samples;
(2) Setting a sampling ratio according to the sample imbalance ratio to determine the sampling magnification N. For each sample x from minority class, randomly selecting several samples from its k neighbors, assuming that the selected neighbor is xn;
(3) For each randomly selected neighbor xn, constructing new sample data with the original sample according to the following formula:

$$x_{new} = x_i + \delta * \left(\widehat{x_i} + x_i\right) \tag{1}$$

where $x_i \in S_{minority}$; $\widehat{x_i}$ represents a sample of the k-nearest neighbors of x_i, $\widehat{x_i} \in S_{minority}$ and δ is a random number with a range [0, 1]. The specific operation is shown in Fig. 1, (a) indicates the k-nearest neighbor (k = 7) of x_i are found, and then the new data is generated according to formula (1), as in (b) The rhombus shows a sample of the newly synthesized data.



(a) (b)

Fig. 1. SMOTE [4]

However, the SMOTE algorithm is prone to excessive generalization and high variance, and it is also easy to generate overlapping data. Besides, the SMOTE method produces the same amount of synthesized data for each original sample in minority class without considering the distribution characteristics of its neighboring sample data. So in order to solve the problems in SMOTE, the Borderline-SMOTE [7] is proposed. Unlike SMOTE, which generates new samples for each of the minority class sample, Borderline-SMOTE only generates new data for the minority class samples close to the boundary. That is to say, a minority class sample is selected, in which the number of majority class samples in its adjacent sample set is greater than half of the total. So the data in such a sample set are used to generate the new synthetic samples to increase classification accuracy.

2.3 GAN

The Generative Adversial Network (GAN) is a deep learning-based generation model. It has been paid more and more attention by academics and industry since it was proposed by Ian Goodfellow et al.

Inspired by the zero-sum game in game theory, GAN regards the generation problem as the game between the two networks of generator and discriminator: the generator continuously produces synthetic data from a given random noise and finally outputs an image, the discriminator is to distinguish whether the output image of the generator is a real image. The former tries to produce data that is closer to the real image, while the latter tries to distinguish between the true and false of the generated data. The basic process of GAN is shown in Fig. 2. As a result, the data obtained by the generator become more and more "perfect" and closer to the statistical distribution of real data. So the generator can generate the data we want, such as images, sequences, videos, and others.

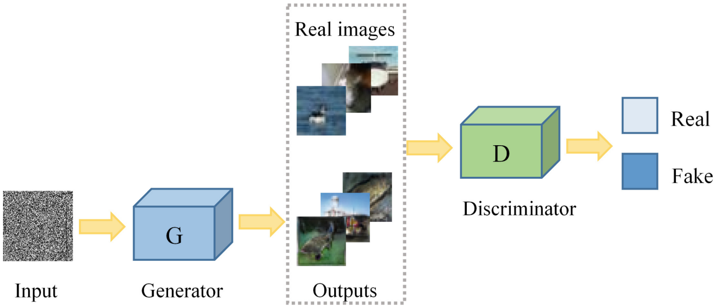


Fig. 2. Processing flow of GAN.

3 Method

The generative adversarial network used in this paper is PG-GAN [8] network. Compared with the previous generative adversarial network like DCGAN [9], WGAN [10] and others [11, 12], PG-GAN has obvious advantages in image generation and can train stably and generate the high-resolution images.

The key idea of PG-GAN is to gradually increase the number of layers of generator and discriminator, that is, to adopt a progressive growing training method. The general process is shown in Fig. 3. Starting from low resolution, the network begins to train and learn. The generator is still used to learn the data mode to generate the corresponding images, while the discriminator is used to judge the authenticity of the generated data. After the low-resolution image is trained, new layers are added to the network structure. And the higher-resolution image is gradually transferred into the training process. Then the network trains the current resolution image stably, and transition to the next higher resolution image by degrees. This new method not only speeds up the training speed of the model, but also greatly stabilizes the training process, so that the model can learn to generate high-quality images, such as the skin lesion images required in this paper.

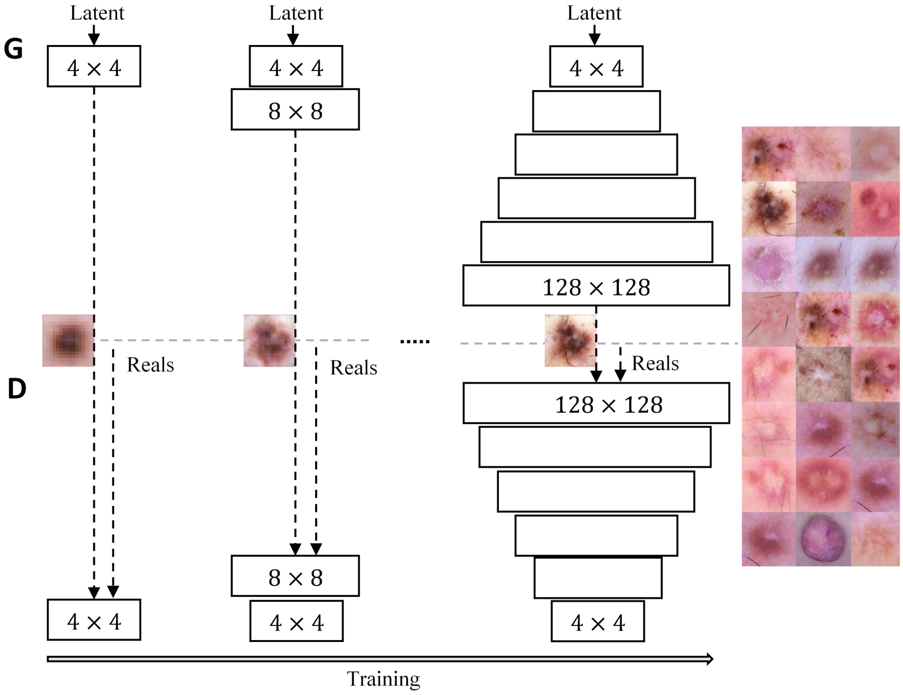


Fig. 3. Training process of PG-GAN on skin lesion images.

Figure 4 shows the operation of PG-GAN network in the growing stage of (a) generator and (b) discriminator. When the generator is in a transitional stage, the image with the resolution 4 × 4 is converted to the output of the same size as the next operation resolution (8 × 8) through resize and convolution. Then the two part outputs make weighted operation. The final output is obtained by to_rgb operation again. The advantage of such a training method is that it can make full use of the results of the previous resolution training, and go through a slow transition (the weight w gradually increases), making the network generated by the training of the next resolution more stable.

The growing stage of the discriminator is shown in Fig. 4(b). The overall detail operation is similar to that in the generator. At the current resolution (8 × 8), the network obtains the output of the same size as that of the next resolution (4 × 4) through pooling and convolution operations. Then the two outputs are weighted, and finally the output is obtained through the to_rgb operation.

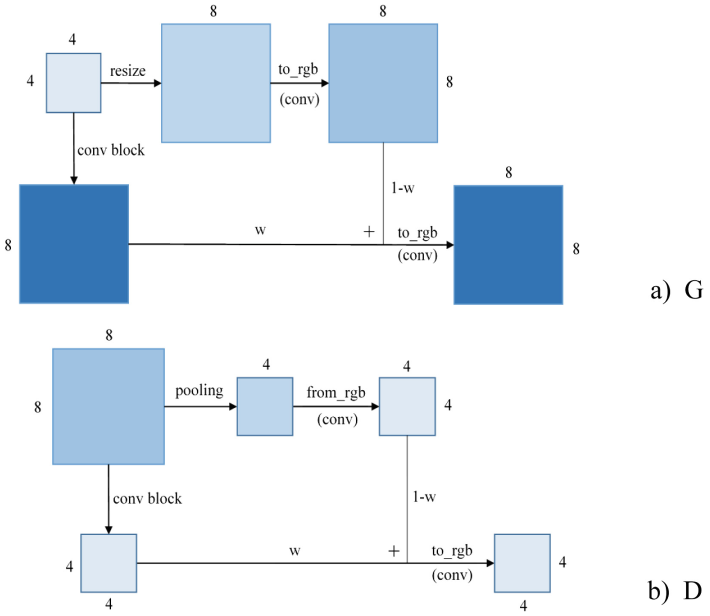


Fig. 4. Growing stage of the generator and discriminator

With the help of growing training, in the early period of the PG–GAN training, the model can keep steady training according to the low resolution images. The whole process of training iteration is mostly done at low resolution, this makes the training time of model greatly shortened. But the generated results are still high-quality. The specific samples can be seen in the experimental part of the paper.

4 Experiments

4.1 Metrics

The following metric methods are used to evaluate the experimental results.

Specificity:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

Sensitivity:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

TP represents the number of True Positive images, that is, in the classification of skin image diseases, the number of sample images that originally belong to the positive class and actually divided into the positive class. *TN* refers to the number of True Negative images, that is, the number of samples that are originally belong to the negative class and actually classified as negative. *FP* refers to the number of False Positive images, that is, the number of samples that are originally belong to the negative class and wrongly classified as positive class. *FN* represents the number of False Negative images, i.e. the number of original positive samples wrongly divided into negative ones.

Confusion matrix:

Confusion matrix		Predicted label	
		Positive	Negative
True label	Positive	TP	FN
	Negative	FP	TN

Precision represents the number of samples in which the model prediction is positive.

$$P = \frac{TP}{TP + FP} \tag{4}$$

Recall indicates the number of correct predictions for the model in the sample whose real label is positive.

$$R = \frac{TP}{TP + FN} \tag{5}$$

F1 measure, a special form of *F*-measure, is the harmonic average of precision and recall. The average value is equal to all the values, while the harmonic average will give more weight to the smaller values, so it can better reflect the effect of the model in the case of data imbalance. *F1* is defined as follows:

$$F1 = \frac{2 * P * R}{P + R} \tag{6}$$

In addition, AUC (Area Under Curve) evaluation metric is introduced in this paper to evaluate classifiers learned from data sets more accurately.

4.2 Result Analysis

The experimental data in this paper mainly come from two places: one is the open source dataset of ISIC 2018 [13, 14], and the other is the dermatology departments of the Shanghai 9th People's Hospital, 6th People's Hospital, and other medical institutions. The types of skin lesions contained in these data include nevus, seborrheic keratosis, melanoma, and so on.

All of the methods mentioned above are implemented with Python and TensorFlow [15]. In addition, an Nvidia Quadro P5000 GPU is used in the experiment to speed up our training on models.

The comparison of the images generated by PG-GAN with real images is shown in Fig. 5, where (a) represent the real images, (b) represent the generated images. It seems that it's not easy to tell them apart. They have great similarity in the shape, color, texture and so on. So the data generated by the GAN can be used as a part of the training set. Then a model can be trained through the data collaboration of real images and generated images. And finally making experiments to verify its feasibility.

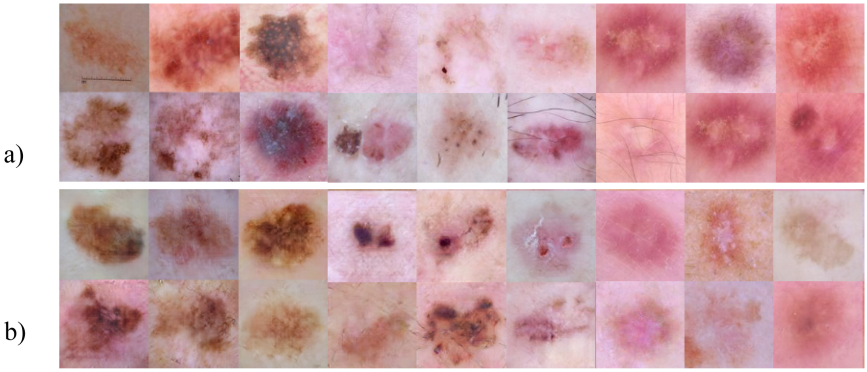


Fig. 5. Real image and generated image

When studying the skin images dataset, it can been easily found that benign nevus and malignant melanoma account for the majority of the dataset. Furthermore, nevus and melanoma have great similarities in appearance, color, etc. see Fig. 6 for details. Therefore, the classification problem of these two classes are mainly analyzed in the experiment, i.e., the classification of nevus and melanoma. In the experiment, our original training set has 5000 images of nevus and 1000 images of melanoma.

Model training is performed on the dataset composed of nevus and melanoma above. The prediction accuracy values of the models obtained by each method on the test set can be seen in Table 1. Among them, Exp-1 refers to the case where 1000 generated images of melanoma are added to the training set, Exp-2 add 4000 generated

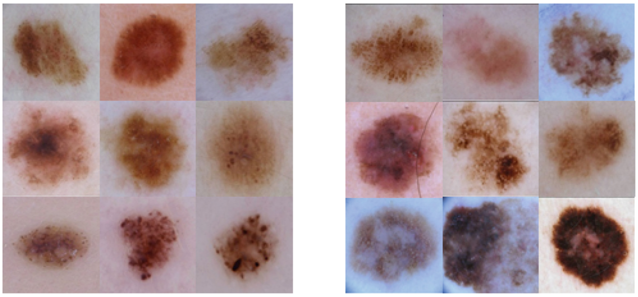


Fig. 6. Nevus and melanoma

melanoma images. It can be found that the accuracy of the random sampling methods, SMOTE and Borderline-SMOTE on the test set is lower than that of the original model which take no methods. The reasons for this may be the under-fitting and over-fitting of the models under these methods. For example, in RandomUnderSampling, the image of nevus is down-sampled to the same number as melanoma, which leads to insufficient training of the model and results in under-fitting problem. In addition, SMOTE and Borderline-SMOTE are methods that are suitable for small size datasets. When they are applied to a large-scale image dataset, there is a high probability that the new data generated by image pixel features is not a skin image. So the accuracy of the model trained by this type of data on the test set will not be high. The numerical results here also prove this.

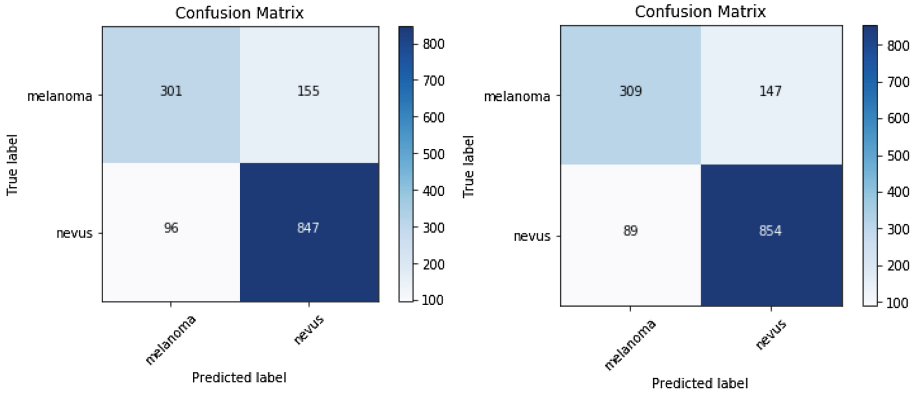
The models trained on the training set containing the generated images have the higher accuracy than the original model. Besides, the Exp-2 model has the highest accuracy value, reaching 83.231%, which is 1.172% higher than the original model. The explains to some extent that the model trained in data collaboration has a higher accuracy. That is to say, the skin images generated by GAN do have an accuracy improvement in the classification of skin lesions.

Table 1. The prediction accuracy on test set.

Method	Test Acc
Origin	0.82059
RandomOverSampling	0.80843
RandomUnderSampling	0.77627
SMOTE	0.78413
Borderline-SMOTE	0.79414
Exp-1	0.82130
Exp-2	0.83231

However, the accuracy above refers to the accuracy of the overall data, namely that the nevus with a larger number of samples has an advantage in model prediction, and its prediction accuracy for malignant melanoma is difficult to judge.

Therefore, the confusion matrix of the model on the test samples are further observed under each method, as shown in Fig. 7. Here only show the confusion matrix about the original model and the Exp-2 model. The values of TP, FN, FP and TN corresponding to the confusion matrix of all models have been recorded in Table 2.



a) Origin

b) Exp-2

Fig. 7. Confusion matrix of the Origin and Exp-2.

Table 2. The values of TP, FN, FP, TN.

Method	TP	FN	FP	TN
Origin	847	96	155	301
RandomOverSampling	849	94	174	282
RandomUnderSampling	723	220	93	363
SMOTE	822	121	181	275
Borderline-SMOTE	851	92	196	260
Exp-1	839	104	146	310
Exp-2	854	89	147	309

At this moment, the corresponding Specificity, Sensitivity, P, R and F1 values can be calculated according to the formulas. The values obtained after their calculation can be seen in Table 3.

In addition to the above metrics, the ROC curves for each method are also plotted, as shown in Fig. 8. The area value on the graph indicates the value of AUC corresponding to the ROC curve. The AUC values of Exp-1 and Exp-2 belong to the highest two of the AUC values of all methods, and the Exp-2's AUC is the largest, which is 0.792.

It can be seen from the above evaluation metrics that the melanoma images generated by GAN do help in improving the overall accuracy of the model and the prediction accuracy of melanoma.

Table 3. The value of metrics.

Method	Sp	Se	P	R	F1
Origin	0.6601	0.8982	0.8453	0.8982	0.8710
RandomOverSampling	0.6184	0.9003	0.8299	0.9003	0.8637
RandomUnderSampling	0.7961	0.7667	0.8860	0.7667	0.8221
SMOTE	0.6031	0.8717	0.8195	0.8717	0.8448
Borderline-SMOTE	0.5702	0.9024	0.8128	0.9024	0.8553
Exp-1	0.6798	0.8897	0.8518	0.8897	0.8703
Exp-2	0.6777	0.9056	0.8531	0.9056	0.8786

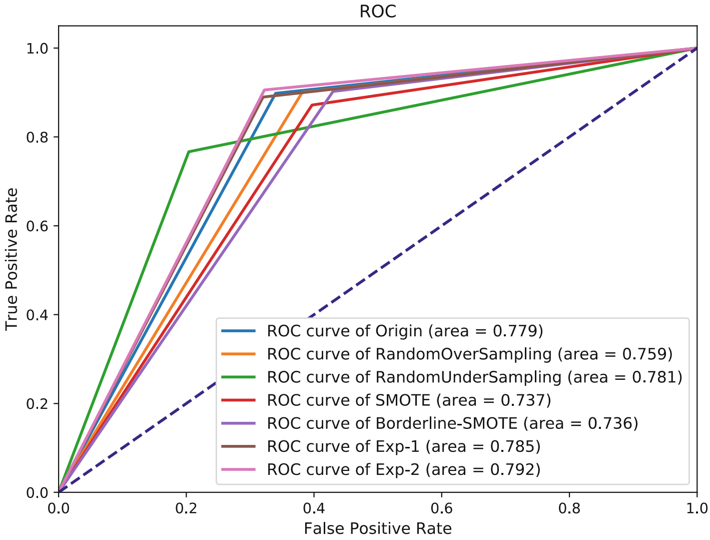


Fig. 8. ROC curves of all methods.

5 Conclusion

The commonly used methods for imbalance data are generally aimed at low-dimensional data, which do not perform well on high-dimensional data such as image. This paper introduces an image generation method based on GAN to realize data expansion for some classes in the training set of skin images. Then a new classifier model is trained by the way of data collaboration. The results show that the high quality skin images generated by the GAN do contribute to improve the overall accuracy of the model and the accuracy of the malignant lesions.

Acknowledgment. This work was funded by Science and Technology Commission of Shanghai Municipality Program (No. 17411952800, No. 18441904500, 18DZ1113400) and Science and Technology Department of Hainan Province (No. ZDYF2018022).

References

1. Chan, P., Stolfo, S.: Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection (1998)
2. Choe, W., Ersoy, O.K., Bina, M.: Neural network schemes for detecting rare events in human genomic DNA. Bioinformatics **16**(12), 1062–1072 (2000)
3. Plant, C., et al.: Enhancing instance-based classification with local density: a new algorithm for classifying unbalanced biomedical data. Bioinformatics **22**(8), 981–988 (2006)
4. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
5. Goodfellow, I.J., et al.: Generative adversarial nets. In: International Conference on Neural Information Processing Systems. MIT Press (2014)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., et al.: SMOTE: synthetic minority over-sampling technique, June 2011. https://doi.org/10.1613/jair.953
7. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). https://doi.org/10.1007/11538059_91
8. Karras, T., et al.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018)
9. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. Computer Science (2015)
10. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017)
11. Berthelot, D., Schumm, T., Metz, L.: BEGAN: Boundary Equilibrium Generative Adversarial Networks (2017)
12. Li, Y., Xiao, N., Ouyang, W.: improved boundary equilibrium generative adversarial networks. IEEE Access 1 (2018)
13. Codella, N.C.F., et al.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC) (2017). arXiv:1710.05006
14. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 180161 (2018). https://doi.org/10.1038/sdata.2018.161
15. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning (2016)