Hindawi
Mobile Information Systems
Volume 2022, Article ID 8572852, 9 pages
https://doi.org/10.1155/2022/8572852

Research Article

A Set of Comprehensive Evaluation System for Different Data Augmentation Methods

Can Zhang,[1,2] **Xu Zhang**,[1] **and Dawei Tu**[1]

[1]*School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China*
[2]*Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia*

Correspondence should be addressed to Can Zhang; can.zhang@student.uts.edu.au

Received 20 January 2022; Revised 14 February 2022; Accepted 19 February 2022; Published 11 March 2022

Academic Editor: Hasan Ali Khattak

Data augmentation is an effective method to prevent model overfitting in deep learning, especially in medical image classification where data samples are small and difficult to obtain. In recent years, different data augmentation methods, such as those based on single data transformation, multiple data mixing, and learning data distribution, have been proposed one after another, but there has never been a systematic system to evaluate various data augmentation methods. An impartial and comprehensive data augmentation evaluation system not only can assess the benefits and drawbacks of existing augmentation approaches in a specific medical image classification but also can provide an effective research direction for the subsequent proposal of new medical image data augmentation methods, thereby advancing the development of auxiliary diagnosis technology based on medical images. Therefore, this paper proposes an objective and universal evaluation system for different data augmentation methods. In this method, different augmented methods are evaluated objectively and comprehensively in terms of classification accuracy and data diversity by using existing large public data sets. The method is universal and easy to operate. To imitate the prevalent small-sized data sets in deep learning, an equal-interval sampling technique based on similarity ranking is presented to select samples from large public data sets and construct a subset that can fully reflect the original set. The augmented data sets are then created using various data augmentation approaches based on the small-sized data sets. Finally, different data augmentation strategies are objectively and fully evaluated based on the comprehensive scores of classification accuracy and data diversity following data augmentation. The validity and feasibility of the suggested sampling method and assessment system in this study are demonstrated by experimental findings on numerous data sets.

1. Introduction

The auxiliary diagnosis technology based on the medical image is in the stage of rapid development, but restricted by the amount of medical image data, the modeling method based on deep learning cannot be explored into more complex models. Data augmentation, a common method to improve model robustness, is widely applied to the training process of various medical image classification models. It is a strategy to increase the amount and diversity of existing data, with the purpose of extracting more useful information by changing existing data or generating "new data," thus improving the generalization ability of the model.

Commonly used image data augmentation technologies [1] include single data transformation (such as geometric transformation, color space transformation, resolution transformation, noise injection, etc.), multiple data mixing, and methods to generate new data by learning data distribution (such as GAN-based data augmentation). Data augmentation based on geometric transformation [2–4], such as flip, rotation, translation, and crop, is equivalent to increasing the perspective and position deviation of the data set, thus enhancing the robustness of the model in these aspects and improving the test accuracy. By altering the brightness in each channel of the original image, the transform based on color space produces new useable and functional data [1]. Its essence is to make the model adapt to

a more complex lighting environment and enhance its robustness under different scenes of lighting by adding various illumination deviations to the samples in the data set. The resolution transform uses $N \times M$ matrix to blur or sharpen the image, so as to help the model better deal with the problem of motion blur encountered in the testing process. At the same time, the sharpened image can highlight more details of objects. A new sample generation technique named "noise injection" [5] is a modern sample generation method that superimposes noise on an image that can be characterized by a random matrix with a given distribution. The image of different quality is simulated; the model's filtering ability of noise interference and redundant information is boosted; and the model's recognition ability is enhanced by artificially applying noise interference to the image and introducing redundant and interference information into the data set. In a single data transformation mode, data augmentation primarily transforms the relevant information of a picture, whereas, in a mixed data mode, data augmentation primarily transforms the relevant information of a picture.

The new training data are achieved by fusing the spatial or feature information of several images. No matter what kind of data augmentation method is mentioned above, they can use little prior information but only the information of these images themselves when generating new image data. Therefore, another new data augmentation method is generated: by means of generative adversarial network and image style transfer, the whole data set is taken as prior knowledge by learning the potential probability distribution of the data set and then sampled in it to generate new data [6–10]. This kind of data augmentation method can theoretically generate infinite kinds of new samples with the same probability distribution as the original data, which is a more excellent data augmentation method.

Although different data augmentation methods [6–10] have been proposed, there is no systematic system to evaluate them. An objective and comprehensive data augmentation evaluation system not only can evaluate the advantages and limitations of the existing augmentation methods in a specific medical image classification but also can provide an effective research direction for the subsequent proposal of new medical image data augmentation methods, to further promote the development of auxiliary diagnosis technology based on medical image. Therefore, in this paper, a set of specific evaluation systems will be given, as shown in Figure 1. The main contributions of this paper are as follows:

(i) To simulate common small-sized data sets in deep learning and prepare data for the subsequent data augmentation evaluation system, an equal-interval sampling algorithm built on similarity ranking is suggested to excerpt samples from large public data sets and generate a subset that can fully characterize the original set.

(ii) A general evaluation system is proposed to objectively and comprehensively evaluate different data augmentation methods based on the extracted small-sized data set, combined with the comprehensive scores of classification accuracy and data diversity. The method is universal and easy to operate as an equal-interval sampling algorithm based on similarity ranking is proposed to extract samples from large public data sets.

(iii) The data augmentation evaluation system proposed in this paper is applied to several data sets, and different data augmentation methods are objectively and comprehensively evaluated.

Section 2 introduces in detail the steps of the equal-interval sampling algorithm based on similarity ranking, which makes data preparation for the construction of subsequent evaluation system. In Section 3, based on the extracted small-sized data set, our data augmentation evaluation system is systematically presented from three dimensions. In Section 4, experiments show that the sampling method proposed in this paper is superior to other sampling methods in terms of classification accuracy and distribution similarity with the original data set. Furthermore, the data augmentation evaluation system proposed in this paper is applied to several data sets to make an objective and comprehensive evaluation of different data augmentation methods. The concluding remarks are provided in Section 5.

2. Equal-Interval Sampling Algorithm Based on Similarity Ranking

To prepare data for the suggested data augmentation method assessment system, a small-sized data set should be taken from the original large data set, which adheres and corresponds to the original data set's distribution and can completely represent it. It is commonly believed that the small data set will be able to remove unnecessary information from the original data set and represent it with as few subsets as feasible while still providing sufficient information. In recent years, a number of scholars have proposed a variety of data sampling methods [11–13], but these sampling methods all have defects such as too much randomness and uncertainty in the process of data set extraction and excessive subjective factors, so they cannot accurately, truly, and objectively reflect the global data information. As a result, this research uses stratified proportional sampling to determine the sampling quantity. The images in each category were then graded according to how similar they were. Finally, according to the previously selected sampling number, equal-interval sampling is performed for this category set, and the Frechet Inception Distance (FID) index [14] between the sampled data set and the original complete data set is applied as a secondary verification. To construct a small-sized data set with similar distribution to the original large-scale data set, which is prepared for subsequent evaluation of various data augmentation methods. The Frechet Inception Distance, or FID, is a metric for assessing the quality of the produced images that were designed specifically to calculate the performance of generative adversarial networks.

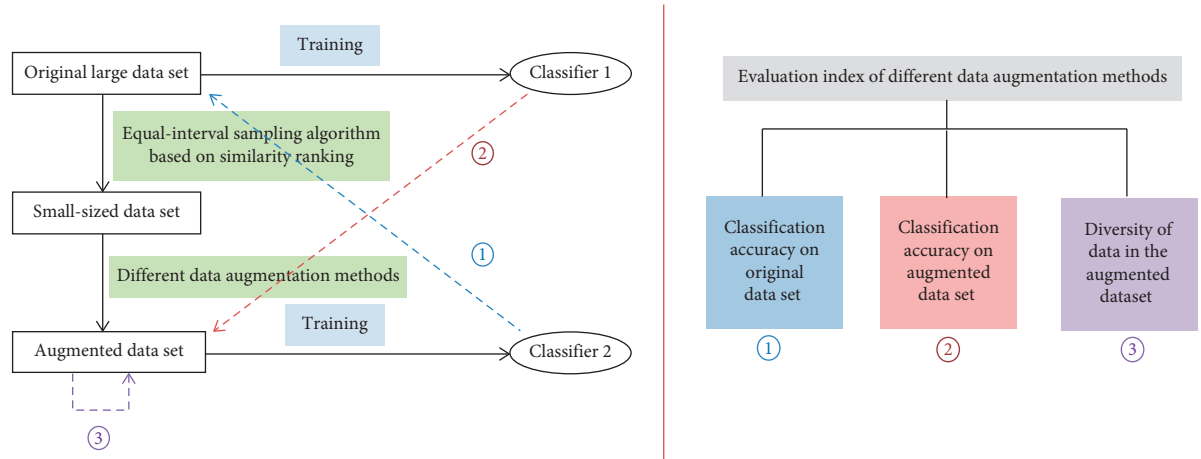

Figure 1: A set of specific evaluation systems for different data augmentation methods.

2.1. Set the Sampling Quantity. For the data of different categories in the original large data set, we adopt the idea of stratified proportional sampling to set the sampling quantity. According to the data proportion of each category in the original data set, we extract the corresponding proportion of samples to form a new small-sized data set. In this way, the problem of focusing on some characteristics or omitting some characteristics caused by simple random sampling can be avoided, and the representativeness of target samples can be improved.

It is assumed that there are k types of data with a total number of M in the original data set, and the data amount of each type is $m_1, m_2, \ldots m_k$ (in general situation, $m_1 \neq m_2 \neq \cdots \neq m_k$). If the total amount of the small-sized data set we want to extract in the end is N (N is typically 20% to 30% of M), the quantity drawn from each category is specified as $n_i = \text{round}(N/Mm_i)$, $i = 1, 2, \ldots, k$.

2.2. Determine the Sample to Be Taken. In Section 2.1, we assume that there are k categories in the original large data set, where the subset of each category is represented as $C_i (i = 1, 2, \ldots, k)$. For the subset of each category C_i, a picture is randomly selected from it, denoted as $X_n, n \in [1, m_i]$, and the similarity between the remaining pictures in this category and picture X_n is calculated successively.

There are many methods to calculate the similarity between pictures. For example, the similarity between two pictures can be determined by obtaining the histogram of two pictures. Although this method is simple and easy to operate and requires little calculation, it can only obtain the similarity of the color information of the two pictures but cannot obtain more information. As long as the color distribution of the two pictures is similar, it is determined that the similarity between the two pictures is high, which is obviously unreasonable; the similarity between two pictures can also be judged by the structural similarity index measurement (SSIM) from the three aspects of brightness, contrast, and structure. Firstly, the image is divided into blocks by using the sliding window and set the total number of blocks is N. Considering the influence of the window shape on the blocks, Gaussian weighting is used to calculate the mean, variance, and covariance of each window. Then the SSIM of the corresponding block is calculated, and finally, the average value is used as the structural similarity measure of the two images. However, this method also has the same disadvantages as using the histogram to judge the similarity. Scholars also propose a method of using deep learning to calculate the similarity of two images [15], which extracts their corresponding feature vectors by putting the two images into a convolutional neural network and then makes a similarity loss function for the two feature vectors at the last layer for network training. Although this method is more objective, it needs to label the data and train the network, which greatly increases the workload. Therefore, to meet the needs of objectivity and a small amount of calculation, the following two methods are used to calculate the picture similarity in this paper:

2.2.1. Calculate the Cosine Distance of the Two Pictures. To calculate the cosine similarity between two images, we first need to represent each image as a vector and then calculate the cosine distance between these two vectors, which is used as an index to measure the difference between two individuals, so as to characterize the similarity of the two images. The closer the cosine value between two vectors is to 1, the closer the included angle is to 0°, that is, the more similar the two vectors (two pictures) are. The specific formula is as follows:

$$\begin{aligned}\cos(\theta) &= \frac{X \bullet Y}{\|X\| \|Y\|} \\ &= \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}},\end{aligned} \tag{1}$$

where X and Y represent two pictures.

2.2.2. Calculate the Euclidean Distance of the Two Pictures. The Euclidean distance is another popular way to determine how similar two images are. Its solution concept is comparable to the concept of cosine similarity. The similarity of two photos can be stated by calculating the actual distance between two locations in N-dimensional space after each image has been represented as a vector. The smaller the Euclidean distance between two vectors, the greater the similarity between two images. The specific formula is as follows:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \quad (2)$$

where X and Y represent two pictures.

For a class subset C_i in the original large data set, after using the above method to successively calculate the distance between the randomly selected picture X_n and all the remaining pictures in the category, we can get a set of specific values, which correspond to the similarity between the remaining pictures in the category and the picture X_n (among them, we set the similarity between picture X_n and ourselves to 0, that is, the most similar). Then, according to the calculated value, the remaining pictures will be from near to far according to the similarity with picture X_n (the closer it is, the more similar it is). Finally, based on the sampling quantity set for each category in Section 2.1, the target pictures are extracted at equal intervals in the sorted picture data set to form a small-sized data set. The "equal-interval extraction" here means that for a class subset C_i, where its total number of pictures before sampling is m_i and the sample size of pictures to be extracted is n_i, we take X_n as the first picture to be extracted and extract every ceil(m_i/n_i) pictures in the picture sequence of sorting number (where ceil $(\cdot)$ means rounding up).

2.3. Assess the Representativeness of the Small-Sized Data Set. In the previous section, we selected two methods to calculate the similarity of images, which are cosine distance and Euclidean distance. As can be seen from Figure 2 below, Euclidean distance measures the absolute distance of each point in space, which is directly related to the position coordinates of each point; cosine distance is a measure of the included angle of space vector, which is more reflected in the difference in direction than position. If the position of point A remains unchanged and point B is away from the origin of the coordinate axis in the original direction, the cosine distance remains unchanged at this time (because the included angle does not change), while the distance between points A and B is obviously changing, which is the difference between Euclidean distance and cosine distance. In conclusion, Euclidean distance can reflect the absolute difference of individual characteristics, while cosine distance is more to distinguish the difference from the direction, but not sensitive to the absolute value.

Euclidean distance and cosine distance have different calculation methods and measurement characteristics. In

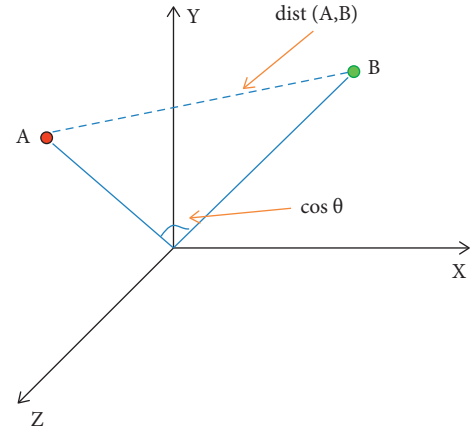

Figure 2: Difference between Euclidean distance and cosine distance.

fact, for different data sets, it is difficult to make a completely fair comparison with a unified standard to measure the similarity between images. Therefore, in order to further determine which small-sized data set extracted by which method can more accurately and truly reflect the original sample set, we perform final screening by evaluating the Frechet Inception Distance (FID) between the small-sized data set after sampling and the original complete set.

FID evaluates the distribution quality of the extracted small-sized data set by calculating the distance between two data sets in the feature space. The formula is as follows:

$$\mathrm{FID}(x,g) = \left\|\mu_x - \mu_g\right\|_2^2 + Tr\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{1/2}\right), \quad (3)$$

where x represents the original large image data set, g represents the small-sized data set after sampling, and μ and $\sum$ represent the mean and covariance, respectively (a probability distribution can be described by its mean and covariance). When two distributions are close, their mean and covariance are also close. Therefore, in this evaluation index, a lower FID value means that the distribution of the small-sized data set after sampling is closer to the distribution of the original large image data set, which is the target small-sized data set we want to obtain.

In general, the steps of equal-interval sampling algorithm based on similarity ranking proposed in this section can be expressed as follows:

Step 1: For a large data set with k categories and the data volume for each category is $sm_1, m_2, \ldots m_k$, the idea of stratified proportional sampling is adopted to set the final sampling quantity $n_i\,(i = 1, 2, \ldots, k)$ for each category

Step 2: Select any picture $X_n, n \in [1, m_i]$ from the subset of each category $C_i\,(i = 1, 2, \ldots, k)$ and calculate the similarity between the remaining pictures in this category and picture X_n by using cosine distance and Euclidean distance, respectively

Step 3: According to the value of similarity, the pictures in each category were arranged in the order of

similarity with picture X_n from near to far (the closer the picture is, the more similar it is)

Step 4: Taking the sampling quantity set for each category in step 1 as the criterion, target images are extracted from the sorted image data set with equal spacing to form a small-sized data set

Step 5: A small-sized data set that corresponds to the distribution of the original data set is finally picked by analyzing the FID between the small sample set sampled by the two similarity solution methods and the original complete set

3. Evaluation Index of Different Data Augmentation Method

After extracting the small-sized data set N_1 that can fully reflect the original data set by the method proposed in the previous section, we expand the data with different data augmentation methods on the basis of the N_1 to generate the augmented data set DA_j (where j represents different data augmentation methods).

In order to objectively and comprehensively evaluate the performance of different augmentation methods, this paper considers the data set after augmentation from three dimensions.

3.1. Classification Accuracy on the Original Data Set Acc_{or}. Accuracy is the most commonly used measure in classification and prediction. It represents the proportion of correctly classified samples in the total samples. In this paper, we take this index as an important evaluation standard to measure the quality of different data augmentation methods.

Specifically, we first select the commonly used classification network with a better classification effect. Next, the augmented data set DA_j obtained by different data augmentation methods is fed into the classification network as training data, and finally, a trained classifier is obtained. Then, the trained classifier is used as the "gold standard" to classify the test set data in the original large data set, and the corresponding classification accuracy is recorded. In order to ensure the reliability of the data, we select three classifiers (VGG16 [3], ResNet50 [16], and Inception-v3 [17]) to classify them and take the average value of the final classification results as $\mathrm{Acc}_{\mathrm{or}}$.

3.2. Classification Accuracy on New Data Set after Augmented Acc_{da}. The second evaluation index in this paper is the same as the first one, which is still the classification accuracy, but the classification accuracy here is for the new data set after the augmentation.

Specifically, we first select the commonly used classification network with a better classification effect. Next, the training sets of the original large data sets are fed into the classification network as training data, and finally, a trained classifier is obtained. Then, the trained classifier is used as the "gold standard" to classify the test set data of DA_j, which is augmented by different data augmentation methods, and the corresponding classification accuracy is recorded. As mentioned above, in order to ensure the reliability of data, we still select three classifiers (VGG16, ResNet50, and Inception-v3) to classify them and average the final classification results and record them as Acc_{da}.

3.3. Diversity of Data in the Augmented Data Set IS. Classification accuracy is indeed an important indicator to evaluate the quality of different augmentation methods. However, for the augmented data sets generated by different augmentation methods, their diversity is also an important index. This paper obtains the clarity and diversity of the augmented data by obtaining the inception score (IS) [18] of the augmented data set.

Inception score generates an N-dimensional output vector y by inputting image x from the augmented data set into a trained model (the Inception-v3 model is used in this paper), where each dimension of the vector represents the probability that the input sample belongs to a certain category.

If an image resolution is very high, it can be more accurately divided into a category, that is, the probability value of belonging to a category should be very high, and the corresponding probability value of belonging to other categories should be very low. At this time, the probability distribution difference between its categories is large, and the entropy of $p(y|x)$ is very small (entropy represents the degree of confusion. The more uniform the distribution, the greater the degree of confusion; on the contrary, the greater the distribution difference, the smaller the degree of confusion).

If the content of pictures in an expanded data set is rich enough, the distribution of pictures in each category should be uniform and will not be biased to the category of a certain feature, that is, the entropy of the marginal distribution $p(y)$ of these pictures in all categories is large.

Combining the above two aspects, the formula of IS is follows:

$$IS(G) = \exp\left(E_{x\sim p_g} D_{KL}(p(y|x)p(y))\right), \tag{4}$$

where D_{KL} represents KL divergence, which is used to measure the distance between two distributions. When the value is larger, the similarity of the two distributions is lower. In IS index, when the distance between $p(y|x)$ and $p(y)$ is large enough, that is, the higher IS value is, the better image quality and richer diversity of the augmented data set will be.

To evaluate the performance of different data augmentation methods, the evaluation index proposed objectively and comprehensively in this paper combines the above three dimensions, and the final total evaluation score TES is as follows:

$$\mathrm{TES} = \mathrm{Acc}_{or} + \mathrm{Acc}_{da} + IS. \tag{5}$$

The higher the value is, the better the data augmentation method is.

4. Experiment Results

This section describes the experimental result of the current work.

4.1. Comparison of Different Sampling Methods. Here, in addition to the two medical imaging data sets, we selected two additional data sets from other fields to verify the broad applicability of our equal-interval sampling algorithm. These data sets are: medical CT image data set (DeepLesion) [19], Pneumonia X-ray image data set [20], ImageNet [21], and CIFAR10 [22], and use six different sampling methods to sample the above two data sets, so as to generate small-sized data sets. The classification accuracy difference between the generated small-sized data set and the original data set, as well as the FID between them, were used as evaluation indexes to evaluate the performance of different sampling methods.

4.1.1. Data Set

(1) DeepLesion. DeepLesion is the largest open data set of multicategory and lesion level labeled clinical medical CT images published by NIH Clinical Center, including 928,020 CT cross-sectional images (512×512 resolution) of 4,427 independent anonymous patients.

(2) Pneumonia X-ray image data set. The Pneumonia X-ray image data set is an open data set from the Guangzhou Women and Children's Medical Center and the University of California team, which contains 2,538 images of bacterial pneumonia and 1,345 images of viral pneumonia.

(3) ImageNet. ImageNet is the world's largest image recognition database currently. It was created by computer scientists at Stanford, California, to imitate a human recognition system. It has around 15 million images and 22,000 categories. In this experiment, we merely use a portion of the ImageNet data as the whole set. As an image database, ImageNet is arranged according to the WordNet hierarchy (currently only the nouns), with hundreds of thousands of pictures depicting each node of the hierarchy. The project has made significant contributions to computer vision and deep learning research. Researchers can use the data for noncommercial purposes for free.

(4) CIFAR10. CIFAR10 is a computer vision data set for pervasive object recognition collected by Alex Krizhevsky et al. It contains 60,000 32×32 RGB color images, with a total of 10 classifications, including 50,000 for training and 10,000 for the test.

4.1.2. Baselines. We compare six sampling methods in this section: (a) random sampling method and (b) stratified sampling method and the sampling algorithm proposed in this paper where (c) histogram, (d) SSIM, € cosine distance, and (f) Euclidean distance are used as indicators in solving image similarity.

4.1.3. Evaluation Metrics

(1) Classification accuracy difference between the generated small-sized data set and the original data set. Firstly, we divide the data in the original large data set into training set $S1$ and test set $T1$. On the one hand, training set $S1$ is used to train a classifier $A1$ (ResNet50 is used in this paper); on the other hand, it is used as the original complete set to generate small-sized data sets (i.e., the small-sized data sets used for data augmentation are sampled from the training set $S1$ in the original large data set). The test set $T1$ prepares for the subsequent verification of classification accuracy.

After making the small-sized data set $N1$ with different sampling methods, we trained another classifier $A2$ with the same classification network (ResNet50). Then the trained $A1$ and $A2$ were tested on test set $T1$ to obtain the two classification accuracy, and their difference was used as an evaluation index. The results are shown in Table 1.

(2) FID between the generated small-sized data set and the original data set. FID evaluates the similarity of two distributions by calculating the distance between two data sets in the feature space. In this experiment, we took the distribution distance of training set $S1$ in the original large data set and small-sized data set $N1$ (we set the data amount of the $N1$ to be 30% of the original sample data amount) by different sampling methods, and the results are shown in Table 2.

The sample method presented in this paper outperforms previous sampling methods in terms of classification accuracy and distribution similarity with the original data set, as shown in Tables 1 and 2. Furthermore, we can see that the methods for resolving picture similarity based on cosine distance and Euclidean distance have various advantages for different data sets. As a result, we undertake final screening by analyzing the FID between the small-sized data set after sampling and the original complete set in order to select which small-sized data set retrieved by the approach can more precisely and truly reflect the original sample set.

4.2. Evaluation Index of Different Data Augmentation Methods. In this section, to prove the universality of our evaluation system, in addition to the medical image data set, we also select another field of data set for the experiment. They are the Pneumonia X-ray image data set [20] and facial expression recognition database (FER2013) [23]. According to the method in Section 3, the data augmentation evaluation system proposed in this paper is used to calculate the overall evaluation scores of different data augmentation methods and then compared them.

4.2.1. Data Set

(1) Pneumonia X-ray image data set. This data set has been described in detail in Section 4.1.

(2) FER2013. The FER2013 data set contains seven expressions: anger, disgust, fear, happiness, sadness, surprise, and

Table 1: Classification accuracy difference between the generated small-sized data set and the original data set.

	DeepLesion	X-ray	ImageNet	CIFAR10
Random sampling method	10.58	8.56	9.08	11.57
Stratified sampling method	9.37	7.39	8.21	9.98
Histogram	8.99	6.85	7.79	8.32
SSIM	7.26	6.04	7.28	8.07
Cosine distance	6.51	5.13	5.17	6.98
Euclidean distance	6.03	4.92	5.93	6.17

Table 2: FID between the generated small-sized data set and the original data set.

	DeepLesion	X-ray	ImageNet	CIFAR10
Random sampling method	74.15	63.58	91.73	84.46
Stratified sampling method	68.34	57.36	88.65	72.53
Histogram	61.27	48.24	73.12	59.87
SSIM	56.75	41.27	68.45	54.13
Cosine distance	46.31	38.14	45.76	41.38
Euclidean distance	48.59	36.83	47.09	39.08

neutrality (marked as 0–6, respectively, during training and testing). Among them, there are 28,708 training images, 3,589 public test images, and 3,589 private test images, and each picture is fixed by size to 48×48 gray image.

4.2.2. Baselines. Firstly, according to the equal-interval sampling algorithm based on similarity ranking proposed in Section 2, we sample a small-sized data set from the X-ray image data set and FER2013, respectively (set the number of samples to 30% of the original data set). Then, for the sampled small-sized data set, different augmentation methods are used to expand the data, and the augmented data set is obtained. According to the data generation method single data transformations (geometric transformations, color space transformations, noise injection, and random erasing), multiple data mixing (the generation of new training data is realized by fusing the spatial or feature information on several images), and learning the data distribution to generate new data are the current data augmentation methods (by learning the potential probability distribution of the data set, taking the whole data set as a priori knowledge, and then sampling it to generate new data by generative adversarial network and image style translation).

In this experiment, several representative data augmentation methods are selected from the above methods, as follows:

(1) Flip. The images in the small-sized data set are flipped along the x-axis or y-axis, respectively, to obtain the image samples with a horizontal or vertical mirror for data augmentation.

(2) Rotate. The image in the small-sized data set is randomly rotated by a certain angle centered on a point (the default is the image center point), and the rotated image samples are obtained for data augmentation.

(3) Color space transformations. Brightness adjustment is performed on each channel of the image in the small-sized data set to generate new available data.

(4) Noise injection. New available data are generated by randomly adding Gaussian noise, gamma noise, and salt and pepper noise to the images in the small-sized data set.

(5) Generate new data through GAN. To augment the data by using the generative adversarial network model [24], an image generation network G needs to be trained based on the original large data set. Then, the trained generation network G is used to generate image samples directly, and the generated samples are added to the small-sized data set, to obtain an augment data set.

(6) Generate new data through CycleGAN. To enlarge tiny data sets using an unsupervised image style translation method, we must first train the cycle generative adversarial network (CycleGAN) model. The CycleGAN is a method for training a deep convolutional neural network to perform tasks of image-to-image translation. Using an unpaired data set, the network learns how to map input and output images. The CycleGAN is trained using two data sets (domain A and domain B) [25]; then take domain A as the source domain and use the trained CycleGAN model to convert it into the image style of domain B, so as to realize the augmentation of domain B data set.

For the X-ray image data set, normal disease-free images (from a hospital CR image database) are used as domain A, and two pneumonia images were used as domain B. After the CycleGAN model was trained, the transformed images were used to expand the small-sized X-ray image data set. Similarly, for the FER2013 data set, we trained the "neutral" class as domain A and the other six classes as domain B (because it is natural to generate emotional faces from nonemotional faces). The trained CycleGAN model is used to expand the other six small-sized data sets. It is worth noting that for the sake of the uniformity of the experimental samples, we removed the "neutral" category from the FER2013 data set in all the experimental tests in this section.

In order to objectively and comprehensively evaluate the performance of the augmented method in different data sets, we use the method proposed in Section 3 to consider the augmented data set from three dimensions: (1) classification accuracy on the original data set Acc_{or}, (2) classification accuracy on new data set after augmented Acc_{da}, and (3) diversity of data in the augmented data set IS. For two data sets, Tables 3 and 4 illustrate a comparison of alternative augmentation strategies. It can be observed that data augmentation methods based on geometric transformations,

Table 3: X-ray image data set.

	Acc_{or}	Acc_{da}	*IS*	TES
Flip	89.86	92.52	0.74	183.12
Rotate	89.15	92.40	0.74	182.29
Color space transformations	90.12	92.65	0.76	183.53
Noise injection	90.36	92.15	0.72	183.23
GAN	90.06	91.35	0.78	182.19
CycleGAN	91.02	91.79	0.81	183.62

Table 4: FER2013 data set.

	Acc_{or}	Acc_{da}	*IS*	TES
Flip	92.47	94.53	0.67	187.67
Rotate	92.58	94.61	0.65	187.84
Color space transformations	93.26	94.35	0.67	188.28
Noise injection	93.12	94.23	0.64	187.99
GAN	92.96	93.85	0.69	187.50
CycleGAN	93.58	94.06	0.69	188.33

such as flipping, rotation, color space transformations, and adding noise, outperform data augmentation methods based on learning data distribution in terms of the accuracy of the new data set after augmentation, but they fall short in terms of the classification accuracy of the original data set and the diversity of data in the augmented data set. As a result, the overall score is slightly lower. As a result, the evaluation system proposed in this paper can not only assess the benefits and drawbacks of existing data augmentation methods in specific medical images or other fields but also provide an effective research direction for new medical image data augmentation methods that are later proposed.

5. Conclusion

This paper proposes an objective and general evaluation system from the two aspects of classification accuracy and data diversity. The evaluation system directs the choice of augmentation strategies for medical images with small data samples and difficult-to-get data. An objective and comprehensive data augmentation evaluation system not only can evaluate the advantages and limitations of the existing augmentation methods in specific medical images classification but also can provide an effective research direction for the subsequent proposed new medical image data augmentation methods. To further promote the development of auxiliary diagnosis technology based on medical image. The experimental results on multiple data sets prove the effectiveness and feasibility of the sampling method and evaluation system proposed in this paper [26–29].

Data Availability

The data supporting this study are from previously reported datasets, which have been cited.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/pdf/1409.1556.

[4] G. Huang, Z. Liu, and L. V. Maaten, "Densely connected convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[5] D. Ma, P. Tang, and L. Zhao, "SiftingGAN: generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1046–1050, 2019.

[6] B. Leng, K. Yu, and J. Qin, "Data augmentation for unbalanced face recognition training sets," *Neurocomputing*, vol. 235, pp. 10–14, 2017.

[7] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 349–360, Springer, Cham, Beijing China, 2018 June, Article ID 10939.

[8] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[9] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, https://arxiv.org/abs/1712.04621.

[10] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," 2017, https://arxiv.org/abs/1711.04340.

[11] A. Choromanska, T. Jebara, H. Kim, M. Mohan, and C. Monteleoni, "Fast scn method," in *Proceedings of the 24th International Conference on Algorithmic Learning Theory*, pp. 367–381, Springer, Columbia, CA, USA, January 2013.

[12] T. A. Hearn and L. Reichel, "Fast computation of convolution operations via low-rank approximation," *Applied Numerical Mathematics*, vol. 75, pp. 136–153, 2014.

[13] M. R. Gajjar, T. V. Sreenivas, and R. Govindarajan, "Fast computation of Gaussian likelihoods using low-rank matrix approximations," in *Proceedings of the IEEE Signal Processing Systems*, pp. 322–327, Beirut, Lebanon, October 2011.

[14] H. Martin, R. Hubert, U. Thomas, N. Bernhard, and H. Sepp, "Gans trained by a two time-scale update rule converge to a nash equilibrium," 2017, https://arxiv.org/abs/1706.08500.

[15] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 23–45, Las Vegas, NV, USA, July 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 34–45, IEEE, Las Vegas, NV, USA, June 2016.

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

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training

GANs," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2226–2234, Barcelona, Spain, December 2016.

[19] Y. Ke, X. Wang, L. Le, R. M. Summers, and K. Yan, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging*, vol. 5, no. 3, pp. 1–12, 2018.

[20] D. K. Kermany and M. Goldbaum, *Labeled Optical Coherence Tomography and Chest X-ray Images for Classification*, Mendeley Data, 2018.

[21] J. Deng, W. Dong, R. Socher et al., "ImageNet: a largescale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, Miami, FL, USA, june 2009.

[22] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Computer Science*, https://www.cs.toronto.edu/%7Ekriz/learning-features-2009-TR.pdf, 2009.

[23] I. J. Goodfellow, D. Erhan, P. L. Carrier et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Neural Information Processing*, pp. 117–124, 2013.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 2672–2680, MIT PRESS, Cambridge, MA, USA, December 2014.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[26] D. Berthelot, N. Carlini, I. J. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: a holistic approach to semi-supervised learning," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, vol. 32, pp. 5050–5060, Vancouver, Canada, December 2019.

[27] V. Verma, A. Lamb, C. Beckham et al., "Manifold mixup: encouraging meaningful on-manifold interpolation as a regularizer," 2018, https://arxiv.org/abs/1806.05236.

[28] Y. Yaguchi, F. Shiratani, and H. Iwaki, "Mixfeat: mix feature in latent space learns discriminative space," 2020, https://openreview.net/forum?id=HygT9oRqFX.

[29] D. Liang, F. Yang, T. Zhang, and P. Yang, "Understanding mixup training methods," *IEEE Access*, vol. 6, pp. 58774–58783, 2018.