



GAN-based one dimensional medical data augmentation

Ye Zhang¹ · Zhixiang Wang² · Zhen Zhang² · Junzhuo Liu¹ · Ying Feng³ · Leonard Wee² · Andre Dekker² · Qiaosong Chen¹ · Alberto Traverso²

Accepted: 27 April 2023 / Published online: 31 May 2023
© The Author(s) 2023

Abstract

With the continuous development of human life and society, the medical field is constantly improving. However, modern medicine still faces many limitations, including challenging and previously unsolvable problems. In these cases, artificial intelligence (AI) can provide solutions. The research and application of generative adversarial networks (GAN) are a clear example. While most researchers focus on image augmentation, there are few one-dimensional data augmentation examples. The radiomics feature extracted from RT and CT images is one-dimensional data. As far as we know, we are the first to apply the WGAN-GP algorithm to generate radiomics data in the medical field. In this paper, we input a portion of the original real data samples into the model. The model learns the distribution of the input data samples and generates synthetic data samples with similar distribution to the original real data, which can solve the problem of obtaining annotated medical data samples. We have conducted experiments on the public dataset Heart Disease Cleveland and the private dataset. Compared with the traditional method of Synthetic Minority Oversampling Technique (SMOTE) and common GAN for data augmentation, our method has significantly improved the AUC and SEN values under different data proportions. At the same time, our method has also shown varying levels of improvement in ACC and SPE values. This demonstrates that our method is effective and feasible.

Keywords Generative adversarial networks · SMOTE · Medical data augmentation · Deep learning · Artificial intelligence

1 Introduction

Radiomics features are widely used by the researchers to quantitative analysis of medical images such as computed tomography (CT) (Bhandari et al. 2021) and magnetic resonance imaging (MRI) (Feng and Ding 2020). However, there are multiple challenges associated with radiomics, in particular, with the radiation pneumonia diagnosis use case. Among them, the most difficult point is the problem of medical data collection.

Collecting medical data is a complex and expensive process, which requires the cooperation of researchers and radiologists, and it often involves some privacy issues.

It is widely recognized that a large number of data samples is required for training a deep learning (DL) model. For example, the ImageNet Dataset (Deng et al. 2009) has 14,197,122 images and the COCO Dataset (Lin, et al. 2014) has about 300,000 images. However, many challenges need to be addressed: a small number of eligible samples in case of a rare disease, the complexity of the data collection process, and the fact that the collected data is often insufficient for model training and may be

Ye Zhang and Zhixiang Wang have contributed equally to this work and share the first authorship.

Qiaosong Chen and Alberto Traverso have contributed equally to this work and share corresponding authorship.

✉ Qiaosong Chen
chenqs@cqupt.edu.cn

✉ Alberto Traverso
alberto.traverso@maastro.nl

¹ Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, Chongqing 400065, People's Republic of China

² Department of Radiation Oncology (Maastr), GROW-School for Oncology, Maastricht University Medical Centre, Maastricht, The Netherlands

³ Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing, China

imbalanced (Wasikowski and Chen 2009; Longadge and Dongre 2015). How to train robust DL models in the presence of limited data is still under investigation.

The oversampling method is one of the most popular methods to address data imbalance. The oversampling technique can increase the number of data sample to achieve a positive and negative sample balance for the training of the model. SMOTE is one of a widely used traditional oversampling methods (Barua et al. 2012; Chawla et al. 2002; Das et al. 2014). It can generate minority class samples along a line segment between the real minority class instances and the nearest minority class neighbors. It is used to effectively expand the few-shot dataset to form a balanced training set for the machine learning (ML) and DL model.

However, under the influence of noisy samples, SMOTE has many limitations (Douzas et al. 2018). For example, the generated data samples may influence the classification accuracy. The process of generating minority class samples using SMOTE does not consider the distribution of majority class samples, and the possibility of overlap between classes increases when samples from class boundaries are used to synthesize new samples.

With the development of computer science, the generative adversarial networks (GANs) (Goodfellow 2014) which consist of a Generator (G) and a Discriminator (D) based on the idea of zero-sum games (Gillies 2016) were proposed. In recent years, many studies have used GANs for medical image data augmentation, such as lung nodule synthesis (Shen et al. 2023; Wang et al. 2022; Tyagi and Talbar 2022) Disease Classification (Chen et al. 2020; Qin et al. 2020; Rashid et al. 2019; Srivastav et al. 2021), and Gastric Cancer Detection (Kanayama et al. 2019). However, GANs are often applied to two-dimensional or three-dimensional medical image data augmentation, but fewer researchers have discussed one-dimensional clinical data augmentation with GANs, like radiomics features (Li et al. 2022) and clinical features. For the 1D clinical data, each individual value has a practical clinical meaning which is significantly different from 2D or 3D image data. Besides, there are still some challenges existing in the 1D clinical data augmentation. GANs tend to overfit the data distribution when augmenting 1D clinical data, ignoring the meaning of each specific value. The radiomics feature extracted from RT and CT images is 1D data. As far as we are concerned, deep learning-based radiomics data augmentation has not been investigated yet, not to mention GAN-based methods. To fill this research gap, we propose a novel 1D Wasserstein Generative Adversarial Network (WGAN) model that can efficiently generate desired samples. The reason is that our model fully takes into account the features of the original data, and it compresses and reconstructs features effectively.

The main contributions of this paper are as follows.

We propose a novel WGAN-GP-based model for augmenting 1D clinical data to address the imbalance problem of the few-shot dataset. Experiments are on the public dataset Heart disease Cleveland (Detrano et al. 1984; Marateb and Goudarzi 2015) and an independent real-world dataset, radiation pneumonitis (Zhang et al. 2022). We demonstrate that our WGAN-GP model can generate data that improves the classification performance of the algorithm compared to SMOTE and common GAN when there are few samples.

2 Related work

It is well known that training a model with unbalanced data can lead to poor performance, especially for predictions in categories with few samples. However, unbalanced data are universal in medical diagnosis. Therefore, it is especially important to choose a proper oversampling method to augment and balance the collected data.

GANs are based on an unsupervised deep learning paradigm. The function of G is to receive random noise Z and generate data samples similar to real data samples, and the D has access to the real and synthetic data instances and tries to tell the difference between them. They are trained by adversarial learning in separate alternating iterations. In the training process, G's goal is to generate as many data samples similar to real data samples as possible to deceive D, while D's goal is to separate the data samples generated by G from the real data samples as much as possible. In this way, G and D constitute a dynamic "game process." GAN was first proposed by Goodfellow et al. (2014). However, the earliest GAN models had shortcomings such as non-convergence, gradient disappearance, training crash, instability, and uncontrollability.

WGAN (Arjovsky and Bottou 2017; Arjovsky et al. 2017) uses Wasserstein distance to measure the distance between the generated distribution and the true distribution, solving the problem of training instability. However, WGAN still has the problems of difficult training and slow convergence. Therefore, Gulrajani et al. proposed WGAN-GP (Gulrajani et al. 2017), which uses gradient penalty instead of weight clipping to achieve the Lipschitz constraint (Cui and Jiang 2017). The training process of WGAN-GP is more stable, and the problem of gradient explosion and disappearance will not occur. The image quality generated by WGAN-GP is also better than WGAN.

GANs have been widely used in the area of medical image data augmentation. Jin et al. (2018) used a 3D GAN to effectively learn lung nodule property distributions in 3D space, enabling GAN to generate lung nodule images.

The robustness of the progressive holistically nested network (P-HNN) model for pathological lung segmentation of CT scans was improved by supplementing the original dataset with the generated images. Bhagat et al. (2019) proposed a data augmentation method using GANs to generate chest X-ray images of pneumonia patients, which significantly improved the classification accuracy of classification models. Kanayama et al. (2019) synthesized gastric cancer images by GANs were used to improve the training set imbalance and trained the gastric cancer detection model using the synthesized images, which led to improved performance of the model. Uzunova et al. (2019) generated large high-resolution 2D and 3D images using GANs. Their scheme enables better image quality and prevents patch artifacts compared to patch-based approaches. In recent years, several studies have proposed the use of GANs for the enhancement of skin disease data to aid in diagnosis (Chen et al. 2020; Rashid et al. 2019; Yang 2021).

GANs are also used to generate one-dimensional medical data or audio. A WGAN model was proposed by Chang et al. (2021). This model can learn the statistical characteristics of the wrist pulse signal and increase the size of the original dataset by generating samples with good fidelity. Lan et al. used the short-time Fourier transform (STFT) to obtain the coefficient matrix from the one-dimensional heart rate signal and then trained the GAN model using different heart rate signal samples to generate samples, which alleviated the problem of insufficient samples for multiple arrhythmias (Lan et al. 2020). A similar study is the WGAN model proposed by Munia et al. (2020) for synthesizing ECG data.

3 Our approach

3.1 Generator

A relatively simple and efficient network structure was designed. The design concept still followed the basic idea of the original WGAN. The generator G was still built with full connection neural networks. The difference is that we reduced the number of channels in each layer, and added Batch Normalization (BN) layer and LeakyRelu activation function between layers. First, the dimension of the input data was expanded to ensure that all of the features can be expressed fully. Then, the features were compressed in order to retain the most effective features, which is the feature extraction process. Sigmoid activation function was used before the final output layer to activate the network output and the network output value was compressed to between 0 and 1. The final network output dimension is

206, which is the characteristic number of our real data samples.

3.2 Discriminator

Similarly, the design of the D follows the idea of original WGAN. However, given the small size of dataset, D employs a simple three-layer full connection neural network, with a LeakyRelu activation function added within layers. It is worth noting that the last layer of our D does not contain any activation function. The dimensions of D's input x and $G(z)$ are 206, while its output dimension is 1. As a result, during the training process, the loss function of the D, i.e., the GP part, was added with a gradient penalty in order to make the D satisfy Lipschitz continuity (Arjovsky and Bottou 2017; Arjovsky et al. 1701; Gulrajani et al. 2017). The advantage of adding the GP part to the loss function is that the D's L2 Norm relative to the original input gradient can be constrained near 1 (bilateral constraint), allowing a better model to be trained. The structure of WGAN-GP network is shown in Fig. 1.

3.3 The training process

In the training process, G's goal is to generate as many false positive data samples similar to real positive data samples as possible to deceive D, while D's goal is to separate the data samples generated by G from the real data samples as much as possible. Therefore, the D was trained first. The specific operation was to feed the D the original real positive data samples and false positive data samples generated by random noise for discrimination and calculated the D's loss value for back-propagation. After that, every five rounds, the G was trained. The specific operation was to feed the false positive sample generated by random noise into the D for discrimination, and then, the G's loss value would be calculated for back-propagation. The model was trained such that the value of the loss function minimizing the D and G is close to 0 or hovering around 0. The training process is shown in Fig. 2.

3.4 Loss function

The Wasserstein distance also called Earth-Mover (EM) distance is used to measure the distance between two distributions. P_r is the distribution of real sample, P_g is the distribution of generated sample. $\Pi(P_r, P_g)$ is the set of all possible joint distributions combined by P_r and P_g . For each possible joint distribution γ , a pair of real data samples x and y can be obtained from it, and the distance $\|x - y\|$ between the two samples can be calculated. Therefore, the expectation value of the real data sample

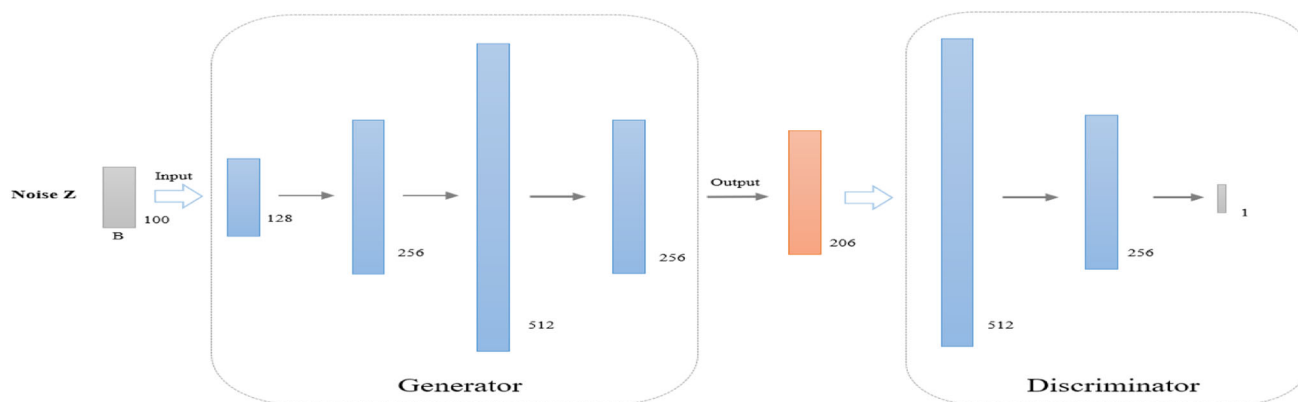


Fig. 1 The structure of WGAN-GP. This structure consists of two parts: Generator and Discriminator. Their composition is full connection neural network, and the number of neurons in each layer can be adjusted dynamically. Generator's input is a vector of random

noise z , and the output is the false positive sample generated by the generator. Then, this output is also used as an input to the discriminator to determine the degree of similarity between the original real data sample



Fig. 2 The training process. This figure depicts our training process from left to right. First, the data are preprocessed. Then, the vector of random noise z was used as input to generate realistic data. The distance between real and generated data was calculated as loss to

train our WGAN-GP model. Finally, the generator of WGAN-GP model can generate a corresponding number of false positive data samples

pair on the distance can be calculated under the γ joint distribution. Among all possible joint distributions, the lower bound of the expectation is the Wasserstein distance. The Wasserstein distance equation is as follows:

$$W(p_r, p_g) = \inf_{\gamma \sim \prod (p_r, p_g)} E_{(x,y) \sim \gamma} [\|x - y\|] \quad (1)$$

The advantage of the Wasserstein distance over the Kullback–Leibler (KL) divergence and the Jensen–Shannon (JS) divergence is that even if the distributions of the two data samples do not overlap or overlap little, it can still reflect their similarity. In this case, JS divergence and KL divergence cannot calculate their similarity in time. According to the Kantorovich–Rubinstein Dual theory, the equivalent form of the Wasserstein distance can be obtained:

$$W(P_r, P_g) = \frac{1}{K} \sup_{\|f\|_L \leq K} E_{x \sim P_r} [f(x)] - E_{x \sim P_g} [f(x)] \quad (2)$$

K is the Lipschitz constant of function $f(x)$. In fact, we don't care about the specific K value, as long as it is not positive infinite, because it will only make the gradient K times larger and will not affect the direction of the gradient. Eq. (2) means that when the Lipschitz constant $\|f\|_L$ of

function f is smaller than K , all possible f are the upper bound of $E_{x \sim P_r} [f_w(x)] - E_{x \sim P_g} [f_w(x)]$, then divided by K . In particular, when using a set of parameters w to define f_w , the above equation can be approximately calculated in the following form:

$$K \cdot W(P_r, P_g) \approx \max_{w: \|f_w\|_L \leq K} E_{x \sim P_r} [f_w(x)] - E_{x \sim P_g} [f_w(x)] \quad (3)$$

In this way, f_w can be defined by using deep neural network and highly approximately satisfies the condition $\sup_{\|f\|_L \leq K}$ required by Eq. (2). Next, when limiting w to a certain range, the discriminative network f_w with parameters w and its last layer is not a nonlinear activation layer can be constructed. Here is the D's loss formula:

$$L = E_{x \sim P_r} [f_w(x)] - E_{x \sim P_g} [f_w(x)] \quad (4)$$

In this way, the loss functions of the G and D are obtained separately:

$$-E_{x \sim P_g} [f_w(x)] \quad (5)$$

$$E_{x \sim P_g} [f_w(x)] - E_{x \sim P_r} [f_w(x)] \quad (6)$$

3.5 The function of the GP part

The weight clipping strategy is to limit the weight of the D to a range, such as $[-0.01, 0.01]$, to ensure that the weight of the D will not change significantly. Therefore, the condition of Lipschitz continuity will be satisfied. However, the weight clipping strategy has two limitations. One is the D's parameters can easily be taken as 0.01 or -0.01 . As a result, the strong fitting ability of D is wasted. The another is that the gradient clipping strategy can easily lead to gradient explosion.

In fact, GP part on D and weight clipping strategy play the same role. The GP part is the gradient constraint. Based on the above, the GP part can be defined as:

$$\lambda E_{x \sim P_x} \left[\left(\|\nabla_x f_w(x)\|_2 - 1 \right)^2 \right] \quad (7)$$

The λ is a hyperparameter and the penalty is on the random samples $x \sim P_x$. $\|\nabla_x f_w(x)\|_2$ is the L2 norm of the x 's gradient, and it will be constrained around 1. For D, the L2 norm of the x 's gradient will be constrained around 1 with the training process. Therefore, the GP part only takes effect in the area where the true and false samples are concentrated and the transition zone in the middle of the true and false samples. As a result, the gradient is very controllable and easy to adjust to the appropriate scale. Therefore, the GP part can significantly improve the training speed and solve the problem of slow convergence of the original WGAN. It should be noted that when D's loss value propagates back, the value of GP part needs to be added to the D's loss value. During the training, the curve of loss function values of GAN and WGAN-GP is shown in Fig. 3. In Fig. 3, the common GAN uses the sum of the loss values of G and D as the reference criterion to judge whether the model converges or not. However, it is different in WGAN-GP, because the loss value of G has no meaning in WGAN-GP algorithm training process. The

measurement standard of model preservation comprehensively considers the sum of the loss value of D and the distance of W. When the absolute value of their sum is the smallest, it is saved as the optimal model.

3.6 Dataset

3.6.1 Heart disease Cleveland

The dataset contains 76 attributes, but 14 of them are mentioned in the published experiments. Up to now, it is also the most used dataset for researchers to predict heart disease. The "target" field indicates whether the patient has heart disease or not. It is an integer value, with 0 indicating a low risk of heart disease and 1 indicating a high risk.

3.6.2 Radiation Pneumonitis dataset

This is a real-world dataset containing 300 patients, of which 66 patients had RP (22%). The two types of data we analyzed were radiomics features extracted from CT images and RD radiotherapy planning dose files. These parameters are considered to have potential predictive power for RP in a clinical setting. Because the original information in the dataset does not meet the format that can be read by the computer, we transformed the original data into numerical features and normalize it as the pre-processing method.

3.7 Evaluation method

Distributed Stochastic Neighbor Embedding (TSNE) (Maaten and Hinton 2008) is a machine learning nonlinear dimensionality reduction algorithm, suitable for reducing high-dimensional data to 2 or 3 dimensions, especially for

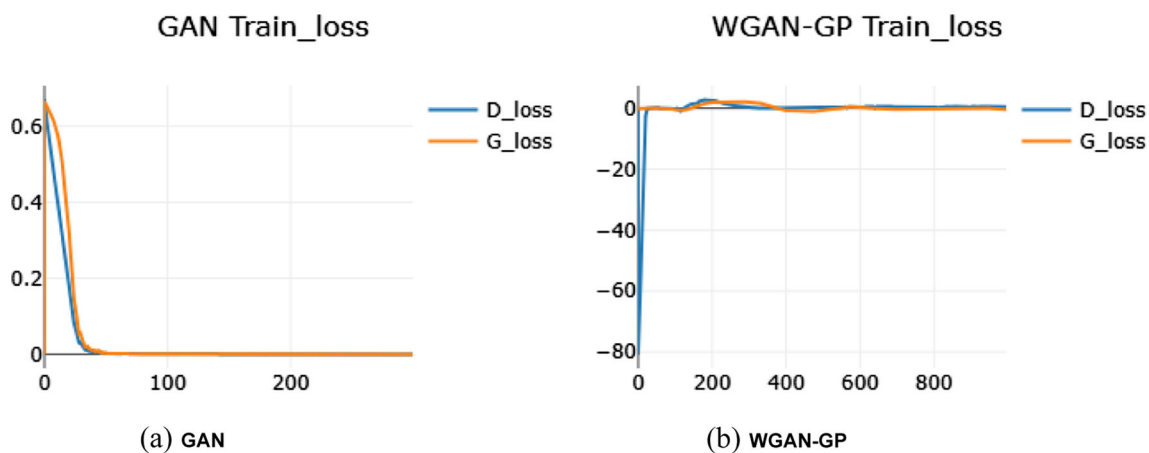


Fig. 3 G and D Loss values

data with different distributions. It can measure their similarity in lower dimensions after performing visualization.

Except for raw data without any processing, there were three data upsampling methods, WGAN-GP, SMOTE, and GAN used for ML tenfold cross-validation. The original dataset was divided into ten equal parts, logistic regression models were trained in different proportions and tested on the corresponding test sets to compute the performance metrics including the Area Under the ROC Curve (AUC), Accuracy (ACC), Sensitivity (SEN), and Specificity (SPE) values for various data samples. Their definitions are as follows.

Suppose we have four types of samples: True positive (TP) is classified as a positive being a real positive sample. True Negative (TN) is determined to be a negative sample, being a real negative sample. False Positive (FP) is determined to be a positive sample, but instead being a negative sample. False Negative(FN) is determined to be a negative sample, but instead being a positive sample. The ACC, SEN, and SPE can be defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$SEN = \frac{TP}{TP + FN} \quad (9)$$

$$SPE = \frac{TN}{TN + FP} \quad (10)$$

For example, if the original data were divided into seven training sets and three test sets, three upsampling methods were applied to train the model on the seven training sets and test on the other three test sets. The above performance evaluation metrics were calculated and compared. The testing process is shown in Fig. 4.

4 Experimental results

4.1 Experimental Setting

In the training process, there was no change in the way training D and G. However, to ensure the synthetic data is similar to the original real data, the G and D training epochs were set to 1000 and 200, respectively. The specific

hyperparameters were set as: Epochs = 1000, Learning rate (Lr) = 0.0002, Batch_size = 16, Latent_dim = 100, and Lambda_gp = 100. (Note that, at the training time, our Epoch here is equal to Batch_size, only one Batch is trained in each Epoch.) The Latent_dim is the number of hidden layer neurons, and the Lambda_gp is the gradient penalty weight coefficient. The sum of the absolute values of the D loss value and the Wasserstein distance was used as the evaluation value. Wasserstein distance is used to measure the distance between two distributions. If the distributions of the two data samples do not overlap or overlap little, it can still reflect their similarity. The model with the best performance on the validation dataset was saved.

4.2 Comparison of ML improvement

Comparison experiments were conducted on the public dataset and the lung dataset, respectively, to verify the performance of the models. The logistic regression models were trained and evaluated by four type datasets. They are the dataset generated by trained WGAN-GP, the dataset generated by the SMOTE, the dataset generated by the trained GAN, and the unprocessed real data. For each of them, the training set and test set are at the ratio of 2:8, respectively. Each logistic regression model was tested 10 times, and the average AUC score was calculated.

In Table 1, the average AUC scores and standard deviation were calculated for each method in two datasets. On the Heart disease Cleveland dataset, the AUC was 0.902 ± 0.016 with WGAN-GP, 0.874 ± 0.019 with SMOTE, 0.877 ± 0.023 with real data, and 0.837 ± 0.023 with GAN. On the radiation pneumonitis dataset, the AUC was 0.606 ± 0.009 with WGAN-GP, 0.585 ± 0.012 with SMOTE, 0.584 ± 0.015 with real data, and 0.572 ± 0.014 with GAN. The standard deviation of WGAN-GP is smaller than other methods, regardless of in Heart disease Cleveland dataset or Radiation Pneumonitis dataset. The statistical test was also conducted in the experiment. The P value was 0.498 with WGAN-GP and SMOTE, 0.232 with WGAN-GP and real data, and 0.440 with WGAN-GP and GAN. Although there is no statistically significant difference, WGAN-GP is significantly better than other methods. Therefore, it can be concluded that the data generated by

Fig. 4 Testing processing

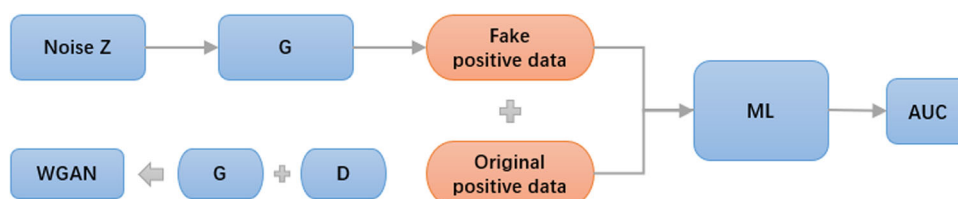
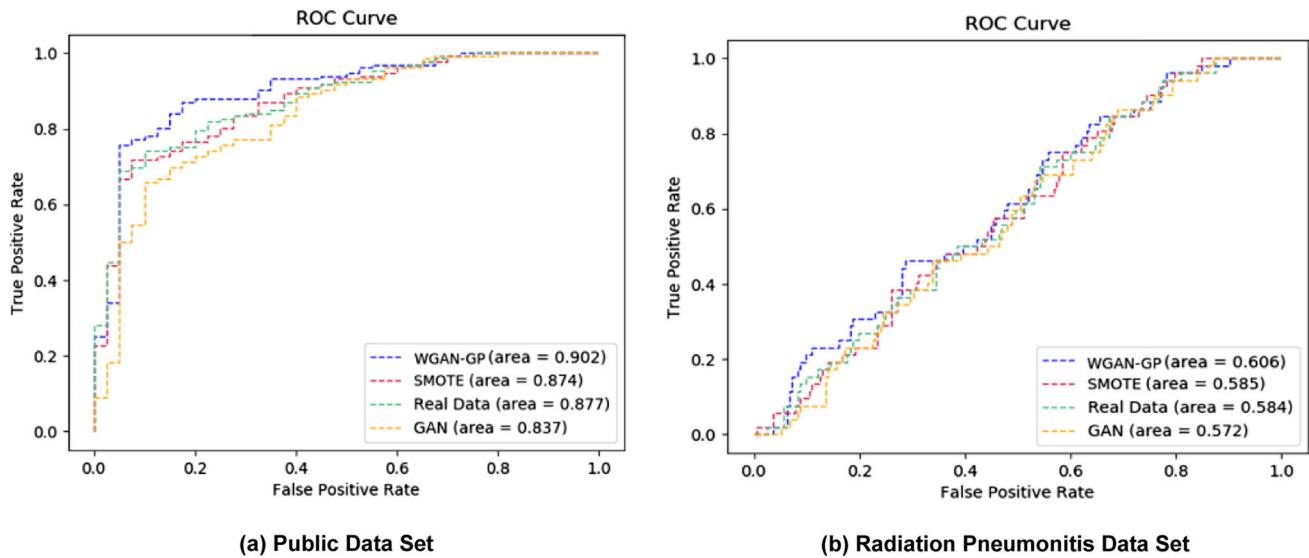


Table 1 The AUC $\pm$ Standard Deviation results of logistic regression model on WGAN-GP, SMOTE, GAN, and unprocessed real data

AUC $\pm$ SD	GAN	SMOTE	WGAN-GP	Real data
Heart disease Cleveland	0.837 ± 0.023	0.874 ± 0.019	0.902 ± 0.016	0.877 ± 0.023
Radiation Pneumonitis	0.572 ± 0.014	0.585 ± 0.012	0.606 ± 0.009	0.584 ± 0.015

**Fig. 5** Comparison of ROC curves under different datasets

WGAN-GP are more stable and have smaller variance and the logistic regression classifier trained on synthetic data generated by WGAN-GP has a better classification performance.

In addition, the ROC curve comparison graph is shown in Fig. 5. In Fig. 5a and b, the ROC curves of the logistic regression models trained with the four methods on the public and lung datasets were plotted, respectively. From Fig. 5a,b, it is significant that the performance of WGAN-GP is better than other methods in both two datasets. Therefore, compared with SMOTE and GAN, the synthetic data generated by WGAN-GP have greater capabilities to improve the performance of ML.

4.3 Synthetic data visualization

The synthetic data generated by WGAN-GP, GAN, and SMOTE were used to train the classification model, with the ratio of training set: test set of 7:3. To illustrate whether the data generated by WGAN-GP can better reflect the real distribution characteristics, TSNE was selected to compare the distribution of real and generated data. The results are shown in Fig. 6.

In the reduced dimension distribution map on the public data, TP denotes the true positive data samples, WGAN-GP_FN denotes the fake negative data samples generated by the WGAN-GP, TN denotes the true negative data samples, and SMOTE_FN denotes the fake negative data samples generated by SMOTE and GAN_FN denotes the fake negative data samples generated by the GAN. From Fig. 6a, it can be observed that the fake negative data generated by the WGAN-GP are closer to the real negative data and have a wider distribution compared with GAN. On the contrary, compared with SMOTE, the data distribution generated by WGAN-GP is more introverted. Therefore, it is easy to conclude that WGAN-GP can better reflect the distribution of real data samples and has a larger potential sample space.

In the reduced dimension distribution map of lung data, TP denotes the true positive data samples, WGAN-GP_FP denotes the fake positive data samples generated by the WGAN-GP, TN denotes the true negative data samples, SMOTE_FP denotes the fake positive data samples generated by the SMOTE and GAN_FP denotes the fake positive data samples generated by the GAN. From Fig. 6b, the fake positive data generated by WGAN-GP are concentrated near the true positive data, while the fake positive

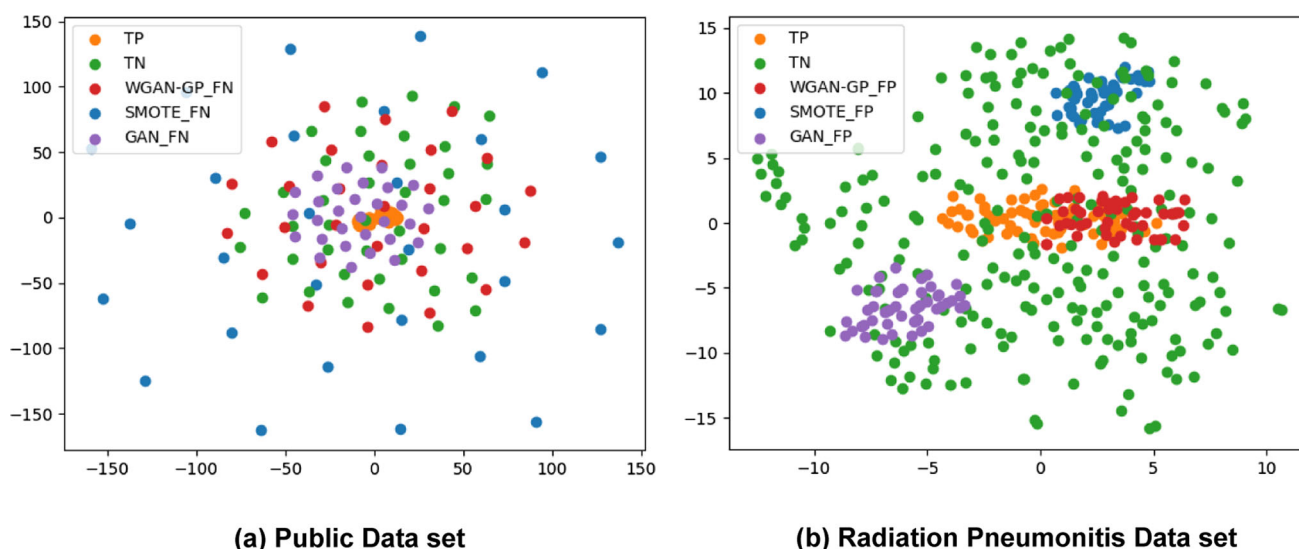


Fig. 6 The generated data distribution comparison chart

data distributions generated by SMOTE and GAN are farther from the distribution of true positive data, which proves that WGAN-GP is more consistent with the distribution close to the original distribution of true positive data samples than SMOTE and GAN.

In short, the distribution characteristics of data samples generated by WGAN-GP are more consistent with the original real data samples, whether on public dataset or private dataset. In contrast, the performance of the other two data augmentation methods is flawed.

4.4 Comparison under different samples sizes

The lung dataset was divided into 10 copies to form the training and test sets in different proportions. Each ratio was trained 10 times using each of the four methods. This was used to verify whether the data generated by the WGAN-GP method have better classification performance. The results are shown in Fig. 7.

In Fig. 7a, the navy-blue line is the AUC score of the logistic regression model trained by synthetic data generated by WGAN-GP. The yellow line is the AUC score of the logistic regression model trained by synthetic data generated by SMOTE. The light blue line is the AUC score of the logistic regression model trained by synthetic data generated by GAN, and the red line is the AUC score of the logistic regression model trained by real data. Similarly, in the three diagrams Fig. 7b–d, they correspond to the performance of ACC, SEN and SPE, respectively.

Figure 7 shows that the improvement of the classification method among WGAN-GP, SMOTE and GAN becomes smaller as the proportion of training data

increases. Because the improvement of upsampling is limited when there is enough data. Moreover, the improvement of WGAN-GP is much larger than that of SMOTE and GAN when the training set is less than 30%. Finally, the AUC and SEN obtained by WGAN-GP are higher than that of SMOTE, No Up-sampling and GAN at all scales. Additionally, the ACC and SPE values also showed varying degrees of improvement under most proportions.

The experimental results show that the data generated by WGAN-GP have better performance in improving the classification performance under various distribution ratios of the training set and test set, which is higher than the traditional method SMOTE and common GAN. It is worth noting that the performance of WGAN-GP is better and more prominent when the training set data are below 30%. It can be concluded that the data generated by WGAN-GP are more stable and have smaller variance and the logistic regression classifier trained on synthetic data generated by WGAN-GP has a better classification performance, especially in the small size of training dataset.

5 Conclusion and future work

5.1 Conclusion

In this paper, a data augmentation method using WGAN-GP for the few-shot imbalance dataset is proposed, which can suit one-dimensional clinical and radiomics data. Compared with SMOTE and common GAN, the WGAN-GP has best performance. Meanwhile, ML cross-validation

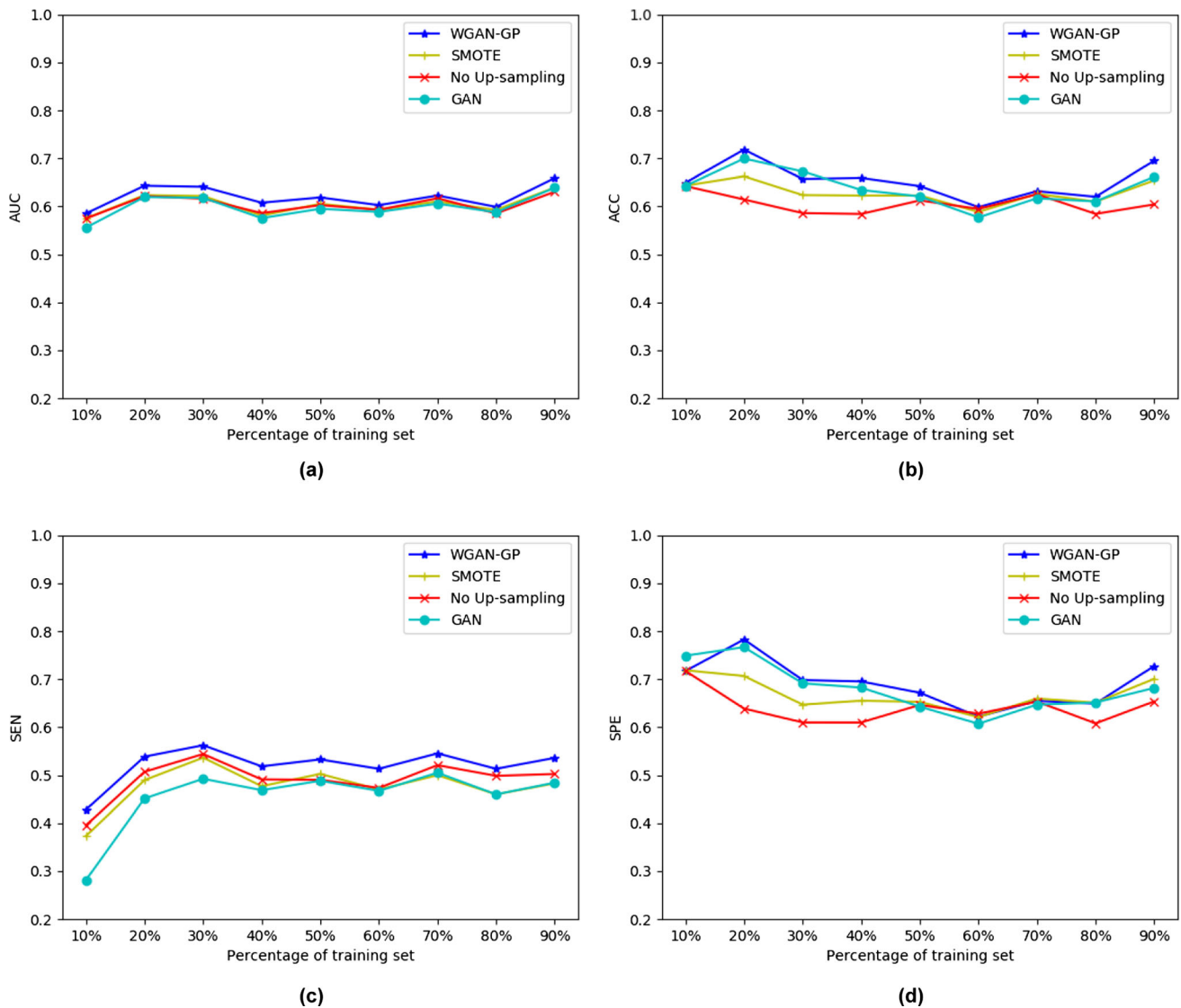


Fig. 7 AUC, ACC, SEN and SPE at different ratio

and TSNE are used for the visualization of data upsampling, and each evaluation method has obtained excellent results.

Therefore, it can be concluded that the synthetic data generated by WGAN-GP can improve the ability of the classification model when the training data are insufficient and unevenly distributed, and the data generated by WGAN-GP are closer to the real sample distribution than the data generated by SMOTE and GAN. It can be considered that WGAN-GP is more suitable for data generated from the few-shot one-dimensional clinical and radiomics datasets.

5.2 Future work

In future, it is worthwhile to further study the application of WGAN-GP on the few-shot one-dimensional dataset expansion and optimize more effectively for the current algorithm to achieve even better performance. Secondly, our experiments suggest that the training of common GANs is really difficult. Especially when the dataset sample is small and the feature dimension is low, the performance of GANs will be significantly reduced. Therefore, future research will try to propose a novel GAN model while improving the performance of WGAN-GP algorithm, which can perfectly solve the above problems.

Acknowledgements We thank the Chinese Scholarship Council (CSC) for their financial support for studying abroad.

Author contributions YZ and ZW designed the study and wrote the article. ZZ, JL, and FY helped to analyze pre-processing the dataset. LW, AD, QC, and AT were the administrative support. All authors read and approved the final manuscript.

Funding No funding.

Data availability Dataset 1 Heart Disease Cleveland is publicly available and can be found in the <http://archive.ics.uci.edu/ml/datasets/Heart> +Release access. Anyone can freely access, use, and share this dataset without the need for additional permission or approval. Dataset 2 Radiation Pneumonitis dataset is a private dataset that requires specific authorization to obtain access permissions. If you are interested in dataset 2, you can apply for access to the dataset by sending an email to the corresponding author. Please explain your research purpose and intended use. The corresponding author will evaluate your request and provide you with access to the dataset after approval. We are committed to protecting the privacy and security of data and encouraging reasonable and responsible data use.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Each author gives permission to submit this article.

Informed consent Each author is informed and agrees to submit this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arjovsky, M. and L. Bottou, Towards Principled Methods for Training Generative Adversarial Networks. *Stat*, 2017. 1050.
- Arjovsky M, Chintala S, Bottou L. (2017) Wasserstein gan. *arXiv* 2017. *arXiv preprint* [arXiv:1701.07875](https://arxiv.org/abs/1701.07875) 30(4).
- Barua S et al (2012) MWMOTE—majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans Knowl Data Eng* 26(2):405–425
- Bhagat V, Bhaumik S (2019) Data augmentation using generative adversarial networks for pneumonia classification in chest X-rays. In: 2019 Fifth international conference on image information processing (ICIIP). IEEE, pp 574–579
- Bhandari A, Ibrahim M, Sharma C et al (2021) CT-based radiomics for differentiating renal tumours: a systematic review. *Abdom Radiol* 46(5):2052–2063
- Chang J, Hu F, Xu H, et al. (2021) Data Augmentation of Wrist Pulse Signal for Traditional Chinese Medicine Using Wasserstein GAN. In: Proceedings of the 2nd international symposium on artificial intelligence for medicine sciences, pp 426–430
- Chawla NV et al (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
- Chen Y, Zhu Y, Chang Y (2020) CycleGAN based data augmentation for melanoma images classification. In: Proceedings of the 2020 3rd international conference on artificial intelligence and pattern recognition, pp 115–119
- Cui S, Jiang Y (2017) Effective Lipschitz constraint enforcement for Wasserstein GAN training. In: 2017 2nd IEEE international conference on computational intelligence and applications (ICCI), IEEE
- Das B, Krishnan NC, Cook DJ (2014) RACOG and wRACOG: two probabilistic oversampling techniques. *IEEE Trans Knowl Data Eng* 27(1):222–234
- Deng J, Dong W, Socher R, et al. (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
- Detrano R, Yiannikas J, Salcedo EE et al (1984) Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation* 69(3):541–547
- Douzas G, Bacao F, Last F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf Sci* 465:1–20
- Feng Q, Ding Z (2020) MRI radiomics classification and prediction in Alzheimer's disease and mild cognitive impairment: a review. *Curr Alzheimer Res* 17(3):297–309
- Gillies DB (2016) Solutions to general non-zero-sum games. In: Contributions to the theory of games (AM-40), Volume IV. Princeton University Press, Princeton pp 47–86
- Goodfellow I et al. (2014) Generative adversarial nets. *Advances in neural information processing systems* 27
- Gulrajani I, Ahmed F, Arjovsky M, et al. (2017) Improved training of wasserstein gans. *Adv Neural Inf Process Syst* 30
- Jin D, Xu Z, Tang Y, et al. (2018) CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 732–740
- Kanayama T, Kurose Y, Tanaka K, et al. (2019) Gastric cancer detection from endoscopic images using synthesis by GAN. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 530–538.
- Kanayama T, Kurose Y, Tanaka K, et al. (2019) Gastric cancer detection from endoscopic images using synthesis by GAN. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 530–538
- Lan T, Hu Q, Liu X, et al. (2020) Arrhythmias Classification Using Short-Time Fourier Transform and GAN Based Data Augmentation. In: 2020 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 308–311
- Li G, Li L, Li Y et al (2022) An MRI radiomics approach to predict survival and tumour-infiltrating macrophages in gliomas. *Brain* 145(3):1151–1161
- Lin TY et al (2014) Microsoft COCO: common objects in context. Springer International Publishing, Cham
- Longadge R, Dongre S (2013) Class imbalance problem in data mining review. *arXiv preprint* [arXiv:1305.1707](https://arxiv.org/abs/1305.1707)
- Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(11)
- Marateb HR, Goudarzi S (2015) A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system. *J Res Med Sci* 20(3):214–223
- Munia MS, Nourani M, Houari S (2020) Biosignal oversampling using wasserstein generative adversarial network. In: 2020 IEEE

- International conference on healthcare informatics (ICHI). IEEE, 1–7
- Qin Z, Liu Z, Zhu P et al (2020) A GAN-based image synthesis method for skin lesion classification. *Comput Methods Programs Biomed* 195:105568
- Rashid H, Tanveer MA, Khan HA (2019) Skin lesion classification using GAN based data augmentation. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 916–919
- Shen Z, Ouyang X, Xiao B et al (2023) Image synthesis with disentangled attributes for chest X-ray nodule augmentation and detection[J]. *Med Image Anal* 84:102708
- Srivastav D, Bajpai A, Srivastava P (2021) Improved classification for pneumonia detection using transfer learning with gan based synthetic image augmentation. In: 2021 11th International conference on cloud computing, data science and engineering (Confluence). IEEE, pp 433–437
- Tyagi S, Talbar SN (2022) CSE-GAN: A 3D conditional generative adversarial network with concurrent squeeze-and-excitation blocks for lung nodule segmentation. *Comput Biol Med* 147:105781
- Uzunova H, Ehrhardt J, Jacob F, et al. (2019) Multi-scale gans for memory-efficient generation of high resolution medical images. In: international conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 112–120
- Wang X, Yu Z, Wang L, et al. (2022) An enhanced priori knowledge GAN for CT images generation of early lung nodules with small-size labelled samples. *Oxid Med Cell Longev*
- Wasikowski M, Chen X-W (2009) Combating the small sample class imbalance problem using feature selection. *IEEE Trans Knowl Data Eng* 22(10):1388–1400
- Yang Y (2021) Data augmentation to improve the diagnosis of melanoma using convolutional neural networks. In: Proceedings of the 2021 international conference on bioinformatics and intelligent computing, pp 151–158
- Zhang Z, Wang Z, Yan M, et al. (2022) Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: A model development study with prospective external validation and decision-curve analysis. *Int J Radiat Oncol Biol Phys*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.