Computers in Biology and Medicine 135 (2021) 104540

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data

Yawen Xiao [a,*], Jun Wu [b], Zongli Lin [c]

[a] Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China
[b] The Center for Bioinformatics and Computational Biology, East China Normal University, Shanghai, 200241, China
[c] Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, 22904-4743, USA

ARTICLE INFO

Keywords:
Cancer diagnosis
Deep learning
Gene expression data
Imbalanced data
Wasserstein generative adversarial networks

ABSTRACT

Background and objective: Cancer is a serious global disease due to its high mortality, and the key to effective treatment is accurate diagnosis. However, limited by sampling difficulty and actual sample size in clinical practice, data imbalance is a common problem in cancer diagnosis, while most conventional classification methods assume balanced data distribution. Therefore, addressing the imbalanced learning problem to improve the predictive performance of cancer diagnosis is significant.
Methods: In the study, we dissect the data imbalance prevalent in cancer gene expression data and present an improved deep learning based Wasserstein generative adversarial network (WGAN) model, which provides a reliable training progress indicator and deeply explores the characteristics of data. The WGAN generates new samples from the minority class and solves the imbalance problem at the data level.
Results: We analyze three publicly available data sets on RNA-seq of three kinds of cancer using the proposed WGAN and compare the results with those from two commonly adopted sampling methods. According to the results, through addressing the data imbalance problem, the balanced data distribution and the expanding sample size increase the prediction accuracy in all three data sets.
Conclusions: Therefore, the proposed WGAN method is superior in solving the imbalanced learning problem of gene expression data, providing significantly better prediction performance in cancer diagnosis.

1. Introduction

Accuracy in disease diagnosis is of vital importance, which is especially so for cancer diagnosis. With the development of computer-aided technology, cancer diagnosis has undergone tremendous improvement. With the aid of computer-aided diagnostic equipment, physicians can identify the medical condition of patients more clearly and make better treatment decisions. Meanwhile, the increasing use of machine learning methods in the clinical field has further improved the accuracy of detection and prediction, which has created an enormous opportunity for applications to cancer diagnosis [1].

Recently, the imbalanced learning problem associated with high performance machine learning based methods for cancer diagnosis is attracting an increasing amount of attention [2,3]. Most conventional machine learning methods assume that the distribution of data across different categories is balanced and do not consider the misclassification cost in a general classification process. However, the problem of imbalanced learning is not negligible in practice. The data imbalance problem generally manifests itself in two scenarios, the proportion of tumor samples is smaller than the proportion of normal samples, and, the other way around, the proportion of normal samples is smaller than the proportion of tumor samples. The imbalance distribution may cause the problem in classification methods that many samples are misclassified into the majority class [4]. In either case, there are adverse effects of misdiagnosis on cancer patients. For instance, if a cancer patient is misdiagnosed as normal, it may lead to delay in correct diagnosis and treatment.

In most of the gene expression data in the TCGA (The Cancer Genome Atlas) database [5], which is one of the most commonly used databases in cancer research, there is also such a data imbalance problem. The problem usually shows that the normal samples are less than the tumor samples, leading to a large number of normal samples being misclassified as tumor samples, causing unnecessary panic and excessive cost and treatment. Therefore, the detection failure caused by the data

[*] Corresponding author.
E-mail addresses: foreverxyw@sjtu.edu.cn (Y. Xiao), junwu302@gmail.com (J. Wu), zl5y@virginia.edu (Z. Lin).

https://doi.org/10.1016/j.compbiomed.2021.104540
Received 22 December 2020; Received in revised form 14 May 2021; Accepted 26 May 2021
Available online 12 June 2021

imbalance problem alone, rather than the detection failure caused by the classifiers, will lead to mistakes in the decision-making of treatment and overwhelming cost. These consequences are particularly serious in cancer diagnosis.

The data imbalance problem has been addressed in related literature and research, and many imbalanced learning methods have been developed, which are discussed in the next section.

2. Literature review

In this study, we focus on the data imbalanced learning problem at the data level. For data imbalanced learning problems, the state-of-the-art solution at the data level is sampling, which uses some mechanisms to re-balance the data distribution, such as undersampling or oversampling [6]. Research shows that for a general classifier, a balanced data distribution provides better predictive performance than an imbalanced data distribution [7].

Among sampling methods, the two most commonly used are random sampling method, and the synthetic minority oversampling technique, or SMOTE [8]. The random oversampling method selects several samples randomly from the minority class, replicates the selected samples and adds them to the original data. The random undersampling, as its name suggests, randomly reduces data or discards samples. The SMOTE method synthesizes new samples and adds them to the minority class. As a learning algorithm, SMOTE randomly selects a k-nearest neighbor of every element of the minority class and superimposes the corresponding feature vector to generate a new sample. Liu and Zeng [9] proposed a breast cancer diagnosis model for mammographic images. They incorporated the oversampling and undersampling methods to rescale the sample distribution between the non-masses and masses to avoid the large effects of data imbalance on the diagnosis accuracy. Shanab et al. [10] analyzed the impact of sampling and the effect of noise on the feature selection stability with six versions of data sampling on four cancer microarray datasets. The results showed that the random oversampling and the SMOTE obtained the best performance on average. Wang and Adrian [11] combined the artificial immune recognition system with the SMOTE to address the imbalanced learning problem by generating minority class. The experiments were conducted on two data sets of breast cancer. Their results revealed enhancements in classification performance of breast cancer as benign and malignant due to the over-sampling process. Bunkhumpornpat et al. [12] proposed an improved SMOTE method, DBSMOTE, which relies on a cluster to better deal with the boundary samples and improve the minarity class detection rates. However, the random sampling method can only increase the sample size but does not have the ability to learn the characteristics and diversity behind the data, which may lead to over-fitting and data skew. The techniques in the SMOTE faminly are generally shallow learning algorithms, whose randomness may lead to insufficient learning and variability of the generated data. Even some advanced SMOTE techniques, such as DBSMOTE and MWMOTE [13], are limited by the SMOTE algorithm itself. For example, they cannot learn enough characteristics from too a small sample size, and need more powerful classifiers to further ensure their effect on classification.

Compared with shallow learning algorithms, deep learning has been considered as an efficient and accurate algorithm and is increasingly used in various fields, including biomedicine [14]. In the deep learning techniques, the generative adversarial networks (GANs), as a sample generation and discriminant algorithm, have sprung up in computer vision [15]. We have noticed, in the previous studies, the powerful self-learning ability of GAN algorithms for data feature and distribution specificity through successive iteration and adversarial learning of the discriminator and the generator. This deep learning ability will bring better results to the imbalanced learning problem. Shin et al. [16] employed a GAN model and synthesized tumor MRI images to address the imbalance problem in the brain MRI data sets, achieving improved performance on tumor segmentation results. Wang et al. [17] explored the cutting-edge Wasserstein GANs to address the data imbalance problem of lung nodules images. In recent years, more and more GAN models have been used for data augmentation [18–20]. However, the general GAN model also has some drawbacks. For example, the loss of the generator and that of the discriminator do not indicate the training progress and the optimization is difficult to achieve. Moreover, in general GANs or WGANs, the convolutional neural network (CNN) architecture is usually used. But it is not appropriate for the numerical cancer gene expression data, which is not as intuitive as the image data. Therefore, the GAN model still needs to be improved to adapt to its application to cancer diagnosis.

In this study, we attempt to apply an improved, deep learning based, GAN to deal with the imbalanced learning problem in cancer diagnosis due to the limited number of cancer sampling groups and the difficulty of sampling in clinical practice. Firstly, we use the Wasserstein distance (W distance) instead of the Jensen-Shannon (JS) divergence in GANs to provide useful optimization gradients in each iteration to serve as a reliable indicator of the training progress, so that GANs can be trained more fully to achieve their optimization. Secondly, this is the first time the GAN algorithm has been applied to cancer gene expression data, which provides a new idea and possibility for deep learning in small sample numerical data and cancer diagnosis. To better apply to gene expression data, we improve the architecture of the WGAN using deep, fully-connected networks instead of CNNs. Thirdly, different from the conventional concept and application of GANs in the field of computer vision, where the GANs are used to expand the quantity of data, we pay more attention to the quality of the data generated as reflected in the follow-up classification tasks. The ability of the generated data for solving the imbalanced learning problem and improving prediction accuracy is evaluated. Moreover, we address the imbalanced learning problem at the data level, which will not affect the running time of the online classification model, making the classification model more efficient and more targeted to deal with the online cancer diagnosis. We compared the improved WGAN method with several commonly used sampling algorithms, more than these two on three sets of RNA-seq data derived respectively from breast, lung and stomach issues. The results show that better prediction performance can be obtained by balancing the data distribution and expanding the sample size, and our proposed WGAN based imbalanced learning method is well suited for gene expression data and shows the superiority in addressing the imbalanced learning problem at the data level in cancer diagnosis.

3. Methods

A flowchart of the proposed deep learning based imbalanced learning cancer diagnosis method is shown in Fig. 1. The data is first preprocessed by differential gene expression analysis. The Wasserstein generative adversarial networks (WGANs) are then applied in the imbalanced learning stage to generate samples to deal with the imbalanced training data. Then the processed balanced training data is used to train and validate classification models, which are provided for test data to obtain the final cancer diagnosis results.

3.1. Data preprocessing

With the advance in gene sequencing technology, more and more gene expression data has been generated. Gene expression data is numerical data, in which each row corresponds to a tissue sample, each column (feature) corresponds to a gene, and each feature value represents the gene expression level. As such, it is structured data with a high dimension. The processing and utilization of such high-dimensional data is an opportunity as well as a challenge. In order to better match the limited sample size of cancer data and to improve performance in classification, feature selection methods are utilized for dimensionality reduction of gene expression data [21]. Differential gene expression analysis is an effective approach to selecting important features.

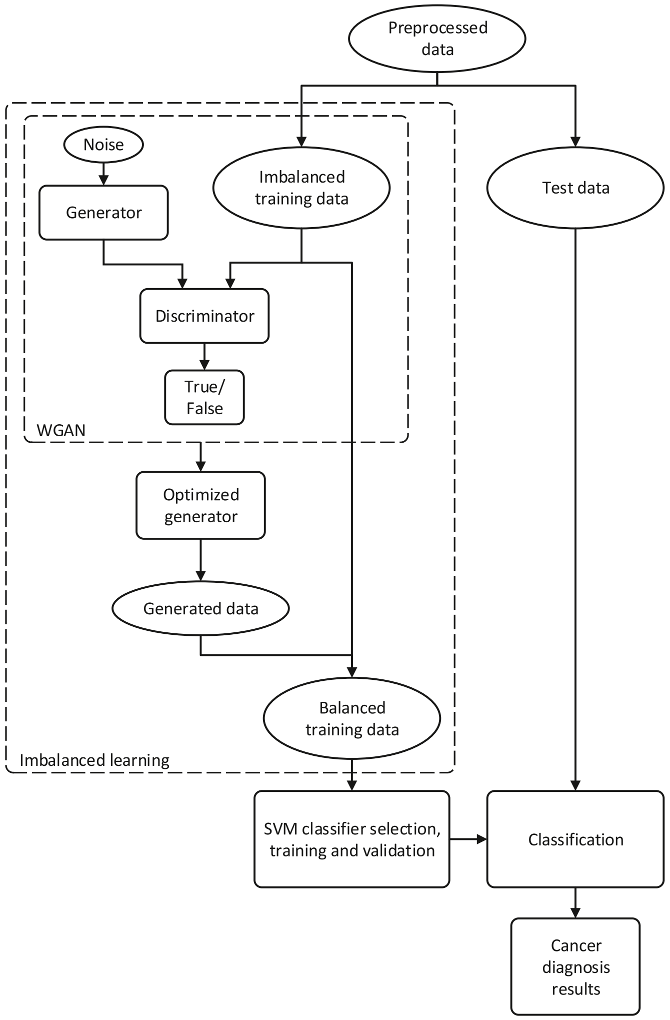


Fig. 1. A flowchart of the imbalanced learning method for cancer diagnosis.

Differential gene expression analysis aims to differentiate genes whose expressions are significantly different under different physiological conditions, which are considered to be important features with more information.

In this study, the DESeq [22] technique is applied to filter the most differentially expressed genes between normal and tumor tissues, providing more information for following classification tasks. The DESeq technique is typically employed to determine if the difference in observable expression level in read count is significant for a given gene, that is, whether the difference is greater than the expected value caused only by natural random variation. In differential gene expression analysis, genes with significantly different expressions are detected and selected by setting the fold change level and the BH-adjusted *p*-value.

3.2. Basic theory of GANs

The main idea of generative adversarial networks (GANs) is derived from the minimax two-player game [23]. A basic GAN consists of a generator *G* that reflects the real data distribution, and a discriminator *D* that aims to discriminate the real data from the data generated by *G* (see Fig. 2). During training, the two models in a GAN optimize themselves and compete with each other simultaneously to improve the ability of both generation and discrimination to find a Nash equilibrium, where *G* is trained towards minimizing the difference between forgeries and real samples, while *D* is trained towards maximizing the confidence in distinguishing the difference between forgeries and real samples. Thus, the minimax two-player game depending on *G* and *D* is evaluated with a cost function *V*(*G*, *D*), that is,

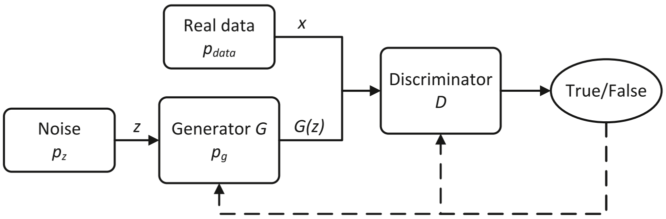


Fig. 2. The structure of a basic GAN.

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))], \tag{1}$$

where x is sampled from the real data distribution $p_{data}(x)$, z is from *a priori* input noise variables $p_z(z)$, and $\mathbb{E}(\cdot)$ is the expectation. We also define $p_g(x)$ as the generated data distribution, $G(z)$ as the data generated by *G*, which is subject to the distribution p_{data}, and $D(x)$ as the probability that x is sampled from p_{data}.

In GAN training process, one model is fixed and the other is optimized [24]. First, the generator is fixed, and the discriminator divides the real samples into positive and the generated samples as negative as much as possible to maximize the discrimination accuracy, thus the optimal solution of the discriminator is obtained as,

$$D^{\star}(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}. \tag{2}$$

Then, for a fixed *D*, the generator is trained by minimizing $\log(1-D(G(z)))$. Considering the saturation situation early in learning, we generally train *G* by maximizing an alternative $\log D(G(z))$. Ideally, the global optimum is reached when $p_g = p_{data}$, which is equivalent to $D(x) = \frac{1}{2}$. In practice, the discriminator may not achieve an ideal optimization for finite data sets. Instead, the optimization process is alternated between several iterations of training *D* and one iteration of training *G*.

3.3. WGANs based deep learning

However, since the proposal of GANs by Goodfellow in 2014 [24], there remain some challenging difficulties and unstable behaviors in training [25]. The difficulties in training GANs arise mainly from dealing with the Jensen-Shannon (JS) divergence and the Kullback-Leibler (KL) divergence, which are generally adopted to measure the generator loss and minimize the loss function of the generator. Specifically, since the generated data and real data are almost impossible to have a non-negligible overlap, the JS divergence between the two data distributions approximates a constant, leading to the gradient vanishing. Moreover, due to the contradiction between different optimal objectives of the KL divergence and the JS divergence in loss function, the gradient update is unstable, and the imbalance of punishment for diversity and accuracy causes mode collapse, namely, lack of diversity.

In order to tackle the above problems, the Wasserstein distance (W distance) [26] is introduced to replace the JS and KL divergences, which is defined as,

$$W(p_{data}, p_g) = \inf_{\gamma\in\prod(p_{data}, p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|], \tag{3}$$

where $\prod(p_{data}, p_g)$ denotes the joint distributions of $\gamma(x, y)$, with margins p_{data} and p_g. In comparison with the KL divergence and the JS divergence, even if the two data distributions do not overlap, the W distance can still reflect their distance, and the loss function of the W distance is continuous and smooth rather than abrupt, so meaningful gradient can be provided for the gradient descent algorithm.

Realizing that the infimum is intractable, we transform equation (3) into the following form:

$$W\left(p_{data}, p_g\right) = \frac{1}{K}\sup_{\|f\|_L \le K} \mathbb{E}_{x\sim p_{data}}\left[f(x)\right] - \mathbb{E}_{x\sim p_g}[f(x)], \tag{4}$$

where the function $f : \mathcal{X}\rightarrow\mathbb{R}$ is K-Lipschitz, which is denoted as $\|f\|_L \le K$, and the supremum is for all function f. Specifically, we use a set of parameters w to define a family of K-Lipschitz functions f_w, which we call the critic functions to represent the discriminator, so that the target problem is transformed into:

$$\max_{w:\|f_w\|_L \le K} \mathbb{E}_{x\sim p_{data}}[f_w(x)] - \mathbb{E}_{x\sim p_g}[f_w(x)]. \tag{5}$$

In deep learning, the parameterized critic functions f_w can be represented by neural networks with the parameters w. On account of the great fitting ability of deep learning, the maximum with the critic functions f_w can accurately approximate the supremum in equation (4). The fact that the functions f_w are K-Lipschitz, namely, the parameters w are constrained in a compact space where $\|f_w\|_L \le K$, we clip all the parameters w to a fixed range, for example, $w \in [-0.01, 0.01]$, after each gradient update in the actual operation. With parameter clipping, there must be a constant K so that the local variation of the function f_w does not exceed it, and the Lipschitz continuity is satisfied.

Therefore, by constructing the critic functions f_w and clipping parameters w within a certain range, we define the value function of WGAN as,

$$V_W = \mathbb{E}_{x\sim p_{data}}[f_w(x)] - \mathbb{E}_{z\sim p_z}[f_w(g_\theta(z))], \tag{6}$$

where w and θ are the corresponding parameters. When the function V_W is maximized, it can be used to approximate the W distance from the generated sample distribution to the real sample distribution. Then, with the optimal discriminator, the generator is trained to approximate the W distance, that is, to minimize the function V_W so as to effectively close the gap between the generated and real samples.

In summary, the W distance used in the WGAN has superior smooth performance in comparison with the KL divergence and the JS divergence, and addresses the vanishing gradient problem. By using the restricted and parameterized discriminator neural networks, the W distance is transformed into a solvable form, thus providing useful optimization gradients for the generator updating. In addition, the approximate W distance can be used to indicate optimization information in each single training iteration, serving as a reliable indicator of the training progress, and is also correlated with the quality of the generated data. So the WGAN model significantly improves training stability compared to GANs and effectively avoids mode collapse.

In addition, the architecture of the WGAN is improved to make the model more suitable for gene expression data. In WGANs, CNNs are generally used as the main model architecture of the generator and the discriminator to complete the task of image generation. For image data, the spatial feature points are highly dependent on each other, and are not affected by spatial translation or rotation. While in gene expression data, the order of genes is arbitrarily adjustable and there is no connectivity between genes. Therefore, CNNs are not suitable for structured gene expression data in spite of its inherent advantages for images. Considering the structural characteristics of gene expression data, the WGAN model is further improved by using deep, fully-connected neural networks, instead of CNNs, to construct the model structure, to explore the complex characteristics of the original data. Therefore, by changing the model structure to deep, fully-connected neural networks and further optimizing the structural parameter settings, the improved WGAN proposed can be well applied to the cancer gene expression data.

3.4. Classification using balanced data

In this work, we have established, based on deep learning, a Wasserstein generative adversarial network method to address the imbalanced learning problem in cancer diagnosis. To be specific, in a WGAN model, given the initial parameters of the generator, the first-generation generated data is the output to the noise input. Then, the discriminator is trained several times using real data and the generated data to update its parameters and gradient up to a set of optimized values. Given the optimized discriminator, the generator is trained once to update its parameters and generate a group of new samples that are closer to the real sample distribution. The entire process iterates multiple times until both the discriminator and the generator are updated to be steady and optimal, according to the training progress indicator provided by WGAN. After that, we utilize the optimized generator to generate a great quantity of new samples that follow the real data distribution, thus expanding the imbalanced training data into balanced training data. In cancer diagnosis, our aim is to accurately distinguish tumor patients from normal persons. By using the balanced training data, the classifier can be trained more precisely, regardless of the impact of sample distribution imbalance.

To confirm the effectiveness of our proposed WGAN based imbalanced learning method and considering that the selection of classifiers has no effect on our imbalanced learning, we apply the support vector machines (SVMs), which is a simple and most commonly used classifier in cancer diagnosis. The SVMs are to find a hyperplane that can dive the sample space. The support vectors refer to the samples closest to the hyperplane, and the margin refers to the distance between two different types of support vectors. SVMs help to obtain the optimal dividing hyperplane corresponding to the maximum margin [27], thus minimizing the cost function of the classification model. By dealing with the imbalanced learning problem off-line at the data level, SVMs can obtain more efficient and accurate on-line classification results in cancer diagnosis by virtue of its superior algorithms in small data sets.

4. Results

4.1. Data collection and preprocessing

In clinical practice, due to the limited number of cancer sampling groups, the difficulty of sampling, patient privacy and other issues, the imbalance of training data is very common, which incurs great limitations to cancer medical research and affects diagnostic accuracy. To specifically demonstrate the sample imbalance problem that prevails in cancer data and the importance of our proposed imbalanced learning method, we used three sets of cancer data from the most commonly used gene expression database in the field of cancer research, the TCGA project web page [5]. The three RNA-seq data sets were obtained from Breast Invasive Carcinoma (BRCA), Lung Adenocarcinoma (LUAD) and Stomach Adenocarcinoma (STAD), respectively, and were derived from subjects with different clinical conditions, cancer stages, ages, and genders. The specifics of these three data sets can be found in Table 1, where N and T represent the number of normal samples and tumor samples, respectively. From the proportion of sample distribution, we observe that the distributions between the normal and tumor samples of the three data sets are quite imbalanced, while the majority of classification methods assume balanced sample distributions between categories. To assess the final classification quality, we extracted the same number of samples from different categories as test data, and formed the training set with the rest data for each data set. In addition, to alleviate

Table 1
Sample distribution of data sets.

Data set	Sample distribution				Sample split	
	Total	Normal	Tumor	Proportion	test set	training set
LUAD	162	37	125	1:3.4	15 N + 15T	22 N + 110T
STAD	271	33	238	1:7.2	15 N + 15T	18 N + 223T
BRCA	878	103	775	1:7.5	30 N + 30T	75 N + 745T

computational burden, we applied the gene differential expression analysis. The DESeq algorithm [22] is used to perform dimensionality reduction and feature selection. Genes that satisfy all the following three conditions were designated as significantly differentially expressed: 1) the fold change threshold is 4, 2) the BH-adjusted *p*-value is less than 0.01, and 3) the mean FPKM is larger than 2. For the LUAD data, the feature dimension of the original data is 20532, and after preprocessing, the dimension is reduced to 1385. For the STAD data, 801 genes were selected from the 29699 original genes. For the BRCA data, the feature dimension is reduced from 20532 to 934. The selected genes were used in the subsequent experimental procedures.

4.2. Performance evaluation metrics

The quality of the diagnosis model is generally assessed by several metrics, including accuracy, precision, recall and F1 score. A confusion matrix, which is given in Table 2, is generally used to illustrate the four possible prediction cases, TP (true positives), FP (false positives), TN (true negatives) and FN (false negatives), where we defined the tumor condition to be positive and the normal condition to be negative.

According to the confusion matrix, the evaluation metrics are calculated as,

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100, \tag{7}$$

$$Precision(\%) = \frac{TP}{TP + FP} \times 100, \tag{8}$$

$$Recall(\%) = \frac{TP}{TP + FN} \times 100, \tag{9}$$

$$F1(\%) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100. \tag{10}$$

In the clinical practice of cancer diagnosis, in addition to the most critical accuracy, precision and recall are also useful indicators. Precision indicates the proportion of the patients who actually have cancer among the patients who are predicted to have cancer, and recall indicates the proportion of the patients who are predicted to have cancer among the patients who actually have cancer. Thus, a prediction model with high precision can prevent misdiagnosis of normal persons into cancer patients, which reduces mental panic and treatment costs, and a prediction model with high recall can avoid missed diagnosis of true cancer patients, which avoids delays in cancer diagnosis and helps to make timely treatment decisions. The F1 score, which denotes a weighted harmonic average to evaluate recall and precision, comprehensively reflects the performance of a classification model.

In addition, the ROC curve (the receiver operating characteristic curve) is used to illustrate the classification results, which is generated by plotting sensitivity ($TP/(TP + FN)$) against specificity ($FP/(FP + TN)$) at different threshold settings, and the AUC (the area under the curve) is a threshold independent indicator, which is provided to compare the model performance.

4.3. Imbalanced learning problem analysis and model comparison

1) The imbalanced learning problem in cancer diagnosis

We analyzed data sets of three different cancers from the most commonly used cancer research database, the TCGA database. It can be seen in the sample distribution that the data imbalance problem in gene expression data in cancer research is very common and serious. However, in most studies of cancer diagnosis methods, the impact of imbalanced data on the results is not considered, which in fact affects the accuracy of prediction to a great extent. To demonstrate the negative effect of data imbalance on cancer prediction results, we used the generally used methods, the random undersampling and random oversampling, to re-balance the original imbalanced training data, and trained classifiers with the adjusted balanced data and original imbalanced data, respectively. The undersampling reduces *T* to be consistent with *N*, and the oversampling randomly samples *N* to coincide with *T*. In the classification stage, we used the SVM classifier, in which the RBF kernel is used as the kernel function, regularization coefficient is set to 1, and gamma is set to 0.000722, 0.001248 and 0.001071 for LUAD, STAD, and BRCA, respectively. By using the balanced test sets for prediction, the results of the three cancers are shown in Tables 3–5. From the results, we observe that the original distribution of the three cancer data sets is imbalanced, which greatly affects the accuracy of cancer prediction in practice. After re-balancing the training sets by random sampling methods, the accuracy is significantly improved. Besides, the prediction performance obtained by the oversampling after expanding the data sets is better than that obtained by the undersampling after reducing the data sets, which is observed by the increased accuracy and F1 score. Specificially, since the sample size becomes too small after undersampling, the model cannot be adequately trained, resulting in under-fitting. Thus, the F1 score was reduced compared to the original data set. Ultimately, after the oversampling of the training samples, both the accuracy and the F1 score have achieved the highest values, which shows that better performance is obtained by the balanced data instead of imbalanced data and that better performance is obtained through oversampling instead of undersampling. By comparing the values of AUC in the tables, the above conclusions are further confirmed. As for the precision and recall, we use the data in Table 3 for specific explanation. For the original data, the training data set is imbalanced and the test data set is balanced. According to the results, we can conclude that half of the samples in the test set, with the label “normal”, are incorrectly predicted as tumor, which is why the recall is 100% and the precision is only 50%. Therefore, a single high recall does not mean that the prediction model performs well, but may cause an excessively high misdiagnosis rate. So in the model assessment, the value of precision and the value of recall should be considered comprehensively, and a good model usually achieves a good trade-off.

Table 2
Confusion matrix.

	Real Positive	Real Negative
Predicted Positive	*TP*	*FP*
Predicted Negative	*FN*	*TN*

2) Impact of sample size in cancer diagnosis

From the above results, we observe that the accuracy resulting from the sample size expansion is higher than from the sample size reduction. Therefore, we further studied the impact of sample size in cancer prediction results after the data has been re-balanced. We consistently used the random undersampling method to gradually increase the sample size for comparison. Shown in Tables 6–8 are the prediction results of different samples on three different data sets. We observe that as the sample size increases, all the metrics are improved. On the basis that data balance has been achieved, the larger the sample size, the better the model performance.

3) Comparison of learning methods from imbalanced data

Table 3
The imbalanced problem on LUAD data.

Training set	Sample size	Acc (%)	Pre(%)	Re(%)	F1 (%)	AUC
Original	$22\,N + 110T$	50.00	50.00	100.00	66.67	0.50
Undersampling	$22\,N + 22T$	70.00	100.00	40.00	57.14	0.70
Oversampling	$110\,N + 110T$	83.33	100.00	66.67	80.00	0.83

Table 4
The imbalanced problem on STAD data.

Training set	Sample size	Acc (%)	Pre(%)	Re(%)	F1 (%)	AUC
Original	$18\ N + 223T$	50.00	50.00	100.00	66.67	0.50
Undersampling	$18\ N + 18T$	73.33	100.00	46.67	63.64	0.70
Oversampling	$223\,N + 223T$	80.00	100.00	60.00	75.00	0.93

Table 5
The imbalanced problem on BRCA data.

Training set	Sample size	Acc (%)	Pre(%)	Re(%)	F1 (%)	AUC
Original	$73\ N + 745T$	50.00	50.00	100.00	66.67	0.50
Undersampling	$73\ N + 73T$	85.00	100.00	70.00	82.35	0.93
Oversampling	$745\,N + 745T$	86.67	100.00	73.33	84.62	0.98

Table 6
The impact of sample size on LUAD data.

Sample size	Acc(%)	Pre(%)	Re(%)	F1(%)	AUC
$10\ N + 10T$	63.33	100.00	26.67	42.11	0.67
$22\ N + 22T$	70.00	100.00	40.00	57.14	0.70

Table 7
The impact of sample size on STAD data.

Sample size	Acc(%)	Pre(%)	Re(%)	F1(%)	AUC
$10\ N + 10T$	70.00	100.00	40.00	57.14	0.69
$18\ N + 18T$	73.33	100.00	46.67	63.64	0.77

Table 8
The impact of sample size on BRCA data.

Sample size	Acc(%)	Pre(%)	Re(%)	F1(%)	AUC
$30\ N + 30T$	66.67	100.00	33.33	50.00	0.63
$50\ N + 50T$	75.00	100.00	50.00	66.67	0.90
$73\ N + 73T$	85.00	100.00	70.00	82.35	0.93

Based on the above experimental results, the effective treatment of the data imbalance problem is to re-balance the data distribution and expand the sample size as much as possible. Therefore, compared to the undersampling method, it is better to use the oversampling method to expand the sample size while achieving balance. In the literature on oversampling, the state-of-the-art solutions at the data level are random oversampling and synthetic sampling with data generation. We applied an improved oversampling method through new sample generation in cancer diagnosis, which is the Wasserstein generative adversarial networks (WGANs) based on deep learning. The model structure of the generator and the discriminator of the improved WGAN is deep, fully-connected neural networks with layers (excluding the output layer) of 100, 128, 256, 512 and 1024 neurons for the generators and layers (excluding the input layer) of 512, 256,1 neurons for the discriminators. The detailed information on layer structure and parameter settings of the WGAN on three cancer gene expression data sets are listed in Table 9. After each gradient update, the absolute values of the parameters are clipped to no more than the value of the clipping parameter. Besides, Leaky ReLU is employed as the activation in hidden layers to avoid gradient sparsity, tanh is employed as the activation in the output layer of the generator, and the sigmoid activation in the output layer of the discriminator is removed to fit the W distance. The optimizer used in the WGAN is the RMSprop instead of the momentum update algorithm to obtain more stable gradient. When setting the epoch, we evaluated the loss values of the generators and the discriminators, and selected the epoch value corresponding to the minimum loss value, so as to ensure that the WGAN model has been fully trained and has achieved the most effective performance in the case of small sample size. Furthermore, we evaluated the final classification results corresponding to the data obtained after five times of training, and the variance of the accuracy in the three data sets are 0.0471, 0.0422 and 0.0218, respectively. This further verifies that the WGAN has been trained properly.

We compared the WGAN method we proposed with the random oversampling, SMOTE and GAN method on the LUAD, STAD and BRCA data sets. To be specific, the size of samples in the minority class, namely the normal class, was increased by these oversampling methods to match the size of samples in the majority class, namely the tumor class, in the three data sets. After re-balancing the data distribution with WGAN, we generated a large number of new balanced samples and the WGANs used are trained in the normal class and the tumor class, respectively. The sample sizes after the further data augmentation process by WGAN are expanded to ($1110\ N + 1110T$), ($2223\ N + 2223\ N$), ($2745\ N + 2745T$) for the LUAD, STAD and BRCA data, respectively. In addition, although the data imbalance problem is addressed at the data level in this study, we do further comparative experiments with a cost sensitive classifier, RUSBoost. RUSBoost is an advanced classifier specially created to solve the problem of imbalanced data classification. We use the RUSBoost in the original imbalanced data sets to compare with the results of other data generated methods.

The experiment results are shown in Tables 10–12, where WGAN expansion indicates that more balanced data is further added. From these tables, we observe that the WGAN method performs significantly better than the other three oversampling methods and the RUSBoost method. To be specific, the WGAN, with its deep model architectures and powerful self-learning and optimization capability, obtains the highest accuracy, at 90.00%, 93.33% and 98.33% on the three cancer data sets, respectively. Also observed in these tables, with the same precision, WGAN achieves the highest recall is, which means that cancer patients can be more fully predicted. Furthermore, the highest F1 score of the WGAN further indicates that the proposed method yields the best performance that is superior to the other three methods as well as the cost sensitive RUSBoost classifier in cancer diagnosis.

Table 10
The prediction results with different imbalanced learning methods on LUAD data.

Training set	Acc(%)	Pre(%)	Re(%)	F1(%)	AUC
RUSBoost	86.67	92.31	80.00	85.71	0.86
Random oversampling	83.33	100.00	66.67	80.00	0.84
SMOTE	86.67	100.00	73.33	84.62	0.86
GAN	86.00	90.00	89.34	87.62	0.88
WGAN	90.00	100.00	80.00	88.89	0.91
WGAN expansion	96.67	100.00	93.33	96.55	0.97

Table 9
Detailed parameters of the improved WGAN.

Data set	Layers of G	Layers of D	Learning rate	Epochs	Batch size	Clipping parameter	No. of the critic
LUAD	6	4	0.00005	800	16	0.01	5
STAD	6	4	0.00005	2000	16	0.01	5
BRCA	6	4	0.00005	1800	16	0.01	5

Table 11
The prediction results with different imbalanced learning methods on STAD data.

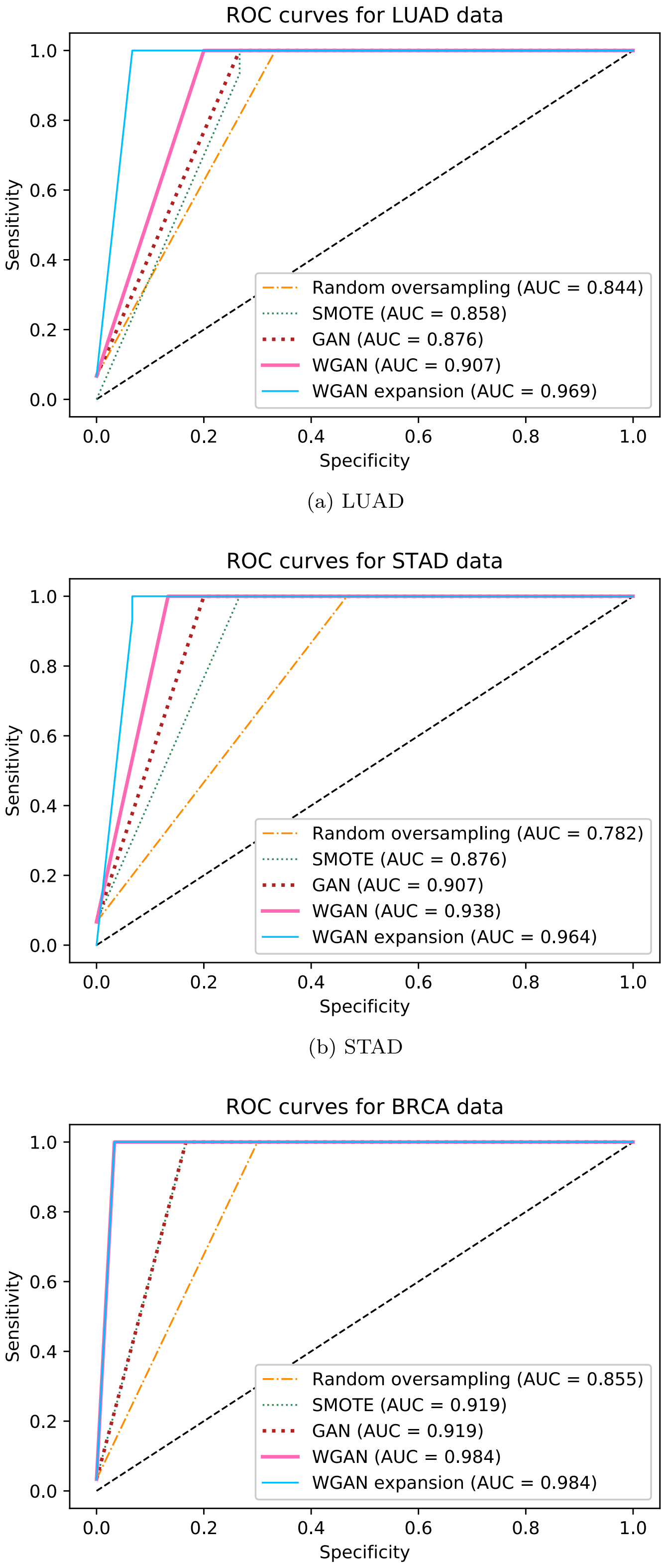
Training set	Acc(%)	Pre(%)	Re(%)	F1(%)	AUC
RUSBoost	93.33	100	86.67	92.86	0.93
Random oversampling	80.00	100.00	60.00	75.00	0.78
SMOTE	90.00	100.00	80.00	88.89	0.88
GAN	90.00	92.86	86.67	89.66	0.91
WGAN	93.33	100.00	86.67	92.86	0.94
WGAN expansion	96.67	100.00	93.33	96.55	0.96

Table 12
The prediction results with different imbalanced learning methods on BRCA data.

Training set	Acc(%)	Pre(%)	Re(%)	F1(%)	AUC
RUSBoost	96.67	96.67	96.67	96.67	0.97
Random oversampling	95.00	100.00	90.00	94.74	0.86
SMOTE	96.67	93.75	100.00	96.77	0.92
GAN	96.67	93.75	100.00	96.77	0.98
WGAN	98.33	100.00	96.67	98.31	0.98
WGAN expansion	98.33	100.00	96.67	98.31	0.98

Compared with WGAN and WGAN expansion, we observe that, on the basis of having achieved the balance of the data, a large expansion of the sample size can further improve the prediction on the LUAD and STAD data sets, which can be seen by comparing the corresponding accuracy and F1 score. For the BRCA data set, since the sample size is large enough to fully train the model after balancing, it is difficult to improve the model performance only by expanding the sample size at the data level after the WGAN has raised the accuracy to 98.33%.

Besides, we illustrated and compared the model performance of the data processed by the four different oversampling methods, random oversampling, SMOTE, GAN and WGAN, on the three data sets by ROC curves, which is shown in Fig. 3. From the curves and results in the figure, we can draw conclusions that are consistent with those obtained from the tables. With the highest AUC scores, 0.907, 0.929 and 0.984 on the three data sets, the model performance of the WGAN processed data is better than SMOTE, which in turn is better than the random sampling. Furthermore, after further expanding the data sets, the WGAN method improves the results of cancer diagnosis.



(a) LUAD

(b) STAD

(c) BRCA

Fig. 3. The ROC curves.

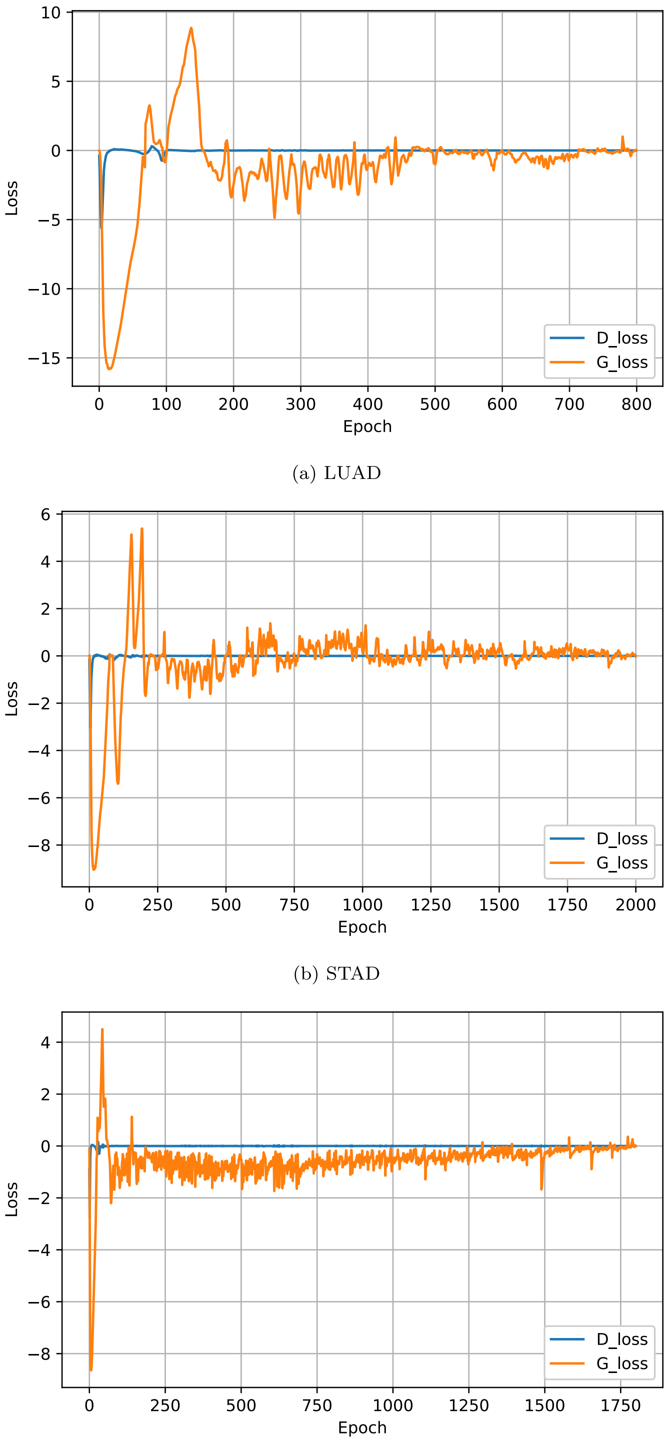
5. Discussion

In the literature on cancer diagnosis, many researchers have focused on improving the algorithms and structures of classification models to improve the prediction performance, while most general classification learning algorithms assume balanced data distribution or negligible misclassification cost due to data imbalance. However, limited by sampling difficulty and actual sample size, data imbalance is a common problem in the databases that can be used in the field of cancer diagnosis. In this study, by observing the data distribution of multiple datasets in the TCGA database, we discussed the imbalanced learning problem of gene expression data for the first time. From the comparison between the original imbalanced data sets and balanced data sets after sampling, we recognize that data imbalance in cancer gene expression data cannot be ignored, and the imbalanced learning problem has a great impact on the accuracy of cancer diagnosis, which entails further discussion and resolution.

Comparisons were made between sample sets after undersampling and oversampling, and between different sample sizes, according to which we notice that better accuracy can be obtained by filling up the minority class sample size through oversampling. The state-of-the-art solutions at the data level of oversampling are random sampling and the SMOTE [2,12,13,28]. Since it only copies the original data, the random oversampling reveals its problem that replicated samples may cause skewing of the prediction results and over-fitting of the training model, and may hinder the learning of the model and data [29]. The SMOTE family methods, as shallow learning algorithms, also have their drawbacks, such as the randomness of the element selection in the nearest neighbors and the blindness of the feature vector superposition, which may also lead to high variability and uncontrollability [30]. As a technique of deep learning algorithms, GANs can effectively learn sample features. Through alternating learning and competing of two deep neural network learners, GANs learn the structure and feature representations behind the training data autonomously and rapidly. Since the introduction of the GAN algorithm, its applications in computer vision have developed rapidly and have been highlighted in various directions of image manipulation and synthesis [26,31]. However, in image processing, the purpose of using GANs is often to generate numerous image samples [18,19], which is only at the level of expanding the sample size for better training, and GANs have not been extensively explored in the downstream analysis and processing after sample generation. The conventional GANs also have drawbacks. GANs cannot indicate the training progress and are difficult to achieve the optimazation, so the model cannot be fully trained. Although some variants of GANs, such as WGANs [20], improve the model optimization process by replacing the commonly used JS divergence and KL divergence, the general model architecture is CNN in the previous research, which has inherent advantages for image data and is not suitable for structured numerical data. Considering the superiority of the GAN method for capturing sample features, as well as its drawbacks, we have improved the GAN method and extended the application of GANs to cancer gene expression data for the first time.

Firstly, we applied the improved WGAN algorithm by replacing the commonly used the JS divergence and the KL divergence with the W distance with a solvable form. The W distance with a continuous and smooth loss function can provide an intuitive and reliable indication of the losses of generator and discriminator during training. Therefore, the improved WGAN can make preliminary judgments and adjustments on the quality of the generated data, avoiding the blindness of conventional GAN training. To further demonstrate the ability that the improved WGAN can indicate the training process, we have plotted the curves of generator loss and discriminator loss after each training on the three cancer sample sets in Fig. 4. When the loss value is near zero, the model has been trained to be approximately optimal. For illustration, we only drew the training process in which the WGAN brought the sample distribution to balance. Based on the loss curves, we set the epoch parameter of the WGAN models used on the three data sets to 800, 2000 and 1800, respectively. It is observed that the training of the models has become steady and optimal when the set epochs are reached. This further demonstrates the advantages of the indicated role of the WGAN and the good quality of the generated samples. Compared to the non-learning random oversampling, the shallow-learning SMOTE and the not-fully-trained GANs, the deep-learning and optimal WGANs undoubtedly provide new samples that are better and more consistent with the original sample distribution, and our experimental comparison results confirm this in various aspects.

Secondly, we improved the architecture of the WGAN to make the model more suitable for gene expression data. The data we used here is a numerical type of structured data, which is different from the image data, so the newly generated numerical samples are not as easy to visually display as images. Meanwhile, the order of genes is arbitrarily adjustable and there is no connectivity between genes. To be applicable to structured gene expression data, the deep, fully-connected neural networks were used, instead of the typically-used CNNs, as the model architecture for the WGAN. Through this improvement, the common properties among the same class samples can be acquired and the redundant and interference information can be eliminated in time. The experimental results have shown that the proposed model can be well trained under few training samples. Through the confrontation learning between the generator and the discriminator, the discriminator will calibrate the generator once after each round of training, so that the loss of the generator after several epochs is gradually reduced. As observed in Refs. [26,32], the training process of GANs uses two adversarial neural networks as training criterion and can be carried out by back-propagation. There is no complex variational lower bound, which greatly reduces the training difficulty and improves the training efficiency. Several studies have also verified the effectiveness of the application of GANs in the case of a small sample size [23,33,34]. Through our further improvement on the WGAN model, namely using fully-connected neural networks instead of CNNs to simplify the deep learning model, the model is more suitable for small sample sizes. Thus, the improved WGAN has achieved better performance in the imbalance learning problem of gene expression data.



(a) LUAD

(b) STAD

(c) BRCA

Fig. 4. The loss curves.

Thirdly, in our study, the purpose of using GANs is no longer to simply generate new samples, but to solve the imbalanced learning problem, thereby improving the accuracy of the subsequent prediction results. Different from the use of discriminators for feature learning, we mainly focus on using generators to generate new samples that are non-repetitive and conform to the original data distribution. Different from considering only the quantity of generated samples without evaluating the impact of sample quality on downstream analysis tasks, we mainly focus on the quality of samples generated by GANs for the performance of the subsequent classifiers. In addition, we visualize the distribution of the data containing the samples generated by different oversampling methods in a two-dimensional space using the t-SNE [35], a visualization technique for high-dimensional data. For illustration, only the distribution of the LUAD data is plotted in Fig. 5. It can be seen in the figure that there are many overlapped scattered points in the sample sets formed by the random oversampling and the SMOTE, and these scattered points cannot be accurately divided. In the sample sets formed by the GAN and the WGAN, different types of data are basically distributed in clusters, which are easily separated by the boundaries or interfaces and, as a result, better classification results can be obtained.

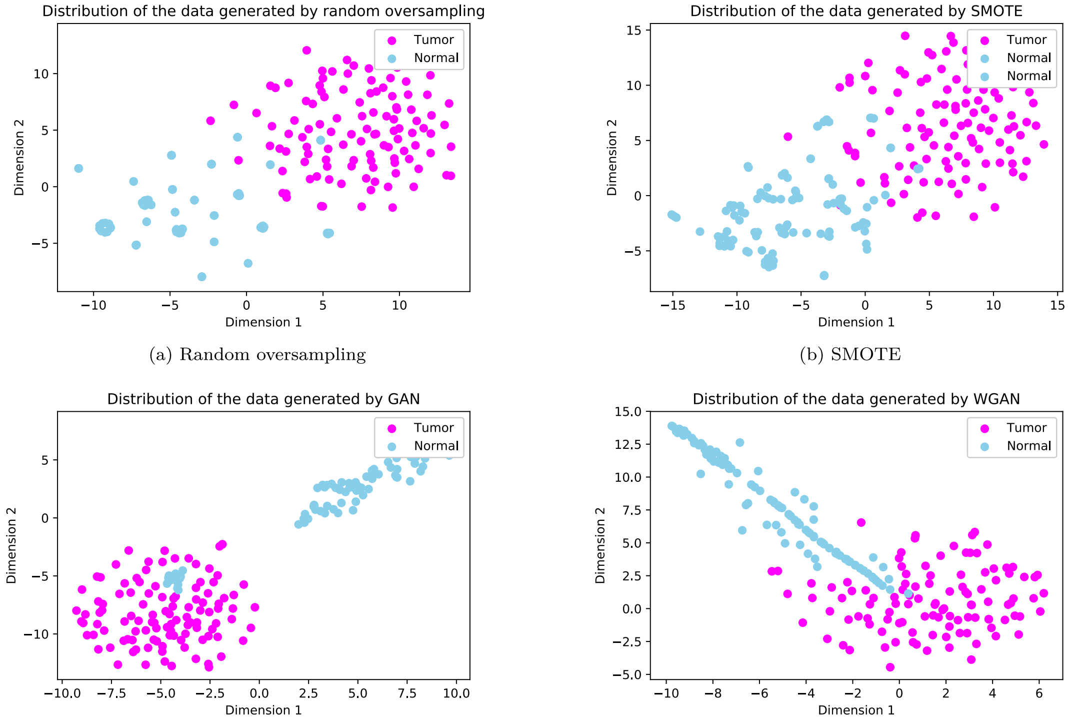
Moreover, we address the imbalanced learning problem at the data level, which will not affect the running time of the online classification model, and make the classification model more efficient and more targeted to deal with the online cancer diagnosis. On the other hand, we concentrate on the imbalanced learning problem at the data level, that is, based on the same classification model in the follow-up analysis without considering the influence of the classifier, and take a different approach by using the improved WGAN algorithm to balance data distributions to improve the prediction accuracy on cancer gene expression data. It provides more possibilities for applying high-performance machine learning methods that require big data in downstream data processing, thereby improving the accuracy and efficiency of cancer diagnosis. In future research, we will add data distribution considerations and processing to the prediction model at the algorithm level. The combination of data level and algorithm level for cancer data prediction will also be worthy of more attention. Besides, in the future work on data biology analysis, the interpretability and relevance of the generated data need to be further explored and traced.

6. Conclusions

In this study, we demonstrated the imbalanced learning problem in cancer gene expression data and proposed a deep learning based Wasserstein generative adversarial network (WGAN) approach for cancer diagnosis. Specifically, we analyzed the sample distributions between categories in gene expression data of lung, stomach and breast tissues and confirmed the prevalence of data imbalance in cancer data. In the specific experimental operation, we first compared the prediction results corresponding to the original imbalanced data and the balanced data after sampling to show that higher accuracy can be obtained by the balanced data. Then, in the case where the data sets are balanced, we gradually expanded the sample size to verify the prediction improvement with the increase of the size of the samples. Finally, we proposed an improved WGAN method using the W distance to provide continuous and reliable indications for the training progress, and employed deep, fully-connected neural networks as the generator and discriminator. The trained optimal generator was used to generate new samples to alleviate data imbalance and further expand the sample size. We compared the proposed WGAN method with the commonly used random oversampling, SMOTE and GAN by combining the imbalanced learning models with the same classifier and evaluated the quality of imbalanced learning by comparing the final prediction performance corresponding to the data sets obtained by different models. The results all suggest that the proposed method outperforms other methods. Therefore, our proposed WGAN method can rely on its superior deep learning ability to generate new samples with both original data characteristics and diversity, and can utilize the reliable training indicators to make the model well suited for gene expression data, thereby providing better prediction performance in cancer diagnosis.

Source code

The source code used in this study can be downloaded from https



(a) Random oversampling

(b) SMOTE

(c) GAN

(d) WGAN

Fig. 5. The distribution of the LUAD data.

://github.com/foreverxyw/WGAN-for-cancer-gene-expression-data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under grant No. 31801118. The authors would like to thank the reviewers in advance for their comments and suggestions.

References

[1] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Canc. Inf. 2 (1) (2006) 59.
[2] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 9 (2008) 1263–1284.
[3] A. Majid, S. Ali, M. Iqbal, N. Kausar, Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines, Comput. Methods Progr. Biomed. 113 (3) (2014) 792–808.
[4] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recogn. Artif. Intell. 23 (2009) 687–719, 04.
[5] The TCGA Database. http://www.cancergenome.nih.gov/. (Accessed 29 April 2017).
[6] S. Fotouhi, S. Asadi, M.W. Kattan, A comprehensive data level analysis for cancer diagnosis on imbalanced data, J. Biomed. Inf. 90 (2019) 103089.
[7] A. Estabrooks, A. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Comput. Intell. 20 (1) (2004) 18–36.
[8] N.V. Chawla, N.V. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[9] X. Liu, X.Z. Zeng, A new automatic mass detection method for breast cancer with false positive reduction, Neurocomputing 152 (2015) 388–402.
[10] A.A. Shanab, T.M. Khoshgoftaar, R. Wald, A. Napolitano, Impact of noise and data sampling on stability of feature ranking techniques for biological datasets, in: 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), IEEE, 2012, pp. 415–422.
[11] K.J. Wang, K.J. Adrian, Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm, Int J Comput Sci Electron Eng (IJCSEE) 1 (3) (2013) 408–412.
[12] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, DBSMOTE: density-based synthetic minority over-sampling technique, Appl. Intell. 36 (3) (2012) 664–684.
[13] S. Barua, M. Islam, X. Yao, K. Murase, MWMOTE–Majority weighted minority oversampling technique for imbalanced data set learning, IEEE Trans. Knowl. Data Eng. 26 (2) (Feb. 2014) 405–425.
[14] P. Mamoshina, A. Vieira, E. Putin, A. Zhavoronkov, Applications of deep learning in biomedicine, Mol. Pharm. 13 (5) (2016) 1445–1454.
[15] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1125–1134.
[16] H.C. Shin, H.C. Tenenholtz, J.K. Rogers, C.G. Schwarz, M.L. Senjem, J.L. Gunter, K. P. Andriole, M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer, Cham, 2018, pp. 1–11.
[17] Q. Wang, X. Zhou, C. Wang, Z. Liu, J. Huang, Y. Zhou, C. Li, H. Zhuang, J.Z. Cheng, WGAN-based synthetic minority over-sampling technique: improving semantic fine-grained classification for lung nodules in CT images, IEEE Access 7 (2019) 18450–18463.
[18] A. Antoniou, A. Storkey, H. Edwards, Data Augmentation Generative Adversarial Networks, arXiv Preprint arXiv:1711.04340, 2017.
[19] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, C. Malossi, Bagan: Data Augmentation with Balancing GAN, arXiv Preprint arXiv:1803.09655, 2018.
[20] M. Marouf, P. Machart, V. Bansal, C. Kilian, D.S. Magruder, C.F. Krebs, S. Bonn, Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks, Nat. Commun. 11 (1) (2020) 1–12.
[21] H. Hijazi, C.A. Chan, Classification framework applied to cancer gene expression profiles, Journal of Healthcare Engineering 4 (2) (2013) 255–283.
[22] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (10) (2010) R106.
[23] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A.A. Bharath, Generative adversarial networks: an overview, IEEE Signal Process. Mag. 35 (1) (2017) 53–65.
[24] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: International Conference on Neural Information Processing Systems, MIT Press, 2014.
[25] T. Salimans, I.J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.
[26] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.
[27] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Their Appl. 13 (4) (2002) 18–28.
[28] Y. Sun, Y. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recogn. 40 (12) (2007) 3358–3378.
[29] D. Mease, A.J. Wyner, A. Buja, Boosted classification trees and class probability/quantile estimation, J. Mach. Learn. Res. 8 (Mar) (2007) 409–439.
[30] B.X. Wang, N. Japkowicz, Imbalanced data set learning with synthetic samples, Proc. IRIS Machine Learning Workshop 19 (2004).
[31] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[32] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, F.Y. Wang, Generative adversarial networks: introduction and outlook, IEEE/CAA Journal of Automatica Sinica 4 (4) (2017) 588–598.
[33] Y. Liu, Y. Zhou, X. Liu, F. Dong, C. Wang, Z. Wang, Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology, Engineering 5 (1) (2019) 156–163.
[34] A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv Preprint arXiv:1511.06434, 2015.
[35] L.V.D. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.