Water Science & Technology

IWA PUBLISHING



Water Science & Technology Vol 88 No 11, 2762 doi: 10.2166/wst.2023.373

# Alleviating sample imbalance in water quality assessment using the VAE–WGAN–GP model

Jingbin Xu [a], Degang Xu [a,*], Kun Wan [a] and Ying Zhang [b]

[a] School of Automation, Central South University, Changsha, Hunan, China
[b] School of Literature, Hunan Normal University, Changsha, Hunan, China
*Corresponding author. E-mail: dgxu@csu.edu.cn

JX, 0009-0004-4133-8128

## ABSTRACT

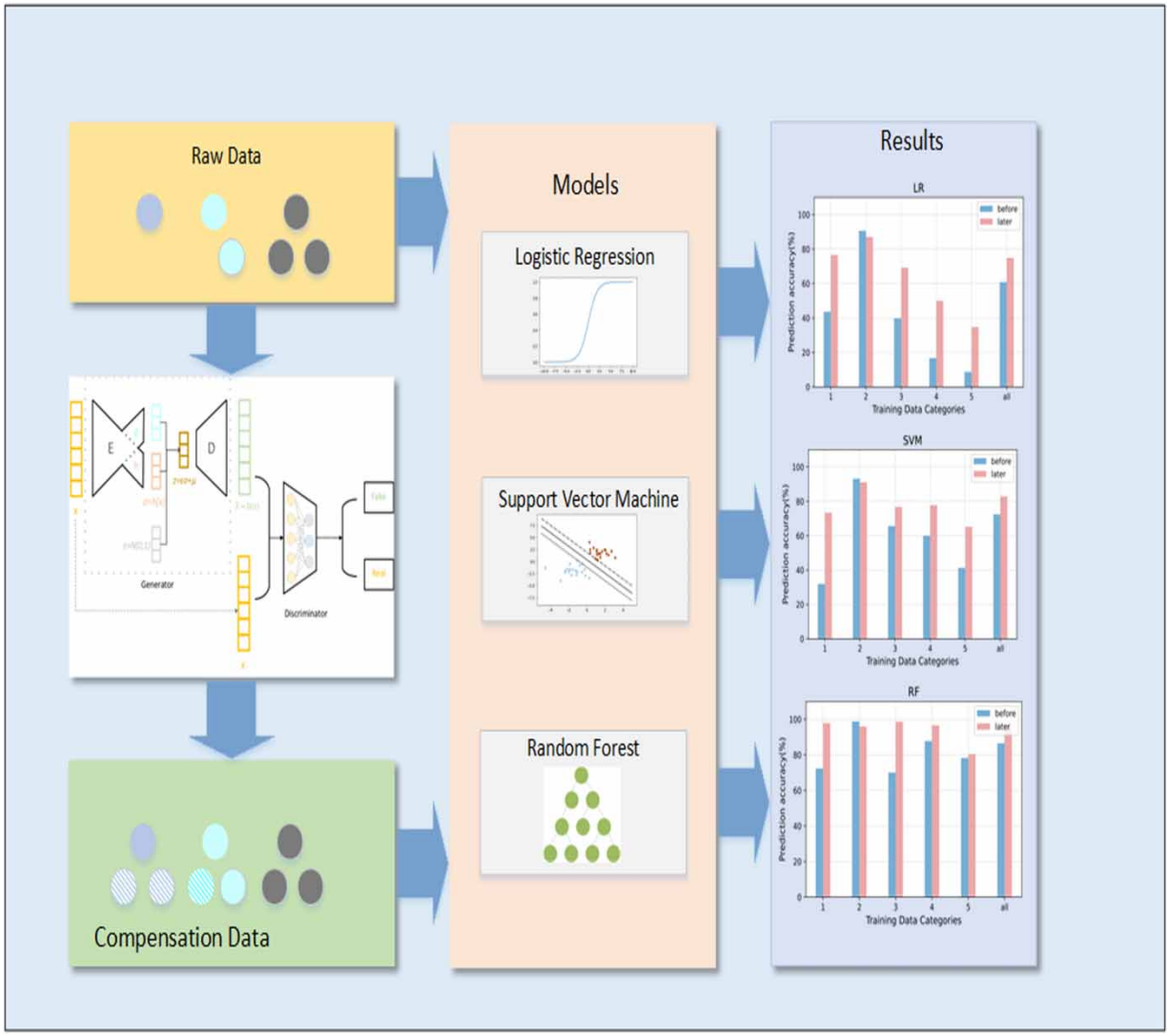
Water resources are essential for sustaining human life and promoting sustainable development. However, rapid urbanization and industrialization have resulted in a decline in freshwater availability. Effective prevention and control of water pollution are essential for ecological balance and human well-being. Water quality assessment is crucial for monitoring and managing water resources. Existing machine learning-based assessment methods tend to classify the results into the majority class, leading to inaccuracies in the outcomes due to the prevalent issue of imbalanced class sample distribution in practical scenarios. To tackle the issue, we propose a novel approach that utilizes the VAE–WGAN–GP model. The VAE–WGAN–GP model combines the encoding and decoding mechanisms of VAE with the adversarial learning of GAN. It generates synthetic samples that closely resemble real samples, effectively compensating data of the scarcity category in water quality evaluation. Our contributions include (1) introducing a deep generative model to alleviate the issue of imbalanced category samples in water quality assessment, (2) demonstrating the faster convergence speed and improved potential distribution learning ability of the proposed VAE–WGAN–GP model, (3) introducing the compensation degree concept and conducting comprehensive compensation experiments, resulting in a 9.7% increase in the accuracy of water quality assessment for multi-classification imbalance samples.

**Key words:** data compensation, deep generative models, imbalanced samples, VAE–WGAN–GP model, water quality assessment

## HIGHLIGHTS

- Novel method: the VAE–WGAN–GP model is introduced to alleviate the problem of imbalanced category distribution in water quality evaluation and improve the accuracy of assessment.
- Water resource management: our research bridges the gap in the distribution of categories in water management by providing deep generative models to compensate for data scarcity in water quality assessments.

## GRAPHICAL ABSTRACT




## INTRODUCTION

Water resources are an indispensable asset for human survival and development, serving as a fundamental pillar of sustainable progress. However, in recent years, rapid urbanization and industrialization have resulted in a gradual decline in the overall availability of freshwater resources. The lack of environmental awareness among numerous small and medium-sized factories exacerbates the pollution of water resources and the deterioration of the ecological environment by directly discharging industrial wastewater into surface rivers. The prevention and control of water resource pollution have become an issue that cannot be overlooked, as they are closely intertwined with global ecological equilibrium and human well-being (Wang & Yang 2016).

Water quality assessment is a fundamental aspect of water resource monitoring. Its objective is to evaluate the degree of water pollution and provide scientific foundations for protecting and managing water resources (Štambuk-Giljanović 1999; Nong *et al.* 2020). Due to the multitude of factors influencing water quality, a complex nonlinear relationship often exists between assessment factors and water quality grades. Therefore, establishing an appropriate evaluation model poses specific challenges. Currently, commonly used methods for water quality assessment include the single-factor method, fuzzy mathematics comprehensive method, support vector machine (SVM), and random forest (RF), and others (Tyagi *et al.* 2013).

The single-factor method involves assigning a water quality grade to each monitoring parameter and selecting the grade of the poorest parameter as the final grade for the water body. This approach does not require mathematical calculations, making it intuitive and straightforward. However, it overlooks the influence of other evaluation factors, resulting in relatively low assessment accuracy. The fuzzy comprehensive evaluation method considers the impact of all evaluation factors on water quality. It emphasizes the role of the most significant pollutant through weighting, making full use of all available information, and partially compensating for the limitations of the single-factor method. However, it faces challenges in reaching consistent conclusions when dealing with large sample sizes (Icaga 2007). The introduction of machine learning methods such as support vector machines and random forests has effectively addressed the limitations of traditional models. These methods have strong nonlinear mapping capabilities and can handle large amounts of feature data, thus providing a more accurate reflection of water quality conditions (Wang *et al.* 2017).

However, in practical situations, water quality conditions in different regions often experience an imbalanced distribution of category samples. In such cases, classifiers tend to classify the majority of samples into the more abundant category, overlooking the less represented categories (More 2016). This leads to a decrease in the accuracy of water quality assessment and may pose significant risks in practical applications. To address the issue of imbalanced samples, researchers have proposed various methods (Dal Pozzolo *et al.* 2015), which can be primarily categorized into two approaches. The first approach focuses on improving the model structure or algorithm of the classifier to mitigate the impact of imbalanced samples. For instance, ensemble learning combines multiple classifiers, enhancing the classifier's performance through voting or weighted averaging. Refining the classifier's classification strategy, assigning a higher cost to misclassifications of minority classes, thus making it more adaptable to the classification task of imbalanced samples (Zhang *et al.* 2004). The second approach utilizes resampling strategies to balance the dataset, including undersampling, oversampling, and hybrid sampling. Undersampling achieves balance by reducing the number of samples from the majority class. This method can reduce computational complexity and noise, but it may lead to information loss and underfitting issues, thereby impacting model performance. Oversampling achieves balance by increasing the number of samples from the minority class. Although it can augment sample size, it may result in overfitting problems and require addressing class overlapping issues (Al-Shabi 2019).

Generative adversarial network (GAN), proposed by Goodfellow *et al.* (2014), has gained tremendous popularity as deep learning models in recent years. The core concept revolves around two neural networks engaging in an adversarial competition to generate highly realistic data samples. Within the GAN framework, a generator network is responsible for synthesizing data samples, while a discriminator network evaluates the authenticity of the generated samples. Through iterative gameplay and learning, these two networks ultimately achieve a state of equilibrium, resulting in remarkably authentic data samples generated by the generator. The introduction of GAN has provided a novel approach to addressing the issue of imbalanced datasets. Specifically, GAN models can generate synthetic data resembling scarce class samples. By blending these synthetic samples with real data, a balanced dataset is formed, leading to improved classifier performance (Bhagwani *et al.* 2021). This methodology effectively alleviates the problem of imbalanced datasets while avoiding the information loss and overfitting issues commonly encountered in traditional sampling methods (Vuttipittayamongkol *et al.* 2021).

Currently, notable advancements have been made in the research field addressing the issue of imbalanced samples through the application of GAN. For instance, Mariani *et al.* (2018) employed GAN to generate imbalanced class image data, thereby effectively enhancing the accuracy of binary image recognition. Similarly, Zhu *et al.* (2022) employed a GAN-based hybrid sampling approach that successfully generated instances aligning with the actual data distribution, significantly reducing the impact of class overlap. Douzas & Bacao (2018) introduced a model based on conditional generative adversarial networks (CGANs), which captured the authentic distribution of the global minority class by incorporating additional conditional information. However, this model suffered from instability during training, leading to pattern collapse and breakdown. Building upon the CGAN model, Xu *et al.* (2021) introduced Wasserstein distance and gradient penalty (GP) strategies, resulting in an improved model capable of generating more realistic data while overcoming the issues of pattern collapse and training instability. Furthermore, researchers have explored combining GAN with other methods, such as adversarial sampling and deep learning ensembles, to further enhance the performance and stability of classifiers (Gao *et al.* 2020; Luo *et al.* 2022; Ding & Cui 2023).

Building upon the aforementioned concerns, this paper presents a novel approach for enhancing water quality data based on the VAE–WGAN–GP model. In this model, the framework of GAN is extended by incorporating the encoding and decoding mechanisms of the VAE, enabling the acquisition of more comprehensive and accurate feature representations through variational inference techniques. Additionally, the Wasserstein distance and GP techniques are introduced, effectively

enhancing the stability and convergence speed of the model. Leveraging VAE, the model reduces the dimensionality of feature representations and extracts latent distribution information such as mean and variance. Through adversarial learning, synthetic samples that closely resemble real samples are generated, thus compensating for scarce water quality data. The main contributions of this paper are as follows:

1. This paper focuses on addressing the issue of imbalanced category samples in water quality evaluation. We propose a method to balance the water quality dataset by compensating for the scarcity of category samples through a deep generative model.
2. We compared the proposed VAE–WGAN–GP generative model and verified its faster convergence speed and improved potential distribution learning ability.
3. For imbalanced multi-classification samples, we introduced the concept of compensation degree and conducted data compensation experiments for each scarcity class sample. We explored the compensation method based on the experimental results and achieved a 9.7% improvement in water quality evaluation accuracy through comprehensive data compensation experiments, validating the effectiveness of the method.

The remaining sections of this paper are organized as follows: Section 'Proposed model' provides detailed information about the proposed model, Section 'Experiment' presents experimental validation and analysis, and Section 'Conclusion' concludes the paper by summarizing the work.

## PROPOSED MODEL

In this section, we provide an introduction to the fundamental principles of VAE and GAN, along with the strategies employed to enhance the training process of GAN. Subsequently, we present our proposed VAE–WGAN–GP model.

### Variational AutoEncoder

The structure of the Variational AutoEncoder (VAE) is similar to that of the AutoEncoder (AE) in that both have two parts: an encoder and a decoder. The encoder extracts rich feature information from the original data, and the decoder reconstructs the input information to train the network by minimizing the reconstruction error. Unlike AE, the encoder part of VAE encodes the original data into a distribution of potential space rather than individual points. Then, the decoder part samples the distribution and reconstructs the information from the sampled data. Figure 1 illustrates the basic network structure of the VAE.

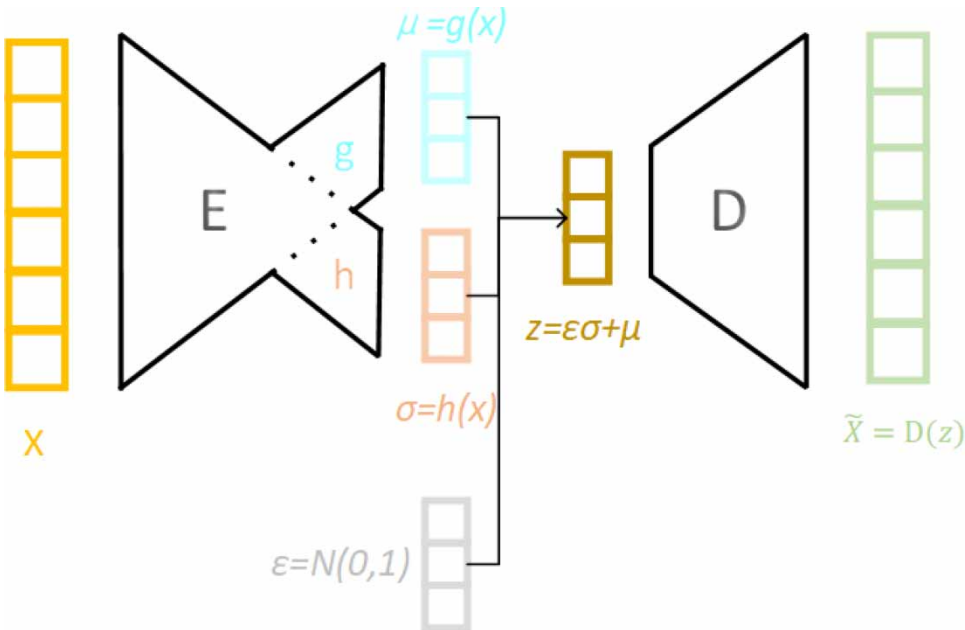



**Figure 1** | VAE network structure diagram.

In Figure 1, $X$ represents the original input vector, the objective of VAE is to learn the latent distribution $q(z|X)$. To achieve this, VAE assumes that the distribution of the latent variables $z$ follows a Gaussian distribution. The process of generating reconstructed data $\tilde{X}$ involves sampling a latent variable z from the latent distribution $p(z)$, with the mean $\mu$ and standard deviation $\sigma$ of the latent distribution $p(z)$ obtained from the encoder network $E$. Subsequently, the decoder network $D$ decodes the latent variable $z$ to generate the distribution $p(X|z)$ of the original data $X$. Additionally, the Kullback–Leibler (KL) divergence is employed as a regularization term to constrain the distance between $q(z|X)$ and $p(z)$.

The optimization objective of VAE is to maximize the lower bound function ELBO (Evidence Lower Bound):

$$\text{ELBO} = E[\log p(X|z)] - KL(q(z|X)||p(z)) \tag{1}$$

where $E[\log p(X|z)]$ represents the reconstruction error of the input data $X$ given the latent variable $z$, and $\text{KL}(q(z|X)||p(z))$ represents the KL divergence of the latent variable. Since the KL divergence is non-negative, maximizing the ELBO is equivalent to minimizing the KL divergence, which is equivalent to maximizing the reconstruction error.

To compute the ELBO, we need to estimate the posterior distribution $q(z|X)$ of the latent variables and the generative distribution $p(X|z)$ of the original data. Specifically, the input data $X$ is mapped to the mean $\mu$ and standard deviation $\sigma$ of the latent variables through the encoder network. Then, a latent variable $z$ is sampled from the distribution of the latent variables, and the decoder network decodes the latent variable z into the distribution $p(X|z)$ of the original data $X$. This process can be represented by the following equation:

$$z \sim q(z|X) = N(\mu, \sigma^2) \sim p(X|z) \tag{2}$$

$$N(\mu, \sigma^2) = \sigma * N(0, 1) + \mu \tag{3}$$

where $\mu$ and $\sigma$ are the outputs of the encoder network. To ensure that the gradient propagation process is continuous a reparameterization of Equation (3) is used.

### GAN and WGAN–GP

Generative adversarial networks (GANs) consist of a generator model and a discriminator model, which engage in a competitive learning process. The generator aims to produce realistic samples to deceive the discriminator, while the discriminator strives to differentiate between real and generated samples, thereby continuously enhancing its accuracy. Figure 2 depicts the fundamental architecture of GAN.

The latent variables $z$ sampled are used by the Generator to produce synthetic samples $\tilde{X}$, which are then fed alongside genuine data into the discriminator for distinguishing between real and fake data. The optimization objective of GAN is to minimize the probability of the discriminator failing to differentiate between generated and real data, as expressed by the following loss function:

$$\min_G \max_D V(D, G) = E_{x\sim p_{\text{data}}(X)}[\log D(X)] + E_{z\tilde{p}_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

where $p_{\text{data}}(X)$ represents the distribution of real data, $p_z(z)$ represents the distribution of latent variables, $D(X)$ represents the probability of the Discriminator classifying them as real data, and $D(G(z))$ represents the probability of the Discriminator classifying generated data as real. $E_{x\sim p_{\text{data}}(X)}$ and $E_{z\tilde{p}_z(z)}$ represent the expectations that are taken over the real data distribution and the distribution of latent variables.

Despite the impressive performance of GAN in many applications, it still faces several issues. For instance, the training processes of GAN are often unstable and prone to problems such as mode collapse and vanishing gradients. To address these issues, Wasserstein distance has been proposed as an improved strategy for GAN.

Wasserstein distance is a measure of the distance between distributions, providing a more accurate assessment of the dissimilarity between two distributions compared to KL divergence and JS divergence. Therefore, incorporating Wasserstein distance as a metric in GAN allows for a better evaluation of the disparities between generated and real data. The Wasserstein

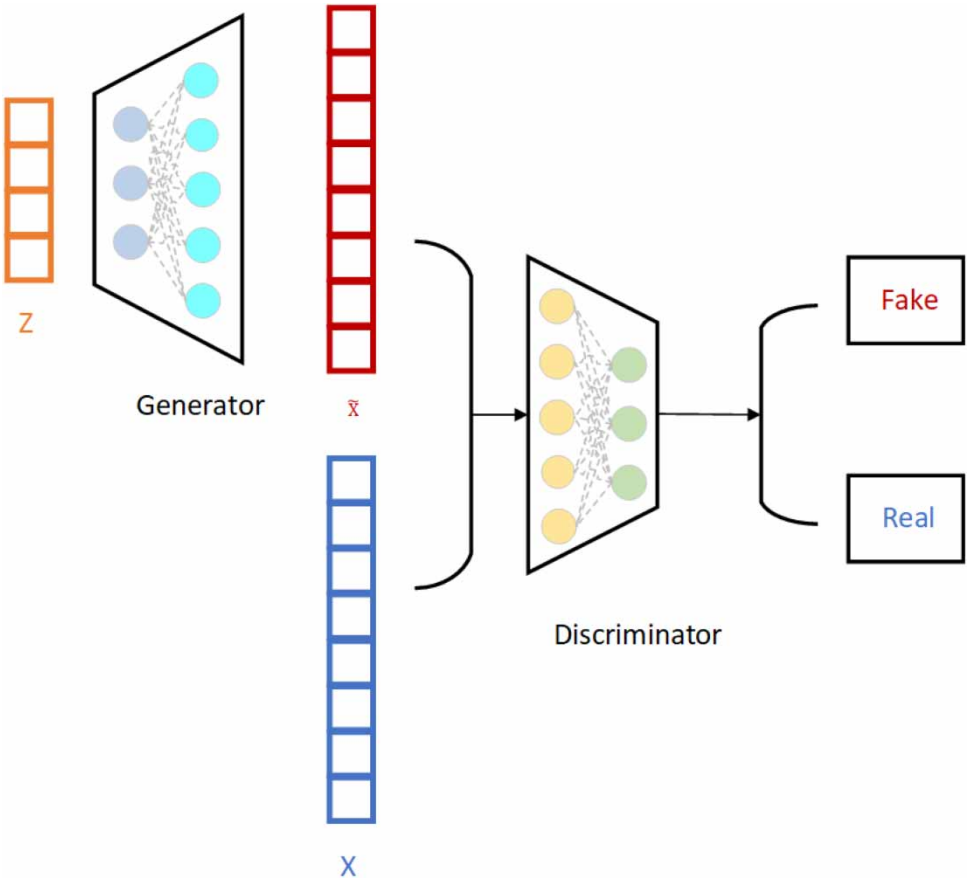



**Figure 2** | GAN network structure diagram.

distance can be represented by the following formula:

$$W(p,q) = \inf_{\gamma \sim \prod(p,q)} E_{x,y\sim\gamma}[\|x-y\|] \tag{5}$$

where $\prod(p,q)$ represents the joint distribution obtained by combining all possible combinations of $p$ and $q$. For each possible joint distribution $\gamma$, a sample $x$ and $y$ can be generated, and $\|x-y$ represents the computation of the distance between this pair of samples. The Wasserstein distance can be understood as the difference between the expected values obtained from one distribution and the other in an optimal scenario. The superiority of the Wasserstein distance over KL divergence and JS divergence lies in its ability to reflect the degree of separation between two distributions, even when they have no overlap.

Therefore, the loss function of WGAN is modified as:

$$\min_G \max_{f\in F} V(f,G) = E_{x\bar{p}}[f(x)] - E_{z\sim p_z(z)}[f(G(z))] \tag{6}$$

Here, $F$ represents the set of all functions that satisfy the Lipschitz continuity constraint, $E_{x\sim p_{\text{data}}(X)}[f(x)]$ represents the expectation of $f(x)$ over real data, and $E_{z\sim p_z(z)}[f(G(z))]$ represents the expectation of $f(x)$ over generated data.

WGAN–GP, based on WGAN, introduces a GP to address the issues of gradient vanishing and mode collapse in WGAN, further enhancing the performance and stability of the model. Specifically, when computing the GP, a set of real samples and generated samples are randomly sampled to ensure the Lipschitz continuity of the discriminator. The sampling procedure is

as follows:

$$\hat{x} = \varepsilon x_r + (1 - \varepsilon)x_g,\ \varepsilon \sim U[0, 1] \tag{7}$$

Here, $\hat{x}$ represents the sampled fake samples, $x_r$ represents the real samples, $x_g$ represents the generated samples, and $\varepsilon$ is a value randomly sampled from a uniform distribution [0,1]. To ensure the Lipschitz continuity of the discriminator, it is required that the gradient norm of the discriminator is equal to 1, which can be expressed as:

$$\|\nabla_{\hat{x}} D_{\hat{x}}\|_2 = 1 \tag{8}$$

Therefore, the loss function of the discriminator is modified as:

$$L_{\mathrm{GP}} = \lambda E_{\hat{x}\sim P(\hat{x})}[(\|\nabla_{\hat{x}} D_{\hat{x}}\|_2 - 1)^2] \tag{9}$$

where $\lambda$ is a hyperparameter that controls the weight of the GP, and $P(\hat{x})$ represents the distribution of fake samples. By incorporating this approach, we can ensure that the discriminator satisfies Lipschitz continuity, thereby addressing the issues of gradient vanishing and mode collapse in WGAN.

### VAE–WGAN–GP

We use a VAE network as the generator model for the GAN network, leveraging the excellent feature extraction capability of VAE to capture the latent space distribution of complex input data. By sampling from the latent space distribution, we reconstruct the data and optimize the reconstruction error and KL divergence to ensure that the generator learns the characteristics of the real data distribution. Additionally, we combine the optimization training strategy of WGAN–GP to enhance the stability of the training process, improve the quality of generated samples, and enhance the model's generalization ability. The overall architecture of the model is illustrated in Figure 3.

The encoder of the VAE generator maps the input data $X$ to a latent vector $z$, while the decoder maps the latent vector $z$ back to the input space, generating similar data $\widetilde{X}$ to the input. The discriminator assesses whether the generated data $\tilde{X}$

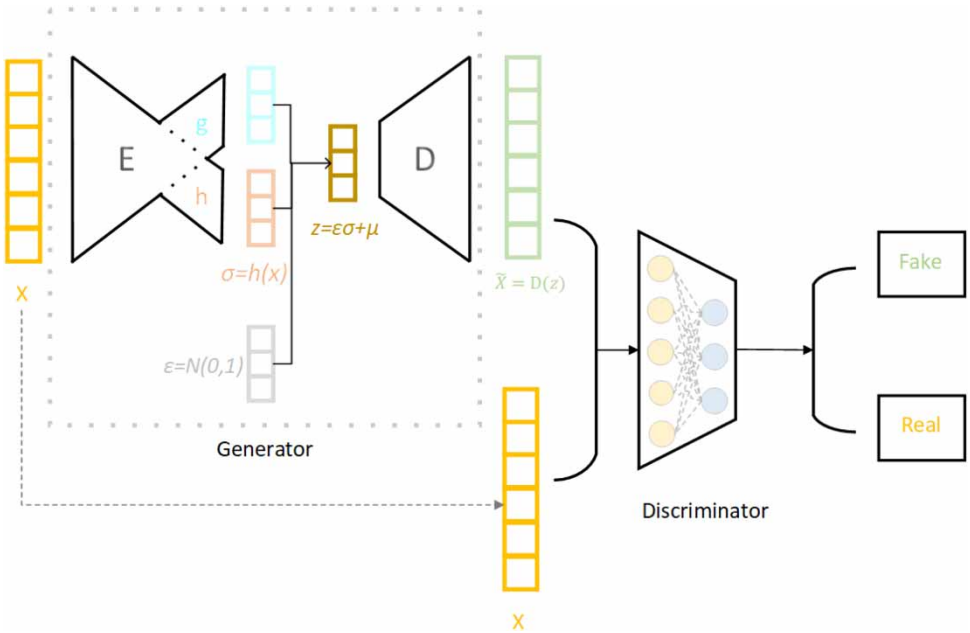



**Figure 3** | VAE–WGAN–GP network structure diagram.

resembles the input data $X$. The specific loss function of the model is as follows:

$$L_{\text{VAE–WGAN–GP}} = L_{\text{VAE}} + L_{\text{WGAN}} + \lambda L_{\text{GP}} \tag{10}$$

$$\begin{aligned} L_{\text{VAE}} &= L_{\text{rec}} + L_{\text{prior}} \\ &= -E_{q_\varphi(z|X)}[\log p_\theta(x|z)] + D_{\text{KL}}(q_\varphi(z|X)||p_\theta(z)) \end{aligned} \tag{11}$$

$$L_{\text{WGAN}} = \max_D E_{x\sim p_{\text{data}}}[D(x)] - E_{z\sim p_z}[D(G(z))] \tag{12}$$

$$L_{\text{GP}} = \lambda E_{\hat{x}\sim P(\hat{x})}[(\|\nabla_{\hat{x}} D_{\hat{x}}\|_2 - 1)^2)] \tag{13}$$

The VAE–WGAN–GP model combines the advantages of VAE and WGAN–GP, ensuring both high-quality generated samples and more stable learning of latent representations. The reconstruction error guarantees the quality of generated samples, the Wasserstein distance promotes training stability, and the GP term further improves the model's performance.

### VAE–WGAN–GP algorithm

The specific training process of the model consists of the following steps:

Step 1: Select samples as the training data for the model and perform preprocessing tasks, such as denoising and normalization.

Step 2: Divide the samples into batches $X^m$ and feed the batched samples $X^m$ into the network to capture the latent distribution information of the data.

Step 3: Initialize the network parameters.

Step 4: Train the VAE generator model and discriminator model in a step-by-step manner, where the discriminator is trained once every five training iterations of the generator.

Step 5: Apply weight clipping in the model using an iterative training approach to update the weights. The model optimization adopts RMSProp and Adam optimization methods until the model converges. From the final training results, select the best-performing model.

Here is the pseudo code:

---

**Require**: Gradient penalty weight $\lambda$, KL loss efficiency $\gamma$, Minimum batch size $m$, learning rate $\alpha$, iterations $i$.

**Require:** $\theta$, $\varphi$, $\omega$ initial parameters。

**For $i$ = 1,2..Repeat until convergence**

$X^m \leftarrow$ Minimum batch division of the dataset

$\mu, \sigma \leftarrow g(X), h(X)$

$\varepsilon \leftarrow$ sample from price $N(0,I)$

$Z \leftarrow \varepsilon^* \sigma + \mu$

$\widetilde{X} \leftarrow Dec(Z)$

$L_{prior} \leftarrow \gamma D_{KL}(q_\varphi(z|X)||p_\theta(z))$

$L_{rec} \leftarrow -E_{q_\varphi(z|X)}[log\, p_\theta(x|z)]$

$\tau \leftarrow$ Sample a random number from price $U[0, 1]$

$\hat{x} \leftarrow \tau X + (1-\tau)\tilde{X}$

$$\begin{aligned} L_{VAE-WGAN-GP} \leftarrow & -E_{q_\varphi(z|X)}[log\, p_\theta(x|z)] + \gamma D_{KL}(q_\varphi(z|X)||p_\theta(z)) \\ & + E_{x\sim p_{data}}[D_\omega(x)] - E_{z\sim p_z}[D_\omega(G(z))] + \lambda E_{\hat{x}\sim P(\hat{x})}[(\|\nabla_{\hat{x}} D_\omega \hat{x}\|_2 - 1)^2] \end{aligned}$$

$\theta^*, \varphi^* \leftarrow -\nabla_{\theta,\varphi} L_{VAE}$

$\omega^* \leftarrow -\nabla_\omega L_{WGAN-GP}$

**End for**

---

## EXPERIMENT

To ensure the rigor and scientific validity of the experiment, we extracted section water quality data from the National Surface Water Real-time Monitoring System. We used these data for conducting model comparison experiments and data prediction experiments. In the following sections, we will provide detailed explanations of the experimental dataset, the experimental content, model setup, and the specific experimental results.

### Experimental dataset

The dataset used in the experimental part of this study is obtained from the National Surface Water Real-time Monitoring System (Ministry of Ecology and Environment; https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html). This system covers water bodies in various geographical locations and environmental conditions, ensuring the diversity and effectiveness of the experimental data.

The National Surface Water Real-time Monitoring System assesses multiple indicators and categorizes water quality into Class I to Class V (Class V being the worst) based on Table 1. These indicators include pH value, ammonia nitrogen, dissolved oxygen, conductivity, turbidity, permanganate index, total phosphorus, and total nitrogen. Among them, indicators such as ammonia nitrogen, dissolved oxygen, permanganate index, total phosphorus, and total nitrogen are crucial for reflecting the eutrophication and organic pollution levels of water bodies. Additionally, water turbidity and conductivity reflect the presence of visible suspended matter and salt content in the water.

Considering the variability of water quality in surface waters due to diverse natural and anthropogenic factors across different regions, this study focused on selecting water quality data from sections within the Yangtze River Basin for investigation. During the process of data collection and processing, strict adherence to relevant standards and protocols was followed. Specifically, the preprocessing procedures encompassed deduplication, screening, imputation of missing values, and treatment of outliers. Moreover, consistent training and testing dataset was consistently employed throughout subsequent model training endeavors, ensuring the reliability and precision of experimental data.

### Experimental content and model setup

The experimental framework of this study primarily consists of two components: comparative experiments of models and data prediction experiments. The first part involves establishing four generative models: VAE, GAN, Wasserstein GAN (WGAN), and VAE–WGA–GP, to generate a new dataset using training samples. Subsequently, the generated data are added to the training set and used to evaluate various classification models. The selected classification models include SVM, logistic regression (LR), and RF. The comparative experiments of models focus on exploring the variations in model loss functions and assessing the convergence of models during the training process. Additionally, the mixed data are fed into different classification models to initially evaluate the performance of the generated models. The second part aims to investigate the influence of generated data from different types of samples on predictive performance. Different proportions of generated data are determined based on the varying deviation values among sample classes. Furthermore, the impact of the generated data on classification results is examined through illustrative representations.

**Table 1** | Surface water environmental quality standards

| Category | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 |
|---|---|---|---|---|---|
| pH | 6 ~ 9 | | | | |
| DO (mg/L) ≥ | 7.5 | 6 | 5 | 3 | 2 |
| $COD_{Mn}$ (mg/L) ≤ | 2 | 4 | 6 | 10 | 15 |
| COD (mg/L) ≤ | 15 | 15 | 20 | 30 | 40 |
| BOD5 (mg/L) ≤ | 3 | 3 | 4 | 6 | 10 |
| NH3-N (mg/L) ≤ | 0.15 | 0.5 | 1.0 | 1.5 | 2.0 |
| TP (mg/L) ≤ | 0.02 | 0.1 | 0.2 | 0.3 | 0.4 |
| TN (mg/L) ≤ | 0.2 | 0.5 | 1.0 | 1.5 | 2.0 |

DO, dissolved oxygen; CODmn. permanganate index; COD, chemical oxygen demand; BOD5, biochemical oxygen demand; NH3-N, ammonia nitrogen; TP, total phosphorus; TN, total nitrogen.

The model configurations primarily involve the design of the generative model network architecture, training parameters, optimization strategies, and the determination of evaluation metrics. The final settings are as follows:

1. VAE: The decoder consists of three convolutional layers, followed by flattening and two fully connected latent variable layers ($\mu$ and $\sigma$). The encoder comprises four convolutional layers, with the final convolutional layer maintaining input–output dimensions.
2. GAN: The generator is a four-layer perceptron neural network, while the discriminator is a three-layer convolutional network. The last layer is flattened and fully connected to provide weighted outputs.
3. WGAN: The generator and discriminator settings are the same as in 2. During each iteration, the training ratio between the generator and discriminator is set to 1:5.
4. VAE–WGAN–GP: The generator adopts the decoder and encoder structures from 1, while the discriminator settings remain the same as in 3. The gradient penalty coefficient $\lambda$ is set to 10, and the divergence loss coefficient KL for VAE is set to 1.

The training was conducted with a batch size of 48 and a learning rate of $\alpha = 0.001$, employing the stochastic gradient descent method for model training. The generator $G$ was optimized using RMSprop, while the discriminator $D$ was optimized using Adam. Evaluation metrics for assessing the performance included precision for each category $P_i$ and the overall prediction accuracy (Accuracy):

$$P_i = \mathrm{TP}_i/(\mathrm{TP}_i + \mathrm{FP}_i) \tag{14}$$

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} P_i}{n} \tag{15}$$

where $\mathrm{TP}_i$ represents the number of correctly predicted instances for each category $i$, $\mathrm{FP}_i$ represents the number of incorrectly predicted instances for each category $i$, and $n$ represents the total number of categories considered.

## Experimental results

In this section, we selected water quality data from sections within the Yangtze River Basin in Hubei Province, obtained from the monitoring system, as our sample dataset. After preprocessing operations, including deduplication, screening, imputation of missing values, and treatment of outliers, we obtained a total of 2,936 usable data instances. Each data instance consists of nine indicator variables: water temperature, pH, dissolved oxygen, conductivity, turbidity, permanganate index, ammonia nitrogen, total phosphorus, and total nitrogen. The last entry in each data instance represents the current water quality evaluation grade. We divided the dataset into 70% for training and 30% for testing, and the composition is shown in Figure 4.

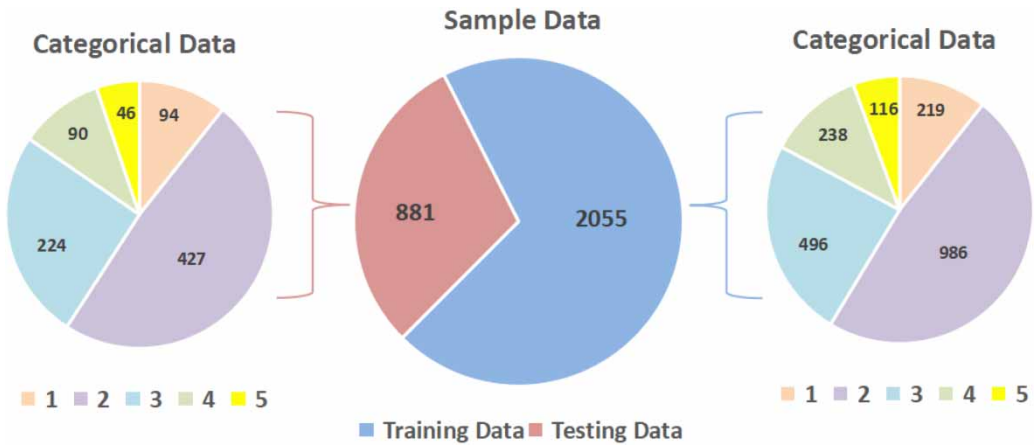



**Figure 4** | Dataset composition.

The middle pie chart illustrates the composition distribution of the training and testing sets within the overall dataset. The left and right pie charts represent the distribution of data belonging to classes 1–5 within the testing and training dataset, respectively. It is evident that there is a significant class imbalance in the data distribution, with class 2 having the highest number of instances and class 5 having the lowest. We trained the four generative models using the training dataset and monitored the loss of the generator and discriminator during the training process, as shown in Figure 5.

Due to the VAE model's exclusive focus on generating sample reconstruction similarity and its lack of engagement in adversarial training with genuine and counterfeit samples, Figure 5(b) discriminator training process does not elucidate its model training procedure. From Figure 5, it can be observed that the generator and discriminator losses in GAN exhibit significant oscillations, and the training process requires a longer duration. On the other hand, WGAN, which employs Wasserstein distance instead of KL divergence for loss calculation and applies the Lipschitz condition as a constraint, demonstrates a noticeable improvement in convergence speed compared to GAN. VAE-GAN-GP, incorporating VAE as the generator to capture latent distribution information of the data, exhibits superior convergence speed and training loss compared to the other three models.

Subsequently, we utilized the trained generative models to generate synthetic samples, while also incorporating the traditional SMOTE oversampling method for comparison. The underlying concept of the SMOTE oversampling method primarily involves 'sampling' a new sample by weighting the variable difference between 0 and 1 randomly for samples of the same category. Since our study focuses on the quality of the synthetic samples generated by each classification model, it is essential to maintain the same class proportions in the synthetic samples as in the training set. In other words, we aim to generate pseudo-training samples using the four generative models in conjunction with the SMOTE oversampling method, such that the proportions and quantities of each class align with those of the original training dataset. This

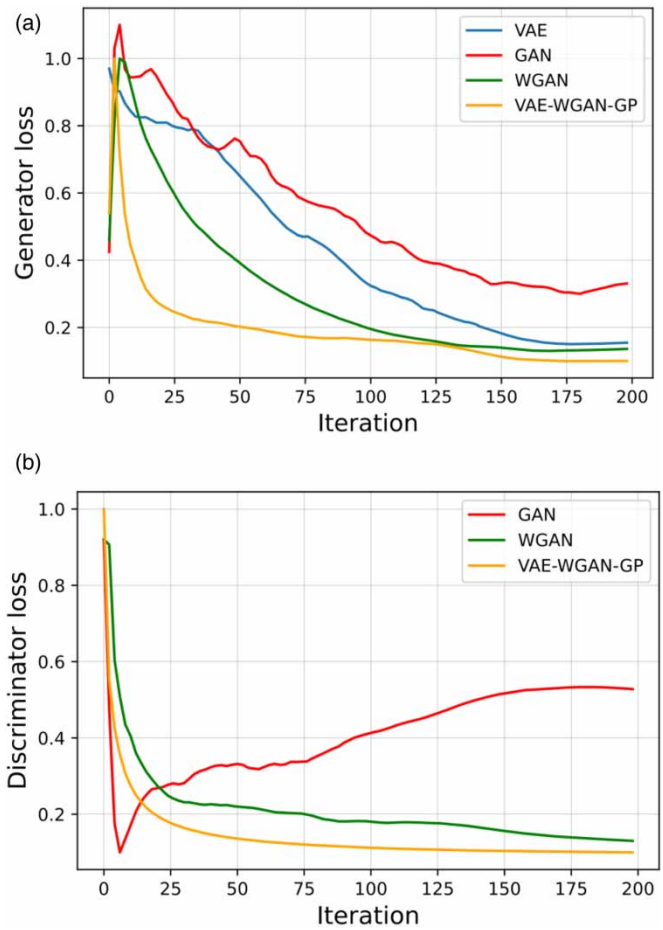



**Figure 5** | Training process diagram. (a) Generator loss and (b) discriminator loss.

pseudo-training dataset was then used for training various classification models, and their performance was evaluated using the testing set. Table 2 presents the dataset used to train the classification models.

To facilitate a comprehensive comparison of the models, we employed overall precision as the evaluation metric for predictions and obtained the prediction precision rates for each classification model using the aforementioned dataset. The results are presented in the graph shown in Figure 6.

Figure 6 clearly shows that the data generated by GAN perform poorly in the SVM, LR, and RF classification models, with only a slight improvement compared to the traditional SMOTE oversampling method. This can be attributed to the inherent difficulty in training GAN, which makes achieving good convergence challenging. The traditional SMOTE oversampling method, which relies on random sampling strategies to augment samples, often struggles to capture the underlying distributional information of the actual data. In fact, the prediction accuracy of this approach is largely attributed to 'imitating' the original samples. Furthermore, the samples generated by GAN are based on random noise and may not accurately capture the depth information in the original data. However, the data generated by VAE show slightly better performance compared to GAN. However, VAE only considers the reconstruction similarity of generated samples and does not differentiate between genuine and fake samples, resulting in lower-quality generated samples. In contrast, the data generated by VAE–WGAN–GP show greater similarity to the real samples. Although the prediction accuracy of SVM and RF classification models is slightly lower than that of the original data (possibly due to the introduction of some noise in the generated data's class information), utilizing the generated data for auxiliary predictions prove to be effective. Furthermore, from the graph, we can observe that the RF classification model consistently achieves the highest predictive accuracy among the three classification models. One plausible explanation for this phenomenon is the ensemble prediction mechanism of RF, which leverages weak learners. During the construction of each tree, random feature selection and sample splitting are employed, effectively reducing model variance, enhancing model robustness, and facilitating the handling of high-dimensional data information. Indeed, when compared to LR and SVM, RF exhibits a superior ability to capture nonlinear relationships and adjust to nonlinear decision boundaries.

**Table 2** | Dataset used for training

| Category | Dataset |
|---|---|
| 1 | Real dataset |
| 2 | SMOTE generated the dataset |
| 3 | VAE generated the dataset |
| 4 | GAN generated the dataset |
| 5 | WGAN generated the dataset |
| 6 | VAE–WGAN–GP generated the dataset |

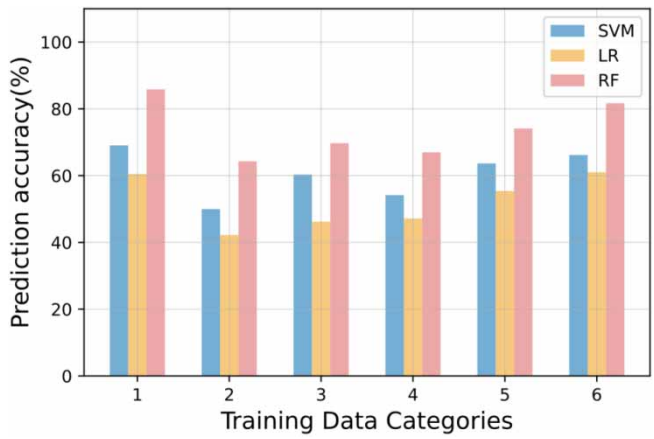



**Figure 6** | Forecast graph for each dataset.

In the second part of the experiment, we selected the VAE–WGAN–GP model as the generative model. By setting different generation ratios for each class of data, we were able to control the quantity of generated data for different classes and further investigate the impact of generated data from different classes on prediction performance. Due to the uneven distribution of different classes in the original dataset, the trained classification models tend to be biased towards the majority class, resulting in poor classification performance for the minority class. Table 3 displays the deviation degree (DD) and compensation degree (CD) for different class samples.

DD measures the degree of category sample imbalance, while CD is the reciprocal of DD and measures the factor by which the category needs to be increased to achieve perfect balance. The calculation is as follows:

$$\mathrm{DD} = M/\mathrm{MAJ} \tag{16}$$

$$\mathrm{CD} = \mathrm{MAJ}/M \tag{17}$$

MAJ represents the number of most categories in the sample set and $M$ represents the number of current categories.

Generally, to achieve balance within the sample set, it is necessary to ensure an equal number of samples for each category. This implies that the current sample count for each category should be compensated by a factor of CD-1. For instance, the category with the highest number of samples would require no compensation, while the category with the lowest sample counts would need to be compensated by a factor of seven times their own count. To facilitate the observation of changes in predictive accuracy during the data compensation process, we divided the compensation procedure for each class into 20 steps and calculated the precision at each step, as illustrated in Figure 7.

Figure 7 illustrates the process of achieving sample balance for the four classes through the generation of model-generated samples. The horizontal axis ranges from 0 to 1 with a step size of 0.05. A value of 0 indicates that the current dataset consists entirely of original data, while a value of 1 indicates that the class has reached complete balance through the generation of additional data at a factor of CD-1 times the original size. The vertical axis represents the improvement rate of the three classification models for the class after incorporating the compensation data into the original dataset, averaged over five computations.

The observed changes in precision for the three classes mentioned above indicate that data augmentation can increase the predictive accuracy of a class to a certain extent. However, as the amount of augmented data continues to increase, the precision gradually declines, and in the case of the RF model, it may even fall below that of the original data. The underlying cause of this phenomenon may be attributed to the increasing proportion of generated samples in the dataset, which leads to the classification model being more influenced by the noise information carried by the generated samples, thereby overlooking the intrinsic mapping relationships within the original samples. The precision of class 5 across the RF classification models does not exhibit significant improvement. Both the RF and LR models demonstrate a sustained downward trend. This could be attributed to the high level of class imbalance within class 5, where the original class samples fail to adequately reflect the internal data distribution. Moreover, the generated data introduce a certain degree of randomness and noise.

We determine the optimal compensation factor for the aforementioned four classes by selecting the compensation factor that yields the highest average precision across the three models during the compensation process. Subsequently, we generate augmented data for the four classes in the original dataset. To facilitate the observation of changes in the class distribution before and after data compensation, we utilize Figure 8 to visualize the results.

Three classification models are trained using the compensated dataset, and the confusion matrix is shown in Figure 9.

Figure 9 presents the confusion matrix for each of the three classification models, depicting their predictions on both the original data and the compensating data. All three models demonstrate an improvement in predictive performance after data

**Table 3** | DD and CD tables

| Category | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 |
|---|---|---|---|---|---|
| DD | 0.222 | 1 | 0.503 | 0.241 | 0.118 |
| CD | 4.502 | 1 | 1.988 | 4.143 | 8.5 |

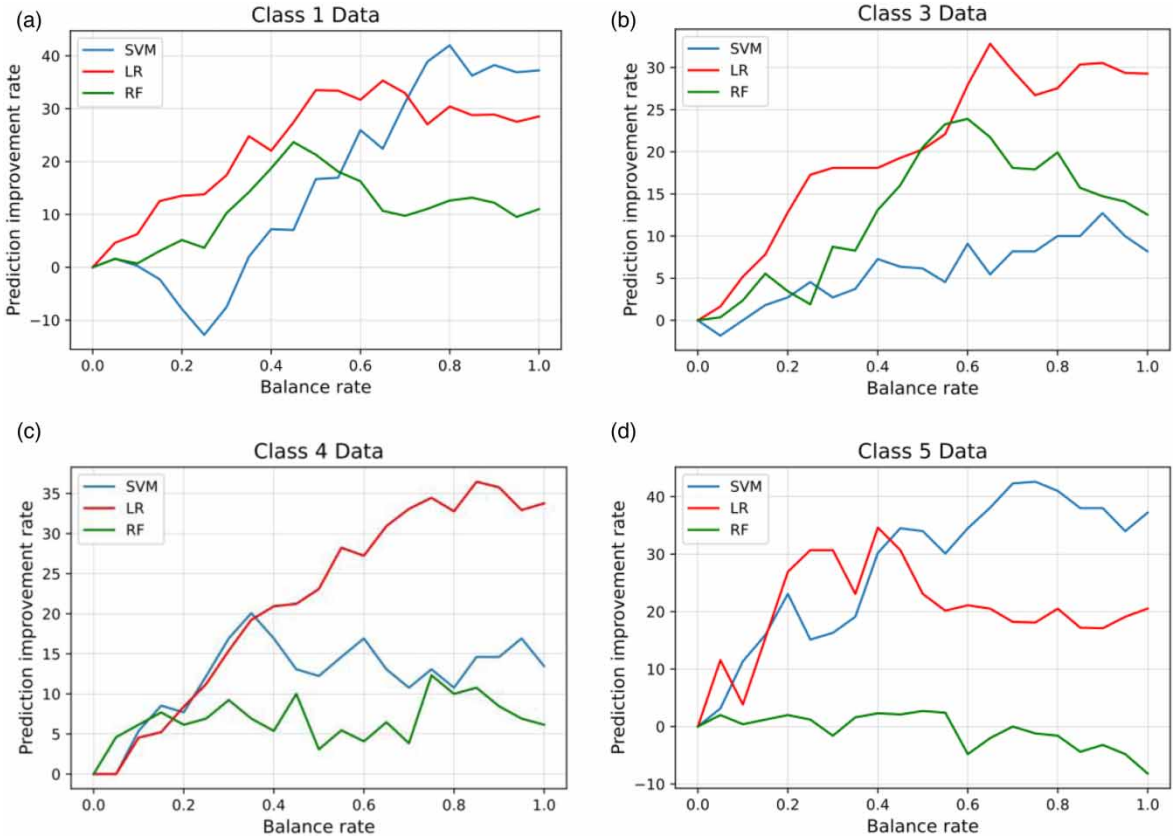



**Figure 7** | Prediction improvement rates by class. (a) Class 1, (b) Class 3, (c) Class 4, and (d) Class 5.

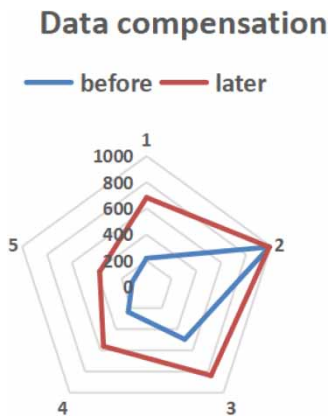



**Figure 8** | Data compensation graph.

compensation. To quantify the effect, we calculate the precision for each class in the three classification models before and after data compensation. These values are shown in Figure 10.

To more clearly observe the predictive performance of the three classification models before and after data compensation, we have selected the overall class as the evaluation criterion and calculated the inference time (IT) and the predictive accuracy (AUC) of the three classification models before and after data compensation, as shown in Table 4.

After compensating for the data, the SVM, LR, and RF models show significant improvements in predicting the accuracy of individual classes. The overall precision rates have increased by 14.1, 12.7, and 9.7%, respectively. From Table 4, it is evident

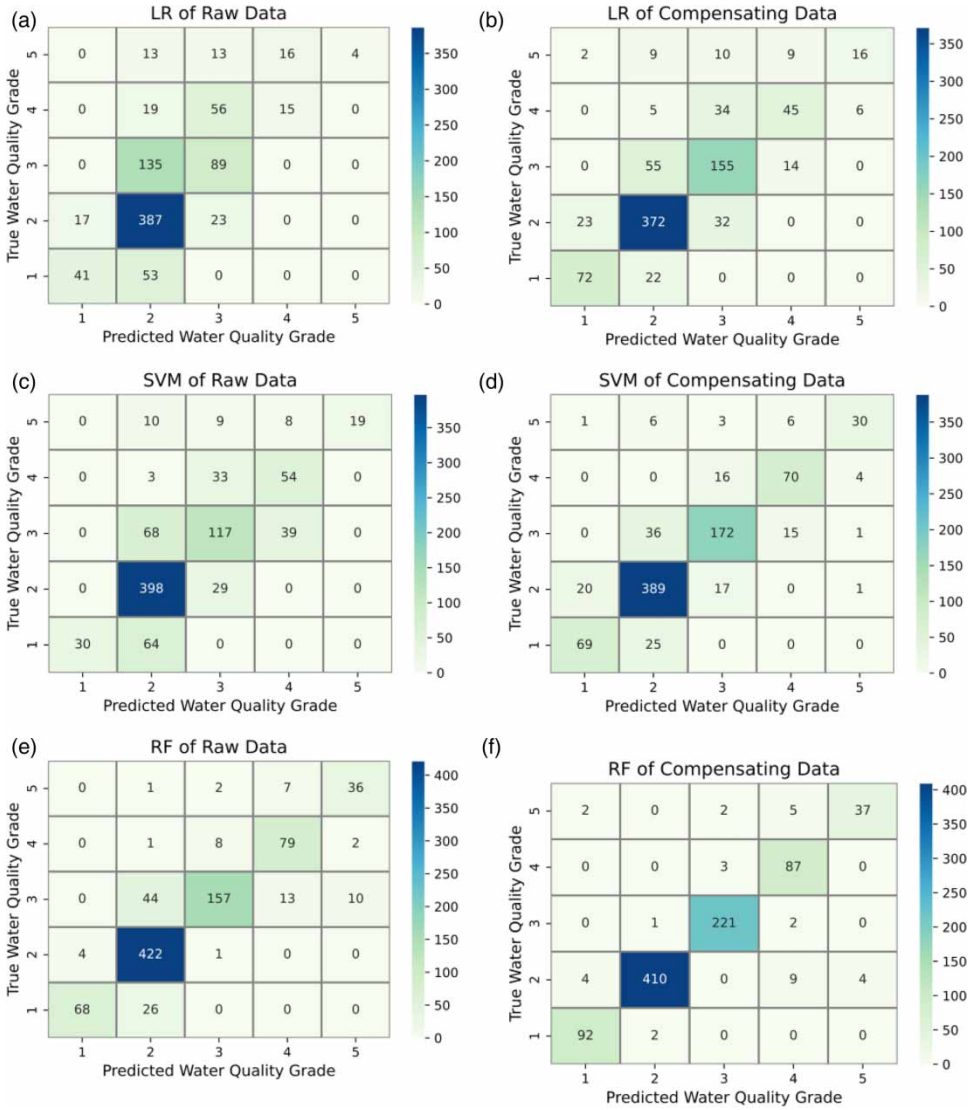



**Figure 9** | Confusion matrix figures. (a) LR of raw data, (b) LR of compensating data, (c) SVM of raw data, (d) SVM of compensating data, (e) RF of raw data, and (f) RF of compensating data.


**Figure 10** | Prediction accuracy before and after data compensation. (a) LR, (b) SVM, and (c) RF.

that the overall inference times rank as LR < SVM < RF, while the overall prediction accuracies follow the order RF > SVM > LR. The LR model, owing to its straightforward inference scheme, achieved the fastest inference speed. However, it struggled to fit well into nonlinear decision boundaries. Furthermore, the imbalanced distribution of sample categories

**Table 4** | Prediction accuracy comparison

| Dataset | Raw data | | Compensation data | |
|---|---|---|---|---|
| Metrics | IT (ms) | AUC (%) | IT (ms) | AUC (%) |
| LR | 193 | 60.8 | 274 | **74.9** |
| SVM | 439 | 70.1 | 601 | **82.8** |
| RF | 593 | 86.4 | 735 | **96.1** |

Bold entries emphasize the enhancement in model prediction accuracy.

led to a bias in decision boundaries toward the majority class, resulting in suboptimal predictive performance. The SVM model, with the introduction of kernel functions, exhibited improved capabilities in handling nonlinear problems. By leveraging kernel functions for dimensional transformations and hyperplane divisions, it achieved further enhancements in prediction accuracy. Nevertheless, the SVM model faces limitations concerning the choice of kernel functions and is susceptible to bias in hyperplane divisions due to uneven sample category distributions. The RF model, thanks to its unique ensemble mechanism of weak classifiers, demonstrated excellent fitting abilities for nonlinear decision boundaries and attained the highest prediction accuracy among the three models. However, this came at the expense of a longer training time, as each weak classifier required individual training. With the inclusion of compensated data, the imbalance in sample distribution was moderately alleviated. As a result, inference times for all three models saw slight increases, accompanied by notable improvements in prediction accuracy. This effectively demonstrates how the VAE–WGAN–GP model mitigates the impact of imbalanced samples by compensating for the data.

## CONCLUSION

This paper focuses on the issue of imbalanced class samples in water quality assessment and proposes a balanced approach that utilizes the deep generative network VAE–WGAN–GP to compensate for the data of scarce categories. This network combines the encoding and decoding mechanisms of a VAE model to learn deep feature representations of the data using variational inference techniques. Simultaneously, it employs the adversarial training mechanism between the generator and discriminator enhancing the richness and diversity of generated samples. Furthermore, during model training, it incorporates Wasserstein distance and GP techniques to highlight the model's stability and convergence speed through experimental comparisons. In the next step, we compare data compensation experiments using different generative models to validate the efficient learning capability and superior feature reconstruction ability of the VAE–WGAN–GP model. Subsequently, we define DD and CD for each class and employ the optimal compensation strategy for each class to achieve data sample balancing. This approach significantly improves the predictive accuracy of the evaluation on the test set and provides valuable guidance for water quality assessment with limited and imbalanced sample data.

In the field of water quality assessment, the issue of imbalanced class samples is a common yet often overlooked problem. Evaluation models constructed based on the original data tend to favor the majority class samples, leading to inaccurate predictions for minority class samples. The primary contribution of this study is the innovative introduction of a generative network approach to compensate for the original sample set. This approach alleviates the imbalance in class sample distribution and improves the predictive accuracy of the model. However, it is important to note that this method has certain limitations. Firstly, since the network employs VAE to learn deep distribution information from the data, it implies that there must be a sufficient quantity of data for rare class samples to accurately represent their internal distribution. Secondly, the distribution information learned by the model is based on statistical class data distribution and cannot address significant temporal variations. In future work, we plan to apply the proposed method to undertake more meaningful tasks in water quality assessment. This may include enhancing predictive accuracy by combining the compensated data with more precise evaluation models or delving deeper into the representation of deep distribution features in the data to acquire more realistic and comprehensive compensated samples.

## DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES


Al-Shabi, M. A. 2019 Credit card fraud detection using autoencoder model in unbalanced datasets. *Journal of Advances in Mathematics and Computer Science* **33** (5), 1–16.

Bhagwani, H., Agarwal, S., Kodipalli, A. & Martis, R. J. 2021 Targeting class imbalance problem using GAN. In: *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*. IEEE, pp. 318–322.

Dal Pozzolo, A., Caelen, O., Johnson, R. A. & Bontempi, G. 2015 Calibrating probability with undersampling for unbalanced classification. In: *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, Cape Town, South Africa, pp. 159–166.

Ding, H. & Cui, X. 2023 A clustering and generative adversarial networks-based hybrid approach for imbalanced data classification. *Journal of Ambient Intelligence and Humanized Computing* **2023**, 1–16.

Douzas, G. & Bacao, F. 2018 Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications* **91**, 464–471.

Gao, X., Deng, F. & Yue, X. 2020 Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. *Neurocomputing* **396**, 487–494.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. & Bharath, A. A. 2014 Generative adversarial networks. arXiv preprint arXiv: **1406**.2661, 1406.

Icaga, Y. 2007 Fuzzy evaluation of water quality classification. *Ecological Indicators* **7** (3), 710–718.

Luo, W., Yang, W., He, J., Huang, H., Chi, H., Wu, J. & Shen, Y. 2022 Fault diagnosis method based on two-stage GAN for data imbalance. *IEEE Sensors Journal* **22** (22), 21961–21973.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C. & Malossi, C. 2018 Bagan: Data augmentation with balancing gan. arXiv preprint arXiv: **1803**.09655.

More, A. 2016 Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv Preprint ArXiv* **1608**, 06048.

Nong, X., Shao, D., Zhong, H. & Liang, J. 2020 Evaluation of water quality in the south-to-north water diversion project of China using the water quality index (WQI) method. *Water Research* **178**, 115781.

Štambuk-Giljanović, N. 1999 Water quality evaluation by index in Dalmatia. *Water Research* **33** (16), 3423–3440.

Tyagi, S., Sharma, B., Singh, P. & Dobhal, R. 2013 Water quality assessment in terms of water quality index. *American Journal of Water Resources* **1** (3), 34–38.

Vuttipittayamongkol, P., Elyan, E. & Petrovski, A. 2021 On the class overlap problem in imbalanced data classification. *Knowledge-based Systems* **212**, 106631.

Wang, Q. & Yang, Z. 2016 Industrial water pollution, water environment treatment, and health risks in China. *Environmental Pollution* **218**, 358–365.

Wang, X., Zhang, F. & Ding, J. 2017 Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific Reports* **7** (1), 12858.

Xu, Y., Zhang, X., Qiu, Z., Zhang, X., Qiu, J. & Zhang, H. 2021 Oversampling imbalanced data based on convergent WGAN for network threat detection. *Security and Communication Networks* **2021**, 1–14.

Zhang, J., Bloedorn, E., Rosen, L. & Venese, D. 2004 Learning rules from highly unbalanced datasets. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*. IEEE, pp. 571–574.

Zhu, B., Pan, X., vanden Broucke, S. & Xiao, J. 2022 A GAN-based hybrid sampling method for imbalanced customer classification. *Information Sciences* **609**, 1397–1411.


First received 25 May 2023; accepted in revised form 8 November 2023. Available online 18 November 2023