Diffusion Model Based on Reverse Guidance of Regional Samples

GUAN YANG, Yunnan University of Finance and Economics, China
ZHIYONG ZENG, Yunnan University of Finance and Economics, China
REN DUAN*, Yunnan University of Finance and Economics, China

When discussing the classification of imbalanced datasets, due to their distribution characteristics, the scarce minority class makes the traditional classification methods biased toward the majority class, reducing minority class recognition. This article mainly starts with the data-level method. It expands the sample size of the minority class by a generative model to improve the classification accuracy and reduce the misclassification cost. Based on the characteristics of the complex distribution of the minority class and the advantages of Diffusion Models, this article proposed a **L**ocal **Re**gional Samples Guidance **D**enoising **D**iffusion **P**robabilistic **M**odel (LReDDPM). The method first divides the sample types of the minority class, takes the gradient information of regional samples as the condition, and then uses the denoising diffusion probabilistic model to generate minority class examples. The generated minority class examples are added to the training set to expand the sample size, enriching the local sample density of the minority class. In addition, we explore diffusion models guided by gradients derived from samples in different regions. The experimental results demonstrate that examples generated by models guided by samples from different regions exhibit varying degrees of improvement in classification performance, with the most significant enhancement observed in the safety and boundary regions. It further indicates that the complex distribution of the minority class plays a crucial role in the classification results. We conduct experiments on ten datasets and compare our results with those of five methods to evaluate the superiority and effectiveness of LReDDPM's method. The final experimental results show that the proposed method can significantly improve classification performance.

JAIR Associate Editor: Ivor Tsang

JAIR Reference Format:
Guan Yang, Zhiyong Zeng, and Ren Duan. 2026. Diffusion Model Based on Reverse Guidance of Regional Samples. *Journal of Artificial Intelligence Research* 85, Article 27 (March 2026), 25 pages. DOI: 10.1613/jair.1.18916

1 Introduction

Challenges in training classifiers on imbalanced datasets have been observed in numerous real-world applications (Chawla, Japkowicz, et al. 2004; He and Garcia 2009; Sun, Wong, et al. 2009). Although there are several avenues for improvement, including algorithmic-level methods (Khan et al. 2017; Krawczyk et al. 2014) and data-level strategies (Barua et al. 2012; Bunkhumpornpat et al. 2009; Cahyana et al. 2019; Xie et al. 2020), the challenge of enhancing classification performance based on the characteristics of complex distributions of the minority class in imbalanced datasets remains a significant issue worthy of exploration.

Recently, diffusion models based on logarithmic likelihood have developed rapidly (Ho et al. 2020; A. Q. Nichol and Dhariwal 2021; Sohl-Dickstein et al. 2015; Song, Durkan, et al. 2021). They demonstrate significant advantages in data distribution modeling and data synthesis, allowing for control over the model generation process and the quality of generated outputs across a range of tasks, including text-to-image generation (A. Nichol et al. 2021),

*Corresponding Author.

Authors' Contact Information: Guan Yang, ORCID: 0009-0002-8144-0472, yangguan@stu.ynufe.edu.cn, Yunnan University of Finance and Economics, Kunming, Yunnan, China; Zhiyong Zeng, zengzhiyong725@163.com, Yunnan University of Finance and Economics, Kunming, Yunnan, China; Ren Duan, duanren@ynufe.edu.cn, Yunnan University of Finance and Economics, Kunming, Yunnan, China.

DOI: 10.1613/jair.1.18916

high-resolution image synthesis (Rombach et al. 2022), image editing (Meng et al. 2021), speech synthesis (M. Jeong et al. 2021), and medical imaging (Song, Shen, et al. 2021). Most of the existing diffusion models are under a uniform data distribution assumption. Unconditional diffusion models tend to generate a disproportionately high number of low-quality samples when the data distribution is non-uniform. However, in the context of generating samples from imbalanced datasets, diffusion models primarily focus on image data (Li et al. 2024; Qin et al. 2023; Shao et al. 2024; Wyatt et al. 2022; Yan et al. 2024). Previous studies have demonstrated that when diffusion models are trained on datasets with uneven class distributions, the diversity and fidelity of the generated samples significantly decrease, particularly for the minority class. For instance, to address the potential degradation of the model caused by skewed distributions, Qin et al. (2023) propose a new equilibrium diffusion model. To tackle the observed issues, the model balances the prior distribution by adjusting the conditional transfer probability during sampling. It effectively transfers common information from head classes to tail classes without compromising the model's ability to represent head classes accurately. Yan et al. (2024) present a diffusion framework incorporating a regularization term for overlap, DiffROP. This framework is proposed to address the issue of class overlap between the probability distributions of the tail class and head class generated by diffusion models. This framework penalizes the overlap between conditional image distributions using probabilistic contrastive learning (PCL) loss, significantly enhancing the image fidelity of training diffusion models on imbalanced datasets. Li et al. (2024) also introduce an iterative online image synthesis (IOIS) framework to tackle the imbalance problem in medical image classification.

While most diffusion model studies focus on unstructured continuous data, such as images, numerous real-world classification datasets are in tabular format. The general Denoising Diffusion Probabilistic Model (DDPM) has now been adapted to tackle issues related to tabular data (Kotelnikov et al. 2023; Mueller et al. 2023; Sattarov et al. 2023; Zamberg et al. 2023). Due to the heterogeneity of individual features and the relatively small size of tabular datasets, training high-quality models for tabular data is more challenging than for computer vision or natural language processing (NLP). Zheng and Charoenphakdee (2022) present a tabular data diffusion model based on conditional scores. The model effectively addresses categorical and numerical variables and is demonstrated to be effective in managing missing values within tabular data. Ouyang et al. (2023) introduce a diffusion framework for learning from datasets with missing values under various mechanisms. Evaluations across multiple tabular datasets confirm the consistency of the proposed method in learning data distribution scores. Jiang et al. (2023) address the issue of declining accuracy in software measurement due to data loss in industrial processes. The enhanced Isolation Forest algorithm is employed to identify regions with missing data, determine their locations and quantities, and construct a data generation model that produces new samples based on the denoising diffusion probabilistic model. Samples that closely resemble the original data distribution are filtered from the newly generated samples, resulting in a filling method based on the diffusion model for the complete dataset. J. Jeong et al. (2023) introduce a novel data augmentation technique that utilizes the improved Deep Convolutional Generative Adversarial Network (DCGAN) model and a new similarity loss function to generate diverse and realistic tabular data, thereby addressing the problem of class imbalance.

In light of imbalanced datasets, we focus on identifying samples from the minority class. For instance, in biomedical applications (He and Garcia 2009), a dataset may consist of 10,923 samples of non-cancer patients (majority class) and 260 samples of cancer patients (minority class). It is essential to develop a classifier that can provide accurate predictions for both the majority and minority classes. In reality, the classifier frequently exhibits a substantial imbalance in accuracy, often achieving nearly 100% accuracy for the majority class while only reaching approximately 10% accuracy for the minority class within the dataset. This discrepancy reveals that 234 minority class are misclassified as majority class, which is analogous to 234 cancer patients being incorrectly diagnosed as non-cancerous. Although traditional learning algorithms can still attain high overall accuracy, the classification model is biased toward the majority class due to the unequal sample sizes. This bias ultimately leads to a higher classification error rate for the minority class. In the medical field, diagnosing a patient with cancer

as not having cancer is more costly than diagnosing a patient without cancer as having cancer. Consequently, in this field, we prioritize improving the recognition rate of the minority class while ensuring that the accuracy of the majority classes is not significantly compromised. As for the application of diffusion models to imbalanced data, most of the existing literature focuses on analyzing the impact of sample size across different classes (i.e., imbalance ratio) on the quality of generated samples. However, there is less discussion regarding the influence of the complex distribution of minority samples on classification. This aspect necessitates a comprehensive extraction of the distribution characteristics of minority samples to enhance the accuracy of minority classification. Traditional oversampling methods, such as SMOTE (Chawla, Bowyer, et al. 2002) and ADASYN (He, Bai, et al. 2008), are often constrained by the distribution of the original data. In cases of high feature dimensions, these methods require the calculation of distances between samples. Consequently, generating new samples through interpolation to balance the dataset increases complexity and reduces sampling efficiency. In contrast, the unique capability of diffusion models to capture high-dimensional dependency relationships effectively addresses various challenges associated with class imbalance. This article primarily focuses on the data-level method of imbalanced learning, explicity addressing the distribution characteristics of underrepresented sample classes. It proposes a method using a conditional diffusion model to generate new samples, thereby enhancing the classification performance of these samples.

This study provides the following contributions:

(1) LReDDPM is a conditional diffusion model based on the local region samples guided. This model can guide the diffusion model's sampling process according to the gradient conditions of various regional types.

(2) Three data quality metrics evaluate the similarity between the generated data and the real data. The results indicate that the distribution of the generated samples closely aligns with that of the real samples, and the local sample information is captured more effectively.

(3) Multiple evaluation metrics are utilized to assess the classification performance on the test set, and the validity of the samples generated from the augmented data is confirmed.

(4) The performance of synthetic examples across different regions was evaluated, and the results demonstrated that this generation method could significantly enhance classification effectiveness to varying degrees. Notably, synthetic examples generated in the safety and boundary areas yielded even more substantial improvements in classification performance. Additionally, it adaptively adjusts the guide information to address the issue of sample scarcity in practical applications, thereby reducing the costs associated with misclassification.

The remainder of this article is organized as follows. Section 2 provides an overview of the work. Section 3 details the methodology for identifying minority classes, offers foundational knowledge of the diffusion model and comprehensively describes the proposed new method. Sections 4 and 5 present the design and results of the experiment and additional relevant experiments. In Section 6, we discuss the samples generated from different regions, the results obtained by training the original DDPM using only regional samples, and the types of regions to which the synthetic samples belong. In Sections 7 and 8, we summarize the thesis and discuss directions for future research.

2 Related Work

2.1 Types of Minority Classes

In classification problems, class imbalance often arises when there are significantly more instances of certain classes than others. In such cases, standard classifiers may become overwhelmed by the larger classes and neglect the smaller ones (Chawla, Japkowicz, et al. 2004). This situation poses challenges most learning algorithms, typically assuming a relatively balanced class distribution (Sun, Wong, et al. 2009). Consequently, it leads to suboptimal classification performance. Some studies have shown that the degree of relative imbalance between classes is not the only factor that hinders learning (Batista et al. 2004; He and Garcia 2009; Japkowicz and Stephen

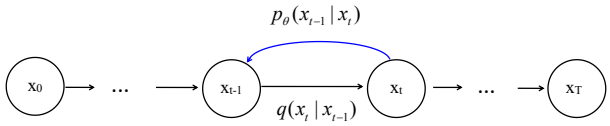


Fig. 1. The directed graphical model of DDPM

2002; Jo and Japkowicz 2004; Prati et al. 2004; Weiss and Provost 2003). It is still possible to learn from minority classes with minimal interference from this imbalance. The complexity of the data distribution is the primary determinant of degradation in classification performance, and the relative imbalance may further exacerbate this decline. We are interested in analyzing the intricate distribution of imbalanced datasets and designing methods to enhance the recognition rate of minority samples based on the characteristics of their complex distribution.

In light of the uneven distribution of a limited number of samples within the class and the complex nature of sample distribution, various scholars have employed different approaches to address these challenges. López et al. (2013) discusses six critical issues related to the data's intrinsic features, including identifying regions with small disjuncts, areas lacking sufficient training data density and information, and noisy data, all of which can diminish classification effectiveness. The discussion focuses on distinguishing minority class regions without utilizing clustering methods, as proposed by Sun, Cai, et al. (2022), and introduces a technique known as Disjuncts-Robust Oversampling (DROS). The author utilizes a light cone to illuminate restricted areas, addressing the small disjunctions of the minority class. Napierala and Stefanowski (2016) propose a method for identifying the type of sample instance by considering the class distribution within the local neighborhood of the instance. The experimental results indicate that different sample types present varying levels of learning difficulty for the classifier. The authors also emphasize leveraging this method to develop new learning algorithms and preprocessing techniques specifically designed to address class imbalances. Among these methods, Sun, Cai, et al. (2022) address the issue of class imbalance through sampling techniques. However, it focuses on a single type of sample distribution, which limits its applicability in solving the problem. Napierala and Stefanowski (2016) emphasized the nature of imbalanced data, the characteristics of minority class distributions, and their impact on classification performance.

According to the relevant studies mentioned above, it is essential to focus on the local characteristics of the complex distribution of minority classes to develop a high-performance classifier.

2.2 Gaussian Diffusion Models

The denoising diffusion probabilistic model (Ho et al. 2020; Sohl-Dickstein et al. 2015) is a hidden variable model that comprises two processes: the diffusion process (forward process) and the denoising process (reverse process). In the forward process, the model gradually perturbs the initial data $x_0 \in R^d$ with varying scales of Gaussian noise ϵ, transforming the noisy sample into a random noise output. In the reverse process, the learned parameters are employed to systematically remove the noise, ultimately reconstructing the original data x_0.

Specifically, the forward process begins at x_0, and the hidden variable $x_1, x_2, \ldots, x_T$ gradually transitions through the Markov chain into pure Gaussian noise $x_T \sim N(0, I)$. Consequently, each Markov transition can be expressed in the following form:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

Where β_t represents the noise level added at time step t. At any time step t, x_t is sampled from the distribution $q(x_t|x_0) = N(x_t; \sqrt{1-\hat{\beta}_t}x_0, \hat{\beta}_t I)$, where $\hat{\beta}_t = 1 - \prod_{i=0}^{t}(1-\beta_i)$.

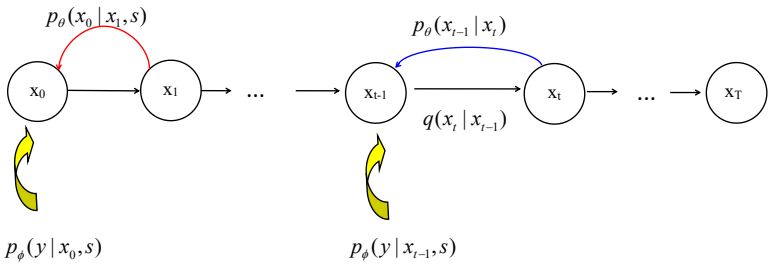


Fig. 2. The directed graphical model considered in this work, known as LReDDPM

In the reverse process, the model gradually denoises the hidden variable x_t to recover the data x_0. This involves sampling from the distribution $q(x_{t-1}|x_t)$ to obtain x_0 during the reverse process (see Figure 1. To approximate this process, we train a neural network $p_\theta(x_{t-1}|x_t)$ with parameters θ to approximate the distribution $q(x_{t-1}|x_t)$,

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

Where, μ_θ and Σ_θ are the mean and variance estimated by $q(x_{t-1}|x_t)$,

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\hat{\alpha}_t}}\epsilon_\theta(x_t, t)) \tag{3}$$

where, $\alpha_t := 1 - \beta_t$, $\hat{\alpha}_t := \prod_{i=0}^{t} \alpha_t$, $\epsilon_\theta(x_t, t)$ represent the predicted noise. In practice, our training objective utilizes the simplified mean squared error sum between the actual noise and the estimated noise.

$$l_t = \mathbb{E}_{x_0,\epsilon,t}\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \tag{4}$$

More specifically, the model $\epsilon_\theta(x_t, t)$ is used to predict the real noise ϵ after minimizing the simplified target loss (4) (Zamberg et al. 2023).

While this framework is effective for continuous data, it cannot be directly applied to discrete data, such as the discrete variables found in tabular datasets. Therefore, when working with tabular data, it is essential to consider appropriate methods for handling discrete variables.

Regarding the treatment of discrete variables within the diffusion model, the literature (Austin et al. 2021; Hoogeboom et al. 2021; Kim et al. 2022; Kotelnikov et al. 2023; Ouyang et al. 2023; Sohl-Dickstein et al. 2015) indicates that the author represented categorical data x_t using one-hot encoding $x_t \in \{0, 1\}^K$(where K denotes the number of categories) and employed a categorical distribution to define the polynomial diffusion process. In addition to one-hot encoding, Zheng and Charoenphakdee (2022) propose two additional techniques for handling categorical variables: analog bit encoding and feature tokenization. The feature tokenization method simultaneously embeds numerical and categorical variables into a single vector of uniform length. Sattarov et al. (2023) also use the technique of embedding categorical variables. In contrast to the approach of Zheng and Charoenphakdee (2022), Zhang et al. (2023), and Sattarov et al. (2023) do not embed numerical variables. In embedding categorical variables, Mueller et al. (2023) utilize score interpolation to extend the diffusion process of categorical data into Euclidean embedding space, thereby applying the same type of noise distribution to both categorical and numerical variables. For other multimodal discrete data, additional processing methods are summarized by Cao et al. (2024). Given the high dimensionality of some dataset variables discussed in this article and the fact that the number of categorical variables is less than the number of numerical variables, considering the calculation efficiency, we do not embed both numerical and categorical variables simultaneously. Instead,

following the approach established by Sattarov et al. (2023), we embed only the categorical variables to complete the diffusion process.

3 Methodology

3.1 Division of Minority Classes

This study primarily focuses on the local characteristics of the complex distribution of minority classes to construct high-performance classifiers. By analyzing the distribution of various complex sample types within the dataset, our method can selectively expand samples in different regions from the perspective of sample generation, making it adaptable to a wide range of complex sample distribution types. It encourages the model to generate data based on local samples of minority classes, thereby improving the recognition rates of these classes and reducing the costs associated with misclassification.

Based on the inherent distribution of the data and independent of the classifier, we categorize the minority class into four distinct regions: safe samples (designated as region 1), border samples (designated as region 2), rare samples (designated as region 3), and outlier samples (designated as region 4). In this article, we refer to the samples located within the region as regional samples. The method for partitioning these regions was adapted from Napierala and Stefanowski (2016) to facilitate calculation and partition. We analyze the K-nearest class labels to identify the type of region represented by the sample. Constructing this local region requires the appropriate selection of the k value and the distance function. In this case, we have fixed the value of k at five and employed the Heterogeneity Difference Measurement (HVDM) (Wilson and Martinez 1997) to calculate the distance between examples. With a k value of 5, the ratio of neighbors from the same class to those from different classes can vary from 5:0 (indicating that all neighbors of the analyzing instance belong to the same class) to 0:5 (indicating that all neighbors belong to different classes), as illustrated below:

$$\text{region}(x|x \in \text{minority class}) = \begin{cases} \text{safe}, & 5{:}0 \text{ or } 4{:}1 \\ \text{border}, & 3{:}2 \text{ or } 2{:}3 \\ \text{rare}, & 1{:}4 \\ \text{outlier}, & 0{:}5 \end{cases} \tag{5}$$

Subsequently, we will train the generative model based on samples from the same regional type, meaning the samples that belong to the same category after the training set has been divided. The results of the sample region division are presented in Appendix B table 12.

3.2 Derivation of the Methodology

This article primarily examines the impact of uneven sample distribution within a class on the performance of imbalanced classification, particularly concerning the minority class of interest. The complex distribution of these samples significantly influences the final classification outcomes. This article begins by examining the intricate regional distribution of minority classes. We first categorize these samples into distinct regions and generate new minority class samples based on regional data. The study employs a conditional denoising diffusion probabilistic model to create these samples. Rather than completely replacing the existing minority class samples, the generated samples are incorporated into the training set to augment the overall sample size.

According to the preliminary experimental results, the enhancement of the final classification performance for the minority class generated by the FinDiff method (Sattarov et al. 2023) does not meet the expected outcomes (see Table 3). It is evident from the results that the samples generated by FinDiff exacerbate category overlap, thereby increasing the difficulty of identifying the distribution of the minority class (see Figure 4). However, We hope the generated samples more accurately reflect the local distribution characteristics of the minority class. In that case, identifying these samples will be facilitated, and their distribution will not interfere with the

majority class. In other words, the generated samples will not affect the identification of the majority class and will effectively reduce the classification interface deviation. Therefore, we propose incorporating information from local regions of the minority class into the sample generation process. This approach aims to generate data that aligns with the distribution characteristics of the minority class through a method we refer to as "guidance". Through the reverse process of Denoising Diffusion Probabilistic Model (DDPM) to achieve guided sampling. In contrast, the forward process continues to adhere to the vanilla Denoising Diffusion Probabilistic Model (DDPM) (see Formula (6) and Appendix A). The model is summarized as follows:

Suppose that x_t and x_{t+1} are denoised samples at times t and $t+1$, respectively. Here, s represents regional samples, and y denotes the class label (Dhariwal and A. Nichol 2021). The distribution assumption under the unconditional denoising diffusion probabilistic model is denoted as q. When s and y are incorporated, the distribution assumption of the new model becomes $\hat{q}$.

$$\begin{aligned}
\hat{q}(x_t|x_{t+1},s,y) &= \frac{\hat{q}(x_t,x_{t+1},s,y)}{\hat{q}(x_{t+1},s,y)} \\
&= \frac{\hat{q}(y|x_t,s,x_{t+1})\hat{q}(x_t|s,x_{t+1})\hat{q}(s|x_{t+1})\hat{q}(x_{t+1})}{\hat{q}(y|s,x_{t+1})\hat{q}(s|x_{t+1})\hat{q}(x_{t+1})} \\
&= \frac{\hat{q}(y|x_t,s,x_{t+1})\hat{q}(x_t|s,x_{t+1})}{\hat{q}(y|s,x_{t+1})} \\
&= \begin{cases} \frac{\hat{q}(y|x_t,s)\hat{q}(x_t|x_{t+1})}{\hat{q}(y|s,x_{t+1})}, & t \geq 1 \\ \frac{\hat{q}(y|x_t,s)\hat{q}(x_t|s,x_{t+1})}{\hat{q}(y|s,x_{t+1})}, & t = 0 \end{cases} \\
&= \begin{cases} Z\hat{q}(y|x_t,s)q(x_t|x_{t+1}), & t \geq 1 \\ Z\hat{q}(y|x_t,s)\hat{q}(x_t|s,x_{t+1}), & t = 0 \end{cases}
\end{aligned} \tag{6}$$

In Formula (6), the denominator $\hat{q}(y|s,x_{t+1})$ is independent of x_t and can be considered a constant. Therefore, we aim to sample from the distributions $Z\hat{q}(x_t|x_{t+1})\hat{q}(y|x_t,s) = Zq(x_t|x_{t+1})\hat{q}(y|x_t,s)(t \geq 1)$ and $Z\hat{q}(x_t|s,x_{t+1})$ $\hat{q}(y|x_t,s)(t=0)$, where $Z = \frac{1}{\hat{q}(y|s,x_{t+1})}$ is considered the normalization constant. If we already have a neural network to approximate $q(x_t|x_{t+1})$, referred to as network $p_\theta(x_t|x_{t+1})$, the next step is to approximate $\hat{q}(y|s,x_t)$ and $\hat{q}(x_0|s,x_1)$. The $\hat{q}(y|s,x_t)$ can be obtained by training a classifier under the local condition s and the noise sample x_t (where x_t is drawn from the marginal distribution $\hat{q}(x_t) = q(x_t)$). Meanwhile, x_0 is derived from $\hat{q}(x_0|s,x_1)$ by combining the information sampled from x_1 and s. The implementation of $\hat{q}(x_0|s,x_1)$ involves two key aspects. First, it does not include a random term during sampling, similar to the reverse sampling process described in the final step of the original DDPM method (as outlined in Algorithm 2 of Ho et al. (2020)). Second, to ensure that the generated samples remain within the range of the regional samples, the mean of the regional samples is added at the end, guided by gradient information.

Similar to Dhariwal and A. Nichol (2021), when inferring the reverse diffusion process, it is only necessary to incorporate the gradient information of $\hat{q}(y|x_t,s)$ into the mean value of the sample distribution. The forward process remains consistent with the Denoising Diffusion Probabilistic Model (DDPM). The reverse process is illustrated in pseudocode (see Algorithm 1). In Algorithm 1, "info(s)" denotes the mean value of the aggregated regional samples. Therefore, we refer to the proposed method as Local Regional Samples Guidance Denoising Diffusion Probabilistic Model (LReDDPM). Regional samples are employed in the reverse process of the LReDDPM.

The specific guidance process of LReDDPM involves training a classifier on regional samples, followed by sampling guided by the classifier's gradient information. During the data restoration process, we assign a specific level of importance to the gradient by including an optional scaling factor τ, as described in Dhariwal and A. Nichol (2021). The optional parameter τ allows us to consider either static or dynamic parameter values. Static parameter values mean that the scaling factor of the gradient remains constant throughout the entire

process, from step $t = T$ to $t = 0$. In contrast, dynamic parameter values indicate that the influence of the gradient on the data recovery process varies at different time steps. If it is necessary to emphasize the role of the gradient more at a specific time step, a larger parameter value can be assigned, and vice versa. For example, at the beginning of reverse sampling, the value of τ should not be set higher. This ensures that the gradient is not overly emphasized during the early stages, allowing the model to focus more on sample quality and ensuring that the samples closely align with the data distribution. As t gradually approaches 0, the scaling factor τ is progressively increased, enabling the model to incorporate gradient information more effectively and generate samples that better represent regional characteristics. In this way, the early stage guarantees sample quality and prevents deviation from the original data distribution, while the later stage ensures that the generated samples reflect the specific characteristics of the regional data. This is precisely the approach that this article primarily adopts. We will provide a more detailed description in Section 4.1 and 5.3.

Algorithm 1 Region Guided Diffusion Sampling

Input: diffusion model $(\mu_\theta, \Sigma_\theta(x_t))$,classifier $p_\phi(y|x_t, s)$,class label y, gradient scale τ, s
Output: x_0
  $x_T \leftarrow$ sample from $\mathcal{N}(0, \boldsymbol{I})$
  for all t from T to 1 **do**
    $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$
    $x_{t-1} \leftarrow$ sample from $\mathcal{N}(\mu + \tau\Sigma\nabla_{x_t} \log p_\phi(y|x_t, s), \Sigma)$ ▷ Contain gradient information obtained from regional samples
    if $t = 1$ **then**
      $x_{t-1} \leftarrow x_{t-1} \bigcup \text{info}(s)$ ▷ Add the mean of the regional samples
    end if
  end for

4 Experiments

4.1 Experimental Setup

Selection of Datasets. The primary objective of this article is to enhance the performance of imbalanced data classification and to analyze and discuss LReDDPM through comprehensive experiments on ten imbalanced binary classification datasets. Among these, segment0, yeast3, page0, and abalone17 are sourced from the KEEL repository, while the remaining datasets are obtained from the UCI repository. The Turkish Music Emotion dataset (abbreviated as Turkish) is originally a four-class classification problem, where the "happy" class is treated as a minority class, and the other classes are combined into a majority class. The crg dataset (short for Statlog German Credit Data) consists of raw data variable types. These datasets are summarized in Table 1. These datasets vary in their size and class proportions, thus offering different domains for LReDDPM.

Selection of Evaluation Metrics and Classifiers. (1) Quantitative Evaluation Indices. To assess the classification performance of the expanded dataset, the following indicators were selected: 1) AUC. AUC (Area Under the Curve) refers to the area under the Receiver Operating Characteristic (ROC) curve. 2) F1-score. The F1-score is the harmonic mean of precision and recall, offering a more effective measure of performance for imbalanced classifiers. 3) G-mean. This metric takes into account both precision and recall values, and the formula is $G\text{-}mean = \sqrt{precision \times recall}$. (2) Qualitative Evaluation Index. The qualitative evaluation index is designed to measure the distribution between the generated sample and the real sample. Three distribution distance indices selected for this purpose include the Jensen-Shannon Distance (abbreviated as JS), Wasserstein Distance (abbreviated as WD), and Maximum Mean Discrepancy (abbreviated as MMD) (Qian et al. 2023). (3)The classifier

Table 1. Datasets detail. The terms "#Train" and "#Test" refer to the number of samples in the training set and the test set, respectively. The terms "#Num" and "#Cat" represent the number of numeric and categorical variables in the dataset, respectively. The term "#Rimb" refers to the imbalanced ratio.

Data	#Train	#Test	#Num	#Cat	#Rimb
segment0	1846	462	19	0	6.02
spambase	3677	920	57	0	1.54
spectfheart	213	52	44	0	3.85
turkish	320	80	50	0	3
yeast3	1187	297	8	0	8.1
page0	4377	1095	10	0	8.8
rice	3048	762	7	0	1.34
crg	800	200	7	13	2.3
abalone17	1870	468	7	1	39.31
parkin	604	152	752	1	2.94

employs decision tree and logistic regression. The AUC values for both the decision tree and logistic regression, as well as the F1-score and G-mean values for the decision tree, are provided respectively.

Comparative Method for Generating Samples. We have selected five oversampling methods for comparison: (1) ADASYN, which stands for Adaptive Synthetic Sampling approach (He, Bai, et al. 2008). The method mitigates the learning bias introduced by the original data distribution by generating additional training data for the minority class. It adaptively adjusts the classification decision boundary to focus more on these difficult-to-learn samples. However, due to the direct invocation of toolkit generation in the imblearn codebase, this method struggles to achieve a specified sample size in accordance with the proportion of regional samples generation. This article presents the classification results based on the generation of the closest sample size for comparison. (2)BorderlineSMOTE (Han et al. 2005). BorderlineSMOTE (Borderline Synthetic Minority Oversampling Technique) is an advanced oversampling method designed to address class imbalance by strategically generating synthetic samples for the minority class. It reinforces the classification boundary where minority instances are most vulnerable to misclassification. This method needs to calculate the K-nearest neighbors of the minority class, then generate new sample points based on the SMOTE (Synthetic Minority Over-sampling Technique (Chawla, Bowyer, et al. 2002)) algorithm after identifying the boundary points. (3)FinDiff (Sattarov et al. 2023). FinDiff is a denoising diffusion probabilistic model that employs embedded coding to represent mixed-modal tabular data. (4)CTGAN (Xu et al. 2019). CTGAN is a conditional generative adversarial network that effectively preserves the underlying structure of actual data, including correlations between columns. It is capable of processing continuous and discrete data. This article utilizes the existing codebase for direct execution. (5) TVAE (Xu et al. 2019). TVAE is a neural network generation model that adapts the Variational Autoencoder (VAE) for tabular data by employing similar preprocessing techniques and modifying the loss functions. This model is referred to as the TVAE model.

Region Type Selection. This article focuses on selecting samples based on the results of regional type division and the distribution characteristics of the minority class. For comparison purposes, we have selected samples from region 2 for the spectfheart and crg datasets to facilitate the reverse process of the diffusion model. Most datasets predominantly select samples from region 1 to guide the reverse process of the diffusion model. Additionally, Section 6 presents the results of guided sample generation across different regions within part of the dataset, including comparisons of guided sample generation in regions 3 and 4.

Parameter Setting. (1) Similar to Ho et al. (2020), we set the forward process variances to constants increasing linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. (2) We set $T = 1000$ for all experiments. (3) Regarding the parameter τ, we treat it as a dynamic parameter value. Since the parameter β_t is exactly equal to the variance and increases with the length of the time step, we consider the reciprocal of the square root of β_t. Additionally, based on experimental results for other parameter values (see Section 5.3), the method presented in this paper primarily uses the parameter value $\tau = 60 * 1/\sqrt{\beta_t}$. That is, the gradient scaling factor in the algorithm diagram (Algorithm 1) is denoted as $\tau = 60 * 1/\sqrt{\beta_t}$. Other values of the parameter τ are discussed in Section 5.3, including: 1)$\tau = 1$; 2)$\tau = 60$; 3)$\tau = 1/\sqrt{\beta_t}$.

Furthermore, to simplify the calculation process, $\hat{q}(y|s, x_t)$ directly uses the gradient of $\hat{q}(y|s)$ to generate the sample. All results in this section are guided by the gradient of $\hat{q}(y|s)$ during generation. All the denoising diffusion probabilistic models presented in this article, the variance Σ_θ is fixed at a constant value, as in Ho et al. (2020).

The different distribution selections of $\hat{q}(y|s, x_t)$ are presented in Section 5.1. For example: 1) Consider also the gradient guidance of s and x_t, i.e., $\hat{q}(y|s, x_t)$; 2) Consider only the gradient guidance of x_t, i.e., $\hat{q}(y|x_t)$. At the same time, in the discussion section of Section 6, the forward process of the model is directly trained using regional samples without considering gradient guidance. In the reverse process, unconditional Denoising Diffusion Probabilistic Models (DDPM) are employed to generate samples. In other words, the diffusion model is equivalent to FinDiff in this context. The only difference is that the training model utilizes regional samples rather than the original dataset.

4.2 Results

Comparison of the Distribution of Generated Samples and Real Samples. Table 2 presents three indicators comparing the generated samples' distribution with real ones. The average distance of the three indicators refers to the distance calculated for each dataset based on the synthetic data and the selected guiding region samples. Since the synthetic data size in some datasets is smaller than that of the selected guiding region, we randomly select an equal number of samples from the guiding region multiple times to calculate the distance. Finally, the average of these distances is computed. The results displayed in Table 2 represent the average ranking of the three metrics across all datasets, with lower values indicating better performance. As shown in Table 2, the synthetic data generated by LReDDPM is more closely aligned with the regional samples distribution of minority classes than other methods, effectively capturing the regional samples distribution of these classes. For specific values of each indicator, please refer to Table 16, Table 17, and Table 18 in Appendix C.

Table 2. Average ranks over all datasets in terms of Jensen-Shannon Distance(JS), Wasserstein Distance(WD), Maximum Mean Discrepancy(MMD). Distances are calculated between generated data and real region data

	JS	WD	MMD
ADASYN	3.8	3.65	3.7
BorderlineSMOTE	4.2	4.35	3.9
FinDiff	2.7	3.3	4.3
TVAE	3.8	3.25	**1.7**
CTGAN	4.2	3.4	4.3
LReDDPM(ours)	**2.3**	**3.05**	3.1

Comparison of AUC Values Across Different Classifiers. We compare five sample generation methods using two classifiers: decision tree and logistic regression. All methods generate approximately the same sample

size. Tables 3 and 4 present the AUC values for the six methods across ten datasets for each classifier. The last row of each table displays the classification results before sample expansion. Table 3 compares classification

Table 3. The AUC values of the dataset under different methods (Decision Tree Classifier)

Method	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
ADASYN	89.540 ± .393	90.615 ± .056	65.844 ± .477	89.424 ± .428	80.430 ± .167	77.490 ± .255	86.993 ± .143	59.410 ± .340	50.537 ± .252	39.956 ± .719
BorderlineSMOTE	89.538 ± .526	91.211 ± .099	52.333 ± 1.018	89.702 ± .213	79.758 ± .098	82.371 ± .300	87.171 ± .103	61.923 ± .275	64.359 ± .865	54.398 ± .364
FinDiff	89.239 ± .471	90.640 ± .121	51.452 ± .402	92.867 ± .205	80.543 ± .357	84.809 ± .218	88.384 ± .060	57.880 ± .174	63.502 ± .786	54.731 ± 1.139
TVAE	89.242 ± .6	90.512 ± .550	60.644 ± 3.109	93.163 ± .811	78.663 ± 3.092	**88.621 ± 2.196**	88.034 ± .859	63.133 ± 1.368	62.817 ± 2.457	47.020 ± 7.709
CTGAN	89.353 ± .195	90.293 ± .268	59.467 ± 5.160	91.692 ± 4.412	82.529 ± 4.436	84.055 ± 2.095	87.950 ± .565	63.269 ± 1.937	62.545 ± 1.321	54.842 ± 6.446
LReDDPM(ours)	**89.577 ± .236**	**91.437 ± .050**	**68.200 ± .747**	**94.244 ± .196**	**82.998 ± .219**	85.822 ± .337	**88.433 ± .066**	**65.031 ± .346**	**65.742 ± .770**	**60.605 ± .508**
Original	89.581 ± .571	90.621 ± .086	67.319 ± .771	90.704 ± .544	75.832 ± .366	87.300 ± .303	86.896 ± .113	60.052 ± .227	61.759 ± .78	47.287 ± 1.565

results (AUC) obtained using decision tree classifiers on the dataset, with the optimal results highlighted in bold. Only the page0 dataset did not yield optimal AUC values among the ten analyzed datasets. However, the results for the remaining nine datasets were the best achieved. Most datasets demonstrated improved classification performance following data expansion. It indicates that the classification performance for the same classifier for specific classes using samples generated by LReDDPM is superior. In contrast, the impact on samples from other classes is less pronounced. Compared to other sampling methods based on sample density distribution, such as ADASYN and BorderlineSMOTE, our method produces superior classification results. The results of experiments suggest that the guidance of locally distributed samples enhances class characteristics and reduces the challenges associated with minority class recognition. Although our method did not achieve optimal results on the page0 dataset, it outperformed both the ADASYN and BorderlineSMOTE methods.

Table 4 presents the AUC values obtained from another classifier. As shown in Table 4, the LReDDPM method achieves higher AUC values across most datasets, indicating superior classification performance. However, under the ADASYN method, the AUC for the abalone17 dataset surpasses that of our method's, likely due to the larger sample size generated by the ADASYN. Additionally, it is noteworthy that while the AUC for the parkin dataset is higher than that of our method under ADASYN, its G-mean and Accuracy do not exceed our results (with the same recall for both methods, as detailed in Table 13 in Appendix C). Under this classifier, the evaluation scores for ADASYN and LReDDPM decreased compared to their values before the dataset expansion, as observed in Table 13. However, the recall value remained unchanged. The complex distribution of the dataset may contribute to this phenomenon. According to Table 3, Table 4, Table 5, and Table 12, the boundary sample in the parkin dataset occupies a relatively high proportion, and the distribution of the minority class may be sparse, resulting in an apparent overlap between the two types of samples.

Although the ADASYN method can achieve the same recall value as LReDDPM (ours) when the sample size is small, the G-mean and Accuracy of the ADASYN method decrease more significantly. These results indicate that ADASYN's recognition for the complex distribution of samples is insufficient. When TVAE and our method generate the same sample size, the Accuracy improves compared to the original dataset. However, the recall value decreases the most of TVAE. Maybe the samples generated by TVAE cause a shift in the classification interface, and this method does not adequately reflect the distribution of the minority class. The analysis presented above shows that the coupling between the samples generated by our method and the classifier is not strong. Performance improvement primarily arises from the generated samples that are more closely aligned with the selected region. Consequently, the impact on the classification of the majority class is minimal, further demonstrating the model's effectiveness in generating samples.

F1-score Comparison. The AUC value only considers the sorting ability of the classifier, not its calibration ability. To achieve a comprehensive performance evaluation, we combined the F1-score and G-mean as two evaluation indicators for thorough consideration. As shown in Table 5, the classification results of our method remain the best across all datasets. In addition, the F1-score across all datasets has improved compared to the

Table 4. The AUC values of the dataset under different methods (Logistic Regression)

Method	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
ADASYN	99.899 ± .000	95.469 ± .000	88.148 ± .000	97.401 ± .000	96.289 ± .000	95.044 ± .000	98.196 ± .000	79.095 ± .000	**91.943 ± .000**	**89.266 ± .000**
BorderlineSMOTE	99.895 ± .000	95.478 ± .000	86.420 ± .000	97.133 ± .000	96.237 ± .000	94.903 ± .000	98.202 ± .000	78.798 ± .000	90.393 ± .000	88.319 ± .000
FinDiff	99.891 ± .000	95.410 ± .000	87.654 ± .000	96.953 ± .000	96.424 ± .000	94.569 ± .000	98.220 ± .000	78.762 ± .000	89.454 ± .000	88.550 ± .000
TVAE	99.897 ± .004	95.443 ± .007	87.358 ± .807	97.616 ± .250	96.335 ± 0.035	95.257 ± .183	98.222 ± .001	78.955 ± .279	90.350 ± .135	88.462 ± .233
CTGAN	99.894 ± .005	95.427 ± .035	86.617 ± 1.412	96.595 ± .595	96.561 ± 0.127	94.233 ± .754	98.221 ± .003	79.369 ± .313	89.895 ± .464	88.721 ± .256
LReDDPM(ours)	**99.907 ± .000**	**95.505 ± .000**	**88.395 ± .000**	**97.670 ± .000**	**96.590 ± .000**	**95.274 ± .000**	**98.223 ± .000**	**80.321 ± .000**	90.568 ± .000	89.220 ± .000
Original	99.891 ± .000	95.419 ± .000	86.914 ± .000	97.133 ± .000	96.310 ± .000	95.060 ± .000	98.224 ± .000	79.143 ± .000	89.978 ± .000	88.873 ± .000

original imbalanced dataset, indicating that our newly designed method can better balance the two metrics of Accuracy and recall. For instance, in the segment0 dataset, the sample expansion methods, except CTGAN and LReDDPM (ours), decrease the F1-score. Below, we comprehensively analyze the Accuracy and recall values under the CTGAN method (see Appendix C, Table 14). The Accuracy of CTGAN increased to 96.222%, while the recall value decreased to 79.795%. These data suggest that although both methods have improved the F1-score, the overall enhancement may come at the expense of classification accuracy for certain classes. Although our method experiences a slight decrease in recall, this decline is relatively minor, while the increase in the F1-score is substantial. Additionally, the BorderlineSMOTE method demonstrates the slightest reduction in recall value (a decrease of 0.031). However, it also incurs the highest drop in Accuracy (a decrease of 0.052), and the F1-score remains unimproved. Consequently, while BorderlineSMOTE enhances the classification accuracy of the minority class, it simultaneously heightens the risk of misclassification for the majority class. Therefore, the segment0 dataset effectively demonstrates the advantages of our new approach.

Under the LReDDPM (our method), the F1-score of the four datasets—yeast3, crg, abalone17, and parkin has a higher improvement rate than other datasets. The results of experiments indicate that the gradient information from regional samples can be utilized more effectively to generate samples that align with the direction of these regional samples, rather than being overly influenced by samples from the majority class. As illustrated by the partitioning information of the abalone17 dataset, the results presented in Table 3- 5 utilize only samples from region 1 to generate three samples. However, this approach can also enhance the classification performance of the classifier. The analysis indicates that the guidance information is pivotal in influencing the movement of the decision boundary. Furthermore, recognizing and focusing on sample information from various regional types can provide additional insights and guidance for our reverse diffusion process (see Tables 7 and 10).

Table 5. The F1-score values of the dataset under different methods (Decision Tree Classifier)

Method	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
ADASYN	85.532 ± .215	89.168 ± .073	41.601 ± .988	79.620 ± .511	65.638 ± .217	38.728 ± .252	85.775 ± .159	43.689 ± .438	3.788 ± .225	23.409 ± .307
BorderlineSMOTE	85.415 ± .157	89.834 ± .108	23.447 ± 1.303	79.913 ± .176	63.965 ± .128	47.670 ± .477	86.066 ± .114	47.157 ± .453	25.822 ± 1.517	32.589 ± 3.584
FinDiff	85.307 ± .123	89.296 ± .142	20.490 ± .609	82.666 ± .183	66.490 ± .659	50.991 ± .710	87.371 ± .064	43.318 ± .188	23.757 ± 1.225	38.288 ± .904
TVAE	85.526 ± .153	89.107 ± .615	34.140 ± 4.031	83.930 ± 1.591	64.419 ± 5.139	61.934 ± 5.372	86.949 ± .987	48.590 ± 1.892	20.059 ± 3.023	30.538 ± 6.380
CTGAN	85.633 ± .091	88.878 ± .313	32.761 ± 5.383	81.084 ± 7.379	68.479 ± 6.771	50.672 ± 4.042	86.910 ± .641	48.926 ± 2.499	20.739 ± 1.012	39.457 ± 5.469
LReDDPM(ours)	**85.737 ± .097**	**90.168 ± .065**	**42.171 ± .803**	**86.450 ± .295**	**70.494 ± .348**	**66.415 ± .481**	**87.397 ± .071**	**51.842 ± .443**	**31.660 ± 1.095**	**43.708 ± .595**
Original	85.578 ± .211	89.242 ± .098	40.102 ± .781	84.616 ± .797	58.983 ± .577	65.879 ± .653	85.620 ± .129	43.149 ± .335	20.570 ± .867	24.698 ± 1.267

Comparison of G-mean Values. Figure 3 presents a comparative graph of the G-mean values across all datasets. The figure demonstrates that the G-mean value achieved by our method remains favorable.

Generate Visual Comparisons of Samples. To examine the influence of the samples generated by our method on classification results, we will observe the locations of the generated samples from a visual perspective. This approach also illustrates how the guided samples address the complex distribution of the minority class. We will employ visualization methods to project multidimensional data into a low-dimensional space while preserving the structural properties of the data. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality

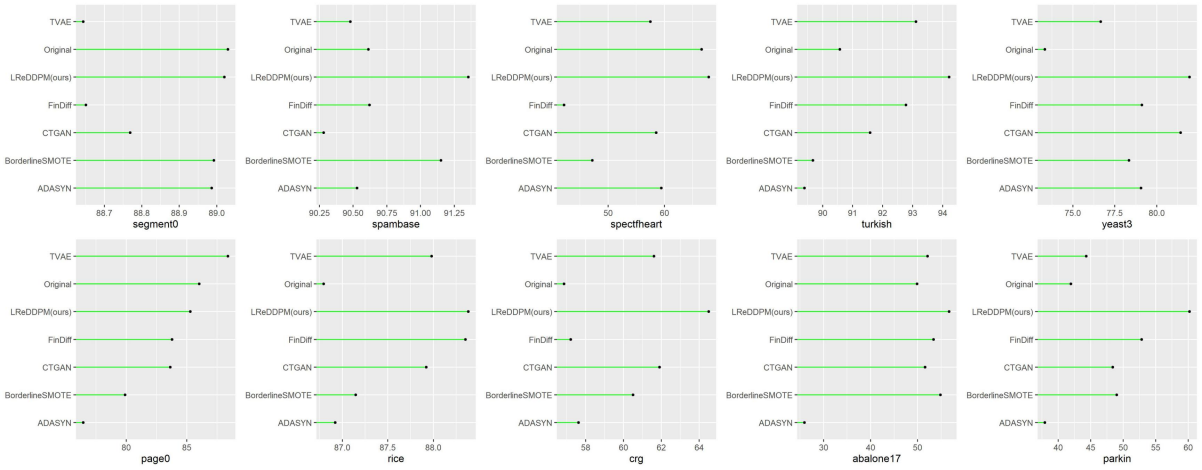


Fig. 3. Comparison of G-mean values across all datasets

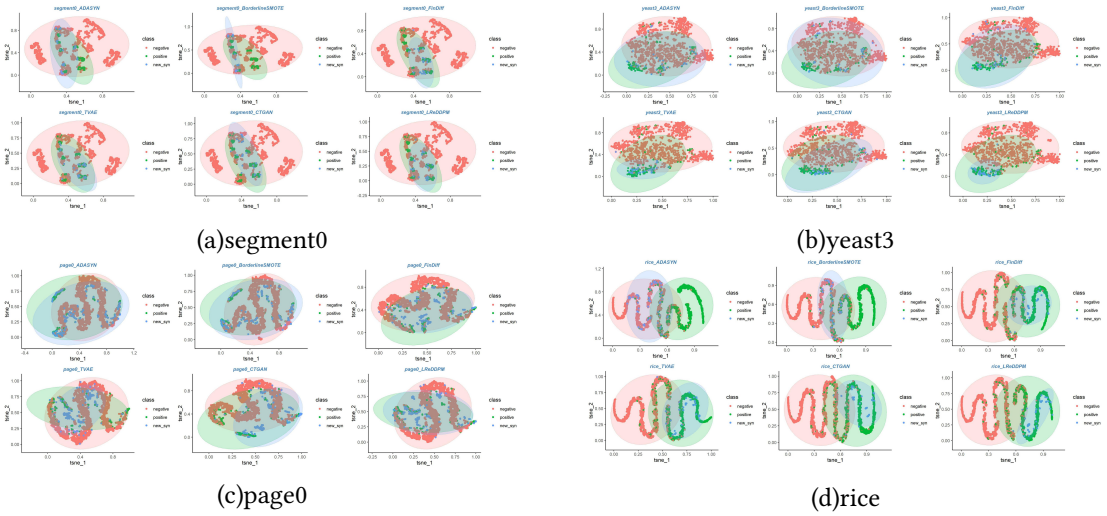


Fig. 4. The location of the region where samples are generated using various methods

reduction technique well-suited for visualizing high-dimensional datasets. It emphasizes the retention of local distances to keep similar instances grouped. This technique enhances the ability to capture the local structure of high-dimensional data while also revealing the global structure (Van der Maaten and Hinton 2008). This article uses t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize samples generated by various methods. The visual image helps us intuitively observe that the locations of these samples are distinct. Figure 4 illustrates the dimensionality reduction graphs of the generated samples across four datasets, while the graphs for the remaining datasets are available in Appendix D.

Notably, the sample size produced by the ADASYN method differs from that of the other five methods. As illustrated in Figure 4, the minority class generated by our new method are more concentrated within the guided

sample region and exhibit less overlap with the majority class. In contrast, CTGAN and FinDiff may deviate from the sample range of the minority class, encroaching upon the majority class region and significantly impacting the classification boundary. The minority samples generated by TVAE are closely aligned with the original minority samples, while the samples produced by BorderlineSMOTE are more concentrated in the border area. Figure 4 provides an alternative perspective that demonstrates our method's effectiveness in improving the classification performance of downstream tasks.

5 Other Experiments

5.1 Different Gradient Guidance

As mentioned in the previous statement, to simplify the calculation, the reverse process sampling $\hat{q}(y|s, x_t)$ directly selects the gradient of the distribution $\hat{q}(y|s)$ for sample generation. This section discusses the method of selecting $\hat{q}(y|s, x_t)$ according to the original formula to generate samples under gradient conditions. It also examines the gradient-guided sample generation method, similar to that described by Dhariwal and A. Nichol (2021), which focuses solely on x_t without considering regional samples. The remaining parameters, including guided sample types, generated sample size, and classifier information for these two methods, are consistent with those presented in Tables 3 and 5. Therefore, the comparisons in Table 6 are made solely concerning Tables 3 and 5. The "+" and "-" in parentheses following the AUC and F1-score in Table 6 indicate the following: the symbol "+" indicates the method outperforms the optimal results found in Tables 3 and 5, "-" means that it is not better than the optimal results in Table 3 and 5. As shown in Table 6, neither $\hat{q}(y|s, x_t)$ nor $\hat{q}(y|x_t)$ surpasses the sample generation method of $\hat{q}(y|s)$. This is also the reason why $\hat{q}(y|s)$ was chosen to initiate the sample generation process discussed earlier.

Table 6. The Results of Different gradients $\hat{q}(y|s, x_t)$

		segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
$\hat{q}(y\|s, x_t)$	AUC	89.831 ± .469(+)	92.009 ± .052(+)	57.044 ± .456(-)	94.020 ± .170(-)	85.554 ± .169(+)	86.675 ± .296(-)	86.659 ± .063(-)	66.193 ± .282(+)	57.412 ± .411(-)	34.208 ± .471(-)
	F1-score	85.721 ± .247(-)	90.841 ± .060(+)	29.974 ± .532(-)	85.770 ± .348(-)	72.984 ± .205(+)	65.446 ± .400(-)	85.370 ± .071(-)	53.056 ± .388(+)	16.138 ± .734(-)	23.150 ± .541(-)
$\hat{q}(y\|x_t)$	AUC	89.871 ± .331(+)	91.267 ± .036(-)	59.230 ± .928(-)	94.001 ± .366(-)	86.123 ± .152(+)	87.533 ± .328(-)	87.713 ± .111(-)	57.312 ± .386(-)	57.688 ± .324(-)	31.047 ± .387(-)
	F1-score	85.611 ± .209(-)	90.009 ± .043(-)	32.482 ± .976(-)	85.959 ± .497(-)	73.415 ± .436(+)	65.666 ± .720(-)	86.582 ± .126(-)	40.884 ± .517(-)	16.660 ± .694(-)	20.230 ± .510(-)

5.2 Different Regional Types

Table 7 presents the results of the new method we developed. The parameters remain identical to those in Table 3. The only modification is the guided regional samples type, which allows us to observe its impact on the classification results. The symbol "-" in parentheses indicates that the value is lower than the result of LReDDPM (ours) as shown in Table 3 and Table 5. As illustrated in Table 7, when samples are generated from a different region within these datasets, the performance results for the three evaluation metrics—AUC, F1-score, and G-mean—are not superior to those from the first or second region types. Specifically, the crg dataset selects the first region, while the remaining three datasets (with the results for abalone17 presented in Table 10) select the second region. Since the second regional samples type is the boundary type sample, the results in Table 7 are further compared to the BorderlineSMOTE method. Tables 3 and 5 display the outcomes of the BorderlineSMOTE method. Figure 3 presents the G-mean value. It is evident that when comparing the classification performance of the samples generated from the second region of the three datasets—yeast3, rice, and abalone17—with the classification performance achieved using BorderlineSMOTE, the results for yeast3 and abalone17 under BorderlineSMOTE are superior to those obtained by our method from the second regional samples. However, as illustrated in Table 5 and Table 10, we can identify a more effective guidance region than that provided by the BorderlineSMOTE

method. The discussion in this section demonstrates that the new method we have developed is highly flexible and not confined to a single guidance region.

This method integrates the distribution characteristics of samples across various dimensions, tailored to different classification tasks and their associated difficulties, thereby offering a broader range of classification strategies. In practical applications, we can assess the sample generation performance guided by different regions to enhance the recognition of the minority class and minimize the likelihood of erroneously deleting samples situated in low-density regions. Furthermore, the guidance information can be adjusted based on the detection results, addressing the issue of sample scarcity in real-world applications and reducing the costs associated with misclassification.

Table 7. The Results of Different Region

	yeast3	rice	crg
AUC	78.679 ± .486(-)	87.668 ± .044(-)	60.700 ± .225(-)
F1-score	63.214 ± .670(-)	86.580 ± .049(-)	44.975 ± .333(-)
G-mean	76.934 ± .598(-)	87.632 ± .045(-)	58.615 ± .283(-)

5.3 Different Hyperparameters of τ

Tables 8 and 9 illustrate how different hyperparameters affect sample generation. We analyze the optimal results of various evaluation metrics for each dataset, where the classifier, the sample size, and the selected guidance region are consistent with those presented in Tables 3 and 4.

(1)In the case of $\tau = 1$ (the first row of Tables 8 and 9), based on the evaluation metrics of AUC and F1 score, only three datasets demonstrated simultaneous improvement in classification performance. The three datasets presented in Tables 8 and 9 are yeast3, page0, and crg.

(2) In the case of $\tau = 60$ (as shown in the second row of Tables 8 and 9), there are four datasets that demonstrate improved classification performance based on the evaluation metric AUC. Additionally, the number of datasets exhibiting an enhanced F1-score is slightly higher. However, it is important to note that these improvements are not fully synchronized, and the number of datasets experiencing a decline in performance exceeds those that show improvement. In specific datasets, we observed that, compared to the results of the method proposed in Section 4 (LReDDPM), while the Accuracy improved, the recall values declined. For instance, the recall for dataset segment0 is 80.133%, for page0 it is 77.497%, and for abalone17 it is 32.867% (see Appendix C, Table 15). The corresponding Accuracy are 96.253%, 91.613%, and 97.011%, respectively. Compared to our method, the recall for the three datasets decreased by 0.185, 0.712, and 0.266, respectively, while the Accuracy increased by 0.022, 0.158, and 0.054, respectively. The experiments' results indicate an improvement in the classification performance of the majority class, suggesting that the samples obtained under this parameter are more conducive to identifying the majority class. However, the classification performance of the minority class, which we focus on, is not adequately represented, resulting in an inability to achieve the desired outcomes. Overall, the comprehensive results across multiple indices show significant fluctuations in model performance, indicating poor robustness under this parameter.

(3) In order to analyze the influence of gradient information on the model as it changes with each time step, we consider the case of $\tau = 1/\sqrt{\beta_t}$ $(0 \le t \le T)$ (as shown in the third row of Tables 8 and 9). Specifically, the closer the data is to the original data x_0, the more significant the gradient information becomes, resulting in a higher weight being assigned to it. From the three different values presented above, the results are not satisfactory. When $\tau = 60$ and $\tau = 1/\sqrt{\beta_t}$, the samples generated by the model demonstrate better classification performance,

as indicated by the AUC and F1-score metrics, compared to when $\tau = 1$ alone. Consequently, the results discussed in Section 4 of this article are all based on $\tau = 60 * 1/\sqrt{\beta_t}$ for training the model and conducting the experimental analysis.

Table 8. The AUC Results of Different hyperparameters τ

Method	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
$\tau = 1$	89.727 ± .443(+)	91.326 ± .058(-)	64.156 ± .812(-)	94.256 ± .168(+)	83.305 ± .297(+)	85.866 ± .185(+)	88.165 ± .133(-)	68.800 ± .288(+)	65.502 ± .892(-)	37.421 ± .518(-)
$\tau = 60$	89.512 ± .299(-)	91.261 ± .083(-)	64.104 ± .824(-)	94.422 ± .190(+)	83.344 ± .318(+)	85.834 ± .094(+)	88.161 ± .051(-)	68.937 ± .297(+)	65.639 ± .708(-)	37.611 ± .323(-)
$\tau = 1/\sqrt{\beta_t}$	89.455 ± .288(-)	91.472 ± .045(+)	68.881 ± 1.300(+)	94.314 ± .217(+)	83.175 ± .355(+)	86.174 ± .091(+)	88.393 ± .121(-)	51.522 ± .611(-)	66.023 ± .328(+)	60.193 ± .934(-)

Table 9. The F1-score Results of Different hyperparameters τ

Method	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
$\tau = 1$	85.677 ± .178(-)	90.044 ± .084(-)	38.206 ± 1.052(-)	86.327 ± .166(-)	70.804 ± .400(+)	66.638 ± .435(+)	87.091 ± .150(-)	56.691 ± .352(+)	31.542 ± 1.617(-)	26.810 ± .488(-)
$\tau = 60$	85.772 ± .131(+)	89.967 ± .088(-)	38.234 ± .962(-)	86.707 ± .282(+)	70.901 ± .520(+)	66.707 ± .354(+)	87.086 ± .058(-)	56.864 ± .368(+)	31.920 ± 1.295(+)	26.950 ± .278(-)
$\tau = 1/\sqrt{\beta_t}$	85.555 ± .188(-)	90.222 ± .053(+)	42.882 ± 1.304(+)	85.865 ± .247(-)	70.552 ± .501(+)	87.093 ± .214(+)	87.357 ± .135(-)	40.608 ± .546(-)	32.295 ± .720(+)	43.216 ± .984(-)

6 Discussion

Discussion of Regions 3 and 4. According to the division results of sample types (see Table 12 in Appendix B), the sample sizes across the four regions are not uniform. Some datasets have insufficient sample sizes in regions 3 and 4. For instance, segment0 contains only one sample belonging to both regions 3 and 4. This article primarily analyzes sample generation based on region 1. Here, we compared the classification results generated by the spambase and abalone17 datasets using samples from region 2, region 3, and region 4. The method employed is LReDDPM (our approach), and the parameters are consistent with those specified in Tables 3 and 5.

It is evident from the results presented in Table 10 that the generated samples from different regions exert varying influences on the downstream classification task. We can selectively choose the most effective regional samples to guide the model based on specific requirements. The model provides insights into the local distribution of various sample types from a different perspective, enhancing the traceability of sample recognition in practical applications and reducing the associated costs. Additionally, sample generation is not confined to a single type. It can facilitate sample generation based on multiple regional information, allowing for the creation of a wider variety of mixed sample generation modes.

Training DDPM Using Regional Samples. We may question whether the forward process of the model can be trained directly using regional samples from the minority class and how this will impact the subsequent classification task when the training set is augmented with the generated samples. Would this approach yield better results? In this regard, we conducted an additional experiment to train DDPM using only regional samples and examined the classification results based on various evaluation metrics across all datasets. The regional samples used here are the same as those employed by LReDDPM for guidance in Tables 3 to 5. As shown in Table 11, the results for the segment0 and abalone17 datasets have improved compared to the optimal outcomes achieved with our previous method. However, the classification results for the other datasets have declined according to the provided evaluation metrics. This method exclusively utilizes regional samples to train the diffusion model. Given the high consistency of the diffusion model with the distribution, the generated samples tend to cluster around the sample type within this region. This clustering may lead to a small sample block, restricting the potential for improvement across the two classes.

Discussion of the Types of Regions in the Synthetic Data. To further verify whether the samples generated by LReDDPM correspond to their respective regional types, we divided all samples produced from the datasets

Table 10. The AUC/F1-score/G-mean Results of spambase and abalone17 (region 2/3/4)

		AUC	F1-score	G-mean
spambase	LReDDPM-r2	91.967 ± .045	90.784 ± .053	91.961 ± .045
	LReDDPM-r3	91.838 ± .058	90.668 ± .067	91.825 ± .058
	LReDDPM-r4	91.993 ± .060	90.838 ± .069	91.983 ± .060
abalone17	LReDDPM-r2	57.912 ± .635	17.694 ± 1.296	41.207 ± 1.634
	LReDDPM-r3	67.424 ± .883	34.680 ± 1.417	59.730 ± 1.501
	LReDDPM-r4	60.148 ± .694	21.269 ± 1.104	46.068 ± 1.697

Table 11. The Results of DDPM. The x_0 is directly using the regional samples

	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
AUC	89.718 ± .182(+)	90.482 ± .060(-)	65.489 ± 1.847(-)	87.772 ± .451(-)	80.488 ± .274(-)	84.707 ± .209(-)	87.713 ± .033(-)	62.970 ± .172(-)	66.965 ± .525(+)	49.307 ± .446(-)
F1-score	86.033 ± .300(+)	89.092 ± .069(-)	38.754 ± 1.856(-)	75.812 ± .499(-)	66.071 ± .379(-)	55.231 ± .321(-)	86.561 ± .040(-)	47.589 ± .212(-)	33.956 ± .686(+)	27.898 ± .461(-)
G-mean	89.174 ± .213(+)	90.471 ± .060(-)	64.627 ± 2.046(-)	87.743 ± .449(-)	79.104 ± .332(-)	84.262 ± .251(-)	87.638 ± .036(-)	60.454 ± .155(-)	58.887 ± .991(+)	45.919 ± .353(-)

accordingly. The categorization method follows Formula (5) described in Section 2.1. The experimental procedure was as follows: we replaced the regional samples used for guidance in the original training set with the samples generated by LReDDPM. Then, we performed sample type categorization on the minority classes within the modified training set using Formula (5). Finally, we verified whether the samples generated by LReDDPM still belonged to their corresponding regional types. The Figure 5 illustrates the regional sample types of the synthetic data. As shown, most datasets consistently generate the same types of regional samples according to their corresponding regional sample guidance. Notably, for the spectfheart and crg datasets, the selected samples used for guidance correspond to the second regional sample. In these two datasets, some sample types generated by the LReDDPM method have been converted into safe regional samples. This transformation has a relatively minor impact on the classifier. Additionally, except for segment0, page0, and crg—which exhibit very few samples transformed into outlier samples—the synthesized samples in the other datasets largely remain within the original safe or boundary regions. These synthesized samples effectively enhance the classifier's ability to learn the decision boundary, thereby improving the classification performance of LReDDPM.

7 Conclusions

In this study, we investigate the binary classification task involving imbalanced datasets. The classes of primary interest exhibit characteristics such as small sample sizes, high dimensionality, and diverse distributions, which pose significant challenges to the classification task in practical applications. However, existing oversampling methods are somewhat limited in their ability to enhance classification performance, and the efficiency of the sampling process is suboptimal in light of the aforementioned issues. On the other hand, the diffusion model is characterized by its strong alignment with data distribution and high sample synthesis efficiency. In order to address the challenges posed by the presence of the minority class, this article first categorizes the types of the minority class and subsequently designs a guided diffusion model based on the minority class of various regional sample types.

We analyze ten imbalanced datasets and evaluate the performance of five different synthetic sampling methods using multiple assessment metrics. We compare the effects of these methods on classifier construction after generating an augmented training set and investigate how various gradients, region types, and hyperparameters influence the model's ability to generate samples. Our analysis reveals that the new method achieves superior downstream classification performance. It can produce diverse generated samples based on different regional

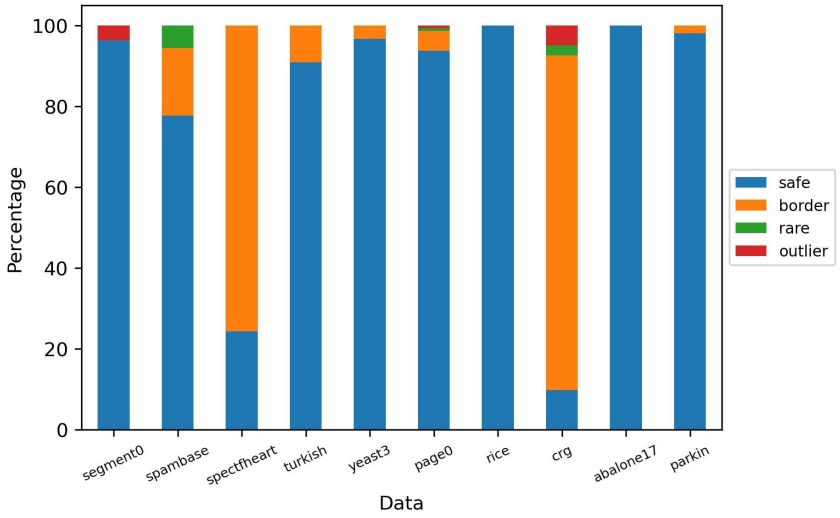


Fig. 5. Percentage stacked bar chart of the division results of synthetic data

sample types, leading to varying improvements in classification performance. This approach offers greater flexibility by providing multiple sample generation modes that are not restricted to a single region. The flexibility of our method improves our classification strategies while accommodating the distribution characteristics of the minority class.

8 Future Work

Because our method relies on partitioning the minority class, which depends on distance calculations, the process becomes slow when the number of feature variables is large (as observed with the parkin dataset), resulting in high time complexity. Therefore, the future directions for work are as follows: (1) We can first reduce the dimensions through methods such as feature selection and then categorize the region types of the minority class after the dimensionality reduction. (2) Guided by the objective of enhancing subsequent classification tasks, the model's forward process is trained directly using either samples from the minority class or the entire training set. This approach further improves both the quality of sample generation and the model's classification performance. (3) Instead of utilizing the simplified version $\hat{q}(y|s)$ of the distribution $\hat{q}(y|s, x_t)$, we will continue to optimize the diffusion model based on the gradient information of $\hat{q}(y|s, x_t)$. In addition, it is important to note that we did not directly balance the sample sizes of the minority and majority classes. Instead, we select the sample size that the model should generate based on the results of the sample division types. Without categorizing the sample division types for the minority class, determining the appropriate sample size for the model remains a significant challenge. These issues will be addressed in future research. Furthermore, our future work will also focus on challenges related to generated samples being transformed into outliers.

Acknowledgments

The authors would like to thank our team members for helpful discussions on topics related to this work. We would like to thank the anonymous reviewers for their helpful remarks, and the editor for their useful feedback that improved this article. This research was partially supported by National Natural Science Foundation of China (NSFC) under grant No.62166045.

References

J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg. 2021. "Structured denoising diffusion models in discrete state-spaces." *Advances in Neural Information Processing Systems*, 34, 17981–17993.

S. Barua, M. M. Islam, X. Yao, and K. Murase. 2012. "MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning." *IEEE Transactions on knowledge and data engineering*, 26, 405–425. doi:10.1109/TKDE.2012.232.

G. E. Batista, R. C. Prati, and M. C. Monard. 2004. "A study of the behavior of several methods for balancing machine learning training data." *ACM SIGKDD explorations newsletter*, 6, 20–29. doi:10.1145/1007730.1007735.

C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. 2009. "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." In: *Advances in knowledge discovery and data mining: 13th Pacific-Asia conference*. Springer Berlin Heidelberg, 475–482. doi:10.1007/978-3-642-01307-2_43.

N. Cahyana, S. Khomsah, and A. S. Aribowo. 2019. "Improving imbalanced dataset classification using oversampling and gradient boosting." In: *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, 217–222. doi:10.1109/ICSITech46713.2019.8987499.

H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. A. Heng, and S. Z. Li. 2024. "A survey on generative diffusion models." *IEEE Transactions on Knowledge and Data Engineering*, 36, 2814–2830. doi:10.1109/TKDE.2024.3361474.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, 16, 321–357. doi:10.1613/jair.953.

N. V. Chawla, N. Japkowicz, and A. Kotcz. 2004. "Special issue on learning from imbalanced data sets." *ACM SIGKDD explorations newsletter*, 6, 1, 1–6. doi:10.1145/1007730.1007733.

P. Dhariwal and A. Nichol. 2021. "Diffusion models beat gans on image synthesis." *Advances in neural information processing systems*, 34, 8780–8794.

H. Han, W. Y. Wang, and B. H. Mao. 2005. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learningr." In: *Advances in Intelligent Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 878–887. doi:10.1007/11538059_91.

H. He, Y. Bai, E. A. Garcia, and S. Li. 2008. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328. doi:10.1109/IJCNN.2008.4633969.

H. He and E. A. Garcia. 2009. "Learning from imbalanced data." *IEEE Transactions on knowledge and data engineering*, 21, 1263–1284. doi:10.1109/TKDE.2008.239.

J. Ho, A. Jain, and P. Abbeel. 2020. "Denoising diffusion probabilistic models." In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling. 2021. "Argmax flows and multinomial diffusion: Learning categorical distributions." *Advances in Neural Information Processing Systems*, 34, 12454–12465.

N. Japkowicz and S. Stephen. 2002. "The class imbalance problem: A systematic study." *Intelligent data analysis*, 6, 429–449.

J. Jeong, H. Jeong, and H. J. Kim. 2023. "BAMTGAN: A Balanced Augmentation Technique for Tabular Data." In: *2023 9th International Conference on Applied System Innovation (ICASI)*. IEEE, 205–207. doi:10.1109/ICASI57738.2023.10179533.

M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim. 2021. *Diff-tts: A denoising diffusion model for text-to-speech*. arXiv: 2104.01409.

D. Jiang, R. Wang, F. Shen, and W. Li. 2023. "Missing data filling in soft sensing using denoising diffusion probability model." *Measurement Science and Technology*, 35, 025117.

T. Jo and N. Japkowicz. 2004. "Class imbalances versus small disjuncts." *ACM Sigkdd Explorations Newsletter*, 6, 40–49. doi:10.1145/1007730.1007737.

S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. 2017. "Cost-sensitive learning of deep feature representations from imbalanced data." *IEEE transactions on neural networks and learning systems*, 29, 3573–3587. doi:10.1109/TNNLS.2017.2732482.

J. Kim, C. Lee, and N. Park. 2022. *Stasy: Score-based tabular data synthesis*. arXiv: 2210.04018.

A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. 2023. "Tabddpm: Modelling tabular data with diffusion models." In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 17564–17579.

B. Krawczyk, M. Woźniak, and G. Schaefer. 2014. "Cost-sensitive decision tree ensembles for effective imbalanced classification." *Applied Soft Computing*, 14, 554–562. doi:10.1016/j.asoc.2013.08.014.

S. Li, Y. Lin, H. Chen, and K. T. Cheng. 2024. "Iterative Online Image Synthesis via Diffusion Model for Imbalanced Classification." In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. doi:10.1007/978-3-031-72086-4_35.

V. López, A. Fernández, S. García, V. Palade, and F. Herrera. 2013. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." *Information sciences*, 250, 113–141. doi:10.1016/j.ins.2013.07.007.

C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Y. Zhu, and S. Ermon. 2021. "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations." In: *International Conference on Learning Representations*.

M. Mueller, K. Gruber, and D. Fok. 2023. *Continuous Diffusion for Mixed-Type Tabular Data*. arXiv: 2312.10431.

K. Napierala and J. Stefanowski. 2016. "Types of minority class examples and their influence on learning classifiers from imbalanced data." *Journal of Intelligent Information Systems*, 46, 563–597. doi:10.1007/s10844-015-0368-1.

A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. 2021. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Model*. arXiv: 2112.10741.

A. Q. Nichol and P. Dhariwal. 2021. "Improved denoising diffusion probabilistic models." In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8162–8171.

Y. Ouyang, L. Xie, C. Li, and G. Cheng. 2023. *Missdiff: Training diffusion models on tabular data with missing values*. arXiv: 2307.00467.

R. C. Prati, G. E. Batista, and M. C. Monard. 2004. "Class imbalances versus class overlapping: an analysis of a learning system behavior." In: *MICAI 2004: Advances in Artificial Intelligence: Third Mexican International Conference on Artificial Intelligence*. Springer Berlin Heidelberg, 312–321. doi:10.1007/978-3-540-24694-7_32.

Z. Qian, B. Cebere, and M. Schaar. 2023. *Synthcity: facilitating innovative use cases of synthetic data in different data modalities*. arXiv: 2301.07573.

Y. Qin, H. Zheng, J. Yao, M. Zhou, and Y. Zhang. 2023. "Class-balancing diffusion models." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18434–18443.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. "High-resolution image synthesis with latent diffusion models." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

T. Sattarov, M. Schreyer, and D. Borth. 2023. "Findiff: Diffusion models for financial tabular data generation." In: *Proceedings of the Fourth ACM International Conference on AI in Finance*, 64–72. doi:10.1145/3604237.3626876.

J. Shao, K. Zhu, H. Zhang, and J. Wu. 2024. *Diffult: How to make diffusion model useful for long-tail recognition*. arXiv: 2403.05170.

J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. 2015. "Deep unsupervised learning using nonequilibrium thermodynamics." In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. PMLR, 2256–2265.

Y. Song, L. Shen, L. Xing, and S. Ermon. 2021. *Solving inverse problems in medical imaging with score-based generative models*. arXiv: 2111.08005.

Y. Song, C. Durkan, I. Murray, and S. Ermon. 2021. "Maximum likelihood training of score-based diffusion models." *Advances in neural information processing systems*, 34, 1415–1428.

Y. Sun, L. Cai, B. Liao, W. Zhu, and J. Xu. 2022. "A robust oversampling approach for class imbalance problem with small disjuncts." *IEEE Transactions on Knowledge and Data Engineering*, 35, 5550–5562. doi:10.1109/TKDE.2022.3161291.

Y. Sun, A. K. Wong, and M. S. Kamel. 2009. "Classification of imbalanced data: A review." *International journal of pattern recognition and artificial intelligence*, 23, 687–719. doi:10.1142/S0218001409007326.

L. Van der Maaten and G. Hinton. 2008. "Visualizing data using t-SNE." *Journal of machine learning research*, 9, 2579–2605.

G. M. Weiss and F. Provost. 2003. "Learning when training data are costly: The effect of class distribution on tree induction." *Journal of artificial intelligence research*, 19, 315–354. doi:10.1613/jair.1199.

D. R. Wilson and T. R. Martinez. 1997. "Improved heterogeneous distance functions." *Journal of artificial intelligence research*, 6, 1–34. doi:10.1613/jair.346.

J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks. 2022. "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 650–656. doi:10.1109/CVPRW56347.2022.00080.

Y. Xie, M. Qiu, H. Zhang, L. Peng, and Z. Chen. 2020. "Gaussian distribution based oversampling for imbalanced data classification." *IEEE Transactions on Knowledge and Data Engineering*, 34, 667–679. doi:10.1109/TKDE.2020.2985965.

L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. 2019. "Modeling tabular data using Conditional GAN." *Advances in Neural Information Processing Systems*, 32.

D. Yan, L. Qi, V. T. Hu, M. H. Yang, and M. Tang. 2024. *Training Class-Imbalanced Diffusion Model Via Overlap Optimization*. arXiv: 2402.10821.

G. Zamberg, M. Salhov, O. Lindenbaum, and A. Averbuch. 2023. *TabADM: Unsupervised Tabular Anomaly Detection with Diffusion Models*. arXiv: 2307.12336.

H. Zhang, J. Zhang, B. Srinivasan, Z. Shen, X. Qin, C. Faloutsos, H. Rangwala, and G. Karypis. 2023. *Mixed-type tabular data synthesis with score-based diffusion in latent space*. arXiv: 2310.09656.

S. Zheng and N. Charoenphakdee. 2022. *Diffusion models for missing value imputation in tabular data*. arXiv: 2210.17128.

A Derivation of the Forward Process in the Diffusion Model

Let

$$\begin{aligned}
&\hat{q}(x_0) := q(x_0) \\
&\hat{q}(x_{t+1}|x_t, s, y) := \hat{q}(x_{t+1}|x_t, s) \\
&\hat{q}(s|x_0) := \text{local region type} \\
&\hat{q}(x_{t+1}|x_t, s) := q(x_{t+1}|x_t) \\
&\hat{q}(x_{1:T}|x_0, s, y) := \prod_{t=1}^{T} \hat{q}(x_t|x_{t-1}, s, y)
\end{aligned} \tag{7}$$

Where, the variable s represents the regional characteristics of the sample, specifically the intrinsic traits of the minority classes within the sample. When we define the forward denoising process $\hat{q}$ under conditions y and s, we can prove that the denoising operation of $\hat{q}$ is just similar to the denoising process of q under conditions y and s. At the same time, we first derive the operation $\hat{q}(x_{t+1}|x_t)$ of the unconditional noise process $\hat{q}$.

First of all,

$$\begin{aligned}
\hat{q}(x_{t+1}|x_t) &= \int_s \hat{q}(x_{t+1}, s|x_t) ds \\
&= \int_s \hat{q}(x_{t+1}|s, x_t)\hat{q}(s|x_t) ds \\
&= \int_s q(x_{t+1}|x_t)\hat{q}(s|x_t) ds \\
&= q(x_{t+1}|x_t) \int_s \hat{q}(s|x_t) ds \\
&= q(x_{t+1}|x_t) \\
&= \hat{q}(x_{t+1}|x_t, s)
\end{aligned} \tag{8}$$

$$\begin{aligned}
\hat{q}(x_{1:T}|x_0) &= \int_{y,s} \hat{q}(x_{1:T}, y, s|x_0) dy ds \\
&= \int_{y,s} \hat{q}(y, s|x_0)\hat{q}(x_{1:T}|x_0, y, s) dy ds \\
&= \int_{y,s} \hat{q}(y, s|x_0) \prod_{t=1}^{T} \hat{q}(x_t|x_{t-1}, y, s) dy ds \\
&= \int_{y,s} \hat{q}(y, s|x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}) dy ds \\
&= \prod_{t=1}^{T} q(x_t|x_{t-1}) \int_{y,s} \hat{q}(y, s|x_0) dy ds \\
&= \prod_{t=1}^{T} q(x_t|x_{t-1}) \\
&= q(x_{1:T}|x_0)
\end{aligned} \tag{9}$$

$$\begin{aligned}
\hat{q}(x_t) &= \int_{x_{0:t-1}} \hat{q}(x_0, \ldots, x_t) dx_{0:t-1} \\
&= \int_{x_{0:t-1}} \hat{q}(x_0)\hat{q}(x_1, \ldots, x_t|x_0) dx_{0:t-1} \\
&= \int_{x_{0:t-1}} q(x_0)q(x_{1:t}|x_0) dx_{0:t-1} \\
&= \int_{x_{0:t-1}} q(x_0, \ldots, x_t) dx_{0:t-1} \\
&= q(x_t)
\end{aligned} \tag{10}$$

As a result,

$$\begin{aligned}
\hat{q}(x_t) &= q(x_t) \\
\hat{q}(x_{t+1}|x_t) &= q(x_{t+1}|x_t)
\end{aligned} \tag{11}$$

We can deduce,

$$\hat{q}(x_t|x_{t+1}) = q(x_t|x_{t+1})(\hat{q}(x_t|x_{t+1}) = \frac{\hat{q}(x_t)\hat{q}(x_{t+1}|x_t)}{\hat{q}(x_{t+1})}) \tag{12}$$

When $t \neq 0$,

$$\begin{aligned}
\hat{q}(x_t|s, x_{t+1}) &= \frac{\hat{q}(x_t, s, x_{t+1})}{\hat{q}(s, x_{t+1})} \\
&= \frac{\hat{q}(x_t)\hat{q}(s|x_t)\hat{q}(x_{t+1}|s, x_t)}{\hat{q}(x_{t+1})\hat{q}(s|x_{t+1})} \\
&= \frac{\hat{q}(x_t)\hat{q}(s)\hat{q}(x_{t+1}|s, x_t)}{\hat{q}(x_{t+1})\hat{q}(s)} \\
&= \frac{\hat{q}(x_t)\hat{q}(x_{t+1}|x_t)}{\hat{q}(x_{t+1})} \\
&= \hat{q}(x_t|x_{t+1})
\end{aligned} \tag{13}$$

In Formula (13), when $t \geq 1$, x_t has not yet been added to region s, therefore $\hat{q}(s|x_t) = \hat{q}(s)$. When $t \geq 0$, $\hat{q}(s|x_{t+1}) = \hat{q}(s)$. And,

$$\begin{aligned}
\hat{q}(y|x_t, s, x_{t+1}) &= \frac{\hat{q}(x_t, s, y, x_{t+1})}{\hat{q}(x_t, s, x_{t+1})} \\
&= \frac{\hat{q}(x_t, s, y)}{\hat{q}(x_t, s, x_{t+1})} \frac{\hat{q}(x_t, s, y, x_{t+1})}{\hat{q}(x_t, s, y)} \\
&= \frac{\hat{q}(y|x_t, s)}{\hat{q}(x_{t+1}|x_t, s)} \hat{q}(x_{t+1}|x_t, s, y) \\
&= \frac{\hat{q}(y|x_t, s)}{\hat{q}(x_{t+1}|x_t, s)} \hat{q}(x_{t+1}|x_t, s) \\
&= \hat{q}(y|x_t, s)
\end{aligned} \tag{14}$$

B Division of Minority Class Examples Types

This section includes a table (see Table 12) that displays the percentage of each of the four sample types within the minority class in the training set.

Table 12. Types of minority class examples (percentage of the training set(%))

	safe(region 1)	border(region 2)	rare(region 3)	outlier(region 4)
segment0	96.5909	2.6515	0.3788	0.3788
spambase	77.1670	17.2657	2.5370	3.0303
spectfheart	6.5217	89.1304	0	4.3478
turkish	73.1707	26.8293	0	0
yeast3	55.5556	24.6032	4.7619	15.0794
page0	68.9342	20.1814	4.0816	6.8027
rice	83.7500	11.4844	1.5625	3.2031
crg	17.0833	63.7500	8.3333	10.8333
abalone17	6.2500	27.0833	25.0000	41.6667
parkin	33.1169	57.1429	5.1948	4.5455

C Additional Results of Evaluation Metrics

This section contains a total of six tables. The first three tables are as follows: Table 13 presents the G-mean, Recall, and Accuracy values for the parkin dataset under two different classifiers; Table 14 displays the Recall and Accuracy values for three methods applied to the segment0 dataset; Table 15 shows the Recall and Accuracy values for three datasets (segment0, page0, and abalone17) evaluated with two methods (LReDDPM-r_var and LReDDPM (ours)). Tables 16 through 18 provide detailed indicator results based on three distribution distance metrics.

Table 13. G-mean, Recall and Accuracy (parkin)

		G-mean	Recall	Accuracy
Decision Tree	ADASYN	37.960	35.333	42.268
	TVAE	44.304	46.684	47.189
	LReDDPM(ours)	60.214	63.474	59.171
	Original	41.947	29.386	56.237
Logistic Regression	ADASYN	79.704	84.211	77.632
	TVAE	80.870	80.526	81.053
	LReDDPM(ours)	81.082	84.211	79.605
	Original	81.988	84.211	80.921

Table 14. Comparison under several different methods (segment0)

	Recall	Accuracy
BorderlineSMOTE	80.369	96.127
CTGAN	79.795	96.222
LReDDPM(ours)	80.318	96.231
Original	80.400	96.179

Table 15. Recall and Accuracy (LReDDPM-r_var, $\tau = 60$)

		Recall	Accuracy
segment0	LReDDPM-r_var	80.133	96.253
	LReDDPM(ours)	80.318	96.231
page0	LReDDPM-r_var	77.497	91.613
	LReDDPM(ours)	78.209	91.455
abalone17	LReDDPM-r_var	32.867	97.011
	LReDDPM(ours)	33.133	96.957

Table 16. The Jensen-Shannon Distance(JS) between synthetic data and real data

	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
ADASYN	0.02813	**0.00994**	0.01730	0.01947	0.02362	0.01517	0.05316	0.02816	0.09180	0.02453
BorderlineSMOTE	0.04164	0.01294	0.01880	0.02167	0.02317	0.01330	0.04507	0.02644	0.10930	0.01540
FinDiff	0.02123	0.01115	0.01580	0.01570	**0.01700**	**0.00570**	0.01833	0.04842	0.11850	0.01570
TVAE	0.03123	0.01666	0.02140	0.02157	0.01702	0.01100	0.02168	0.03120	0.10490	0.03080
CTGAN	0.03959	0.02817	0.03190	0.03147	0.02168	0.00940	0.03418	**0.01567**	0.08480	0.03980
LReDDPM(ours)	**0.02067**	0.01765	**0.01480**	**0.01557**	0.01842	0.01200	**0.01689**	0.03318	**0.07290**	**0.01420**

Table 17. The Wasserstein Distance(WD) between synthetic data and real data

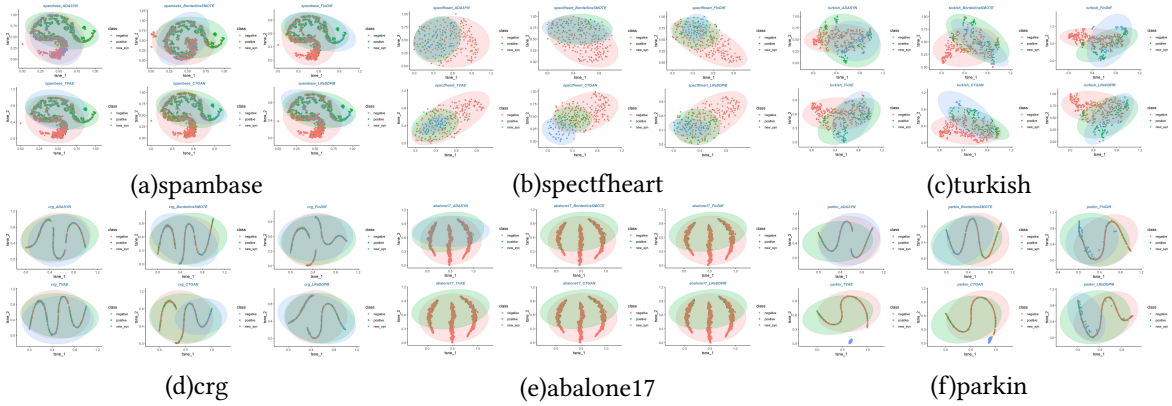
	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
ADASYN	**0.04991**	**0.02669**	**0.04880**	**0.04253**	0.02335	0.00950	0.03450	0.05710	**0.31130**	0.04980
BorderlineSMOTE	0.07564	0.05664	**0.04880**	0.04717	0.02717	0.00780	**0.03330**	**0.04880**	0.43260	**0.02500**
FinDiff	0.07179	0.05609	**0.04880**	0.04683	0.00627	**0.00770**	**0.03330**	**0.04880**	0.58700	0.02920
TVAE	0.07198	0.05617	**0.04880**	0.04650	0.00713	**0.00770**	**0.03330**	**0.04880**	0.41840	0.03040
CTGAN	0.07179	0.05633	**0.04880**	0.04650	0.01905	**0.00770**	**0.03330**	**0.04880**	0.50320	0.02690
LReDDPM(ours)	0.07218	0.05623	**0.04880**	0.04683	**0.00507**	**0.00770**	**0.03330**	**0.04880**	0.34210	0.02650

Table 18. The Maximum Mean Discrepancy(MMD) between synthetic data and real data

	segment0	spambase	spectfheart	turkish	yeast3	page0	rice	crg	abalone17	parkin
ADASYN	0.84321	1.47890	**0.41020**	**0.98880**	0.31133	0.15530	0.71806	6.09152	3.27410	1339.67
BorderlineSMOTE	1.66386	5.70968	0.60470	1.25690	0.29127	0.14310	0.42271	5.93876	4.92830	454.394
FinDiff	1.12516	6.19049	1.45150	2.08837	0.22997	0.10320	0.15326	9.77733	3.61780	1127.85
TVAE	**0.47303**	**1.16262**	1.09630	1.54597	**0.14137**	**0.07790**	**0.11333**	**5.05607**	2.10750	732.476
CTGAN	1.39771	34.9747	2.47600	3.09820	0.21422	0.12370	0.40861	5.89510	3.27180	1246.28
LReDDPM(ours)	1.12414	1.48309	1.44720	2.04013	0.15987	0.17740	0.13806	7.73394	**1.12640**	**24.2620**

D Sample t-SNE Graph of Partial Datasets

This section includes a t-SNE graph (see Figure 6) that supplements Section 4.2. Figure 6 illustrates the locations of the regions where the remaining six datasets generate samples using different methods.



(a)spambase (b)spectfheart (c)turkish

(d)crg (e)abalone17 (f)parkin

Fig. 6. The location of the region where samples are generated using various methods (The remaining six datasets)

E Some Symbols and Terms

LReDDPM-r: This abbreviation refers to the samples generated by our method through region guidance. The number indicates the corresponding region. For example, LReDDPM-r1 signifies guidance by region 1. When not specified, it denotes guidance by region 1, while the spectfheart and crg datasets are guided by region 2.
LReDDPM(ours)-r_var: This abbreviation represents the hyperparameter τ, which takes the value $\tau = 60$.
Regional Samples: The letter s in the formula denotes the regional samples, while the letter r represents the regional samples in the method notation. The regional samples were not explicitly drawn from the training set. Instead, it was conceptually divided into distinct regions to illustrate the various types of samples. This concept enhances the understanding of the sample's local distribution information, which is why the term "region" was employed.

Received 20 April 2025; accepted 12 February 2026