Beyond Distribution Matching: Contrastive-Enhanced Diffusion for Imbalanced Tabular Data Augmentation

Yuehang Ma
College of Computer Science
Sichuan University
Chengdu, China
mayuehang@stu.scu.edu.cn

Siying Li
College of Computer Science
Sichuan University
Chengdu, China
lisiying2024@stu.scu.edu.cn

Linna Wang
College of Computer Science
Sichuan University
Chengdu, China
lenawang@stu.scu.edu.cn

Li Lu*
College of Computer Science
Sichuan University
Chengdu, China
luli@scu.edu.cn

***Abstract*—Class imbalance in tabular data is a fundamental problem in real-world machine learning scenarios, often leading to suboptimal performance and generalization of downstream models on minority class. While existing data augmentation techniques partially alleviate sample scarcity, most fail to explicitly characterize intra-class consistency and inter-class discrepancy in data synthetic modeling. Consequently, ensuring that synthesized minority samples possess genuine discriminative power remains a formidable challenge. Concurrently, contrastive learning has been extensively validated in the vision and language domains for its ability to align positive pairs while segregating negative ones. Therefore, we propose CDTab, a contrastive-enhanced diffusion framework that models the feature distribution of minority samples while simultaneously capturing discriminative structures. Our method not only enhances the fidelity of generated data but also bolsters the semantic distinguishability of minority samples. Systematic experiments across six benchmark datasets using five mainstream classifiers demonstrate that our framework significantly outperforms competitive augmentation baselines in both generation quality and downstream predictive performance. The code is available at https://github.com/ningzuow/CDTab.**

***Index Terms*—Tabular Data, Data Augmentation, Class Imbalance, Contrastive Learning, Diffusion Models**

I. Introduction

Learning from imbalanced tabular data remains a fundamental challenge in machine learning [1], [2]. When minority samples are severely under-represented, standard empirical risk minimization is dominated by majority-class statistics, often leading to distorted decision boundaries that systematically overlook rare but critical patterns [3]. In high-stakes applications such as fraud detection and medical diagnosis, such minority-class failures commonly incur unacceptable costs [4], [5].

Existing research primarily addresses class imbalance through paradigms at the algorithmic and data levels [6], [7]. Algorithm-driven approaches adjust model weights to emphasize the importance of minority samples [8], but they often suffer from hyperparameter-sensitive and unstable training problems. Data-driven strategies are therefore more widely adopted due to their universality across models [6]. Given that undersampling methods often discard critical information, synthetic oversampling has become the predominant approach, particularly methods like SMOTE and its variants [9], [10], which generate new minority instances via local interpolation. However, these methods implicitly assume linear neighborhood structures and struggle to preserve semantic consistency in high-dimensional, heterogeneous tabular data [11]. Recent advances replace interpolation with generative modeling [12], [13], where diffusion-based approaches have shown promise due to their stable training dynamics and strong distributional coverage [14], [15]. Nevertheless, existing generative augmentation methods remain fundamentally distribution-centric: they only focus on matching marginal minority distributions while being largely agnostic to the discriminative information induced by class boundaries.

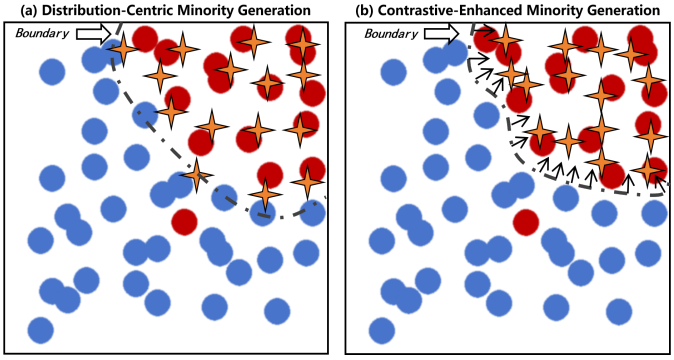

Fig. 1. Distribution-Centric and Contrastive-Enhanced minority instances distribution. The blue dots represent the majority class samples, the red dots represent the minority class samples, and the orange four-pointed stars represent the newly generated minority class instances. The gray dotted lines indicate the decision boundary. The solid black arrows point to the minority class cluster space, representing the discrimination information that makes the decision boundary clearer.

As shown in Fig. 1, this distribution–discrimination mismatch leads to a critical failure mode: although generated samples may appear statistically plausible, they frequently populate boundary-adjacent or semantically ambiguous regions, limiting their effectiveness or even harming perfor-

* Corresponding author.

mance in downstream classifiers. Crucially, this issue cannot be resolved by improving generative fidelity alone. Effective minority augmentation requires not only recovering the minority distribution, but doing so within a representation space that explicitly encodes intra-class consistency and inter-class separation. Contrastive learning emerges as a self-supervised principled paradigm to enforce such structural constraints by aligning positive pairs and segregating negative ones in the latent manifold [16]–[18]. By integrating this discriminative objective, we can bridge the gap between statistical fidelity and semantic distinguishability.

Motivated by this observation, we introduce a contrastive diffusion perspective in tabular minority data generation. Our key insight is that minority synthesis should be guided by a discriminatively structured latent space, where generative modeling and class-aware representation learning are tightly coupled. Rather than treating generation and discrimination as decoupled stages, we jointly optimize them within a unified feature space, allowing the discriminative structure to directly constrain the generative process.

In this paper, we propose CDTab, a contrastive-enhanced diffusion framework for tabular minority class generation. CDTab jointly models minority class generation and inter-class discrimination within a unified feature representation space. It achieves high-fidelity fitting to minority class sample distributions while explicitly preserving their semantic consistency and distinctiveness from majority class samples. CDTab directly generates target minority samples without requiring additional conditional guidance mechanisms. Additionally, this paper introduces a structure-aware sample filtering strategy within the contrastive subspace, further retaining high-quality synthetic data with stronger discriminative consistency, thereby enhancing the stability of generation. The main contributions of this paper are summarized as follows:

- We propose CDTab, a contrastive-enhanced diffusion framework for tabular minority class generation that jointly optimizes diffusion generation and contrastive discrimination objectives within a unified feature space. It enables the generation process to learn minority class distributions while being constrained by the discriminative structure of categories, supporting the direct generation of minority class samples.
- We design a contrastive, latent space-based, discriminatively structure-aware sample filtration mechanism to selectively retain generated samples with stronger semantic consistency, enhancing the stability and effectiveness of data augmentation.
- We evaluate CDTab against 7 baseline methods, including SMOTE, its variants and generative models, across 6 widely used datasets with imbalance ratios ranging from 3.18 to 20.02. Each method-balanced dataset is used for binary classification with 5 ML models. Extensive experiments demonstrate that CDTab improves generation quality and downstream classification performance over existing oversampling and generative baselines.

II. Related work

A. SMOTE-based Oversampling Methods

SMOTE (Synthetic Minority Over-sampling Technique) [19] is one of the most commonly used oversampling methods for addressing the class imbalance problem in tabular datasets. It generates new minority instances by interpolating the existing minority samples in the feature space. Based on this framework, various variants have been proposed, such as ADASYN [20] which introduces an adaptive sampling strategy to assign higher generation weights to minority samples located in sparse or difficult areas; Borderline-SMOTE [21] only generates minority instances close to the class boundary; SMOTEENN [22] combines SMOTE with the edit nearest neighbor (ENN) rule, where oversampling is performed first, and then noise samples are deleted based on local neighborhood consistency.

B. Deep Generative Models for Tabular Data

We particularly note that the latest advancements in deep generative models have enabled the generation of high-quality synthetic data [23]–[25]. Earlier methods mainly relied on variational autoencoders [24] and generative adversarial networks to simulate heterogeneous feature distributions. Among them, TVAE [26] is a representative table synthesizer based on variational autoencoders, specifically designed for mixed numerical and categorical attributes. CTGAN [27] introduces conditional generation for table features, while CTAB-GAN [28] and CTAB-GAN+ [29] further improve the stability of training and the modeling ability of discrete features through frequency-aware sampling strategies, making them common benchmark models for table data augmentation. In recent years, diffusion-based models have demonstrated superior generation quality when dealing with table data [14]. Representative methods include TabDDPM [30] and TabSyn [31], and the recent TabDiff [32] has achieved state-of-the-art generation quality through its innovative training and sampling mechanisms.

Although these existing table data synthesizers have certain effectiveness, most do not explicitly address data augmentation for handling class imbalance, and they mostly focus on feature distribution modeling and do not show the problem of modeling the discriminative boundaries of classes. Our method models the minority class under contrastive augmentation, can expand high-quality minority class for imbalanced datasets, and thereby effectively improve the performance of downstream classifiers.

III. Problem Definition

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote a tabular dataset for binary classification, where each instance $\mathbf{x}_i$ is composed of heterogeneous attributes, including numerical features and categorical features, and $y_i \in \{0, 1\}$ is the corresponding class label. We consider the class-imbalanced setting in which the minority class ($y = 1$) is significantly underrepresented compared to the majority class ($y = 0$), i.e., $|\mathcal{I}_{min}| < |\mathcal{I}_{maj}|$, where $\mathcal{I}_{min} = \{i : y_i = 1\}$ and $\mathcal{I}_{maj} = \{i : y_i = 0\}$.

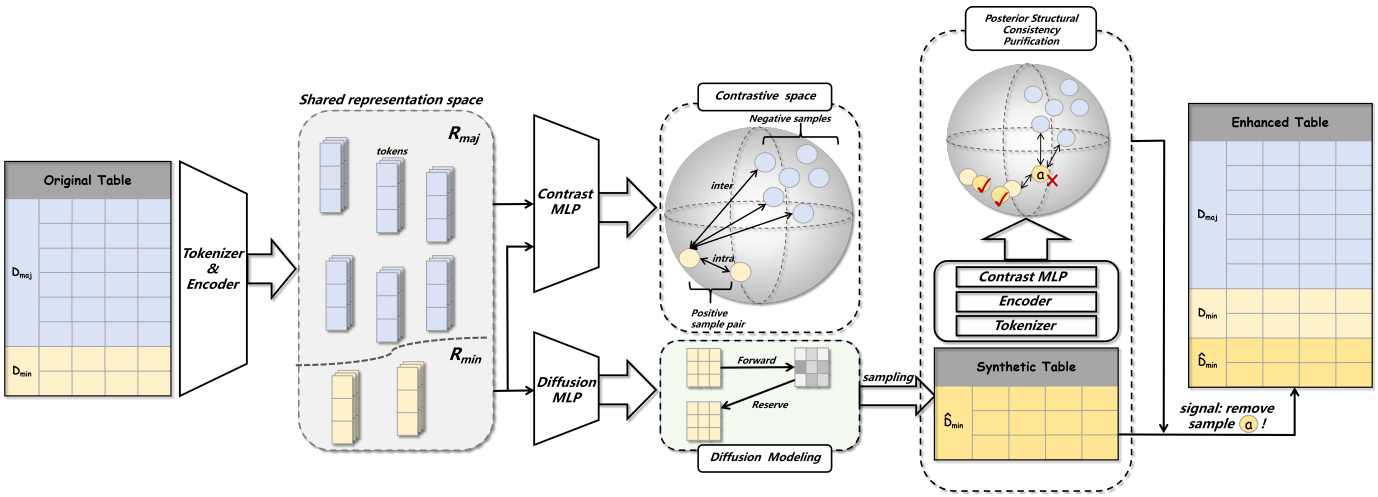

Fig. 2. A high-level overview of the CDTab. This model generates minority class samples with clear discriminative semantics by simultaneously conducting contrastive learning and diffusion modeling in a shared representation space. During the posterior process, we designed a structural consistency check to filter out the noise in the generated samples.

The objective of this work is to synthesize additional minority-class samples to alleviate class imbalance, thereby reducing the bias induced in downstream classification tasks. Formally, we aim to generate a set of synthetic instances $\hat{\mathcal{D}}_{min} = \{\hat{\mathbf{x}}_j\}$ such that the induced distribution $P_G(\mathbf{x} \mid y = 1)$ approximates the true minority conditional distribution $P(\mathbf{x} \mid y = 1)$. Beyond distributional fidelity, the generated samples are required to exhibit clear minority-class semantics and remain well separated from the majority class in a discriminative representation space $\mathcal{H}$. In the following section, we will provide a detailed description of the method we have proposed.

IV. Method

This section details the architecture of CDTab, a contrastive-enhanced diffusion framework, designed to address class imbalance in tabular data from a unified discriminative perspective. It achieves this by jointly modelling the distribution of minority class and explicitly implementing class separability within a shared latent space. As illustrated in Fig. 2, the workflow comprises two tightly coupled components:

(i) Contrastive-enhanced diffusion training and synthesis: This framework first embeds the table samples into a shared latent space through a common encoder, ensuring that the generation and contrast branches learn the same representation. On top of this representation, diffusion modeling is applied to capture the minority-class distribution, while contrastive learning enforces class-aware discriminative structure. Minority samples are finally generated through the reverse diffusion sampling process in the learned latent space.

(ii) Posterior Structural Consistency Purification: We perform posterior structure consistency optimization on the generated minority class samples. We project the minority class samples and the real training set samples onto the discriminative space learned by contrastive learning, and the minority class samples dominated by the majority class neighbors will be discarded, thereby suppressing semantic noise and effectively improving the performance of downstream classification tasks.

A. Contrastive-enhanced Diffusion Training and Synthesis

In this section, we provide a detailed explanation of how the contrastive-enhanced diffusion network is trained and how it generates samples of the minority class. Since we aim to generate the minority class samples not only to fit the distribution characteristics, but also to have a clear inter-class discrimination, the model first maps training samples $\mathbf{x}$ into a shared latent representation space $\mathcal{R}$ through a common representation learning module. Formally, each sample is projected as $\mathbf{z} = f(\mathbf{x})$, where $\mathbf{z} \in \mathcal{R}$, and $f(\cdot)$ denotes the shared encoding function. Both the diffusion branch and the contrastive branch are learned upon this shared latent space, such that their respective learning signals are jointly reflected in $\mathcal{R}$.

1) Minority-Targeted Diffusion Modeling: Training a diffusion model on the entire dataset under class imbalance may cause the generative process to be dominated by majority-class statistics, which can bias the synthesis of minority samples and blur inter-class decision boundaries. To avoid such majority-induced interference and preserve clear minority semantics, we restrict the diffusion modeling process to minority-class samples only, focusing explicitly on learning their intrinsic feature distribution.

Let $\mathcal{R}_{\min} \subset \mathcal{R}$ denote the minority-class latent subspace induced by the shared representation module. The forward diffusion process is defined as

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t\,;\, \sqrt{\alpha_t}\,\mathbf{z}_0,\, (1-\alpha_t)\mathbf{I})\,, \quad \mathbf{z}_0 \in \mathcal{R}_{\min}, \quad (1)$$

where $\{\alpha_t\}_{t=1}^{T}$ controls the noise schedule.

To accommodate the heterogeneous characteristics of different feature types, we adopt a feature-adaptive noise injection scheme inspired by TabDiff [32]. For a minority latent

representation $\mathbf{z}_0 \in \mathcal{R}_{\min}$, numerical features are corrupted by additive Gaussian noise, while categorical features are perturbed via stochastic masking:

$$\begin{aligned} \mathbf{z}_t^{\text{num}} &= \mathbf{z}_0^{\text{num}} + \boldsymbol{\sigma}^{\text{num}}(t) \odot \boldsymbol{\epsilon}, \\ \mathbf{z}_t^{\text{cat}} &\sim \text{Mask}\big(\mathbf{z}_0^{\text{cat}};\, \boldsymbol{\alpha}^{\text{cat}}(t)\big), \end{aligned} \tag{2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\sigma}^{\text{num}}(t)$ and $\boldsymbol{\alpha}^{\text{cat}}(t)$ denote feature-wise noise schedules for numerical and categorical components, respectively.

The noise schedules are parameterized independently for each feature dimension and learned during training:

$$\begin{cases} \sigma_{\rho_i}^{\text{num}}(t) = \big(\sigma_{\min}^{1/\rho_i} + t(\sigma_{\max}^{1/\rho_i} - \sigma_{\min}^{1/\rho_i})\big)^{\rho_i}, & \textbf{(numerical)} \\ \alpha_{k_j}^{\text{cat}}(t) = 1 - t^{k_j}, & \textbf{(categorical)} \end{cases} \tag{3}$$

where ρ_i and k_j are learnable parameters controlling the noise schedules of the i-th numerical and j-th categorical features, respectively. This learnable noise scheduling addresses the inherent feature heterogeneity of tabular data. Unlike image generation where pixels share a homogeneous domain, each dimension in a table possesses unique physical semantics and distributions. It achieves feature-wise customized modeling, allowing the diffusion branch to adaptively calibrate the corruption intensity for numerical and categorical attributes, thereby preserving the intricate correlations within high-dimensional tabular structures. Specifically, the reverse denoising process is parameterized by a multi-layer Transformer network, which effectively captures the complex inter-feature dependencies within the minority-class latent space. The diffusion branch is trained by minimizing a denoising objective defined exclusively on minority-class representations.

2) Discriminable Contrastive Structure Learning: Although the diffusion branch can generate a certain amount of samples that are highly consistent with the underlying distribution, in order to clearly enhance semantic consistency, we introduce a discriminative learning objective from a comparative perspective, thereby enabling the model to be encouraged to achieve intra-class compactness among the minority class samples, while expanding the representation gap between the minority class and the majority class, thereby promoting clearer class boundaries.

Contrastive learning requires first determining the positive and negative samples. Specifically, given a mini-batch of training samples $\mathcal{B} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{B}|}$, we first identify, for each sample $\mathbf{x}_i$, its Euclidean nearest neighbor $\mathbf{x}_j$ of the same class in the original data space according to the Euclidean distance. Based on the shared latent representation space $\mathcal{R}$, each sample and its class-consistent nearest neighbor then form a positive pair $(\mathbf{z}_i, \mathbf{z}_j)$, where $\mathbf{z}_i = f(\mathbf{x}_i)$ and $\mathbf{z}_j = f(\mathbf{x}_j)$. For each positive pair, all remaining samples in the batch whose labels satisfy $y_k \neq y_i$ are regarded as negative samples. To clearly highlight the boundaries between classes, this procedure is applied symmetrically to both the minority and majority classes.

To enable stable learning of discriminative structure in the shared latent space, we adopt a momentum-based contrastive network architecture. Given the positive pairs' latent representations $\mathbf{z}_i, \mathbf{z}_j \in \mathcal{R}$, they are first obtained via the shared representation function $f(\cdot)$. These latent features are then projected into a contrastive embedding space $\mathcal{S}$ through a projection head $g(\cdot)$:

$$\mathbf{h}_i = g(\mathbf{z}_i), \quad \mathbf{h}_j = g(\mathbf{z}_j), \quad \mathbf{h} \in \mathcal{S}. \tag{4}$$

To construct stable positive pairs, we further introduce a momentum encoder f_ξ, whose parameters ξ are updated as an exponential moving average of the online encoder parameters θ:

$$\xi \leftarrow m\xi + (1-m)\theta, \tag{5}$$

where $m \in [0, 1)$ denotes the momentum coefficient. The momentum encoder maps samples from the same mini-batch into momentum features $\mathbf{h}_j^m = g(f_\xi(\mathbf{x}_j))$.

Based on the constructed positive and negative samples, the contrastive loss is defined using the InfoNCE [33] objective. Let $\mathcal{P}$ denote the set of positive sample pairs in a mini-batch, and $\mathcal{N}_i$ denote the set of negative samples associated with anchor $\mathbf{h}_i$. The contrastive loss is formulated as

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} -\log \frac{\exp\big(\text{sim}(\mathbf{h}_i, \mathbf{h}_j^m)/\tau\big)}{\sum\limits_{k\in\mathcal{N}_i\cup\{j\}} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k^m)/\tau)}. \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. By normalizing over the number of positive pairs, the loss consistently accounts for all class-consistent neighborhoods within the batch.

3) Joint Optimization and Generative Sampling: Unlike traditional generative augmentation that treats distribution fitting and discrimination as decoupled tasks, our joint optimization objective facilitates manifold alignment between the generative and discriminative pathways. By imposing contrastive constraints, we force the diffusion process to evolve on a discriminative manifold in the latent space, ensuring that the synthesized minority samples are not only statistically plausible but also structurally coherent with the class boundaries.

In order to enable the model to simultaneously capture the unique distribution characteristics of minority groups as well as the structural information for class discrimination, we adopt an end-to-end approach to jointly optimize the diffusion loss and the contrast loss. Both of these objectives are back-propagated through the shared representation learning module, allowing the generation and discrimination learning signals to jointly shape the shared latent space. The entire training objective is defined as

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0\in\mathcal{R}_{min}} \left[\underbrace{\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|_2^2}_{\text{Distribution Fitting}} + \lambda \cdot \underbrace{\mathcal{L}_{con}(\phi(\mathbf{z}))}_{\text{Manifold Alignment}} \right] \tag{7}$$

where λ controls the relative importance of discriminative structural learning. After training, synthetic minority samples are generated via the reverse diffusion process operating in the shared latent representation space. Specifically, starting from

Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the learned diffusion model iteratively denoises the latent variable according to

$$\mathbf{z}_{t-1} = p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t), \quad t = T, \ldots, 1, \tag{8}$$

resulting in a synthesized minority latent representation $\mathbf{z}_0 \in \mathcal{R}_{\min}$. The generated latent sample is then mapped back to the original data space through the decoder $d(\cdot)$ associated with the shared representation module, $\hat{\mathbf{x}} = d(\mathbf{z}_0)$ yielding a synthetic minority-class instance $\hat{\mathbf{x}}$, which is able to preserve high feature fidelity while maintaining clear and consistent class semantics.

B. Posterior Structural Consistency Purification

Since the essence of generative models is to fit the probability distribution, even though category constraints are imposed through a comparative perspective, the synthesized minority class dataset $\hat{\mathcal{D}}_{\min}$ may still inevitably contain samples that deviate from the expected class discriminative structure. Therefore, it is necessary to perform posterior refinement on the generated dataset based on the category semantics.

Since the contrastive objective shapes a discriminative latent space $\mathcal{H}$ with intra-class compactness and inter-class separability, neighborhood relations measured in $\mathcal{H}$ provide a geometry-aware criterion for assessing whether a synthesized sample is semantically consistent with the learned class structure. Therefore, we transfer the refinement process to $\mathcal{H}$ and perform a posterior structural consistency filtering on the generated minority samples.

Concretely, let $\hat{\mathcal{D}}_{\min}$ denote the generated minority dataset and $\mathcal{D}$ denote the original training set. We first project all samples into the contrastive space $\mathcal{H}$ via the learned mapping $\phi(\cdot)$:

$$\mathbf{h} = \phi(\mathbf{x}) \triangleq g(f(\mathbf{x})), \quad \mathbf{h} \in \mathcal{H}, \tag{9}$$

where $f(\cdot)$ is the shared representation function and $g(\cdot)$ is the projection head. For any two samples $\mathbf{x}_a, \mathbf{x}_b$, we measure their proximity in $\mathcal{H}$ using the cosine distance

$$d_{\mathcal{H}}(\mathbf{x}_a, \mathbf{x}_b) = 1 - \frac{\phi(\mathbf{x}_a)^\top \phi(\mathbf{x}_b)}{\|\phi(\mathbf{x}_a)\|_2 \, \|\phi(\mathbf{x}_b)\|_2}. \tag{10}$$

Based on $d_{\mathcal{H}}(\cdot, \cdot)$, for each synthesized sample $\hat{\mathbf{x}}_i \in \hat{\mathcal{D}}_{\min}$, we retrieve its K nearest neighbors $\mathcal{N}_K(\hat{\mathbf{x}}_i)$ from the projected pool $\mathcal{D}$. Following the Edited Nearest Neighbor (ENN) principle, we discard $\hat{\mathbf{x}}_i$ if its neighborhood is dominated by majority-class samples:

$$\hat{\mathcal{D}}_{pure} = \{\hat{\mathbf{x}}_i \in \hat{\mathcal{D}}_{min} \mid \Psi_{\mathcal{H}}(\phi(\hat{\mathbf{x}}_i)) = 1\}, \tag{11}$$

$$\Psi_{\mathcal{H}}(\mathbf{h}) = \mathbb{I}\left[\sum_{\mathbf{h}_j \in \mathcal{N}_K(\mathbf{h})} \mathbb{I}(y_j = y_{min}) > \frac{K}{2}\right], \tag{12}$$

where $\hat{\mathcal{D}}_{pure}$ represents the refined minority subset, and $\Psi_{\mathcal{H}}(\cdot)$ acts as a purification operator that validates the structural integrity of each synthetic instance within the contrastive latent space $\mathcal{H}$.

TABLE I
DATASETS. INCLUDING THE TYPES AND NUMBERS OF FEATURES, THE NUMBER OF SAMPLES, AND THE IMBALANCE RATIO (IR).

Dataset	Continuous	Categorical	Samples	IR
adult	6	9	48842	3.18
default	14	11	30000	3.52
shoppers	10	8	12330	5.46
yeast	8	0	1484	8.10
winequality	11	0	735	12.85
magic	10	0	7022	20.02

V. EXPERIMENTS

A. Datasets

We selected six widely used datasets for research from the UCI ML repository: adult [34], default of credit card clients (default) [35], online shoppers purchasing intention dataset (shoppers) [36], yeast [37], wine-quality-red (winequality) [38], and MAGIC gamma telescope (magic) [39]. These datasets cover a range of imbalance issues from low to high, different sample quantities, and various feature compositions. To build a binary classification task, we applied some preprocessing strategies to a portion of the dataset. For the yeast dataset, we apply One-vs-Rest (OVR) strategy, which means one class is designated as the minority class, while the remaining classes are merged into the majority class; for the winequality dataset, we apply One-vs-One (OVO) strategy, which means a class pair is treated as separate binary tasks; for the magic dataset, we artificially first remove some of the majority class to create a balanced dataset, and then remove some of the minority class to achieve an extremely imbalanced state.

B. Experiment Setup

Baselines. We compared the classification performance of CDTab with seven baseline methods, including four widely used synthetic oversampling techniques (SMOTE, ADASYN, Borderline-SMOTE and SMOTEENN) and three deep learning generative models: TVAE, CTAB-GAN+ and TabDiff.

ML classifiers. Experiments are conducted using 5 commonly used classifiers: Logistic Regression (LR), Random Forest (RF), XGBoost, AdaBoost, and Support Vector Machine (SVM).

Evaluation Protocol and Metrics. We first applied each augmentation method to the imbalanced dataset, then trained and evaluated the downstream classifier. Specifically, each original dataset was randomly split into training and test sets at a 7:3 ratio. We applied different methods to the training set for minority class synthesis enhancement, and set the enhanced imbalance ratio to 2:1. Then we trained the classifier with the enhanced dataset, and finally evaluated the classification performance on the reserved test set using the F1 score and the area under the precision-recall curve (AUPRC).

C. Performance of downstream tasks

Overall Comparison. Fig. 3 summarizes the average performance of 5 ML models on original and augmented datasets,

TABLE II
CLASSIFICATION PERFORMANCE COMPARISON IN TERMS OF F1 ACROSS DIFFERENT DATASETS.

Datasets	adult	default	shoppers	yeast	winequality	magic	Average
IR	3.18	3.52	5.46	8.10	12.85	20.02	8.855
Original	0.6587	0.4391	0.5802	0.7131	0.0321	0.3814	0.4674
SMOTE	0.6835	0.4763	0.6407	0.7214	0.2034	0.4530	0.5297
ADASYN	0.6675	0.4743	0.6435	0.7010	0.1810	0.4245	0.5153
Borderline-SMOTE	0.6857	0.4814	0.6409	0.7009	0.2411	0.4736	0.5373
SMOTEENN	0.6872	0.4937	0.6368	0.7205	0.2433	0.4222	0.5339
TVAE	0.6734	0.4640	0.5929	0.7348	0.2412	0.4738	0.5300
CTAB-GAN+	0.6809	0.4721	0.6262	0.6976	0.1622	0.4641	0.5172
TabDiff	0.6892	0.4951	0.6419	0.6709	0.1040	0.5026	0.5173
CDTab	**0.6919**	**0.4988**	**0.6565**	**0.7498**	**0.3275**	**0.5083**	**0.5721**

along with their average ranks. Overall, CDTab significantly outperforms the traditional resampling techniques and the deep generative baseline model in terms of F1 scores across all six datasets. Moreover, unlike the Baseline method, CDTab consistently maintains a high F1 score on these datasets with varying imbalances in ratios, sample sizes, and feature compositions. Next, we will specifically analyze the performance of our method in the experiments.

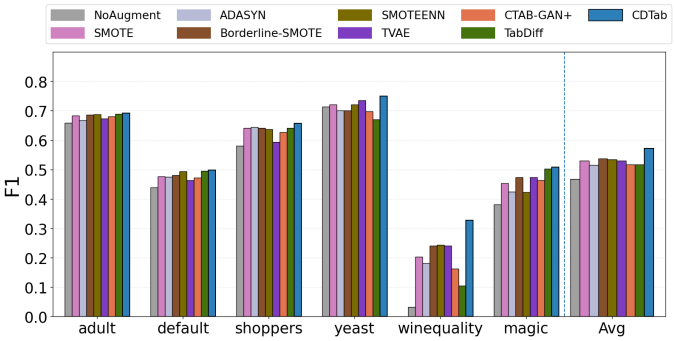

Fig. 3. F1 rankings of all augmentation models across 6 datasets. "Avg" denotes the average score of each metric over the 6 datasets.

Comparison of classification performance. Table II summarizes the average F1 performance across five classifiers on six datasets with different imbalance ratios (IR). CDTab achieves the highest overall F1 score and shows consistently competitive performance across all datasets. Compared with SMOTE and its variants, our method demonstrates significant advantages across all datasets. For large-scale complex datasets with less impact of imbalance, such as adult and default, CDTab can directly outperform by 1-2 percentage points, reaching 0.6919 and 0.4988 respectively. On the winequality dataset, which has a relatively severe class imbalance (IR = 12.85) and less sample information, our method significantly outperforms SMOTE (0.3275 vs. 0.2034) and ADASYN (0.3275 vs. 0.1810), indicating its superior ability in handling extreme imbalance situations. Although SMOTEENN shows competitiveness in some moderately imbalanced datasets, its performance declines in highly imbalanced datasets, while our method remains stable.

Compared with the deep generative model baselines including TVAE, CTAB-GAN+ and TabDiff, our method consistently achieved higher F1 scores on all six datasets. It is worth noting that on the yeast and winequality datasets, CTAB-GAN+ and TabDiff might have suffered significant performance degradation due to the small size of the datasets and high imbalance rate. However, our method still performed well, with F1 scores reaching 0.7498 and 0.3275 respectively, showing a significant improvement compared to other methods.

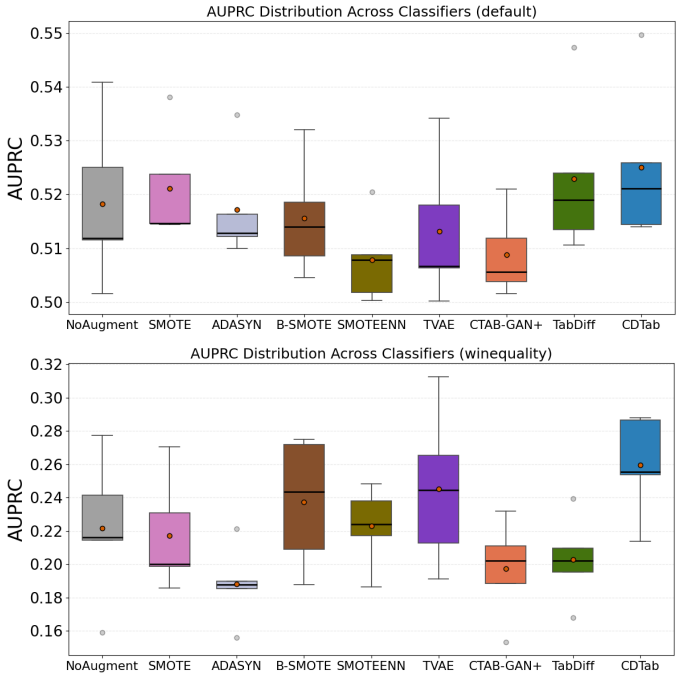

Fig. 4. Boxplots of AUPRC scores across different classifiers under various data augmentation methods on the default (top) and winequality (bottom) datasets. Each box summarizes the performance distribution over multiple classifiers. B-SMOTE stands for Borderline-SMOTE.

We further evaluated the effectiveness and stability of CDTab by examining the AUPRC across various category models. The results on the default and the winequality dataset of CDTab and baselines was reflected in the Fig. 4. As shown,

CDTab consistently achieved high median AUPRC values, and its distribution was relatively concentrated, indicating that the performance improvement was consistent across different classifiers. In contrast, some oversampling methods and deep generative baselines exhibited greater variability, suggesting that they are more sensitive to the choice of classifiers, e.g., Borderline-SMOTE shows a relatively wide interquartile range across classifiers, while TVAE demonstrates noticeable performance dispersion, particularly on the winequality dataset. These observations align with our overall predictive comparisons and further confirm that the proposed contrastive-enhanced method generates robust minority samples with strong generalization across various downstream classifiers.

D. Data-oriented distribution verification

To demonstrate the effectiveness of our method, we designed an experiment to model and generate minority class samples in the case of missing minority class information, and then detected the feature distribution, discriminative semantics, and diversity. Specifically, we employ CDTab (excluding the subsequent data filtering module) together with baseline deep generative models to model the magic dataset and generate synthetic minority-class samples. The generated minority samples are then compared against the complete set of real minority samples from the raw magic dataset. To quantitatively assess the quality of the generated samples, we adopt three complementary metrics: Shape (statistical consistency), C2ST gap (distributional fidelity via classification), and NN-distance (point-wise similarity and diversity), which jointly provide a comprehensive evaluation of the fidelity and structural consistency of the generated minority samples.

TABLE III
QUALITY OF MINORITY CLASS GENERATION.

Method	Shape ↑	C2ST gap↓	NN-distance↓
TVAE	0.6643	0.3707	1.2603
CTAB-GAN+	0.9019	0.4539	1.3814
TabDiff	0.9406	0.4649	1.2715
CDTab	**0.9463**	**0.0995**	**1.0395**

The relevant results are shown in Table III. CDTab achieved the best overall performance on all three indicators. In terms of Shape, our method obtained the highest score (0.9463), indicating that the generated samples closely match the actual distribution of the minority class. More importantly, compared with all baseline models, the C2ST gap produced by our method (0.0995) was significantly lower, suggesting that the generated samples are significantly more difficult to distinguish from the actual minority class samples, thus demonstrating better discriminative consistency. Additionally, our method yields the smallest NN distance, suggesting that real minority samples can be well represented by the generated ones, without exhibiting severe mismatch or distributional gaps. In summary, CDTab generated high-quality, distributionally faithful, and semantically consistent minority class samples, which is also an important reason why we were able to effectively improve the performance of the downstream classifier.

E. Ablation Study

The high-quality and semantically coherent samples generated by the CDTab are the result of the combined effect of contrastive learning-enhanced diffusion modeling and post-processing filtering. To distinguish the individual roles of these components, we conducted an ablation study by selectively removing the contrastive learning module and the filtering mechanism. This analysis aims to explore whether each component plays different and complementary roles in improving classification performance and the quality of generated samples.

TABLE IV
ABLATION STUDY.

	shoppers		yeast	
	AUPRC	F1	AUPRC	F1
CDTab	**0.6974**	**0.6565**	**0.7924**	**0.7498**
w/o posterior purification	0.6829	0.6475	0.7731	0.6958
w/o contrastive module	0.6834	0.6333	0.7704	0.7290

The results of the ablation study are reported in Table IV. When both the contrastive module and the filtration strategy are removed, the performance drops noticeably on both datasets, indicating that the generative process alone is insufficient to guarantee robust downstream performance. Removing only the filtration module results in a performance drop; however, the remaining contrastive learning still preserves a portion of the performance gains (e.g., in the shoppers dataset, its AUPRC reaches 0.6829 and its F1 score reaches 0.6475), suggesting its effectiveness in improving the discriminative quality of the generated samples. However, the full model consistently achieves the best performance in terms of both AUPRC and F1 on the shoppers and yeast datasets, demonstrating that the contrastive constraint and the post-hoc filtration are complementary and jointly responsible for the observed performance gains.

VI. CONCLUSION

In this paper, we introduce CDTab, which is a contrastive-enhanced diffusion generation framework designed to address the class imbalance problem in binary classification tasks. CDTab marries joint diffusion modeling with contrastive learning to recover the original feature distributions while enabling the generation of samples with explicit and robust minority-class semantics. We conducted extensive experiments using classical machine learning models on multiple benchmark datasets. The results show that: (i) Using CDTab to synthesize minority class augmented datasets can significantly improve the prediction performance of downstream classifiers; (ii) The contrast module can learn the class semantics, thereby effectively constraining the generation in terms of discrimination; (iii) Posterior Structural Consistency Purification can

effectively filter out the noise samples synthesized, thereby improving the performance of downstream classifiers; (iv) CDTab outperforms baseline enhancement techniques including TabDiff in various evaluation metrics. Overall, CDTab provides a novel and competitive solution for addressing the class imbalance problem by systematically applying the discriminative perspective of contrastive learning to enhance minority class diffusion modeling. This work expands the understanding and practical application of the contrastive perspective in imbalanced learning and data augmentation scenarios.

REFERENCES

[1] M. Altalhan, A. Algarni, and M. T.-H. Alouane, "Imbalanced data problem in machine learning: A review," *IEEE Access*, 2025.

[2] W. Ren, T. Zhao, Y. Huang, and V. Honavar, "Deep learning within tabular data: Foundations, challenges, advances and future directions," *arXiv preprint arXiv:2501.03540*, 2025.

[3] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," *Journal of Big Data*, vol. 12, no. 1, p. 71, 2025.

[4] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial intelligence review*, vol. 57, no. 10, p. 273, 2024.

[5] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, "Adversarial attacks for tabular data: Application to fraud detection and imbalanced data," *arXiv preprint arXiv:2101.08030*, 2021.

[6] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing*, vol. 143, p. 110415, 2023.

[7] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, p. 122778, 2024.

[8] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: review of methods and applications," in *IOP conference series: materials science and engineering*, vol. 1099, no. 1. IOP Publishing, 2021, p. 012077.

[9] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing class imbalance of health data: a systematic literature review on modified synthetic minority oversampling technique (smote) strategies," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1310–1318, 2024.

[10] M. S. Santos, P. H. Abreu, N. Japkowicz, A. Fernández, C. Soares, S. Wilk, and J. Santos, "On the joint-effect of class imbalance and overlap: a critical review," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6207–6275, 2022.

[11] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," *IEEe Access*, vol. 10, pp. 47 643–47 660, 2022.

[12] A. X. Wang, S. S. Chukova, C. R. Simpson, and B. P. Nguyen, "Challenges and opportunities of generative models on tabular data," *Applied Soft Computing*, vol. 166, p. 112223, 2024.

[13] T. Liu, Z. Qian, J. Berrevoets, and M. van der Schaar, "Goggle: Generative modelling for tabular data by learning relational structure," in *The Eleventh International Conference on Learning Representations*, 2023.

[14] Z. Li, Q. Huang, L. Yang, J. Shi, Z. Yang, N. van Stein, T. Bäck, and M. van Leeuwen, "Diffusion models for tabular data: Challenges, current progress, and future directions," *arXiv preprint arXiv:2502.17119*, 2025.

[15] H. Koo and T. E. Kim, "A comprehensive survey on generative diffusion models for structured data," *arXiv preprint arXiv:2306.04139*, 2023.

[16] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, "A comprehensive survey on contrastive learning," *Neurocomputing*, vol. 610, p. 128645, 2024.

[17] S. B. Rabbani, I. V. Medri, and M. D. Samad, "Attention versus contrastive learning of tabular data: a data-centric benchmarking," *International Journal of Data Science and Analytics*, vol. 20, no. 4, pp. 3069–3091, 2025.

[18] C. Niu, D. Chen, J. Zhou, J. Wang, X. Luo, Q.-H. Liu, Y. Li, and J. Lv, "Neural boneprint: Person identification from bones using generative contrastive deep learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7609–7618.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 2008, pp. 1322–1328.

[21] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[22] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[23] R. Sauber-Cole and T. M. Khoshgoftaar, "The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey," *Journal of Big Data*, vol. 9, no. 1, p. 98, 2022.

[24] M. C. Stoian, E. Giunchiglia, and T. Lukasiewicz, "A survey on tabular data generation: Utility, alignment, fidelity, privacy, and beyond," *arXiv preprint arXiv:2503.05954*, 2025.

[25] A. Kiran and S. S. Kumar, "A comparative analysis of gan and vae based synthetic data generators for high dimensional, imbalanced tabular data," in *2023 2nd International Conference for Innovation in Technology (INOCON)*. IEEE, 2023, pp. 1–6.

[26] H. Ishfaq, A. Hoogi, and D. Rubin, "Tvae: Triplet-based variational autoencoder using metric learning," *arXiv preprint arXiv:1802.04403*, 2018.

[27] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[28] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," in *Asian conference on machine learning*. PMLR, 2021, pp. 97–112.

[29] Z. Zhao, A. Kunar, R. Birke, H. Van der Scheer, and L. Y. Chen, "Ctab-gan+: Enhancing tabular data synthesis," *Frontiers in big Data*, vol. 6, p. 1296508, 2024.

[30] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "Tabddpm: Modelling tabular data with diffusion models," in *International conference on machine learning*. PMLR, 2023, pp. 17 564–17 579.

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

[32] J. Shi, M. Xu, H. Hua, H. Zhang, S. Ermon, and J. Leskovec, "Tabdiff: a unified diffusion model for multi-modal tabular data generation," in *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[34] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[35] I.-C. Yeh, "Default of Credit Card Clients," UCI Machine Learning Repository, 2009, DOI: https://doi.org/10.24432/C55S3H.

[36] C. Sakar and Y. Kastro, "Online Shoppers Purchasing Intention Dataset," UCI Machine Learning Repository, 2018, DOI: https://doi.org/10.24432/C5F88Q.

[37] K. Nakai, "Yeast," UCI Machine Learning Repository, 1991, DOI: https://doi.org/10.24432/C5KG68.

[38] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine Quality," UCI Machine Learning Repository, 2009, DOI: https://doi.org/10.24432/C56S3T.

[39] R. Bock, "MAGIC Gamma Telescope," UCI Machine Learning Repository, 2004, DOI: https://doi.org/10.24432/C52C8B.