Uncertainty-Aware Generative Oversampling Using an Entropy-Guided Conditional Variational Autoencoder

Amirhossein Zare[1], Amirhessam Zare[1], Parmida Sadat Pezeshki[1], Herlock (SeyedAbolfazl) Rahimi[2], Ali Ebrahimi[3], Ignacio Vázquez-García[4,5], Leo Anthony Celi[6,7,8]

[1]School of Medicine, Tehran University of Medical Sciences, Tehran, Iran [2]Department of Electrical and Computer Engineering, Yale University, New Haven, CT, USA [3]SDU Health Informatics and Technology, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark [4]Department of Pathology and Krantz Family Center for Cancer Research, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA [5]Broad Institute of MIT and Harvard, Cambridge, MA, USA [6]Laboratory for Computational Physiology, MIT, Cambridge, MA, USA [7]Division of Pulmonary, Critical Care & Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA [8]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract

Class imbalance remains a major challenge in machine learning, especially for high-dimensional biomedical data where nonlinear manifold structures dominate. Traditional oversampling methods such as SMOTE rely on local linear interpolation, often producing implausible synthetic samples. Deep generative models like Conditional Variational Autoencoders (CVAEs) better capture nonlinear distributions, but standard variants treat all minority samples equally, neglecting the importance of uncertain, boundary-region examples emphasized by heuristic methods like Borderline-SMOTE and ADASYN.

We propose Local Entropy-Guided Oversampling with a CVAE (LEO-CVAE), a generative oversampling framework that explicitly incorporates local uncertainty into both representation learning and data generation. To quantify uncertainty, we compute Shannon entropy over the class distribution in a sample's neighborhood: high entropy indicates greater class overlap, serving as a proxy for uncertainty. LEO-CVAE leverages this signal through two mechanisms: (i) a Local Entropy-Weighted Loss (LEWL) that emphasizes robust learning in uncertain regions, and (ii) an entropy-guided sampling strategy that concentrates generation in these informative, class-overlapping areas.

Applied to clinical genomics datasets (ADNI and TCGA lung cancer), LEO-CVAE consistently improves classifier performance, outperforming both traditional oversampling and generative baselines. These results highlight the value of uncertainty-aware generative oversampling for imbalanced learning in domains governed by complex nonlinear structures, such as omics data.

1 Introduction

The class imbalance problem, characterized by a severely skewed distribution of samples across classes, remains a critical challenge in machine learning Chen et al. (2024). Standard learning algorithms, optimized for overall accuracy, tend to develop a strong predictive bias towards the majority class (Das et al., 2018). Consequently, instances from the minority class, which are often of greatest interest in high-stakes domains like medical diagnosis, fraud detection, and industrial fault prediction, are frequently misclassified (Buda et al., 2018; Makki et al., 2019; Malhotra and Kamal, 2019).

To mitigate this issue, a variety of techniques have been developed, broadly categorized into algorithm-level and data-level approaches (Gao et al., 2025; Yang et al., 2024; Buda et al., 2018). Data-level methods, which involve rebalancing the dataset prior to training, are particularly popular due to their model-agnostic nature (Chen et al., 2024; Buda et al., 2018). Among these, oversampling techniques, which increase the representation of the minority class, are widely used (Azhar et al., 2023). The field has evolved from simple replication (Japkowicz, 2000) to the landmark Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), which generates new, synthetic samples via linear interpolation between neighboring minority instances.

✉ amhosseinzare@gmail.com, amir.hessam.zare@gmail.com, pezeshkiparmida@gmail.com, herlock.rahimi@yale.edu, aleb@mmmi.sdu.dk, ivazquez-garcia@mgh.harvard.edu, leoanthonyceli@yahoo.com

Source code available at: https://github.com/Amirhossein-Zare/LEO-CVAE

Despite its widespread adoption, SMOTE and its variants suffer from fundamental limitations. Their local, interpolative nature can generate noisy samples in regions of class overlap and fails to capture the global, often complex, distribution of the minority class (Kamalov et al., 2025; Hong et al., 2024; Tang et al., 2025; Wang et al., 2025a). Crucially, these methods typically ignore the rich distributional information present in the majority class, limiting the diversity and fidelity of the generated data (Ai et al., 2023).

These shortcomings have motivated a paradigm shift towards deep generative models, which can learn global data distributions to generate novel and consistent samples (Liu et al., 2007). While Denoising Diffusion Models have achieved state-of-the-art performance in high-fidelity image synthesis (Ho et al., 2020), their application to tabular data remains a challenging area of research due to complex, non-Gaussian feature distributions (Li et al., 2025b). For tabular settings, they are also computationally intensive to train and sample from (Shi et al., 2025).

By contrast, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (VAEs) (Kingma and Welling, 2013) have been more widely explored for tabular data generation. GANs, however, often suffer from training instability (Sampath et al., 2021), whereas VAEs provide a more stable and tractable alternative. Among them, the Conditional VAE (CVAE) (Sohn et al., 2015) offers a principled framework for class-conditioned oversampling and has been successfully applied in imbalanced learning (Fajardo et al., 2021). However, a key limitation of standard CVAEs is that they are agnostic to sample importance, treating all minority instances as equally informative. This uniform approach overlooks a critical insight from heuristic methods like Borderline-SMOTE (Han et al., 2005) and the Adaptive Synthetic Sampling (ADASYN) (Haibo et al., 2008): not all samples are equally valuable for refining a classifier's decision boundary. Instances located deep within a class's feature space are less informative than those situated in regions of high class overlap, where the boundary is most ambiguous. These "borderline" or "hard-to-learn" samples are strategically crucial, as they provide the most challenging examples for a classifier to learn (Japkowicz and Stephen, 2002; Haibo et al., 2008).

Consequently, a significant gap remains in integrating this principle of uncertainty-awareness into the powerful distributional learning capabilities of a deep generative model. This gap is particularly pronounced in clinical genomics, which is the central focus of the current study. We hypothesize that the high dimensionality and intricate, nonlinear relationships in this domain cause the core assumption of linear interpolation used by SMOTE to break down, resulting in biologically implausible synthetic data (Blagus and Lusa, 2012). Furthermore, the inherent biological heterogeneity and the continuum-like nature of disease progression mean that class boundaries are rarely sharply defined, creating regions of high predictive uncertainty for a classifier. To leverage this insight, we turn to information theory to develop a formal measure of this local uncertainty. We quantify the degree of class overlap using Shannon entropy, the canonical measure of uncertainty, allowing us to identify high-entropy regions as a quantitative proxy for a sample's 'hard-to-learn' status.

In essence, this combination of a complex, nonlinear data manifold and inherently ambiguous class boundaries establishes clinical genomics as a uniquely challenging domain. This setting reveals the limitations of two key approaches: traditional oversampling methods, which struggle with complex, non-linear data, and standard generative models, which are agnostic to uncertain and hard-to-learn regions in the feature space. Consequently, it provides an ideal setting to validate a generative framework specifically engineered to target and learn from these high-uncertainty regions.

To this end, we propose the Local Entropy-Guided Oversampling with a CVAE (LEO-CVAE). We formalize the notion of a "hard-to-learn" or "uncertain" region using local Shannon entropy, a measure that quantifies the mixture of class labels within a sample's local neighborhood. The LEO-CVAE framework leverages this local entropy signal in two synergistic ways: first, it guides the CVAE's training process through a weighted loss function that prioritizes learning in the high-entropy regions; second, it directs the synthetic data generation by preferentially selecting high-entropy instances as seeds. By concentrating both the learning and generative processes on these ambiguous areas, LEO-CVAE directly reinforces the classifier's decision boundary where it is weakest. The primary contributions of this work are:

- **A Novel Uncertainty Metric:** We introduce the Local Entropy Score (LES) to formally quantify sample-level uncertainty, identifying the most informative, class-overlapping regions for oversampling.

- **An Uncertainty-Aware Generative Framework:** We propose LEO-CVAE, which integrates the LES signal through two core components: a Local Entropy-Weighted Loss (LEWL) and an entropy-guided sampling strategy.

- **Empirical Validation:** We demonstrate the effectiveness of LEO-CVAE on challenging imbalanced clinical genomics datasets for both binary and multiclass classification through a systematic comparison against a suite of traditional and generative oversampling methods.

2 Related Works

2.1 Traditional Oversampling and its Limitations

Data-level approaches to class imbalance are dominated by oversampling techniques. The simplest method, Random Oversampling (Japkowicz, 2000), mitigates imbalance by duplicating minority class instances but is highly prone to overfitting. The Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) provided a significant improvement by generating new samples via linear interpolation between a minority instance and its k-nearest minority neighbors. This approach has become a foundational baseline and has inspired a large family of variants (Fernández et al., 2018; Li et al., 2025a; Douzas et al., 2018; Douzas and Bação, 2019; Kunakorntum et al., 2020; Wang et al., 2025b). Borderline-SMOTE (Han et al., 2005), for example, focuses generation on minority samples near the class boundary, which are deemed more critical for classification. The Adaptive Synthetic Sampling (ADASYN) (Haibo et al., 2008) method generates more synthetic data for minority samples that are harder to learn, based on the proportion of majority class instances in their local neighborhood.

Despite these advances, traditional methods share critical weaknesses. Their reliance on local, linear interpolation restricts the diversity of generated samples, often confining them to the convex hull of the original minority data (Dai et al., 2019; Wang et al., 2025a). Furthermore, by focusing exclusively on minority class neighbors, they ignore the global data structure and the valuable information contained within the majority class, which can lead to the generation of noisy samples in regions of class overlap (Batista et al., 2004; Ai et al., 2023).

2.2 Deep Generative Models for Oversampling

To overcome the limitations of local interpolation, research has shifted towards deep generative models, which learn an approximation of the entire data distribution, $p(x)$, enabling the generation of novel and globally consistent samples (Liu et al., 2007). The two most prominent architectures are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (VAEs) (Kingma and Welling, 2013). While GANs can produce highly realistic samples, they often suffer from training instability and mode collapse (Sampath et al., 2021). VAEs, by contrast, offer a stable, probabilistic framework for learning a smooth, continuous latent space from which to generate diverse data (Kingma and Welling, 2013).

The Conditional VAE (CVAE) (Sohn et al., 2015) is a natural fit for oversampling, as it allows for targeted, class-specific data generation by conditioning the model on class labels. A CVAE is comprised of two networks: an encoder, $q_\phi(z|x, c)$, which learns to map data points into a probabilistic latent space conditioned on a class label c, and a decoder, $p_\theta(x|z, c)$, which learns to reconstruct data from that latent space. This provides a principled foundation for generating new minority samples by sampling from the latent space and decoding with the desired class label (Fajardo et al., 2021).

2.3 A Taxonomy of CVAE-Based Innovations

Recent research has extended the CVAE framework for imbalanced learning along several distinct axes of innovation. Understanding these helps to precisely situate the contribution of LEO-CVAE.

Loss-Function Modification: Drawing inspiration from discriminative learning, some approaches apply a focal loss to the CVAE's reconstruction objective to assign greater weight to samples with high reconstruction error (Lin et al., 2017). This innovation modifies how the model learns from existing data.

Latent Space Structuring: A second, powerful approach focuses on engineering a more discriminative latent space where class manifolds are well-separated. The Discriminative VAE (DVAE) (Guo et al., 2019) learns a latent mixture-of-Gaussians prior, which forces minority and majority samples into distinct clusters. This explicitly models the class boundary, enabling the generation of samples near this critical region. Similarly, the Contrastive Tabular VAE (CTVAE) (Wang et al., 2025a) integrates a contrastive loss term into the training objective. This actively pulls latent representations of same-class samples together while pushing different-class samples apart, resulting in a highly structured latent space that improves the quality of generated samples. A key aspect of these methods is that after carefully structuring the latent space, they still typically sample randomly from the prior distribution for generation.

Knowledge Transfer: For scenarios of extreme data scarcity, Majority-Guided VAE (MGVAE) (Ai et al., 2023) uses a transfer learning paradigm. It pre-trains on the abundant majority class to learn a robust feature representation and then fine-tunes on the few minority samples, effectively inheriting the diversity of the majority distribution. This approach excels at solving the specific problem of learning from few samples.

Our proposed LEO-CVAE method contributes a distinct and novel perspective to this line of research. While the previously discussed methods focus on modifying the training loss, structuring the latent space, or implementing knowledge transfer, LEO-CVAE is the first to use local entropy as a direct signal to quantify

sample-level uncertainty. Instead of treating all minority samples as equally important, LEO-CVAE identifies those in ambiguous, class-overlapped regions and intensifies both the model's learning and its generative process on these specific points. This allows LEO-CVAE to strategically reinforce the decision boundary where it is most contested, offering a new, uncertainty-aware approach to generative oversampling.

3 Methods

This section details our proposed oversampling method, Local Entropy-Guided Oversampling with a Conditional Variational Autoencoder (LEO-CVAE). We first provide the problem formulation, followed by an overview of the CVAE, which serves as our generative foundation. We then introduce our core contribution: the Local Entropy Score (LES), a metric for quantifying sample-level uncertainty to identify high-entropy regions within the feature space. Finally, we describe how LES is integrated into the CVAE through our two novel mechanisms: the Local Entropy-Weighted Loss (LEWL) for model training and the Entropy-Guided Sampling strategy for data generation.

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a training dataset of N samples, where $x_i \in \mathbb{R}^D$ is a D-dimensional feature vector and $y_i \in \{c_1, c_2, \ldots, c_C\}$ is its corresponding class label. The dataset $\mathcal{D}$ is imbalanced if the class distribution is skewed, i.e., there exists a majority class c_{maj} and a minority class c_{min} such that the number of samples $N_{maj} \gg N_{min}$. The goal of oversampling is to generate a new set of synthetic minority samples, $\mathcal{D}_{syn}$, such that when combined with the original data ($\mathcal{D}' = \mathcal{D} \cup \mathcal{D}_{syn}$), the resulting dataset is balanced or near-balanced, leading to improved performance of a classifier trained on $\mathcal{D}'$.

3.2 CVAE Foundation

Our method is built upon a CVAE (Sohn et al., 2015), a generative model that learns a latent representation of data conditioned on class labels. A CVAE consists of two neural networks: an encoder and a decoder.

The encoder network, parameterized by ϕ, learns to approximate the intractable true posterior distribution $p(z|x, c)$. It maps a data point x and its class condition c to the parameters of a diagonal Gaussian distribution, $q_\phi(z|x, c) = \mathcal{N}(z|\mu, \text{diag}(\sigma^2))$. The mean vector μ and variance vector σ^2 are the direct outputs of the encoder network.

The decoder network, parameterized by θ, reconstructs the data by modeling the distribution $p_\theta(x|z, c)$. To generate a reconstructed sample $\hat{x}$, a latent vector z is first sampled from the encoder's output distribution using the reparameterization trick: $z = \mu + \epsilon \odot \sigma$, where $\epsilon \sim \mathcal{N}(0, I)$. This sample z is then concatenated with the one-hot class vector c_{oh} and passed as input to the decoder.

The standard CVAE is trained by minimizing a loss function derived from the negative of the Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathcal{L}_{\text{CVAE}} = & -\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)] \\ & + \beta \cdot D_{KL}(q_\phi(z|x, c)||p(z|c))\end{aligned} \tag{1}$$

The first term is the reconstruction loss, which measures how well the model reconstructs the input data, and the second is the Kullback-Leibler (KL) divergence, which regularizes the latent space to follow a prior distribution $p(z|c)$.

3.3 Local Entropy Score (LES)

The core novelty of our method is the introduction of LES to guide the CVAE. LES quantifies the complexity of the feature space surrounding a given sample. For each sample x_i, we first identify its k nearest neighbors, $\mathcal{N}_k(x_i)$. The neighborhood is determined using the standard Euclidean distance between feature vectors. We then compute the probability distribution of classes within this neighborhood. Let k_j be the count of neighbors belonging to class c_j. The local probability of class c_j is $p(c_j|x_i) = k_j/k$. The LES for sample x_i, denoted as $H(x_i)$, is then calculated using the Shannon entropy formula:

$$H(x_i) = -\sum_{j=1}^{C} p(c_j|x_i) \log_2 p(c_j|x_i) \tag{2}$$

For this calculation, we follow the standard convention that the contribution of any class c_j with $p(c_j|x_i) = 0$ is taken to be 0, as $\lim_{p\to 0^+} p \log p = 0$.

The resulting score, which we denote $H_i = H(x_i)$, provides a quantitative measure of the heterogeneity of the sample's local feature space. The score ranges from a minimum of 0, signifying a perfectly homogenous neighborhood where all neighbors belong to the same class, to a theoretical maximum of $\log_2(C)$, where C is the total number of classes. This maximum value corresponds to a state of maximum entropy, representing the highest possible uncertainty where neighbors are uniformly distributed across all classes. A high LES, therefore, directly identifies a sample located in a complex, class-overlapping region of the feature space. This score becomes the critical signal for guiding both the learning and generation processes in our LEO-CVAE framework.

3.4 The LEO-CVAE Method

LEO-CVAE leverages the calculated local entropy scores in two critical stages: modifying the CVAE loss function to focus learning on high-entropy regions and guiding the generation of new synthetic samples.

3.4.1 The Local Entropy-Weighted Loss (LEWL)

To guide the CVAE's training, our primary innovation is a novel loss function, the LEWL. It instills uncertainty-awareness by modifying the CVAE's reconstruction component, while the standard KL divergence term is retained for its conventional regularization role. The complete LEWL objective is a composite function defined as:

$$\mathcal{L}_{\text{LEWL}} = \mathcal{L}_{\text{W-Recon}} + \beta \cdot \mathcal{L}_{\text{KLD}} \tag{3}$$

Here, β is a hyperparameter that weights the contribution of the Kullback-Leibler (KL) divergence term. We detail each component below.

Weighted Reconstruction Loss ($\mathcal{L}_{\text{W-Recon}}$): The innovation of our training paradigm is captured in this component. We replace the CVAE's standard reconstruction error with a Weighted Mean Squared Error (MSE). This loss component strategically prioritizes samples that are most informative for learning a robust model: those belonging to minority classes and those located in complex, class-overlapping regions identified by a high LES. The loss for a single sample x_i with class label y_i and pre-calculated local entropy H_i is defined as:

$$\mathcal{L}_{\text{W-Recon},i} = w_{\text{class}}(y_i) \cdot w_{\text{entropy}}(H_i) \cdot \|x_i - \hat{x}_i\|_2^2 \tag{4}$$

The two weighting factors are:

- **Class-Imbalance Weight (w_{class}):** This weight is calculated from the inverse frequency of each class in the training data. By assigning a higher weight to samples from underrepresented classes, it compels the model to overcome the inherent bias of the imbalanced dataset.
- **Entropy-Focus Weight (w_{entropy}):** This factor directs the model's attention toward samples in ambiguous, high-entropy regions. It is defined as $w_{\text{entropy}}(H_i) = (1 + H_i)^\gamma$, where the hyperparameter $\gamma \geq 0$ controls the intensity of the focus. When $\gamma = 0$, this guidance is inactive. For $\gamma > 0$, the model is more heavily penalized for failing to accurately reconstruct samples located in high-entropy regions, forcing it to learn a more discriminative representation of the decision boundary.

The final reconstruction loss, $\mathcal{L}_{\text{W-Recon}}$, is the mean of these individually weighted losses calculated across all samples in a mini-batch.

KL Divergence Loss ($\mathcal{L}_{\text{KLD}}$): To ensure the latent space remains well-structured, we include the standard KL Divergence Loss. This is the unmodified, conventional regularization term from the CVAE's ELBO. It measures the divergence between the encoder's output distribution and a class-conditional prior and is defined as:

$$\mathcal{L}_{\text{KLD}} = D_{\text{KL}}(q_\phi(z|x,c)||p(z|c)) \tag{5}$$

To counteract the common CVAE training issue of posterior collapse, where the KLD term vanishes and the decoder learns to ignore the latent code z, we enforce a minimum threshold on this loss component during optimization. This ensures that the latent variables continue to encode meaningful information throughout the training process.

3.4.2 Entropy-Guided Sample Generation

After training the LEO-CVAE, we use it to generate synthetic samples for each minority class until it reaches parity with the majority class. Applying the same core principle used in the LEWL, we focus the creation of new samples on high-entropy regions. The procedure for a given minority class c_{min} is as follows:

1. **Calculate Generation Count:** Determine the number of new samples to generate, $N_{gen} = N_{maj} - N_{min}$.
2. **Select Seed Samples:** The seeds for generation are chosen from the original minority class instances in the training set. To prioritize the synthesis of new data in high-entropy regions, we employ a non-uniform selection strategy guided by LES. A sampling probability distribution, P, is established over the minority class samples where, instead of uniform selection, the probability of selecting a given sample x_i is made proportional to its entropy-focus weight, the same transformation of its LES used in the LEWL:

$$P(x_i) \propto (1 + H_i)^\gamma, \quad \text{for all } (x_i, y_i) \text{ where } y_i = c_{\text{min}} \tag{6}$$

   The hyperparameter γ again controls how strongly this selection process favors samples in high-entropy regions. From this entropy-weighted distribution, N_{gen} seed samples are drawn with replacement to initiate the data generation process. This ensures that samples residing in high-entropy neighborhoods are more likely to be chosen as templates for creating new synthetic data.
3. **Generate Synthetic Data:** For each selected seed sample x_{seed}, a new synthetic sample $\hat{x}_{\text{new}}$ is generated. This is achieved by first encoding the seed to its latent distribution, then sampling a new latent vector $z_{\text{new}} \sim q_\phi(z|x_{\text{seed}}, c_{\text{min}})$ using the reparameterization trick. Finally, the resulting vector z_{new} is passed to the decoder to produce the synthetic sample $\hat{x}_{\text{new}}$.

Algorithm 1 LEO-CVAE Framework

Input: Training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, k for k-nearest neighbors, hyperparameters γ, β.
Output: Resampled dataset $\mathcal{D}' = \mathcal{D} \cup \mathcal{D}_{syn}$.

Step 1: Quantifying Sample-Level Uncertainty
for each sample $x_i \in \mathcal{D}$ **do**
  Find k-nearest neighbors $\mathcal{N}_k(x_i)$.
  Calculate local entropy score (LES) $H(x_i)$ using Eq. (2).
end for
Step 2: Training with Entropy-Weighted Loss
Initialize LEO-CVAE model (Encoder$_\phi$, Decoder$_\theta$).
for number of training epochs **do**
  for each mini-batch $\{(x_b, y_b, H_b)\}_{b=1}^{B}$ **do**
    Compute $\mu_b, \log \sigma_b^2 = \text{Encoder}_\phi([x_b, c_{b\_oh}])$.
    Sample $z_b \sim \mathcal{N}(\mu_b, \text{diag}(\sigma_b^2))$.
    Reconstruct $\hat{x}_b = \text{Decoder}_\theta([z_b, c_{b\_oh}])$.
    Calculate loss $\mathcal{L}_{\text{LEWL}}$ using Eq. (3).
    Update ϕ and θ via gradient descent.
  end for
end for
Step 3: Entropy-Guided Generation
$\mathcal{D}_{syn} \leftarrow \emptyset$
Identify minority classes $\mathcal{C}_{min}$ and majority count N_{maj}.
for each class $c_j \in \mathcal{C}_{min}$ **do**
  Let $\mathcal{D}_j$ be the set of samples in class c_j.
  Calculate sampling probabilities $P(x_i)$ for all $x_i \in \mathcal{D}_j$ using Eq. (6).
  $N_{gen} \leftarrow N_{maj} - |\mathcal{D}_j|$.
  Select N_{gen} seed samples $\{x_{seed}\}$ from $\mathcal{D}_j$ with replacement using probabilities P.
  Generate N_{gen} synthetic samples $\{\hat{x}_{new}\}$ from $\{x_{seed}\}$.
  $\mathcal{D}_{syn} \leftarrow \mathcal{D}_{syn} \cup \{(\hat{x}_{new}, c_j)\}$.
end for
return Resampled dataset $\mathcal{D}' = \mathcal{D} \cup \mathcal{D}_{syn}$.

This ensures that the synthetic data is generated around the most informative minority samples located in high-entropy regions, effectively reinforcing the decision boundary in contested regions of the feature space. The complete LEO-CVAE oversampling process is summarized in Algorithm 1.

4 Experimental Setup

Experimental validation was conducted on two challenging clinical genomics datasets: The Cancer Genome Atlas (TCGA) lung cancer and Alzheimer's Disease Neuroimaging Initiative (ADNI). These datasets were specifically selected, as they are characterized by the high dimensionality and complex, nonlinear relationships inherent to genomic data, necessitating a powerful generative model. Moreover, the inherent biological heterogeneity among patients and the gradual progression of disease mean that class boundaries are not sharp, distinct lines. Instead, these factors create a continuum where samples from different classes intermingle, forming complex, overlapping decision boundaries. This results in the exact high-entropy regions that our entropy-guided ('LEO') mechanism is designed to target and resolve, providing a perfect testbed to rigorously evaluate our method's capacity to handle these real-world clinical data challenges.

4.1 Datasets

TCGA Lung Cancer Dataset: Gene expression (RNA-Seq) and corresponding clinical data for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) were obtained from TCGA via the UCSC Xena (Tomczak et al., 2015). A total of 817 patient samples with complete records for both expression profiles and Progression-Free Survival (PFS) were included in this study. From an initial 17,738 gene expression features, the 64 most informative ones were selected using mutual information. The classification task was to predict patient PFS, which was categorized as 'short' (≤ 1 year) or 'long' (> 1 year). The final cohort consisted of 147 'short' PFS samples (the minority class) and 670 'long' PFS samples (the majority class).

ADNI Dataset: Blood-based gene expression data were obtained from the ADNI database (`adni.loni.usc.edu`) for a multiclass classification task. ADNI, launched in 2003 as a public–private partnership led

by Principal Investigator Michael W. Weiner, was designed to evaluate whether imaging, biological markers, and clinical/neuropsychological assessments could be combined to track the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) (Petersen et al., 2010). Data from 744 participants were included in this analysis. Using a mutual information–based feature selection strategy, the initial 49,386 probe sets per sample were reduced to the 64 most informative features. The objective was to classify participants into Cognitively Normal (CN), MCI, or AD, with the dataset distributed as follows: 246 CN, 382 MCI, and 116 AD .

4.2 Comparison Methods

We benchmarked our proposed LEO-CVAE against a diverse suite of seven comparison methods. The specific configurations for each method are detailed in Supplement s.1 to ensure full reproducibility.

1. **No Oversampling**: The baseline performance of the classifier on the original, imbalanced data.
2. **Random Oversampling (Japkowicz, 2000)**: The simplest approach, which randomly duplicates samples from the minority class.
3. **SMOTE (Chawla et al., 2002)**: The classic Synthetic Minority Over-sampling Technique.
4. **Borderline-SMOTE (Han et al., 2005)**: An advanced SMOTE variant that focuses on the class boundary.
5. **ADASYN (Haibo et al., 2008)**: An adaptive approach that generates more data for harder-to-learn minority samples.
6. **Standard CVAE (Sohn et al., 2015)**: A baseline CVAE model with an identical architecture to LEO-CVAE but using a standard, unweighted reconstruction loss.
7. **CVAE with Focal Loss (Lin et al., 2017)**: A CVAE trained with a focal loss-inspired reconstruction objective to focus on samples with high reconstruction error.

4.3 Evaluation Protocol and Implementation Details

All experiments followed a consistent and rigorous evaluation protocol to ensure a fair comparison.

- **Experimental Design:** We employed a 5-fold stratified cross-validation strategy. In each fold, oversampling methods were applied only to the training set. The validation set remained untouched to provide an unbiased estimate of performance. For reproducibility, all experiments were conducted with a fixed random seed of 42.
- **Oversampling Models and Baselines:** To ensure a fair comparison, baseline models were configured with consistent parameters. The k-NN-based methods (i.e., SMOTE, Borderline-SMOTE, and ADASYN) used the same task-specific neighbor parameters. Similarly, our proposed LEO-CVAE and all other CVAE-based baselines shared an identical network architecture and core training parameters. Full details on the architectures and hyperparameters for all models are provided in Supplements s.1 and s.2.
- **Downstream Classifier:** We used a Multi-Layer Perceptron (MLP) with a consistent architecture across all experiments to evaluate the quality of the data generated by each oversampling method. The MLP was trained using the Adam optimizer (learning rate: 1×10^{-4}, weight decay: 1×10^{-3}) with an early stopping patience of 20 epochs. Performance was monitored using the validation Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for binary tasks and the micro-averaged AUC-ROC for multiclass tasks. A detailed description of the classifier's architecture is available in Supplement s.3.
- **Evaluation Metrics:** Model performance was evaluated using distinct sets of metrics for binary and multiclass classification tasks. For binary tasks, performance was assessed using the AUC-ROC, the Area Under the Precision-Recall Curve (AUPRC), and the F1-Score. For multiclass tasks, these same metrics were aggregated using both macro and micro averaging (macro/micro AUC-ROC, macro/micro AUPRC, and macro/micro F1-Score) to provide a more nuanced view of the model's performance across all classes.

5 Results

5.1 Performance on TCGA Lung Cancer Dataset (Binary Classification)

The training dynamics of LEO-CVAE on the binary TCGA lung cancer dataset are illustrated for a representative fold in Figure 1. The model exhibits stable convergence with no evidence of significant overfitting, as the validation and training loss curves closely track one another. The healthy stabilization of the KL Divergence and a final reconstruction correlation approaching 1.0 collectively affirm the successful and well-regularized training of the generative model.

The comparative classification results are presented in Table 1. LEO-CVAE achieved the highest AUC-ROC (0.661 ± 0.030) and AUPRC (0.889 ± 0.021), indicating improved ranking ability and better precision–recall trade-off for the minority class compared with all other

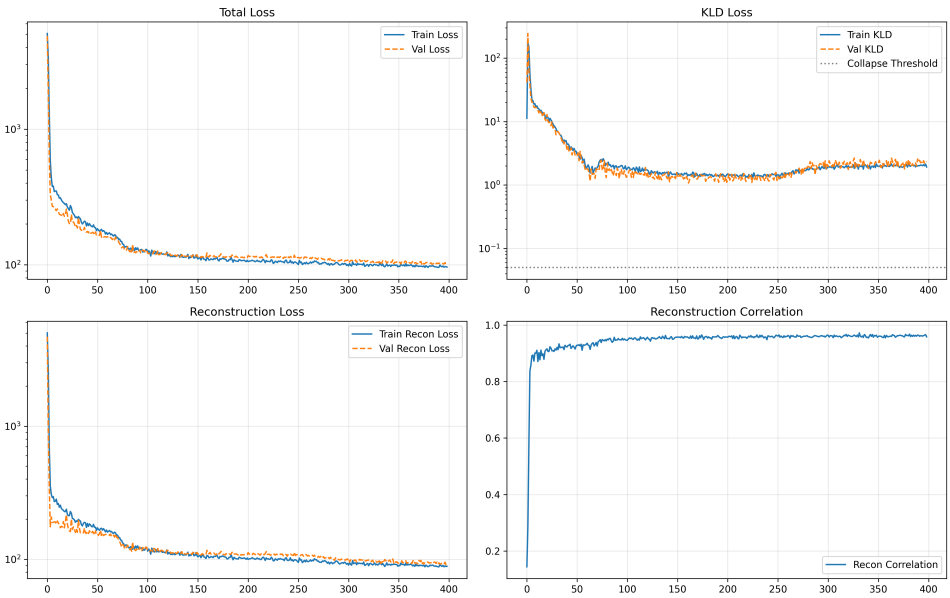

Figure 1: Training history of the LEO-CVAE model for a representative fold on the TCGA lung cancer dataset.

Table 1: Performance on the TCGA Lung Cancer Dataset

Model	**AUC-ROC**	**AUPRC**	**F1-Score**
No Oversampling	0.620 ± 0.040	0.868 ± 0.028	$\mathbf{0.903} \pm 0.006$
Random Oversampling	0.613 ± 0.065	0.861 ± 0.040	0.767 ± 0.021
SMOTE	0.611 ± 0.023	0.868 ± 0.025	0.807 ± 0.023
Borderline-SMOTE	0.630 ± 0.044	0.874 ± 0.022	0.797 ± 0.026
ADASYN	0.617 ± 0.052	0.869 ± 0.031	0.786 ± 0.026
Standard CVAE	0.614 ± 0.074	0.869 ± 0.040	0.883 ± 0.015
CVAE + Focal Loss	0.645 ± 0.043	0.885 ± 0.016	0.871 ± 0.028
LEO-CVAE	$\mathbf{0.661} \pm 0.030$	$\mathbf{0.889} \pm 0.021$	0.881 ± 0.012

Note: Values are presented as mean $\pm$ standard deviation over 5 folds. The best result in each column is highlighted in bold.

methods. The No Oversampling baseline obtained the highest reported F1-score (0.903 ± 0.006).

Non-generative oversampling strategies, such as Borderline-SMOTE and ADASYN, showed modest increases in AUPRC compared to the baseline, but these were often accompanied by minimal or no gains in AUC-ROC, reflecting a common trade-off in which minority-class precision improves at the expense of overall discriminative ability. In contrast, CVAE-based approaches generally outperformed non-generative methods, with LEO-CVAE standing out as the only method to deliver notable, simultaneous gains in both AUC-ROC and AUPRC over the baseline.

5.2 Performance on ADNI Dataset (Multiclass Classification)

The pattern of stable convergence was replicated on the multiclass ADNI dataset (Figure 2), establishing the reliability of the model's training process prior to performance evaluation.

For the multiclass ADNI dataset (Table 2), the comparative results are more nuanced but again highlight the strengths of LEO-CVAE in achieving balanced performance across classes. The proposed model obtained the highest macro-averaged AUC-ROC (0.587 ± 0.025), macro-AUPRC (0.412 ± 0.015), and micro-AUPRC (0.500 ± 0.014), indicating improved per-class discrimination and a stronger precision–recall trade-off. The No Oversampling baseline performed strongly on micro-averaged metrics, such as micro-AUC-ROC (0.690 ± 0.017) and micro-F1 (0.500 ± 0.027), which weight performance by class size and are often dominated by populous classes. In contrast, LEO-CVAE's superior macro-averaged metrics (unweighted averages that treat all classes equally) highlight its ability to distribute performance gains more evenly, benefiting minority classes without sacrificing overall discrimination.

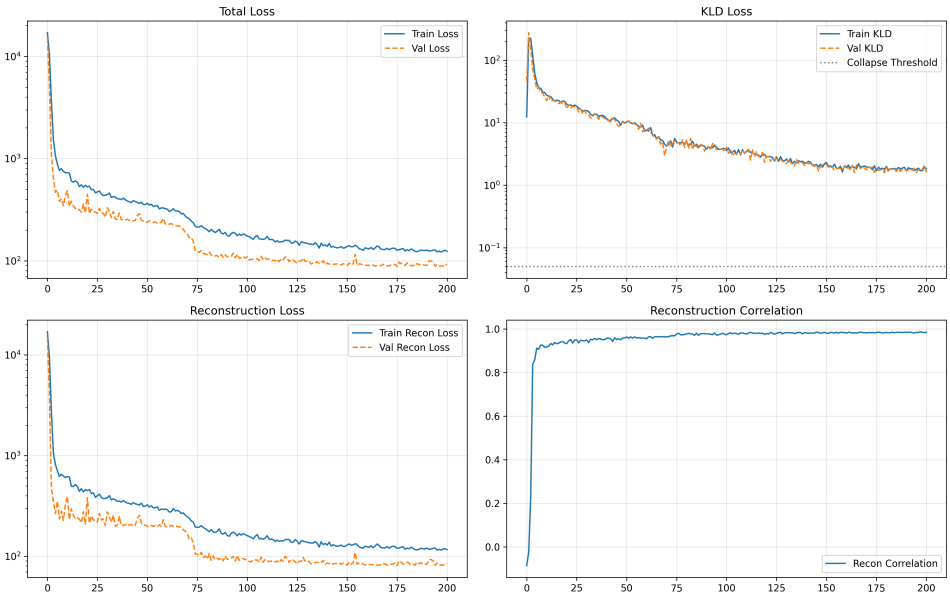

Figure 2: Training history of the LEO-CVAE model for a representative fold on the ADNI dataset.

Table 2: Performance on the ADNI Dataset

	AUC-ROC		AUPRC		F1-Score	
Model	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**
No Oversampling	0.570 ± 0.033	**0.690** ± 0.017	0.398 ± 0.022	0.492 ± 0.021	0.311 ± 0.026	**0.500** ± 0.027
Random Oversampling	0.564 ± 0.029	0.588 ± 0.020	0.394 ± 0.027	0.393 ± 0.018	**0.381** ± 0.012	0.402 ± 0.020
SMOTE	0.561 ± 0.041	0.606 ± 0.030	0.403 ± 0.035	0.416 ± 0.035	0.377 ± 0.043	0.410 ± 0.040
Borderline-SMOTE	0.558 ± 0.049	0.609 ± 0.030	0.392 ± 0.040	0.409 ± 0.027	0.373 ± 0.037	0.417 ± 0.028
ADASYN	0.556 ± 0.037	0.612 ± 0.032	0.383 ± 0.034	0.420 ± 0.026	0.367 ± 0.042	0.431 ± 0.021
Standard CVAE	0.563 ± 0.032	0.665 ± 0.027	0.386 ± 0.025	0.469 ± 0.038	0.359 ± 0.024	0.474 ± 0.033
CVAE + Focal Loss	0.562 ± 0.027	0.667 ± 0.022	0.395 ± 0.029	0.485 ± 0.039	0.347 ± 0.029	0.469 ± 0.030
LEO-CVAE	**0.587** ± 0.025	0.683 ± 0.013	**0.412** ± 0.015	**0.500** ± 0.014	0.375 ± 0.043	0.484 ± 0.020

Note: Values are presented as mean $\pm$ standard deviation over 5 folds. The best result in each column is highlighted in bold.

CVAE-based methods generally delivered stronger micro-metrics, likely due to generating a more diverse set of synthetic samples, but LEO-CVAE distinguished itself with a balanced macro/micro profile.

6 Ablation Study

To rigorously evaluate the individual contributions of the core components within our proposed LEO-CVAE framework, a comprehensive ablation study was conducted. One or more of the model's three main mechanisms were systematically disabled:

1. Entropy-weighted loss
2. Entropy-guided sampling
3. Inverse frequency class weighting

The tested model variants are defined in Table 3. All experiments followed the evaluation protocol outlined in Section 4.

6.1 Results and Analysis

The results for the binary TCGA lung cancer dataset are reported in Table 4, and for the multiclass ADNI dataset in Table 5. Across both datasets, the full LEO-CVAE consistently achieved balanced, high performance, demonstrating the value of integrating all three mechanisms.

On the TCGA lung cancer dataset (Table 4), the full LEO-CVAE achieved the highest AUC-ROC (0.661 ± 0.030) and AUPRC (0.889 ± 0.021). Removing entropy-weighted loss (Ablation 1) or entropy-guided sampling (Ablation 2) reduced both metrics, confirming the utility of each entropy-based component. While the Standard CVAE achieved a high F1-score, this was accompanied by a comparatively low AUC-ROC.

On the ADNI dataset (Table 5), the full LEO-CVAE achieved the highest macro-averaged AUC (0.587 ± 0.025), indicating balanced performance across all classes. Ablation 2 (no entropy-guided sampling) achieved the highest micro-averaged AUC ($0.686 \pm$

Table 3: Ablation Study Experimental Design

Model Variant	Entropy-Weighted Loss	Entropy-Guided Sampling	Class Weighting	Purpose
Full LEO-CVAE	✓	✓	✓	The complete proposed model.
LEO-CVAE (w/o Ent-Weighted Loss)	✗	✓	✓	Isolates the effect of the entropy-weighted loss.
LEO-CVAE (w/o Ent-Guided Sampling)	✓	✗	✓	Isolates the effect of entropy-guided sampling.
LEO-CVAE (w/o Class Weights)	✓	✓	✗	Isolates the effect of class weighting.
CVAE + Class Weights	✗	✗	✓	A simpler, non-entropy based model.
Standard CVAE	✗	✗	✗	The foundational generative baseline.

Table 4: Ablation Study Results on the TCGA Lung Cancer Dataset

Model Variant	AUC-ROC	AUPRC	F1-Score
LEO-CVAE (Full Model)	**0.661** $\pm$ 0.030	**0.889** $\pm$ 0.021	0.881 $\pm$ 0.012
LEO-CVAE (w/o Entropy-Weighted Loss)	0.600 $\pm$ 0.054	0.863 $\pm$ 0.036	0.861 $\pm$ 0.025
LEO-CVAE (w/o Entropy-Guided Sampling)	0.631 $\pm$ 0.032	0.875 $\pm$ 0.025	0.861 $\pm$ 0.026
LEO-CVAE (w/o Class Weights)	0.616 $\pm$ 0.062	0.873 $\pm$ 0.023	0.865 $\pm$ 0.030
CVAE + Class Weights	0.617 $\pm$ 0.026	0.865 $\pm$ 0.025	0.868 $\pm$ 0.027
Standard CVAE	0.614 $\pm$ 0.074	0.869 $\pm$ 0.040	**0.883** $\pm$ 0.015

Note: Values are presented as mean $\pm$ standard deviation over 5 folds. The best result in each column is highlighted in bold.

0.019) and macro F1-score (0.379 ± 0.013), though differences with the full model were small and within standard deviations. This highlights the LEWL as an especially impactful component of the LEO-CVAE.

In summary, for the binary TCGA lung cancer dataset, all three components contribute to robust gains in AUC and AUPRC, with the entropy-based mechanisms having the largest impact. In the multiclass ADNI setting, the LEWL emerges as particularly effective, while removing entropy-guided sampling (Ablation 2) does not substantially harm performance and, in some cases, even slightly improves certain metrics. This validates the overall framework design by underscoring the power of the core entropy-weighted loss, while also highlighting that the optimal configuration of complementary components, such as class weighting and entropy-guided sampling, can be application-dependent.

7 Conclusion

In this work, we addressed the critical challenge of class imbalance in complex tabular data, with a focus on its application to clinical genomics data. Traditional oversampling methods, which rely on local, linear interpolation, are often ill-suited for such data. Their core assumption, that a straight line between two minority samples represents plausible data, frequently fails within the complex, non-linear manifolds characteristic of the biological processes underlying genomics data. While deep generative models like the Conditional Variational Autoencoder (CVAE) can capture these complex global distributions, a key limitation of standard CVAEs is that they learn the global distribution of a class by implicitly treating all training samples as equally informative. This uniform approach overlooks a critical insight: not all samples are equally valuable for refining a classifier's decision boundary. Instances located deep within a class's feature space are less informative than those situated in regions of high class overlap, where the boundary is most ambiguous. These "hard-to-learn" or "borderline" samples are strategically crucial, as they provide the most challenging examples for a classifier. While heuristic methods like Borderline-SMOTE and ADASYN pioneered the strategy of focusing on such instances, a significant gap remains in integrating a formal principle of uncertainty-awareness into a powerful distributional learning framework.

To bridge this gap, we introduce the Local Entropy-Guided Oversampling with a CVAE (LEO-CVAE), a novel generative oversampling framework that quantifies sample-level uncertainty using local Shannon entropy. By synergistically integrating this signal through a Local Entropy-Weighted Loss (LEWL) and an entropy-guided sampling strategy, LEO-CVAE focuses both its learning and generative processes on the most informative, class-overlapping regions of the feature space. Our empirical evaluation demonstrated that LEO-CVAE achieves superior and more balanced performance on challenging genomics datasets, delivering notable gains in AUC-ROC and AUPRC. The ablation study further revealed that the LEWL was the most impactful component, underscoring the benefit of compelling the model to learn a more robust representation of the contested decision boundary.

7.1 Limitations and Future Directions

This work lays the foundation for a new uncertainty-aware generative framework for imbalanced learning. We outline three promising avenues for future research based on the principles established in this study.

Table 5: Ablation Study Results on the ADNI Dataset

	AUC-ROC		AUPRC		F1-Score	
Model Variant	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**
LEO-CVAE (Full Model)	**0.587** $\pm$ 0.025	0.683 $\pm$ 0.013	**0.412** $\pm$ 0.015	**0.500** $\pm$ 0.014	0.375 $\pm$ 0.043	0.484 $\pm$ 0.020
LEO-CVAE (w/o Entropy-Weighted Loss)	0.544 $\pm$ 0.050	0.655 $\pm$ 0.031	0.380 $\pm$ 0.033	0.459 $\pm$ 0.043	0.322 $\pm$ 0.037	0.446 $\pm$ 0.031
LEO-CVAE (w/o Entropy-Guided Sampling)	0.579 $\pm$ 0.040	**0.686** $\pm$ 0.019	**0.412** $\pm$ 0.016	0.491 $\pm$ 0.023	**0.379** $\pm$ 0.013	**0.495** $\pm$ 0.025
LEO-CVAE (w/o Class Weights)	0.533 $\pm$ 0.019	0.646 $\pm$ 0.013	0.377 $\pm$ 0.021	0.445 $\pm$ 0.018	0.350 $\pm$ 0.034	0.458 $\pm$ 0.030
CVAE + Class Weights	0.566 $\pm$ 0.040	0.667 $\pm$ 0.031	0.387 $\pm$ 0.036	0.478 $\pm$ 0.040	0.322 $\pm$ 0.034	0.458 $\pm$ 0.044
Standard CVAE	0.563 $\pm$ 0.032	0.665 $\pm$ 0.027	0.386 $\pm$ 0.025	0.469 $\pm$ 0.038	0.359 $\pm$ 0.024	0.474 $\pm$ 0.033

Note: Values are presented as mean $\pm$ standard deviation over 5 folds. The best result in each column is highlighted in bold.

Applicability and Broader Evaluation

Our study's focus on clinical genomics data was deliberate, as our framework is specifically designed for data with complex, non-linear characteristics. The principles of LEO-CVAE may also be well-suited for other high-dimensional omics data, such as proteomics and metabolomics, which share similar characteristics of complex, non-linear relationships. For simpler tabular data lacking these complex characteristics, traditional methods may still perform adequately. Future work should therefore extend this evaluation to a wider variety of tabular datasets, accompanied by formal statistical significance testing, to better characterize the regimes where different oversampling strategies are most effective.

Extensions to Alternative Generative Models

The core concepts of entropy-guided learning and generation could be generalized to other data modalities, such as images, or adapted for other generative architectures like diffusion models and GANs. One particularly promising direction is adapting the LEO framework to diffusion-based synthesis. The denoising loss could be reweighted by the Local Entropy Score (LES) to counter majority-class gradient domination, while LES could also modulate class-conditional guidance during reverse diffusion, steering generation toward higher-quality minority samples. This integration of uncertainty-guided sampling with state-of-the-art synthesis offers a compelling path forward. Research also suggests that hybrid strategies combining traditional and generative models may yield further improvements (Wang et al., 2025a).

Advancing the Uncertainty Metric

The mechanism for calculating local entropy warrants deeper investigation. Our choice of k-NN with Euclidean distance was a direct extension of the principles in SMOTE (Chawla et al., 2002), providing a robust baseline. However, the effectiveness of this uncertainty metric could be enhanced by exploring alternative distance metrics better suited for high-dimensional data, such as Manhattan distance, cosine similarity, and manifold-based distance, or by moving beyond distance-based methods with techniques like Kernel Density Estimation (KDE). A particularly novel direction would be to leverage the probability distributions from a pre-trained model to guide the entropy calculation, potentially creating a more nuanced, model-aware measure of uncertainty.

Acknowledgments

Gene expression (RNA-Seq) and corresponding clinical data for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) were obtained from TCGA via the UCSC Xena.

Data collection and sharing for this project was also funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability Statement

The TCGA-LUAD and TCGA-LUSC datasets are publicly available via the UCSC Xena platform (`https://xenabrowser.net/`). The ADNI dataset is available to qualified researchers upon application through the Laboratory of Neuro Imaging (LONI) website (`https://adni.loni.usc.edu/`). Access to both datasets is subject to the data use agreements of their respective repositories. The source code for the LEO-CVAE framework is available at `https://github.com/Amirhossein-Zare/LEO-CVAE`.

References

Qingzhong Ai, Pengyun Wang, Lirong He, Liangjian Wen, Lujia Pan, and Zenglin Xu. *Generative Oversampling for Imbalanced Data via Majority-Guided VAE*. 2023.

Nur Azhar, M. S. Mohd Pozi, Aniza Mohamed Din, and Adam Jatowt. An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35:6651–6672, 2023. doi: 10.1109/TKDE.2022.3179381.

Gustavo Batista, Ronaldo Prati, and Maria-Carolina Monard. A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. *SIGKDD Explorations*, 6:20–29, 2004. doi: 10.1145/1007730.1007735.

Rok Blagus and Lara Lusa. Evaluation of smote for high-dimensional class-imbalanced microarray data. volume 2, 12 2012. doi: 10.1109/ICMLA.2012.183.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, October 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.07.011.

Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. SMOTE: Synthetic Minority Oversampling Technique. *J. Artif. Intell. Res. (JAIR)*, 16: 321–357, 2002. doi: 10.1613/jair.953.

Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and C. Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-024-10759-6.

W. Dai, K. Ng, K. Severson, W. Huang, F. Anderson, and C. Stultz. Generative Oversampling with a Contrastive Variational Autoencoder. pages 101–109, November 2019. ISBN 2374-8486. doi: 10.1109/ICDM.2019.00020.

Swagatam Das, Shounak Datta, and Bidyut Chaudhuri. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81, 03 2018. doi: 10.1016/j.patcog.2018.03.008.

Georgios Douzas and Fernando Bação. Geometric smote a geometrically enhanced drop-in replacement for smote. *Information Sciences*, 501, 06 2019. doi: 10.1016/j.ins.2019.06.007.

Georgios Douzas, Fernando Bação, and Felix Last. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, 465, 06 2018. doi: 10.1016/j.ins.2018.06.056.

Val Andrei Fajardo, David Findlay, Charu Jaiswal, Xinshang Yin, Roshanak Houmanfar, Honglei Xie, Jiaxi Liang, Xichen She, and D. B. Emerson. On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169, 2021. ISSN 09574174. doi: 10.1016/j.eswa.2020.114463.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh Chawla. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 04 2018. doi: 10.1613/jair.1.11192.

Xinyi Gao, Dongting Xie, Yihang Zhang, Zhengren Wang, Conghui He, Hongzhi Yin, and Wentao Zhang. A comprehensive survey on imbalanced data learning. 02 2025. doi: 10.48550/arXiv.2502.08960.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 3, 2014. doi: 10.1145/3422622.

Ting Guo, Xingquan Zhu, Yang Wang, and Fang Chen. *Discriminative Sample Generation for Deep Imbalanced Learning*. 2019.

He Haibo, Bai Yang, E. A. Garcia, and Li Shutao. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. pages 1322–1328, June 2008. ISBN 2161-4407. doi: 10.1109/IJCNN.2008.4633969.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. pages 878–887.

Springer Berlin Heidelberg, 2005. ISBN 978-3-540-31902-3.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 06 2020.

Sungchul Hong, Seunghwan An, and Jong-June Jeon. Improving smote via fusing conditional vae for data-adaptive noise filtering. 05 2024. doi: 10.48550/arXiv.2405.19757.

Nathalie Japkowicz. The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence ICAI*, 2000.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6:429–449, 11 2002. doi: 10.3233/IDA-2002-6504.

Firuz Kamalov, Salah Choutri, and Amir Atiya. Analytical formulation of synthetic minority oversampling technique (smote) for imbalanced learning. *Gulf Journal of Mathematics*, 19:400–415, 01 2025. doi: 10.56947/gjom.v19i1.2639.

Diederik Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 12 2013.

Intouch Kunakorntum, Woranich Hinthong, and Phond Phunchongharn. A synthetic minority based on probabilistic distribution (symprod) oversampling for imbalanced datasets. *IEEE Access*, PP:1–1, 06 2020. doi: 10.1109/ACCESS.2020.3003346.

Ying Li, Yali Yang, Peihua Song, Lian Duan, and Rui Ren. An improved smote algorithm for enhanced imbalanced data classification by expanding sample generation space. *Scientific Reports*, 15, 07 2025a. doi: 10.1038/s41598-025-09506-w.

Zhong Li, Qi Huang, Lincen Yang, Jiayang Shi, Zhao Yang, Niki van Stein, Thomas Bäck, and Matthijs van Leeuwen. Diffusion models for tabular data: Challenges, current progress, and future directions, 2025b. URL https://arxiv.org/abs/2502.17119.

T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. pages 2999–3007, October 2017. ISBN 2380-7504. doi: 10.1109/ICCV.2017.324.

Alexander Liu, Joydeep Ghosh, and Cheryl Martin. Generative oversampling for mining imbalanced datasets. pages 66–72, 01 2007.

Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Said Hacid, and Hassan Zeineddine. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access*, PP:1–1, 2019. doi: 10.1109/ACCESS.2019.2927266.

Ruchika Malhotra and Shine Kamal. An Empirical Study to Investigate Oversampling Methods for Improving Software Defect Prediction using Imbalanced Data. *Neurocomputing*, 343, 2019. doi: 10.1016/j.neucom.2018.04.090.

Ronald Petersen, P.S. Aisen, L.A. Beckett, Michael Donohue, A.C. Gamst, D.J. Harvey, Clifford Jack, W.J. Jagust, Leslie Shaw, A.W. Toga, J.Q. Trojanowski, and Michael Weiner. Alzheimer's disease neuroimaging initiative (adni): Clinical characterization. *Neurology*, 74:201–9, 01 2010. doi: 10.1212/WNL.0b013e3181cb3e25.

V. Sampath, I. Maurtua, J. J. Aguilar Martín, and A. Gutierrez. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *J Big Data*, 8(1):27, 2021. ISSN 2196-1115 (Print) 2196-1115. doi: 10.1186/s40537-021-00414-0.

Dingyuan Shi, Lulu Zhang, Yongxin Tong, and Ke Xu. Understanding and mitigating the high computational cost in path data diffusion, 2025. URL https://arxiv.org/abs/2502.00725.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

Xi Tang, Wenhai Li, Weichao Sun, Tianzhu Wen, and Kai Guo. *Unbalanced Data Oversampling Method Based on Improved VAE-CGAN*. 2025.

Katarzyna Tomczak, Patrycja Czerwinska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): An immeasurable source of knowledge. *Contemp Oncol (Pozn)*, 19:A68–A77, 01 2015.

Alex X. Wang, Minh Quang Le, Huu-Thanh Duong, Bay Nguyen Van, and Binh P. Nguyen. CTVAE: Contrastive Tabular Variational Autoencoder for imbalance data. *Knowledge and Information Systems*, 67 (6):5335–5354, 2025a. ISSN 0219-1377 0219-3116. doi: 10.1007/s10115-025-02377-7.

Alex Xing Wang, Viet-Tuan Le, Hau Nguyen Trung, and Binh Nguyen. Addressing imbalance in health data: Synthetic minority oversampling using deep learning. *Computers in biology and medicine*, 188: 109830, 02 2025b. doi: 10.1016/j.compbiomed.2025.109830.

Kaixiang Yang, Zhiwen Yu, Wuxing Chen, Zefeng Liang, and C. Chen. Solving the imbalanced problem

by metric learning and oversampling. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–14, 12 2024. doi: 10.1109/TKDE.2024.3419834.

s Model Architectures and Hyperparameters

This supplement provides detailed information on the network architectures and hyperparameter settings used in our experiments to ensure full reproducibility.

s.1 Baseline Model Configurations

Non-generative baselines were implemented using the `imbalanced-learn` Python library. All CVAE-based baselines use the identical network architecture as LEO-CVAE (Supplement s.2). Specific configurations are in Table s1.

Table s1: Baseline Model Hyperparameters

Model	Library/Base	Key Parameter	Value
SMOTE	`imbalanced-learn`	'k_neighbors'	7 (Binary) / 15 (Multiclass)
Borderline-SMOTE	`imbalanced-learn`	'kind'	'borderline-1'
		'k_neighbors'	7 (Binary) / 15 (Multiclass)
		'm_neighbors'	7 (Binary) / 15 (Multiclass)
ADASYN	`imbalanced-learn`	'n_neighbors'	7 (Binary) / 15 (Multiclass)
Standard CVAE	Our implementation	KLD Weight (β)	1.0
		Minimum KLD	0.1
CVAE with Focal Loss	Our implementation	KLD Weight (β)	1.0
		Minimum KLD	0.1
		Focusing Param. (γ)	1.0

s.2 LEO-CVAE Architecture and Hyperparameters

Our proposed LEO-CVAE model and all CVAE-based baselines share the identical network architecture detailed below, with an input dimension of 64.

- **Encoder Architecture**:
  1. Input: Concatenation of feature vector ($D_{in} = 64$) and one-hot class label (C), size = $64 + C$.
  2. 'Linear'($64 + C \to 64$) $\to$ 'LeakyReLU'(0.2) $\to$ 'Dropout'($p = 0.1$).
  3. 'Linear'($64 \to 32$) $\to$ 'LeakyReLU'(0.2) $\to$ 'Dropout'($p = 0.1$).
  4. Two parallel 'Linear' output heads from the 32-neuron layer:
     - 'Linear'($32 \to 16$) for the latent mean (μ).
     - 'Linear'($32 \to 16$) for the latent log-variance ($\log \sigma^2$).
- **Decoder Architecture**:
  1. Input: Concatenation of latent vector ($D_z = 16$) and one-hot class label (C), size = $16 + C$.
  2. 'Linear'($16 + C \to 32$) $\to$ 'LeakyReLU'(0.2) $\to$ 'Dropout'($p = 0.1$).
  3. 'Linear'($32 \to 64$) $\to$ 'LeakyReLU'(0.2) $\to$ 'Dropout'($p = 0.1$).
  4. 'Linear' output layer mapping from $64 \to 64$ to reconstruct the original feature vector.

The hyperparameters for training LEO-CVAE are detailed in Table s2.

Table s2: LEO-CVAE Training Hyperparameters

Hyperparameter	Symbol	Value	Description
Optimizer	-	Adam	-
Learning Rate	-	1×10^{-3}	-
Weight Decay	-	1×10^{-5}	-
Batch Size	-	32	-
Max Epochs	-	500	-
Early Stopping Patience	-	25	-
Gradient Clip Norm	-	1.0	Maximum norm for gradient clipping.
Latent Dimension	D_z	16	Dimensionality of the latent space.
Entropy k-nearest neighbors	k	7 (Binary) / 15 (Multiclass)	Neighbors for local entropy calculation.
Entropy Focus	γ	0.5 (Binary) / 2.5 (Multiclass)	Weighting for high-entropy samples.
KLD Weight	β	4.0	Weight of the KL divergence term.
Minimum KLD	-	0.1	Floor to prevent posterior collapse.

s.3 MLP Classifier Architecture

The Multi-Layer Perceptron (MLP) used as the downstream classifier has an input dimension of 64 and the following single-hidden-layer architecture:

- **Hidden Layer:** 'Linear' layer $(64 \to 32) \to$ 'BatchNorm1d' $\to$ 'ReLU' $\to$ 'Dropout' ($p = 0.5$).
- **Output Layer:** 'Linear' layer mapping from $32 \to 1$ (Binary) or $32 \to 3$ (Multiclass).

The MLP was trained using the hyperparameters listed in Table s3.

Table s3: MLP Classifier Training Hyperparameters

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-3}
Batch Size	32
Max Epochs	200
LR Scheduler	ReduceLROnPlateau
LR Scheduler Patience	5
LR Scheduler Factor	0.7
Early Stopping Patience	20
Early Stopping Metric	AUC (Binary) / AUC Micro (Multiclass)
Gradient Clip Norm	0.5
Label Smoothing	0.1