This version of the article has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s10489-026-07532-3. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

Balanced Synthetic Data Generation with Variational AutoEncoders: A Trade-Off Analysis of Balance, Fairness, Utility and Privacy

Marco Rondina (0009-0008-8819-3623)[1*], Antonio Vetrò (0000-0003-2027-3308)[1], Luca Gilli (0000-0001-7334-5227)[2], Shalini Kurapati (0000-0003-1492-4100)[2], Juan Carlos De Martin (0000-0002-7867-1926)[1]

[1*]Politecnico di Torino, Turin, 10129, TO, Italy.
[2]ClearboxAI, Turin, 10129, TO, Italy.

*Corresponding author(s). E-mail(s): marco.rondina@polito.it;
Contributing authors: antonio.vetro@polito.it; luca@clearbox.ai;
shalini@clearbox.ai; juancarlos.demartin@polito.it;

Abstract

The growing adoption of ML/AI systems in high-stakes contexts has raised concerns about the quality, fairness and privacy of training data. Although synthetic data offers a promising way to address these concerns, the imbalance of sensitive categorical features still represents a source of bias. This paper presents two alternative approaches for a Variational AutoEncoder (VAE) training loss function that penalise deviations from a balanced distribution of sensitive feature classes. These are an additive balance component ($\mathcal{BL}_{term}$) and a Simpson-index-based reweighting component ($\mathcal{BL}_{weight}$). In comparison with other approaches in literature, which alter the architecture of the VAE, our method is a lightweight and tunable mechanism to steer the generation towards balance. An empirical evaluation was conducted on 1,260 synthetic datasets generated from three widely used benchmark datasets (Adult, South German Credit, and COMPAS), across two batch-epoch configurations, seven loss function configurations and thirty train/evaluation splits. The results show that $\mathcal{BL}_{term}$ reduces sensitive feature imbalance and improves statistical parity, while preserving utility and privacy at levels comparable to the baseline. The $\mathcal{BL}_{weight}$ formulation, although capable of reducing imbalance, exhibits a weaker impact on fairness and may degrade

utility and privacy at high magnitudes. The proposed contribution offers practitioners a simple, tunable mechanism for improving the balance and fairness of synthetic tabular data generation without compromising utility or privacy.

Keywords: synthetic data, fairness, bias, privacy, data imbalance

1 Introduction

Applications based on statistical algorithms are becoming dominant in the landscape of the digital revolution. Machine learning (ML) and artificial intelligence (AI) are increasingly being used to analyse data, make predictions, classify records, generate text and images, and more. The output of such systems is increasingly being used to make decisions that affect humans, posing significant risks and challenges.

The quality of the results depends, by definition, on the utility of training data [1]. The principle of transparency, which is required by norms [2] and good practices [3], is limited by the need to respect the right to privacy [4, 5]. In the case of personal data, fair treatment of all subpopulations is an important objective in order to avoid discrimination [6, 7].

Synthetic data is emerging as a potential solution to preserve utility and privacy while improving fairness [8–10]. Synthetic data has been used to generate very different types of data, from computer vision [11] to text corpus [12]. In this research project, our focus is on tabular synthetic data generation. While there are various techniques available for the generation of synthetic tabular data, we are centring our attention on Variational AutoEncoders (VAE) [13]. Unlike Generative Adversarial Networks (GANs) [14], where the generator maps random noise to data without an explicit representation of the input, VAEs encode the data into a structured latent space that captures the joint distribution of all features; this interconnected representation makes it natural to guide the generation process by acting directly on the loss function, which is the approach we adopt in this work. The objective of this study is to enhance the generation of data and to optimize the balance of selected categorical sensitive features, while assessing the impacts on utility and privacy. To this end, we propose two different formulations of the loss function terms: a summed independent term and a multiplicative term. Both terms are based on the distance between the actual class distribution and the ideal class distribution (perfectly balanced across all classes, the uniform distribution). These terms are designed to guide data generation in a tunable way and represent a useful tool for managing imbalance in sensitive features: as also reported in Section 2, this is also a more lightweight mechanism in comparison with other approaches in literature, which alter the architecture of the VAE or its latent space structure. To evaluate the quality of the proposed loss function terms, we performed an empirical evaluation on three source datasets with different train-test splits, for a total of 1,260 synthetic datasets generated from the three sources. However, modifying the loss function to steer the generation towards balance may introduce unintended trade-offs. For example, penalising deviations from uniformity could distort the joint distribution of the data, thereby reducing the utility of

the synthetic dataset for downstream tasks. Alternatively, it could push the model to generate atypical records that resemble real training samples, thereby degrading privacy. A balanced synthetic dataset that sacrifices utility or privacy would be of limited practical value; therefore, we assess all four dimensions — balance, fairness, utility, and privacy — to provide a comprehensive evaluation of the proposed modifications.

In Section 2 we describe the background and related work. Section 3 details the starting model and the loss function modifications. Section 4 is dedicated to the research question. Section 5 describes the methodology used for the empirical evaluation. In Sections 6 and 7 we present the obtained results and discuss the relative implications. Finally, Section 8 outlines the limitations of our findings, Section 9 presents possible directions for future work and in Section 10, we present our conclusions.

2 Background and Related Work

In this section, we provide some background on synthetic data generation and Variational AutoEncoders (VAEs) and we present some related work in the area of synthetic data generation to solve imbalance problems.

2.1 Synthetic data and Variational AutoEncoders

In this work, synthetic data refers to data produced by a generative model that takes the original dataset as input. The quality of a synthetic dataset is usually determined by how well it preserves the statistical properties of the original dataset while presenting different rows in order to protect the privacy of individuals. There are several methods available today for generating synthetic data, each with its own peculiarities and limitations [15, 16]. Variational AutoEncoder by Kingma and Welling [13] is a neural network architecture based on a probabilistic latent space. The latent space is typically characterised by a Gaussian distribution and by a lower dimensionality than the input data. Dimensionality reduction forms the basis of preserving privacy by enabling the network to discard certain details about the input data while retaining its general statistical properties. This network is based on the pairing of two different networks: the encoder and the decoder. The encoder is dedicated to translating the input into the latent space. The decoder generates the output by sampling from the latent space. The training objective is to maximise the Evidence Lower BOund (ELBO), which decomposes into two terms as shown in (1): a reconstruction term, measuring how faithfully the decoder reproduces the input, and a Kullback–Leibler (KL) divergence term, which regularises the latent space by keeping the learned posterior $q_\phi(\mathbf{z}|\mathbf{x})$ close to the standard Gaussian prior $p(\mathbf{z})$.

$$\mathcal{L}_{\text{VAE}} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))}_{\text{regularisation}} \tag{1}$$

This network has become popular thanks to its applications in the computer vision field [17, 18]. Because tabular data combines categorical and continuous features, whose structure differs markedly from that of image pixels, the VAE has been adapted for use with tabular datasets. TVAE has emerged as a pre-processing solution for

modelling rows of tabular data containing a mix of continuous and discrete columns [19]. For categorical features, VAEs typically employ cross-entropy loss to measure the reconstruction error, quantifying the discrepancy between the predicted class probabilities and the actual class distribution. Given a categorical feature with m classes, the cross-entropy loss is defined as $\text{CE} = -\sum_{i=1}^{m} y_i \log(\hat{y}_i)$, where y_i is the true class indicator and $\hat{y}_i$ is the reconstructed probability for class i. In practice, however, a Gaussian negative log-likelihood (NLL) reconstruction term is sometimes used as a drop-in replacement, even when the features are categorical. Although theoretically grounded only for continuous data under the assumption of a Gaussian decoder [13, 20], this formulation is nonetheless widely adopted as a practical, heuristic choice even for categorical features [20]: it applies a quadratic penalty on deviations from the target (whose gradient is linear in the error), it avoids the numerical instability of logarithmic terms in cross-entropy (i.e., $\log 0$ when predicted probabilities collapse) and tends to produce well-conditioned gradient updates in complex, multi-column tabular pipelines. In this formulation, for a batch of N samples and all categorical features jointly — with m denoting the total number of one-hot encoded dimensions across all categorical features — the reconstruction loss is defined as:

$$\mathcal{L}_{\text{cat}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{m} \frac{(y_{ni} - \hat{y}_{ni})^2}{2\sigma_c^2} \tag{2}$$

where y_{ni} is the true one-hot encoded class indicator for sample n and class i, $\hat{y}_{ni}$ is the reconstructed probability produced by the softmax output layer, and σ_c^2 is a hyperparameter controlling the weight of the reconstruction term relative to the KL divergence. This corresponds to a scaled MSE: the squared errors are averaged over the batch and summed over all one-hot encoded dimensions, then normalised by $2\sigma_c^2$. In the context of this work, the baseline model adopted in our industrial setting uses this Gaussian NLL formulation as the reconstruction loss for categorical features. We tested our proposed loss function modifications in an industrial setting based on this design choice: we provide more details in Section 3.

2.2 Synthetic data generation to address data imbalance

Several approaches have been proposed to address data imbalance through synthetic data generation, ranging from simple resampling techniques to fairness-aware deep generative models.

The earliest and most widely adopted techniques operate directly on the feature space without training a generative model. The *Synthetic Minority Over-sampling Technique* (SMOTE) [21] addresses class imbalance by interpolating new samples between existing minority-class instances, combined with majority-class under-sampling. The *Adaptive Synthetic Sampling* approach (ADASYN) [22] extends this idea by assigning a higher generation density to minority-class samples that are harder to learn, thus focusing the augmentation effort where the class boundary is most uncertain. While effective in many settings, these methods operate locally in feature space and do not model the global joint distribution of the data, which limits their ability to capture complex inter-feature dependencies.

A second line of work addresses imbalance and fairness through Generative Adversarial Networks (GANs). *FairGAN* [23] is one of the first GAN-based architectures explicitly designed to generate data satisfying group fairness constraints. *CTAB-GAN* [24] proposes a conditional GAN for tabular data that handles class imbalance through conditional generation, jointly modelling continuous, categorical and mixed-type features. *TabFairGAN* [25] extends this line by incorporating fairness constraints directly into the GAN training objective to produce data that satisfies statistical parity for sensitive features. *DECAF* [26] takes a causally-aware perspective, exploiting the causal graph of the data to generate counterfactually fair synthetic samples. Systems such as *GenEthos* [27] combine synthetic data generation with automated bias detection and mitigation.

Approaches more specifically focused on VAE architectures address imbalance from different angles. Louizos et al. [28] promote statistical independence between the sensitive attribute and the latent representation using the *Maximum Mean Discrepancy* [29], obtaining a latent space that is invariant to the sensitive feature. Grari et al. [30] employ a causal VAE to achieve fairness without requiring the sensitive attribute at inference time, separating fair and sensitive information in the latent space. Other works directly target minority-class under-representation: Wan et al. [31] use a VAE to synthesise new minority-class instances, while *ADA-INCVAE* [32] proposes an adaptive reconstruction loss weighted by the local density of majority samples, preventing posterior collapse for minority classes. Hong et al. [33] use a conditional VAE to estimate the density of minority-class samples in the latent space, using this information to filter out noisy interpolations before applying SMOTE in the original feature space. Güldal et al. [34] propose a VAE-based approach that generates synthetic data to balance medical datasets, showing improvements in downstream classification performance. Chadebec et al. [35] propose a data augmentation strategy that modifies the sampling procedure of the VAE decoder by exploiting the geometry of the latent manifold, enriching underrepresented regions of the input distribution.

We do not alter the architecture of the VAE: instead, we propose two modifications of the loss function that act on the reconstructed distribution of the sensitive categorical features, providing a lightweight and tunable mechanism to steer the generation towards balance. With respect to the approaches based on latent representation [28, 30], our modifications act on the reconstructed distribution of the observed categorical features via the loss function. In comparison to data augmentation approaches [31, 33, 35], we generate a substitute synthetic dataset. While the main principle of *ADA-INCVAE* [32] is similar to ours, in that both reweight the reconstruction loss, the formulation is different. Their per-sample weight is computed based on the Euclidean distance between minority and majority samples, whereas we focus on the class frequencies of the sensitive features.

3 Proposed approach

Herein we provide some details about the underlying model (identified as *vanilla*), then we present our proposed modifications of the loss function. To enhance reproducibility, the code related to the proposed loss function modifications, along with the materials

needed to reproduce the empirical validation study, is available online at https://github.com/RondinaMR/VAE-Balance-Loss and has been archived at https://doi.org/10.5281/zenodo.22348572.

3.1 The Vanilla Variational AutoEncoder

This experiment is based on the VAE architecture implemented in a commercial application for synthetic data generation. This section provides a high-level description of the architecture and the loss function of the "vanilla" VAE, which is the baseline model for our proposed modifications. In terms of preprocessing, the model applies a quantile transformation mapping continuous features onto a standard normal distribution and a one-hot encoding to categorical features.

The **decoder** is type-aware, meaning that it uses a different output layer and loss function for continuous and categorical features: for continuous features, the output layer is linear and the loss is Mean Squared Error (MSE), while for categorical features, the output layer is a softmax and the loss is another MSE.

The final "vanilla" loss function is as follows:

$$\mathcal{L}_{\text{vanilla}} = \mathcal{L}_{\text{ord}} + \mathcal{L}_{\text{cat}} + \mathcal{L}_{L2} + \mathcal{L}_{KLD} \tag{3}$$

The term $\mathcal{L}_{\text{ord}}$ is a scaled MSE (fixed-variance Gaussian NLL) over numerical features, where the squared errors of all numerical features are divided by $2\sigma^2_{\text{ord}}$ with σ_{ord} a fixed hyperparameter: this sets the weight of the continuous reconstruction term relative to the other components of the loss, which is consistent with all continuous features sharing a common scale after the quantile transformation.

The term $\mathcal{L}_{\text{cat}}$ operates on categorical features represented as one-hot encodings: it is likewise an MSE (fixed-variance Gaussian NLL) between the softmax probabilities produced by the decoder and the original one-hot vectors, divided by $2\sigma^2_{\text{cat}}$, and with a penalty term to discourage the decoder from producing degenerate probability distributions concentrated on a single class.

The term $\mathcal{L}_{L2}$ is an L2 regularization on the network weights, whose coefficient is set to zero in the experiments reported here.

The term $\mathcal{L}_{\text{KLD}} = -\frac{\alpha}{2}\,\mathbb{E}\big[1 + \log\sigma_z^2 - \mu_z^2 - \sigma_z^2\big]$ is the Kullback-Leibler divergence between the approximate posterior $q_\phi(z|x) = \mathcal{N}(\mu_z, \sigma_z^2)$ and the prior $p(z) = \mathcal{N}(0, I)$, weighted by the coefficient $\alpha = 0.1$; it regularizes the latent space by encouraging a continuous and well-structured distribution amenable to sampling.

3.2 Proposed Loss Function Balancing Terms

The proposed modifications of the loss function focus on solving the problem of class imbalance. The intuitive idea is to add a term that guides the training phase to minimise the distance between the actual distribution and the ideal distribution of categorical sensitive features. We consider a categorical feature as "sensitive" if it operationalizes one of the protected attributes defined in Article 21 of the Charter of Fundamental Rights of the European Union [36]. Further details on the selection of sensitive features are provided in Section 5.

We take the uniform distribution, in which all class frequencies are equal, as the ideal distribution of a sensitive feature (Eq. 4), where m is the number of classes.

$$U = \frac{1}{m} \tag{4}$$

The rationale for targeting the uniform distribution lies in the generative nature of the VAE. By penalising deviations from uniformity during training, the balance loss steers the decoder (and, indirectly through backpropagation, the encoder and the structure of the latent space) to produce balanced class distributions for sensitive features.

We propose two formulations for the balanced loss:

- an additive term $\mathcal{BL}_{term}$ based on the distance between the reconstructed class probabilities vector and the uniform distribution U;
- a re-weight term $\mathcal{BL}_{weight}(D)$ based on the Simpson Index D.

The additive term $\mathcal{BL}_{term}$ is presented in (6). It consists of the sum, over the batch size N and over the sensitive features s, of the squared difference between the reconstructed class probabilities (reconstruction_k) and the uniform distribution U. The weight of the balance magnitude, denoted by $\mathcal{BM}$, functions as the knob employed to adjust the intensity of the balance correction.

$$\mathcal{BM} \in [0, 1] \tag{5}$$

The values of $\mathcal{BM}$ that were tested during the validation experiments are described in Section 5.

$$\mathcal{BL}_{term} = \mathcal{BM} \cdot \sum_{n=1}^{N} \sum_{k=1}^{s} (\text{reconstruction}_k - U)^2 \tag{6}$$

At the end of each train step, we compute the mean of the numerical loss term (not affected by our modification), the vanilla loss term (not affected by our modification) and the balance loss. In the manuscript, we refer to the combination of the additive term $\mathcal{BL}_{term}$ with the specific magnitude $\mathcal{BM}$ as $t_{\mathcal{BM}}$ (e.g. $t_{0.1}$, $t_{0.5}$ and $t_{1.0}$).

The weight term $\mathcal{BL}_{weight}(D)$ is based on the Simpson index D loss (7), i.e. a quantity representing the distance between the perfectly balanced distribution and the actual distribution of the sensitive features in the generated synthetic dataset. Viewed from another perspective, $\mathcal{L}_D$ represents a Simpson's index that takes values between 0 (maximum balance) and 1 (minimum balance).

$$\mathcal{L}_{\text{D}} = 1 - D \in [0, 1] \tag{7}$$

$\mathcal{BL}_{weight}(D)$ is then defined as follows:

$$\mathcal{BL}_{\text{weight}}(D) = e^{\mathcal{M}_{exp} \cdot \mathcal{BM} \cdot (\mathcal{L}_D)} \tag{8}$$

where $\mathcal{M}_{exp}$ represents an exponential magnitude term. $\mathcal{M}_{exp}$ fixes the scale of the exponent, while $\mathcal{BM}$ is the knob used to adjust the intensity of the balance correction. Since only their product affects the loss, the two parameters are not identifiable

separately. Therefore fixing $\mathcal{M}_{exp}$ and varying $\mathcal{BM}$ is sufficient to explore the space of balance corrections, with $\mathcal{M}_{exp}$ setting the upper end of the range spanned by $\mathcal{BM}$. Following preliminary evaluations, we set $\mathcal{M}_{exp} = 10.0$ for all the datasets and configurations. Lower values of $\mathcal{M}_{exp}$ (e.g., 1.0) produced a weaker effect on the balance loss, while higher values (e.g., 100.0) led to instability in the training process.

The Simpson index D is computed as detailed in (9). In the manuscript, we refer to the combination of the re-weight term $\mathcal{BL}_{weight}$ with the specific magnitude $\mathcal{BM}$ as $w_{\mathcal{BM}}$ (e.g. $w_{0.1}$, $w_{0.5}$ and $w_{1.0}$).

4 Research Question

In this section we provide a description of the goal that drives this research. We follow the Goal-Question-Metric approach [37]: first we describe the goal, then we continue with the questions and finally we present the metrics that allow us to answer the defined research questions.

The aim of the research is to propose a loss function modification to guide the synthetic data generation by rebalancing the imbalanced sensitive features. The empirical investigation based on the proposed loss functions is defined by the following objective.

We **analyze** the synthetic dataset generated with an ad-hoc balance loss term; **for the purpose of** improving balance of sensitive features; **with respect to their** balance, fairness, utility and privacy; **from the point of view** of researchers and developers; **in the context of** tabular data containing sensitive features generated with Variational Autoencoders.

Based on the above objective, we defined the following three research questions:

RQ1 Does the proposed loss function modification improve the **balance** and **fairness** of sensitive features?
RQ2 Does the proposed loss function modification degrade the **utility**?
RQ3 Does the proposed loss function modification degrade the **privacy**?

In order to evaluate the statistical reliability of the results, we defined the following null hypotheses to be refuted:

$H1_0$ The proposed loss function modification does not improve the **balance** and **fairness** of sensitive features;
$H2_0$ The proposed loss function modification does not degrade the **utility**;
$H3_0$ The proposed loss function modification does not degrade the **privacy**.

To assess the research questions, we measured and evaluated the following metrics (described in Section 4.1):

RQ1 Balance loss $\mathcal{L}_{\text{imbalance}}$, Statistical Parity Difference (SPD)
RQ2 Train on Synthetic, Test on Real (TSTR) Accuracy
RQ3 Distance to Closest Record (DCR) Synthetic-Holdout

In order to evaluate the impact of the proposed loss function modifications on the presented metrics, we examine how the metrics are distributed across the different

training splits and loss function configurations. We then analyze the paired differences between these distributions and those of the vanilla baseline configuration.

4.1 Measures for utility, privacy, balance and fairness

To measure the performance of our work in terms of utility, balance and privacy, we base our findings on a number of metrics. For measurement purposes, we split the original dataset O into the training dataset T and the validation dataset V (namely 80%-20%). The training dataset T is then used to generate the synthetic dataset S, while the validation dataset V is not. Figure 1 presents the experiment design.

4.1.1 Balance

The chosen balance measure is a diversity measure used in ecology: the Simpson index [38, 39]. The Simpson index is computed for each (sensitive) feature and assumes values between 0 (lowest balance) and 1 (highest balance, class frequencies are uniformly distributed). It is presented in (9): m represents the number of classes, while f_i is the relative frequency of class i. Previous studies have identified fairness implications when the index falls below a 0.30 threshold [38].

$$D = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1 \right) \tag{9}$$

We used the Simpson index to obtain a *Simpson Imbalance Loss of the Dataset Index* (SILDI), a single value for a given synthetic dataset. SILDI is computed as the sum of the differences between 1 and the Simpson index, for each sensitive feature (Eq. 10).

$$\text{SILDI} = \sum_{f \in S} (1 - D_f) \tag{10}$$

4.1.2 Fairness

To measure the fairness of the generated synthetic dataset, we adopt the *Statistical Parity Difference* (SPD) [7]. SPD quantifies the difference in the probability of receiving a positive outcome between an unprivileged group and a privileged group, as defined in (11), where Y denotes the predicted outcome and D denotes the sensitive attribute. A value of 0 indicates perfect parity, meaning the two groups are equally likely to receive a positive outcome. Positive values indicate that the unprivileged group is favoured, while negative values indicate that the privileged group is favoured.

$$P_r(Y = 1 | D = \text{unprivileged}) - P_r(Y = 1 | D = \text{privileged}) \tag{11}$$

In the context of our empirical study, for each dataset and for each sensitive feature, we identify the privileged and unprivileged groups according to the criterion described in Section 5. We then compute the SPD for the original dataset O and for the synthetic dataset S to evaluate how well the synthetic data preserves or improves the statistical parity of the sensitive feature r.

$$SPD_r = SPD_{S,r} - SPD_{O,r} \tag{12}$$

4.1.3 Utility

In order to measure the data utility we use the *Train on Synthetic, Test on Real* (TSTR) approach [40]. According to the taxonomy proposed by Koubeissy et al. [16], this approach falls under the "realism" category, as it evaluates the potential of synthetic data to serve as a functional substitute for real data in downstream tasks, such as training a machine learning model. In our case, an XGBoost classifier [41] is trained to discriminate between real samples drawn from T and synthetic samples from S, with detection performance on V serving as an inverse proxy for data quality. In our empirical evaluation, we compute the TSTR accuracy for both the synthetic dataset S and the training dataset T, and then we calculate the difference between the two to assess how much utility is preserved in the synthetic data compared to the original training data.

4.1.4 Privacy

We measure the preservation of privacy in the synthetic dataset, with respect to the original one, using the *Distance to Closest Record* (DCR) [42], based on Gower's distance [43]. Gower's distance (13) measures the distance between two mixed-type (numerical, categorical) data rows s and o, each consisting of p features (columns). w_{sok} represents the weight assigned to the k-th feature (we consider each feature equally important, so $w_{sok} = 1$), while d_{sok} represents the distance between the k-th feature of the records s and o.

$$D_{so} = \frac{\sum_{k=1}^{p} d_{sok} \cdot w_{sok}}{\sum_{k=1}^{p} w_{sok}} \tag{13}$$

Feature-wise distances d_{sok} are computed differently depending on whether each feature is categorical (14) or numerical (15). In (15) R_k represents the range for the k-th feature.

$$Dcat_{sok} = \begin{cases} 0 \text{ if } x_{sk} = x_{ok} \\ 1 \text{ if } x_{sk} \neq x_{ok} \end{cases} \tag{14}$$

$$Dord_{sok} = \frac{|x_{sk} - x_{ok}|}{R_k} \tag{15}$$

The Distance to Closest Record is defined as the minimum distance between the given synthetic record s in S and every original record o in O:

$$\text{DCR}(s) = \min d(s, o) \text{ for each } o \in O \tag{16}$$

A value of $\text{DCR}(s) = 0$ indicates that s replicates at least one record in the original dataset O. The DCR values for all synthetic instances were calculated against both the training dataset (T) and the validation dataset (V). To measure privacy risk, we calculated the percentage of synthetic instances that are closer to an instance of the training dataset T than to an instance of the validation dataset V. The minimum risk

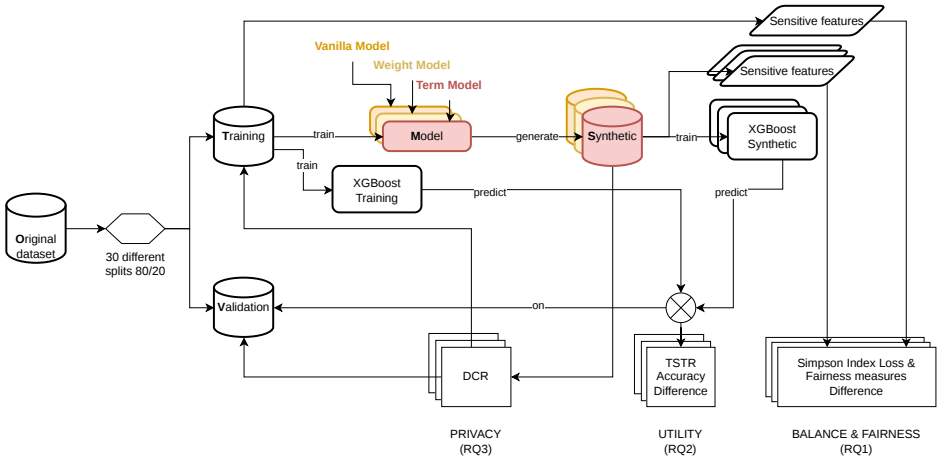


Figure 1: Experiment design schema. The original dataset O is split into a training dataset T and a validation dataset V. The training dataset T is used to train the VAE model, which generates the synthetic dataset S. The validation dataset V is used to evaluate the utility and privacy of the synthetic dataset. The synthetic dataset S is evaluated for balance and fairness of sensitive features, utility, and privacy.

level is represented by a value of 50%: this indicates that the synthetic dataset does not disclose any information that could be used by an attacker to determine whether a specific row of the training dataset T or of the validation dataset V was used to train the model. For this reason, in the paired differences presented in Table 5 we consider the excess over 50% as a measure of leakage risk.

5 Experiment Design

To evaluate the proposed modifications to the loss function, we performed an empirical evaluation based on the research questions discussed above. Figure 1 summarizes the experimental design. The different RQs test different parts of the process:

RQ1 compare the SILDI computed on the sensitive features of the Training dataset T and of the Synthetic dataset S, and compute the Statistical Parity Difference (SPD);

RQ2 train two XGBoost models, the first on the synthetic dataset S and the second on the training dataset T. Then compare their performance by classifying records from the validation dataset V;

RQ3 compute the distance to the closest record (DCR) from the synthetic dataset S to both the training dataset T and the validation dataset V.

The **treatment** involves the proposed modifications to the loss function, which are used to train the VAE model. We tested the two different modifications of the loss function, $\mathcal{BL}_{weight}$ and $\mathcal{BL}_{term}$. For both the loss modifications, we tested three different **balance magnitudes** $\mathcal{BM}$: 0.1, 0.5 and 1.0. These values sample the domain

of $\mathcal{BM}$ at its lower end, middle and upper bound, to characterise the trade-off between balance, fairness, utility and privacy across the range of corrections. Naturally, the balance loss is not applied in the vanilla configuration, which serves as a baseline for comparison.

We tested two different configurations of **batch size** (and consequently of number of **epochs**). The first configuration is the mono-batch one. Since the loss and the balance loss are calculated batch by batch, this solution provides a more precise evaluation of the balance loss, as it takes into account all classes in the entire dataset. The downside of this solution is that the entire dataset must be loaded into memory at once, which may be infeasible for large datasets without appropriate hardware. This is why we also include a batch size equal to 512: this ensures that memory requirements are kept to a minimum while remaining large enough to reliably estimate the balance loss. In summary, we selected two configurations: a mono-batch with 1,250 epochs and a batch size of 512 with 50 epochs. All the other hyperparameters were kept constant across all configurations and datasets.

Table 1: Sensitive features considered by the loss function for balancing synthetic data generation. For each sensitive feature considered in the fairness analysis, the privileged and unprivileged groups are defined.

Dataset	**Sensitive Feature**	**Fairness evaluation**	
		Privileged Group	**Unprivileged Group**
Adult	*sex*	Male	Female
	race	White	All other classes
	native_country	United States	All other classes
S. German Credit	*famges* (personal status&sex)	Male, married/widowed	All other classes
	gastarb (foreign worker)	Non-foreign worker	Foreign worker
	laufkont (checking account status)	×	×
	sparkont (savings account status)	×	×
	verm (property)	×	×
	wohn (housing)	×	×
COMPAS	*sex*	Male	Female
	race	Caucasian	All other classes
	age_cat	Greater than 45	25–45 and Less than 25

We tested the treatment on three different **datasets**, which are commonly used in the literature for testing synthetic data generation techniques and for fairness evaluation: Adult [44], South German Credit [45] and COMPAS [46]. The Adult dataset, extracted from the 1994 U.S. Census database, contains socio-demographic attributes such as *age*, *education*, *occupation*, *sex*, *race* and *native_country*, with the binary task of predicting whether an individual's annual income exceeds $50K. The South German Credit dataset describes applicants for credit from a German bank, including features related to financial status, personal characteristics and employment, with the

binary task of classifying creditworthiness as good or bad. The COMPAS dataset, released by ProPublica, contains criminal history and demographic information — including *race* and *sex* — used by the COMPAS recidivism risk assessment tool to predict the likelihood of reoffending within two years. Unlike the other two datasets, the target variable in the COMPAS dataset is a three-class variable (low, medium, and high risk). To maintain consistency across the datasets, we binarised the COMPAS dataset's target variable by treating the low-risk class as positive and the medium- and high-risk classes as negative.

We treat as **sensitive** the attributes that directly encode one of the protected attributes defined in Article 21 of the Charter of Fundamental Rights of the European Union [36] and that the dataset provides in categorical form. The restriction to categorical features stems from the operationalization of the Simpson index (and the balance loss), based on frequency distributions. We decided to not introduce a discretization of continuous features to avoid altering the balance objective with an arbitrary threshold. This is the reason why the *age* feature is excluded from Adult and South German Credit datasets, where it is provided as a continuous variable, while in COMPAS it is included as *age_cat*, already discretized by ProPublica. The proposed loss function modifications are applied to all sensitive features of the datasets, while the fairness analysis is performed only on a subset of these features, specifically those related to demographic characteristics. In order to compute the statistical parity difference (see equation 11) it is necessary to select a reference group, i.e. the privileged group. For each attribute, the **privileged group** comprises a single observed category, while the unprivileged group comprises its complement. The privileged category is the most common one, accounting for six of the eight attributes reported in Table 1. The two exceptions are the race and age attributes of COMPAS. For the race attribute, we retain the baseline of the originating study rather than the modal category. For the age attribute, to avoid offsetting effects, we select an extreme of the order ("Greater than 45"). As the comparison group pools the remaining categories, the metric quantifies the aggregate gap with respect to the baseline and does not provide information about individual categories.

Table 1 summarises the sensitive features and the privileged/unprivileged groups for each sensitive feature in the 3 datasets.

The **layer size**, which also determines the dimensionality of the latent space, was set as dataset-dependent. For lower-dimensionality datasets (Adult and South German Credit), we adopted a two-layer architecture with a latent space dimension approximately equal to half the input dimensionality after one-hot encoding. For the higher-dimensionality COMPAS dataset, a series of pilot experiments over double-layer configurations revealed posterior collapse across all tested architectures and all loss models (including the vanilla baseline), indicating architecture-induced instability rather than a loss-specific artifact. We therefore adopted, for this dataset, a single-layer architecture with latent dimension $\approx 71\%$ of the input dimensionality (i.e., $\lfloor d/1.4 \rfloor$, where d is the input size after preprocessing), which was the minimum intervention necessary to obtain stable generation across all seeds.

For each dataset and for each configuration of the loss function, we generated 30 different splits 80%-20% of the original dataset. For each split we trained three VAE

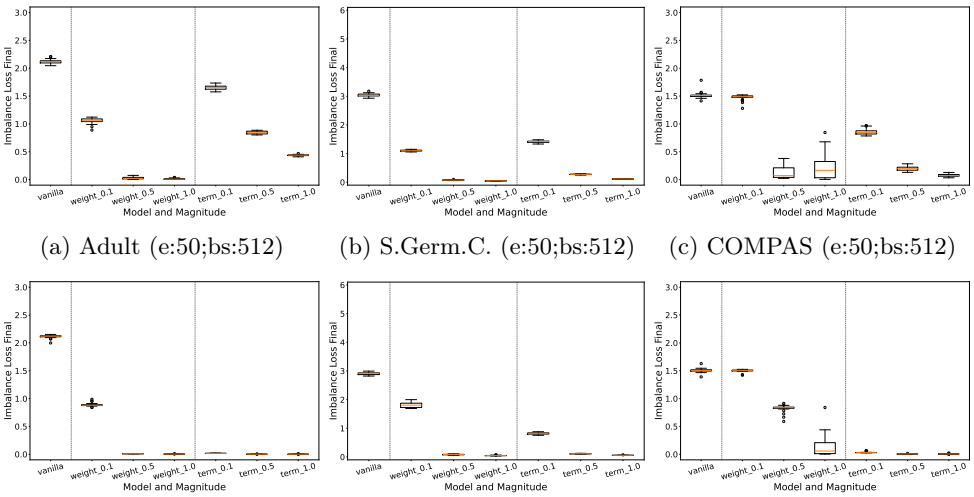
models to generate a synthetic dataset: vanilla, i.e. the original VAE model of the commercial application we based our work on, and the two proposed modifications of the loss function ($\mathcal{BL}_{term}$ and $\mathcal{BL}_{weight}$). This resulted in a total of 420 synthetic datasets (30 splits $\times$ 2 batch sizes and epochs configurations $\times$ 7 loss function and balance magnitudes configurations) for each dataset under investigation. The total number of synthetic datasets generated is therefore 1,260 (420 for each of the 3 datasets).

6 Experimental Results

In this Section we present the results of our empirical evaluation, which are based on the research questions defined in Section 4. The results are firstly presented in the form of boxplots, which show the distribution of the measured metrics for each configuration of the loss function and for the vanilla configuration. The boxplots enable us to visually compare the performance of the different configurations and to identify differences in terms of balance, fairness, utility and privacy. The main goal of the figures is to evaluate the trade-off between these three aspects: we expect that the proposed modifications of the loss function will improve balance, but we want to verify if this improvement comes at the cost of utility and/or privacy. Fairness results are presented in a separate figure, since they are measured for each sensitive feature and not as a single value for the entire dataset as the other metrics. In addition to the boxplots, we present the result of a statistical analysis of the paired differences between the proposed loss modifications and the vanilla baseline, using the metrics defined in Section 4.1. The paired differences are computed by comparing the same training-validation split of the original dataset, in such a way that the differences are not affected by the variability of the splits. This part of the analysis is presented in the form of four tables (Tables 2, 3, 4 and 5), one for each metric. In each table we present the reference average (μ) and standard deviation (σ) of the vanilla configuration. Then, for each configuration and for each dataset, we present: the percentage of splits for which the proposed modification improves (RQ1) or degrades (RQ2 and RQ3) the vanilla configuration; the average (μ) and standard deviation (σ) of the paired differences of the metric under analysis. The statistical analysis does not include significance tests, since the splits (80%-20%) are not fully independent, as they are drawn from the same original dataset, and p-values would overstate the strength of the evidence.

6.1 RQ1: Balance and Fairness

The results for the balance measurement are shown in Figure 2 and Table 2. The proposed modifications of the loss function show a consistent improvement in terms of balance, as measured by the *Simpson Imbalance Loss of the Dataset Index* (SILDI), for all the three datasets under investigation. All the configurations presents a 100% of splits for which the proposed modification improves the vanilla version on the same split, with the exception of the $w_{0.1}/e50$ and $w_{0.1}/e50$ configurations for the COMPAS dataset (namely 86.7% and 50.0%). In the mono-batch configurations (e1250) $\mathcal{BL}_{term}$ reaches a low SILDI at much smaller magnitudes than $\mathcal{BL}_{weight}$, while with a batch size of 512 (e50) the ordering is reversed and $\mathcal{BL}_{weight}$ is the more effective of the two. $\mathcal{BL}_{term}$ is monotone in the magnitude across all datasets and both regimes, whereas



(a) Adult (e:50;bs:512) (b) S.Germ.C. (e:50;bs:512) (c) COMPAS (e:50;bs:512)

(d) Adult (e:1250;bs:mono-batch) (e) S.Germ.C. (e:1250;bs:mono-batch) (f) COMPAS (e:1250;bs:mono-batch)

Figure 2: Final *Simpson Imbalance Loss of the Dataset Index* (SILDI) for the three datasets under investigation. The first row shows the results for a batch size of 512 and 50 epochs, while the second row shows the results for a mono-batch and 1,250 epochs. The vertical dashed lines subdivide the different configurations of the loss function. From left to right, these are the vanilla configuration; the $\mathcal{BL}_{weight}$ configuration, with magnitudes of 0.1, 0.5 and 1.0; and the $\mathcal{BL}_{term}$ configuration, with magnitudes of 0.1, 0.5 and 1.0. The SILDI is computed as the sum of the differences between 1 and the Simpson index for each sensitive feature, and it represents a single value for a given synthetic dataset. The lower the SILDI, the better the balance of the sensitive features in the generated synthetic dataset. The upper limit of South German Credit dataset is 6 because it has 6 sensitive features, while the other two datasets have 3 sensitive features, as indicated in Table 1.

$\mathcal{BL}_{weight}$ is not. In the mono-batch configurations, the $\mathcal{BL}_{term}$ configuration is able to reduce the SILDI for the Adult and COMPAS dataset even with a magnitude of 0.1, while for the South German Credit dataset a magnitude of 0.5 is required to achieve a comparable, though not equally low, value.

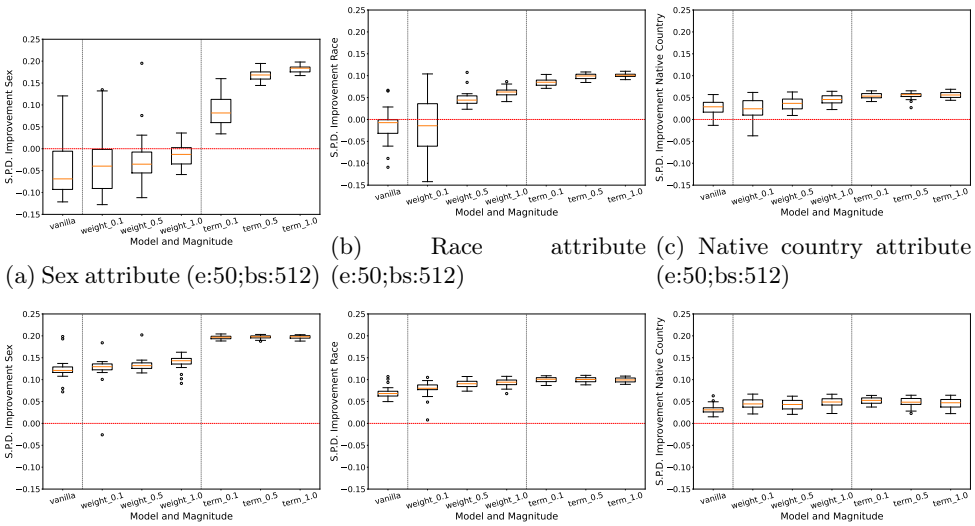
To observe the **fairness** improvement we analysed the statistical parity difference (SPD), as described in Section 4.1. Starting from the Adult dataset (Figure 3), we can see that the $\mathcal{BL}_{term}$ configuration shows an improvement in terms of statistical parity difference for all the sensitive features, while the $\mathcal{BL}_{weight}$ configuration appears to perform similarly to the vanilla model. In the case of the *sex* sensitive feature, the $\mathcal{BL}_{term}$ configuration shows also a gradual improvement with increasing magnitudes. This is confirmed by the statistical analysis of the paired differences (Table 3), which presents a percentage of splits for which the proposed modification improves the vanilla

Table 2: Balance (RQ1). Paired differences between the proposed loss modifications and the vanilla baseline, using the SILDI (the lower, the better). The % represents the ratio of split seeds for which the modification **improves** the vanilla version on the same split. A configuration with no systematic effect splits the seeds evenly, so 50% means no difference. SILDI scale is the number of those features: 3 for Adult and COMPAS, 6 for South German Credit. The means below are therefore not comparable within columns. The reference row is the vanilla baseline, w is the $\mathcal{BL}_{weight}$, t is the $\mathcal{BL}_{term}$.

Configuration	Adult $[0, 3]$	S. German C. $[0, 6]$	COMPAS $[0, 3]$
Ref. e50	$+2.1205 \pm 0.0379$	$+3.0421 \pm 0.0524$	$+1.5167 \pm 0.0580$
$w_{0.1}$/e50	100.0% (-1.0672 ± 0.0541)	100.0% (-1.9374 ± 0.0520)	86.7% (-0.0434 ± 0.0778)
$w_{0.5}$/e50	100.0% (-2.0938 ± 0.0441)	100.0% (-2.9648 ± 0.0517)	100.0% (-1.3941 ± 0.1314)
$w_{1.0}$/e50	100.0% (-2.1092 ± 0.0403)	100.0% (-2.9930 ± 0.0552)	100.0% (-1.2941 ± 0.2435)
$t_{0.1}$/e50	100.0% (-0.4681 ± 0.0461)	100.0% (-1.6284 ± 0.0473)	100.0% (-0.6615 ± 0.0782)
$t_{0.5}$/e50	100.0% (-1.2759 ± 0.0400)	100.0% (-2.7644 ± 0.0540)	100.0% (-1.3185 ± 0.0661)
$t_{1.0}$/e50	100.0% (-1.6833 ± 0.0410)	100.0% (-2.9257 ± 0.0523)	100.0% (-1.4392 ± 0.0676)
Ref. e1250	$+2.1154 \pm 0.0281$	$+2.9113 \pm 0.0423$	$+1.5070 \pm 0.0345$
$w_{0.1}$/e1250	100.0% (-1.2277 ± 0.0392)	100.0% (-1.1024 ± 0.0853)	50.0% (-0.0053 ± 0.0415)
$w_{0.5}$/e1250	100.0% (-2.1056 ± 0.0285)	100.0% (-2.8274 ± 0.0450)	100.0% (-0.6837 ± 0.0655)
$w_{1.0}$/e1250	100.0% (-2.1116 ± 0.0282)	100.0% (-2.8614 ± 0.0454)	100.0% (-1.3637 ± 0.1852)
$t_{0.1}$/e1250	100.0% (-2.0985 ± 0.0285)	100.0% (-2.0888 ± 0.0460)	100.0% (-1.4742 ± 0.0360)
$t_{0.5}$/e1250	100.0% (-2.1132 ± 0.0281)	100.0% (-2.7981 ± 0.0410)	100.0% (-1.5008 ± 0.0342)
$t_{1.0}$/e1250	100.0% (-2.1132 ± 0.0280)	100.0% (-2.8521 ± 0.0425)	100.0% (-1.5011 ± 0.0351)

version on the same split higher than 92% for all the $\mathcal{BL}_{term}$ configurations, while the $\mathcal{BL}_{weight}$ configurations span between 51.1% ($w_{0.1}/e50$) and 86.7% ($w_{1.0}/e1250$).

The results for the South German Credit dataset are shown in Figure 4. In this case we can observe different results for the two sensitive features under investigation, namely *famges* and *gastarb*. *Famges* is a feature representing the combination of the family status and the *sex* of the applicant, while *gastarb* is a feature representing whether the applicant is a foreign worker or not (values: yes or no). We



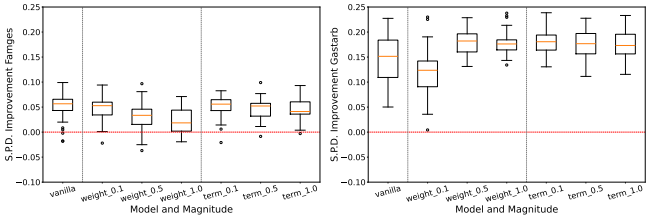
(a) Sex attribute (e:50;bs:512) (b) Race attribute (e:50;bs:512) (c) Native country attribute (e:50;bs:512)

(d) Sex attribute (e:1250;bs:mono-batch) (e) Race attribute (e:1250;bs:mono-batch) (f) Native country attribute (e:1250;bs:mono-batch)

Figure 3: Fairness measurements improvement for the Adult dataset. The fairness improvement is computed as the difference between the statistical parity difference (SPD) of the synthetic dataset and the SPD of the original dataset for each sensitive feature. A positive value indicates an improvement in fairness (w.r.t. the original dataset), while a negative value indicates a worsening of fairness. The first row shows the results for a batch size of 512 and 50 epochs, while the second row shows the results for a mono-batch and 1250 epochs. The vertical dashed lines subdivide the different configurations of the loss function. The horizontal dashed line at 0 indicates no change in fairness.

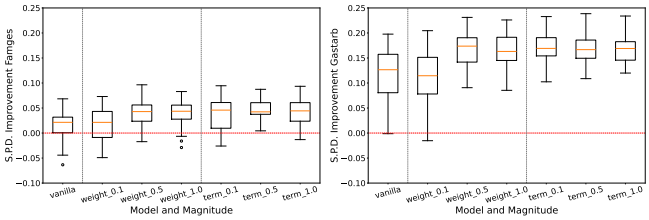
can observe an improvement in terms of statistical parity difference for the *gastarb* feature, while the improvement for the *famges* feature is less consistent across the different configurations of the loss function. From Table 3 we can see that both $w_{0.1}/e50$ and $w_{1.0}/e1250$ configurations present a percentage of splits for which the proposed modification improves the vanilla version on the same split lower than 50% (namely 35.0% and 43.3%). For all other configurations, the difference between the 50-epoch and 1,250-epoch configurations is notable, with the latter showing a higher percentage of splits for which the proposed modification improves upon the vanilla version (consistently above 70%).

The results for the COMPAS dataset are shown in Figure 5. The sensitive features under investigation in the fairness analysis are *sex*, *race* and *age* category, as shown in Table 1. The $\mathcal{BL}_{weight}$ configurations present a pitfall for the *sex* sensitive feature, in the case of magnitude equal to 1.0, showing a marked decrease in statistical parity difference improvement compared to the vanilla model. This behaviour is observed also for the *age* feature, although in this case it is not a matter of the average results but of the distribution of the results, which is more spread out and includes more



(a) Famges attribute (e:50;bs:512)

(b) Gastarb attribute (e:50;bs:512)

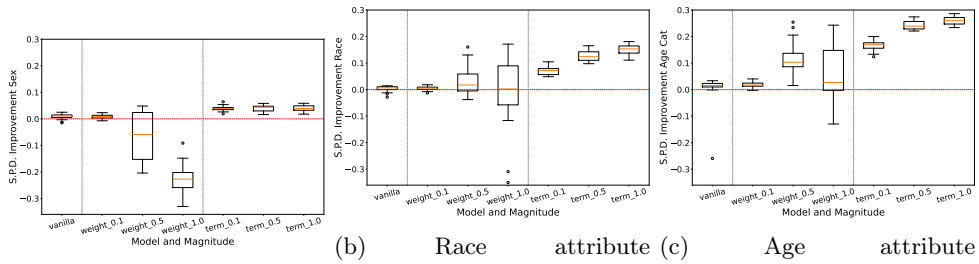


(c) Famges attribute (e:1250;bs:mono-batch)

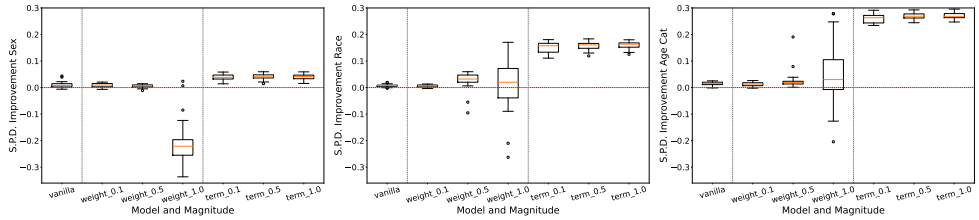
(d) Gastarb attribute (e:1250;bs:mono-batch)

Figure 4: Fairness measurements improvement for the South German Credit dataset. The fairness improvement is computed as the difference between the statistical parity difference (SPD) of the synthetic dataset and the SPD of the original dataset for each sensitive feature. A positive value indicates an improvement in fairness (w.r.t. the original dataset), while a negative value indicates a worsening of fairness. The first row shows the results for a batch size of 512 and 50 epochs, while the second row shows the results for a mono-batch and 1250 epochs. The vertical dashed lines subdivide the different configurations of the loss function. The horizontal dashed line at 0 indicates no change in fairness.

negative values compared to the vanilla model. The $\mathcal{BL}_{term}$ configurations, on the other hand, show a relevant improvement in terms of statistical parity difference for all the sensitive features. This is confirmed by Table 3: the $\mathcal{BL}_{term}$ configurations present a percentage of splits for which the proposed modification improves the vanilla version on the same split higher than 98%, for both mono-batch (e1250) and multi-batch (e50) settings. The $\mathcal{BL}_{weight}$ configurations present a growing percentage of improved splits with increasing magnitude between 0.1 and 0.5, but the percentage drops for the magnitude of 1.0. Decomposing the $w_{1.0}$ cells by sensitive feature, we can see that $w_{1.0}$ worsens relevantly the SPD for the *sex* feature, while continuing to improve the SPD for the *age_cat* feature. At magnitude 1.0 both formulations rebalance the marginal distribution of *sex*, by splitting the data between the two groups almost equally. The difference lies in what happens to the conditional distribution of the outcome. Under $w_{1.0}$, the rate of favourable outcomes for males increases, while for females, it decreases. In one split the configuration also removed the privileged category of *race* from the synthetic data entirely, leaving the statistical parity difference undefined. This means that the $w_{1.0}$ boxplots in Figures 5b and 5e are computed over 29 seeds



(a) Sex attribute (e:50;bs:512) (b) Race attribute (e:50;bs:512) (c) Age attribute (e:50;bs:512)



(d) Sex attribute (e:1250;bs:mono-batch) (e) Race attribute (e:1250;bs:mono-batch) (f) Age attribute (e:1250;bs:mono-batch)

Figure 5: Fairness measurements improvement for the COMPAS dataset. The fairness improvement is computed as the difference between the statistical parity difference (SPD) of the synthetic dataset and the SPD of the original dataset for each sensitive feature. A positive value indicates an improvement in fairness (w.r.t. the original dataset), while a negative value indicates a worsening of fairness. The first row shows the results for a batch size of 512 and 50 epochs, while the second row shows the results for a mono-batch and 1250 epochs. The vertical dashed lines subdivide the different configurations of the loss function. The horizontal dashed line at 0 indicates no change in fairness.

instead of 30. In Table 3 the mean and standard values are computed by dividing by 30 to avoid overestimating the effect of the 29 seeds.

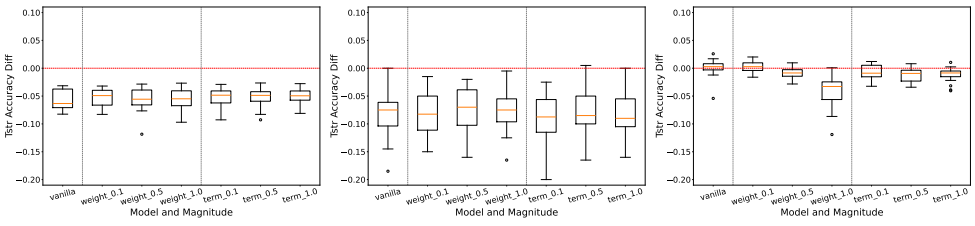
Answer to RQ1. On **balance**, the evidence contradicts $H1_0$: the modification lowers the SILDI on all 30 splits in 34 of the 36 dataset–configuration cells, with two exceptions being $w_{0.1}$ on COMPAS, which improves 86.7% of the splits in the multi-batch setting (e50) and only 50.0% in the mono-batch one (e1250). On **fairness**, the answer depends on the configuration: $\mathcal{BL}_{term}$ improves the SPD in almost all the splits on Adult and COMPAS and in 58.3% to 78.3% on South German Credit. $\mathcal{BL}_{weight}$ never exceeds 86.7% and falls below 50% in 8 of its 18 cells. Therefore, $H1_0$ does not hold for balance on all three datasets, nor for fairness under the $\mathcal{BL}_{term}$ formulation, although the latter improves a modest majority of the splits on South German credit. For $\mathcal{BL}_{weight}$ the fairness effect is consistent on Adult but irregular elsewhere, and at magnitude 1.0 on COMPAS it reverses.

Table 3: Fairness (RQ1). Paired differences between the proposed loss modifications and the vanilla baseline, using the SPD. The % represents the ratio of split seeds for which the modification **improves** the vanilla version on the same split. The μ and σ pool the paired differences of all sensitive attributes of the dataset (3 for Adult and COMPAS, 2 for South German Credit), so the σ includes the variation between attributes as well as between splits. A configuration with no systematic effect splits the seeds evenly, so 50% means no difference. The reference row is the vanilla baseline, w is the $\mathcal{BL}_{weight}$, t is the $\mathcal{BL}_{term}$.

Configuration	Adult	S. German Credit	COMPAS
Ref. e50	-0.0115 ± 0.0529	$+0.0960 \pm 0.0611$	$+0.0068 \pm 0.0304$
$w_{0.1}$/e50	51.1% ($+0.0035 \pm 0.0566$)	35.0% (-0.0141 ± 0.0395)	46.7% ($+0.0038 \pm 0.0305$)
$w_{0.5}$/e50	74.4% ($+0.0299 \pm 0.0486$)	48.3% ($+0.0089 \pm 0.0439$)	65.6% ($+0.0196 \pm 0.1056$)
$w_{1.0}$/e50	83.3% ($+0.0425 \pm 0.0539$)	48.3% ($+0.0045 \pm 0.0460$)	35.6% (-0.0631 ± 0.1636)
$t_{0.1}$/e50	97.8% ($+0.0866 \pm 0.0610$)	68.3% ($+0.0200 \pm 0.0341$)	100.0% ($+0.0851 \pm 0.0637$)
$t_{0.5}$/e50	97.8% ($+0.1187 \pm 0.0872$)	58.3% ($+0.0156 \pm 0.0376$)	100.0% ($+0.1299 \pm 0.0908$)
$t_{1.0}$/e50	98.9% ($+0.1246 \pm 0.0928$)	58.3% ($+0.0133 \pm 0.0370$)	100.0% ($+0.1424 \pm 0.0964$)
Ref. e1250	$+0.0756 \pm 0.0411$	$+0.0665 \pm 0.0670$	$+0.0109 \pm 0.0085$
$w_{0.1}$/e1250	78.9% ($+0.0074 \pm 0.0256$)	43.3% (-0.0045 ± 0.0395)	42.2% (-0.0016 ± 0.0059)
$w_{0.5}$/e1250	86.7% ($+0.0132 \pm 0.0165$)	73.3% ($+0.0377 \pm 0.0550$)	62.2% ($+0.0091 \pm 0.0290$)
$w_{1.0}$/e1250	86.7% ($+0.0180 \pm 0.0217$)	71.7% ($+0.0367 \pm 0.0523$)	41.1% (-0.0606 ± 0.1501)
$t_{0.1}$/e1250	93.3% ($+0.0405 \pm 0.0286$)	75.0% ($+0.0371 \pm 0.0521$)	98.9% ($+0.1386 \pm 0.0905$)
$t_{0.5}$/e1250	92.2% ($+0.0395 \pm 0.0306$)	78.3% ($+0.0400 \pm 0.0516$)	98.9% ($+0.1449 \pm 0.0928$)
$t_{1.0}$/e1250	92.2% ($+0.0387 \pm 0.0307$)	78.3% ($+0.0395 \pm 0.0537$)	98.9% ($+0.1456 \pm 0.0932$)

6.2 RQ2: Utility

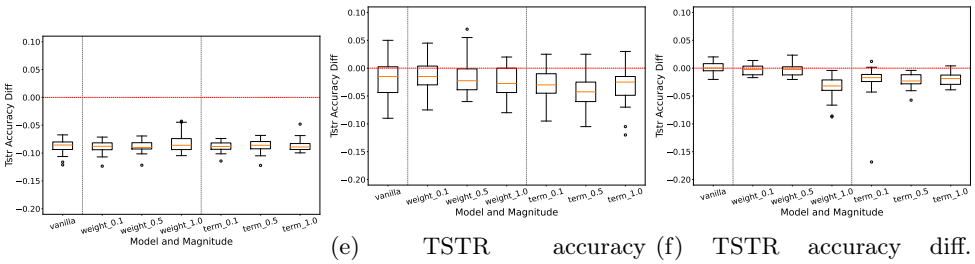
The results for the utility measurement are shown in Figure 6 and Table 4. The measure used to evaluate utility is the TSTR (Train on Synthetic, Test on Real) accuracy difference between the synthetic dataset and the original training dataset. As detailed in Section 4.1, the value of 0 represents the threshold at which the synthetic dataset achieves the same classification accuracy as the original training dataset. During the



(a) TSTR accuracy diff. Adult (e:50;bs:512)

(b) TSTR accuracy diff. S. German Credit (e:50;bs:512)

(c) TSTR accuracy diff. COMPAS (e:50;bs:512)



(d) TSTR accuracy diff. Adult (e:1250;bs:mono-batch)

(e) TSTR accuracy diff. S. German Credit (e:1250;bs:mono-batch)

(f) TSTR accuracy diff. COMPAS (e:1250;bs:mono-batch)

Figure 6: TSTR accuracy difference (synthetic vs. training) for the three datasets under investigation. The first row shows the results for a batch size of 512 and 50 epochs, while the second row shows the results for a mono-batch and 1,250 epochs. The vertical dashed lines subdivide the different configurations of the loss function. From left to right, these are the vanilla configuration; the $\mathcal{BL}_{weight}$ configuration, with magnitudes of 0.1, 0.5 and 1.0; and the $\mathcal{BL}_{term}$ configuration, with magnitudes of 0.1, 0.5 and 1.0. The horizontal dashed line at 0 represents the threshold at which the synthetic dataset achieves the same classification accuracy as the original training dataset. Values below 0 indicate a decrease in utility with respect to training on the original data.

generation of synthetic datasets we expect a decrease in utility with respect to training on the original data, especially for the configurations of the loss function with higher balance magnitudes, which are expected to generate more balanced synthetic datasets but with a greater deviation from the original data distribution.

The distribution of values (Figure 6) shows that, on Adult and South German Credit, the results for all configurations of the loss function remain comparable to those of the vanilla model, whereas on COMPAS, some configurations stand out from the rest. By observing the analysis of the paired differences with the vanilla baseline (Table 4), we can see that for Adult the percentage of splits for which the proposed modification does not degrade the utility spans from 43.3% to 60.0% (5 out of 12 configurations below 50%), for South German Credit from 26.7% to 60.0% (7 out of 12 $\leq 50\%$), and for COMPAS from 3.3% to 53.3% (11 out of 12 $\leq 50\%$). On Adult, the largest average loss is -0.0004 ± 0.0138 ($\mu \pm \sigma$ of the paired differences, $w_{0.1}$/e1250). In

Table 4: Utility (RQ2). Paired differences between the proposed loss modifications and the vanilla baseline, using the TSTR. The % represents the ratio of split seeds for which the modification performs **no worse** than the vanilla version on the same split. A configuration with no systematic effect splits the seeds evenly, so 50% means no difference. The reference row is the vanilla baseline, w is the $\mathcal{BL}_{weight}$, t is the $\mathcal{BL}_{term}$.

Configuration	Adult	S. German Credit	COMPAS
Ref. e50	-0.0562 ± 0.0170	-0.0803 ± 0.0398	$+0.0016 \pm 0.0140$
$w_{0.1}$/e50	46.7% ($+0.0040 \pm 0.0210$)	43.3% (-0.0003 ± 0.0271)	53.3% ($+0.0011 \pm 0.0128$)
$w_{0.5}$/e50	53.3% ($+0.0013 \pm 0.0229$)	56.7% ($+0.0030 \pm 0.0377$)	13.3% (-0.0106 ± 0.0141)
$w_{1.0}$/e50	46.7% ($+0.0009 \pm 0.0232$)	60.0% ($+0.0060 \pm 0.0381$)	3.3% (-0.0433 ± 0.0292)
$t_{0.1}$/e50	60.0% ($+0.0047 \pm 0.0233$)	26.7% (-0.0087 ± 0.0294)	33.3% (-0.0081 ± 0.0148)
$t_{0.5}$/e50	53.3% ($+0.0046 \pm 0.0231$)	46.7% (-0.0002 ± 0.0323)	10.0% (-0.0136 ± 0.0135)
$t_{1.0}$/e50	53.3% ($+0.0064 \pm 0.0210$)	56.7% (-0.0018 ± 0.0323)	6.7% (-0.0131 ± 0.0169)
Ref. e1250	-0.0882 ± 0.0118	-0.0182 ± 0.0337	$+0.0005 \pm 0.0101$
$w_{0.1}$/e1250	43.3% (-0.0004 ± 0.0138)	60.0% ($+0.0045 \pm 0.0234$)	33.3% (-0.0031 ± 0.0085)
$w_{0.5}$/e1250	56.7% (-0.0001 ± 0.0100)	60.0% ($+0.0023 \pm 0.0330$)	40.0% (-0.0038 ± 0.0136)
$w_{1.0}$/e1250	56.7% ($+0.0072 \pm 0.0228$)	36.7% (-0.0063 ± 0.0361)	3.3% (-0.0344 ± 0.0181)
$t_{0.1}$/e1250	46.7% ($+0.0002 \pm 0.0118$)	30.0% (-0.0105 ± 0.0256)	10.0% (-0.0230 ± 0.0303)
$t_{0.5}$/e1250	56.7% ($+0.0013 \pm 0.0109$)	26.7% (-0.0245 ± 0.0331)	3.3% (-0.0222 ± 0.0113)
$t_{1.0}$/e1250	43.3% ($+0.0023 \pm 0.0186$)	36.7% (-0.0160 ± 0.0325)	6.7% (-0.0205 ± 0.0106)

the mono-batch configuration (e1250), the South German Credit dataset shows a decrease in utility for the $\mathcal{BL}_{term}$ configuration at all magnitudes. The worst performance is observed for the $t_{0.5}$/e1250, which shows a decrease in utility of -0.0245 ± 0.0331. In the case of the COMPAS dataset, the worst performance is observed for the $w_{1.0}$, which shows a decrease in about 97% of the splits, with a mean difference ($\mu \pm \sigma$) of -0.0433 ± 0.0292 for multi-batch (e50) and -0.0344 ± 0.0181 for mono-batch (e1250). For this dataset, the only configuration whose mean difference is not negative is the $w_{0.1}$/e50, with a $\mu \pm \sigma$ difference of $+0.0011 \pm 0.0128$, a value that lies well within its own dispersion. All the other configurations stand under 15% of splits for which the proposed modification does not degrade the utility, with the exceptions of $t_{0.1}$/e50

(33.3%), $w_{0.1}$/e1250 (33.3%) and $w_{0.5}$/e1250 (40.0%). We can observe that not necessarily higher magnitudes of the balance correction lead to a higher degradation of utility: a strictly monotone decrease relationship in the percentage of non-degraded splits is observed only on COMPAS dataset, in multi-batch setting (e50), for both loss variants.

Answer to RQ2. The evidence does not allow for a consistent verdict across the three datasets. On Adult, with 43.3% to 60.0% of splits showing no degradation in utility and the largest average loss is $\mu = -0.0004$, so the modifications are almost neutral. On South German Credit, the configurations that lose at least one accuracy point on average are the three $\mathcal{BL}_{term}$ ones in the mono-batch setting (e1250). On COMPAS we can observe a more widespread degradation of utility, with 11 out of 12 configurations showing a decrease in utility for more than 50% of the splits, and the worst performance can reach a decrease of -0.0433 ± 0.0292 for the $w_{1.0}$/e50 configuration. Therefore, the answer to RQ2 is dataset-dependent.

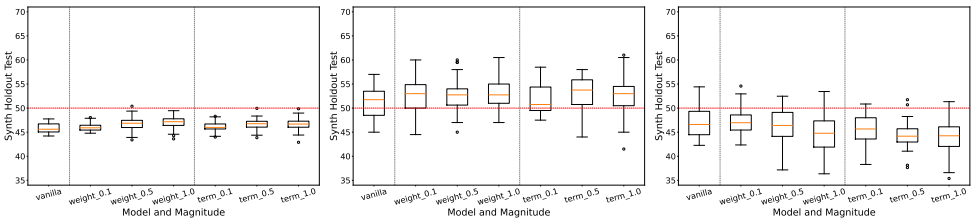
6.3 RQ3: Privacy

The results for the privacy measurement are shown in Figure 7 and Table 5. The privacy observation is based on the DCR S-H metric, which measures the percentage of synthetic rows whose distance from the nearest training record is less than their distance from the nearest holdout record. In Figure 7, the horizontal dashed line at 50% represents the privacy risk threshold: values above 50% indicate that synthetic records are systematically closer to training records than holdout records, revealing potential privacy leaks. Table 5 provides a comparison of the paired differences of the excess over 50 between the proposed loss modifications and the vanilla baseline. The % represents the ratio of split seeds for which the modification performs no worse than the vanilla version on the same split. A negative mean μ indicates that the modification carries less leakage than vanilla.

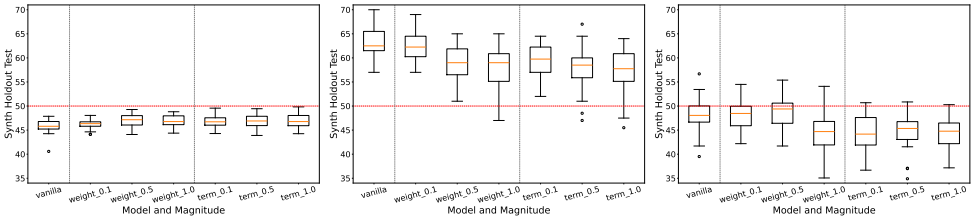
On Adult and COMPAS, the vanilla baseline is positioned below the 50% threshold: in these cases, the splits are considered worse than vanilla only if they fall above the threshold. For example, a split with a DCR S-H value of 49 is considered equal (i.e. not worse) if the vanilla one is also below 50 on the same split, while a split with a value of 51 is considered worse than vanilla (with difference equal to 1.0, i.e. the excess over 50).

This is the reason why on Adult 11 out of 12 configurations show 100% of splits that are no worse than vanilla and with $\mu \pm \sigma = 0.0 \pm 0.0$. However, this uniformity records the absence of memorisation on Adult rather than a benefit of the modification: no configuration (vanilla included) ever exceeds the threshold, with the single exception of one split at 50.39 under $w_{0.5}$/e50, which alone produces the 96.7% and $\mu \pm \sigma = +0.01 \pm 0.07$.

On South German Credit, the vanilla baseline is 51.52 ± 3.32 for multi-batch (e50) and 63.20 ± 3.49 for mono-batch (e1250). A characteristic of this dataset that is worth noting is that the holdout dataset is composed of a very small number of records (200), implying higher variability in the DCR S-H metric. All the configurations show a percentage of no worse splits $\geq$ 50.0%, with the exception of $t_{0.5}$/e50 (46.7%).



(a) DCR S-H Adult (e:50;bs:512) (b) DCR S-H S. German Credit (e:50;bs:512) (c) DCR S-H COMPAS (e:50;bs:512)



(d) DCR S-H Adult (e:1250;bs:mono-batch) (e) DCR S-H S. German Cr. (e:1250;bs:mono-batch) (f) DCR S-H COMPAS (e:1250;bs:mono-batch)

Figure 7: DCR Synthetic-Holdout (S-H) metric for the three datasets under investigation. The first row shows the results for a batch size of 512 and 50 epochs, while the second row shows the results for a mono-batch and 1,250 epochs. The vertical dashed lines subdivide the different configurations of the loss function. From left to right, these are the vanilla configuration; the $\mathcal{BL}_{weight}$ configuration, with magnitudes of 0.1, 0.5 and 1.0; and the $\mathcal{BL}_{term}$ configuration, with magnitudes of 0.1, 0.5 and 1.0. The horizontal dashed line at 50% represents the minimum privacy risk threshold: values above 50% indicate that synthetic records are indistinguishable from real records.

The strongest results are observed for the mono-batch configuration, since this is the only setting in which the vanilla model shows a substantial leak (63.20). There every proposed modification reduces it, by up to 5.62 points on average ($t_{1.0}$/e1250). The multi-batch configurations, conversely, show an increased μ in all configurations (from 0.68 to 1.22), with the exception of $t_{0.1}$/e50 (-0.10).

On COMPAS, the vanilla baseline is 47.16 ± 3.23 for multi-batch (e50) and 48.19 ± 3.60 for mono-batch (e1250). The results show that the $\mathcal{BL}_{term}$ consistently improves the privacy preservation, with a percentage of no worse splits $\geq 93.3\%$ and negative μ values for all configurations. $\mathcal{BL}_{weight}$ shows a more mixed behaviour, with a percentage of no worse splits $\geq 73.3\%$ and negative μ values for configurations with magnitudes equal to 0.5 and 1.0.

A relevant note about the validity of the results: on COMPAS, the variability is never lower than the mean difference between the proposed modifications and the vanilla baseline. On South German Credit we have $|\mu| \leq |\sigma|$ for $w_{1.0}$/e1250, $t_{0.5}$/e1250 and $t_{1.0}$/e1250.

Table 5: Privacy (RQ3). Paired differences using the DCR-S-H (excess over 50). The % represents the ratio of split seeds for which the modification performs **no worse** than the vanilla version on the same split. The %, the μ and the σ corresponds to the paired difference of the excess over 50, so a negative mean means the modification carries less leakage than vanilla. A seed where neither the modification nor vanilla goes above 50 contributes exactly zero, so a dataset that never leaks shows a column of zeros. The reference row is the vanilla baseline, w is the $\mathcal{BL}_{weight}$, t is the $\mathcal{BL}_{term}$.

Configuration	Adult	S. German Cr.	COMPAS
Ref. e50	$+45.88 \pm 1.06$	$+51.52 \pm 3.32$	$+47.16 \pm 3.23$
$w_{0.1}$/e50	100.0% ($+0.00 \pm 0.00$)	60.0% ($+0.68 \pm 3.64$)	86.7% ($+0.01 \pm 0.58$)
$w_{0.5}$/e50	96.7% ($+0.01 \pm 0.07$)	53.3% ($+0.97 \pm 3.45$)	96.7% (-0.20 ± 0.73)
$w_{1.0}$/e50	100.0% ($+0.00 \pm 0.00$)	56.7% ($+0.93 \pm 3.46$)	93.3% (-0.23 ± 1.07)
$t_{0.1}$/e50	100.0% ($+0.00 \pm 0.00$)	60.0% (-0.10 ± 3.39)	100.0% (-0.38 ± 0.87)
$t_{0.5}$/e50	100.0% ($+0.00 \pm 0.00$)	46.7% ($+1.22 \pm 3.45$)	100.0% (-0.33 ± 0.76)
$t_{1.0}$/e50	100.0% ($+0.00 \pm 0.00$)	50.0% ($+1.07 \pm 3.96$)	100.0% (-0.37 ± 0.82)
Ref. e1250	$+45.84 \pm 1.42$	$+63.20 \pm 3.49$	$+48.19 \pm 3.60$
$w_{0.1}$/e1250	100.0% ($+0.00 \pm 0.00$)	56.7% (-0.88 ± 4.04)	80.0% ($+0.01 \pm 1.37$)
$w_{0.5}$/e1250	100.0% ($+0.00 \pm 0.00$)	80.0% (-4.10 ± 4.58)	73.3% (-0.10 ± 1.28)
$w_{1.0}$/e1250	100.0% ($+0.00 \pm 0.00$)	83.3% (-4.85 ± 4.76)	100.0% (-0.54 ± 1.11)
$t_{0.1}$/e1250	100.0% ($+0.00 \pm 0.00$)	76.7% (-3.92 ± 4.82)	100.0% (-0.63 ± 1.47)
$t_{0.5}$/e1250	100.0% ($+0.00 \pm 0.00$)	90.0% (-5.50 ± 4.81)	93.3% (-0.63 ± 1.57)
$t_{1.0}$/e1250	100.0% ($+0.00 \pm 0.00$)	83.3% (-5.62 ± 5.34)	100.0% (-0.66 ± 1.50)

Answer to RQ3. The results show that only in 1 configuration-dataset combination out of 36 the percentage of not worse splits is below 50%. In all the other cases, it is $\geq 50\%$, reaching 100% on Adult and COMPAS. The mean differences are positive (more leakage) in 8 out of 36 combinations (5 in South German Credit multi-batch configuration), while on Adult almost every difference is exactly zero because no configuration (vanilla included) ever exceeds the threshold. The only setting in which the vanilla model shows a substantial leak is South German Credit in the mono-batch setting, and there every proposed modification reduces it. We can therefore conclude that the proposed modifications, on the tested datasets, do not degrade privacy relative to the vanilla baseline, and improve it where a leak is actually present.

7 Discussion

The results of the empirical evaluation, presented in the previous section, allow us to answer the research questions defined in Section 4.

With respect to **RQ1**, the results contradict the null hypothesis $H1_0$: the proposed modifications of the loss function produce a relevant improvement in balance, as measured by the *Simpson Imbalance Loss of the Dataset Index* (SILDI). In terms of fairness, we observe that $\mathcal{BL}_{term}$ configurations contradict $H1_0$. This is not the case for $\mathcal{BL}_{weight}$ configurations, which show a more mixed and dataset-dependent behaviour. With respect to **RQ2** and **RQ3**, the research questions ask whether the proposed modifications degrade, respectively, utility and privacy. In terms of utility, the results show that the proposed modifications affect the TSTR accuracy difference in a dataset-dependent manner: neutral on Adult, mixed on South German Credit, and mostly negative on COMPAS. In terms of privacy, the proposed modifications neither degrade nor improve privacy relative to the vanilla baseline. However, they do improve privacy in the only setting where a leak is present: the South German Credit mono-batch configuration. This configuration is also the one most prone to overfitting in the study due to the small size of the dataset. We hypothesize that the rebalancing of loss components encourages the model to generate synthetic records in sparsely populated areas of the data space. This increases the distance between synthetic and real records, thereby improving privacy. The inherent tension between fairness and privacy has been explored in other ML contexts [47]; in our setting the two dimensions move together rather than in opposition, since $\mathcal{BL}_{term}$ is the better formulation on both.

The comparison of the two configurations of the loss function, namely $\mathcal{BL}_{weight}$ and $\mathcal{BL}_{term}$, shows that the $\mathcal{BL}_{term}$ configuration has a more consistent impact on fairness. On balance the two are equivalent overall, winning nine head-to-head comparisons each at matched magnitude, but the outcome is governed by the training regime rather than by the formulation: $\mathcal{BL}_{weight}$ prevails in seven of the nine multi-batch comparisons and $\mathcal{BL}_{term}$ in seven of the nine mono-batch ones. The additive nature of the $\mathcal{BL}_{term}$ model, which penalises as a separate loss term the squared distance of the reconstructed categorical distribution from the uniform one, provides a more direct mechanism for encouraging the model to learn balanced representations. The exponential reweighting of the reconstruction loss used by the $\mathcal{BL}_{weight}$ model has a more diluted influence on the learning process, especially at lower magnitudes. The choice between the two is not free of cost: $\mathcal{BL}_{term}$ is preferable for fairness, while $\mathcal{BL}_{weight}$ preserves utility better in four of the six dataset and epoch settings, and on South German Credit the three $\mathcal{BL}_{term}$ configurations in mono-batch are the only ones losing at least one accuracy point. On privacy the ordering is reversed: $\mathcal{BL}_{term}$ reduces the leakage more than $\mathcal{BL}_{weight}$ in every setting where the metric moves at all, most clearly on South German Credit in mono-batch, where the average reduction is 5.01 against 3.28 points.

These findings suggest that modifying the loss function of a VAE-based generative model reliably improves balance, improves fairness under the $\mathcal{BL}_{term}$ formulation, and leaves privacy essentially untouched, at a utility cost that is negligible on Adult,

limited to the mono-batch $\mathcal{BL}_{term}$ configurations on South German Credit, and substantial on COMPAS. The results also emphasise the importance of carefully selecting the magnitude of the balance hyperparameter, as excessively high values can result in a degradation of utility. For practitioners, these results can be summarised as follows: when fairness is the priority and the utility budget allows it, use $\mathcal{BL}_{term}$; when utility must be preserved, use $\mathcal{BL}_{weight}$; when privacy leakage must be reduced, use $\mathcal{BL}_{term}$ again, which lowers the excess over the reference more than $\mathcal{BL}_{weight}$ in every setting where the baseline leaks at all. With $\mathcal{BL}_{weight}$ the desired balance should be achieved with the lowest possible magnitude, since its exponential reweighting concentrates the cost at the highest one. Privacy, however, pulls the magnitude the other way: in the only setting where the vanilla model leaks substantially, the reduction grows monotonically with the magnitude for both formulations, so a practitioner bound by both constraints has to pick a point on that trade-off rather than a default.

8 Threats to validity

In this section, we outline the major threats to validity as classified in the existing literature for empirical software engineering research [48].

8.1 Internal Validity

Internal validity concerns the extent to which the experimental design and execution allow us to draw valid conclusions about causal relationships within the study.

8.1.1 Selection of sensitive features for balance and fairness measurement

The choice of which features to consider as sensitive when measuring balance may affect the results. In our experiments, we selected features that are widely recognised as sensitive in the fairness literature, such as gender, *race* and nationality in the Adult dataset, primarily following the provisions of Article 21 of the EU Charter of Fundamental Rights [36]. However, different contexts or applications may require different sets of sensitive features, and alternative selections may lead to different balance and fairness outcomes.

8.1.2 Selection of privileged and unprivileged classes for fairness measurement

The selected fairness metrics require the identification of privileged and unprivileged groups within sensitive features. These designations follow the operational rule described in Section 5: the privileged group is the most common category of the attribute, with two exceptions in the COMPAS dataset. For example, in the Adult dataset, male was designated as privileged for the *sex* feature and white for the *race* feature. While the rule makes the assignment reproducible, the two exceptions in the COMPAS dataset reflect specific contextual choices that may not be applicable to all domains of application or cultural contexts. Different privilege assignments could lead to different fairness measurements.

8.2 External Validity

External validity refers to the extent to which our findings can be generalised to other contexts, populations or settings.

8.2.1 Type of generative model

Our study exclusively focuses on Variational AutoEncoders (VAEs) as the generative model architecture. Although VAEs have been shown to be effective for generating synthetic data, other generative models, such as Generative Adversarial Networks (GANs), normalised flows or diffusion models, may behave differently with regard to balancing sensitive features. Moreover, we considered only a specific implementation of a VAE, and different architectural choices (e.g. different encoder/decoder structures, latent space dimensionality, or alternative loss function formulations) could yield different results. In particular, the reconstruction term for categorical features is a fixed-variance Gaussian NLL rather than the cross-entropy more commonly adopted in the VAE literature (see Section 2). Since all our comparisons are paired against the same vanilla baseline, this choice does not affect the internal validity of the results, but the magnitude of the effects we report may not carry over to implementations adopting a different reconstruction loss. The proposed loss function modifications have not yet been tested with other architectures. Consequently, our findings cannot be generalised to other types of generative models without further investigation.

8.2.2 Type of dataset

All experiments were conducted on tabular datasets containing a mixture of categorical and continuous features. The generalisability of our approach to other data types, such as images, text, time series or graph-structured data, remains unexplored. The categorical nature of sensitive features in tabular data makes the balance problem particularly pertinent in this area. However, applying our approach to other data types would require adaptation and separate validation.

8.2.3 Number and diversity of datasets

Our empirical evaluation was conducted using three datasets: Adult, COMPAS and South German Credit. Although these datasets are widely used as benchmarks in fairness and machine learning research, they represent only a limited range of potential application domains. The fact that all three datasets share certain characteristics (binary classification tasks, presence of demographic-sensitive features and moderate size) may limit the generalisability of our findings. To fully assess the broader applicability of our approach, additional experiments on datasets from different domains with different characteristics (e.g. multi-class classification, regression tasks, larger-scale datasets and different types of imbalance) would be necessary. Furthermore, the specific imbalance patterns present in these datasets may not be representative of all the imbalance scenarios that practitioners might encounter in real-world applications. Raising the number of datasets under investigation would also allow a more robust statistical analysis, enabling the use of statistical tests that require independence of

observations. The low dimensionality of South German Credit implies higher noise in the results of the metric, especially the DCR S-H, as discussed in Section 6.3.

8.2.4 Batch and epoch configuration

The mono-batch configuration computes the balance loss for the entire dataset simultaneously. This requires all the records to be held in memory, so it is not feasible for large datasets without the appropriate hardware, as noted in Section 5. Consequently, the multi-batch configuration is the most realistic option at scale, and the two regimes do not behave identically: on balance, $\mathcal{BL}_{weight}$ prevails in seven of the nine multi-batch comparisons and $\mathcal{BL}_{term}$ in seven of the nine mono-batch ones. Therefore, practitioners working on datasets that do not fit in memory should refer to the multi-batch results.

9 Future work

The results of this study suggest several directions for future research. A first set of directions directly addresses the **threats to validity** identified in Section 8. Our proposed loss function modifications were evaluated exclusively on a specific VAE implementation; applying and validating the same modifications to other VAE architectures, including implementations that adopt a cross-entropy reconstruction term for categorical features, would allow assessing the generality of the approach beyond the architectural choices made in our industrial baseline. A broader evaluation should also extend to other classes of generative models, notably Generative Adversarial Networks (GANs), normalised flows and diffusion models, to determine whether the balance-oriented loss modifications can be adapted or whether equivalent mechanisms are more naturally expressed within those frameworks. Similarly, the empirical evidence presented here is based on three tabular benchmark datasets with binary classification tasks and demographic sensitive features. Replicating the study on datasets with different characteristics (including multi-class classification, regression targets, larger scale and heterogeneous types of imbalance) would strengthen the external validity of the findings and provide clearer guidance on the scope of applicability of the method. A more specific question concerns the dependence of the utility cost of the re-weighting formulation on the dataset: we hypothesise that this cost depends on the extent to which the predictive structure of the target is shared with the sensitive attributes. Testing this on datasets with a controlled degree of dependence between the two would clarify it.

A second direction concerns the **extension of the balance objective** itself. In the current formulation, the loss penalises imbalance independently for each individual sensitive feature. A natural extension would address the joint distribution of a sensitive feature and the target variable. This intersectionality approach could enhance the balance, but the impact on utility must be thoroughly studied.

A third direction concerns the **documentation and governance** of the synthetic data generation process. As synthetic datasets are increasingly used as drop-in replacements for or augmentations of real data in sensitive decision-making contexts, the question of how to communicate the provenance, generation parameters, balance

choices and known limitations of a synthetic dataset becomes practically important. Future work could investigate standardised reporting frameworks — analogous to model cards or datasheets for datasets — adapted to the specific characteristics of generated synthetic data, including the disclosure of the balance loss configuration and its effects on the distribution of sensitive features.

10 Conclusion

This paper addressed the problem of sensitive feature imbalance in tabular synthetic data generation based on Variational AutoEncoders. Motivated by the widespread use of VAE-based generative models in industrial settings and the well-documented risks of perpetuating or amplifying imbalance in downstream applications, we proposed two lightweight modifications to the VAE loss function: an additive balance term ($\mathcal{BL}_{term}$) and a Simpson-index-based reweighting term ($\mathcal{BL}_{weight}$). Both modifications penalise deviations from the uniform distribution of sensitive feature classes during training, without altering the architecture of the VAE or the structure of its latent space.

We evaluated the proposed modifications on three publicly available benchmark datasets — Adult, South German Credit and COMPAS — measuring their impact along four dimensions: balance (Simpson-based *Simpson Imbalance Loss of the Dataset Index* - SILDI), fairness (Statistical Parity Difference), utility (TSTR accuracy difference) and privacy (DCR Synthetic-Holdout). For each dataset, we generated 420 synthetic datasets across two batch-epoch configurations (mono-batch with 1,250 epochs and batch size 512 with 50 epochs) and seven loss function configurations (vanilla baseline, plus three magnitudes for each of the two proposed modifications), for a total of 1,260 synthetic datasets.

Both proposed modifications consistently reduce the SILDI, which decreases on all 30 splits in 34 of the 36 dataset–configuration cells, whereas the improvement in statistical parity is consistent under $\mathcal{BL}_{term}$ and irregular under $\mathcal{BL}_{weight}$ (**RQ1**). The $\mathcal{BL}_{term}$ configuration proved more robust and effective, zeroing the SILDI even at low magnitudes in the mono-batch regime, and improving fairness on all three datasets, although only on a modest majority of the splits on South German Credit. The $\mathcal{BL}_{weight}$ configuration, while effective in reducing imbalance, showed a weaker and less consistent impact on fairness, and exhibited instability at high magnitudes (magnitude 1.0) on the COMPAS dataset. Moreover, the proposed modifications leave privacy essentially untouched (**RQ3**), while the cost in utility is dataset-dependent (**RQ2**): negligible on Adult, limited to the mono-batch $\mathcal{BL}_{term}$ configurations on South German Credit, and substantial on COMPAS. On privacy, $\mathcal{BL}_{term}$ reduces the leakage more than $\mathcal{BL}_{weight}$ in every setting in which the vanilla baseline leaks at all.

Overall, these findings demonstrate that it is possible to steer a VAE-based synthetic data generator towards balanced sensitive feature distributions (and the associated fairness benefits) by means of a simple and tunable loss function modification, at a negligible cost in terms of privacy and at a cost in terms of utility that has to be assessed on the specific dataset. The $\mathcal{BL}_{term}$ formulation is the more effective choice where fairness is the priority, while $\mathcal{BL}_{weight}$ preserves utility better, so that the selection depends on which of the two constraints prevails in the specific setting.

Declarations

- **Funding**: This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

- **Competing interests**: L. Gilli and S. Kurapati are co-founders of Clearbox AI (Turin, Italy), in which they hold a financial interest. The work presented in this article builds upon a codebase originally developed by Clearbox AI. The remaining authors have no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate**: Not applicable.
- **Consent for publication**: Not applicable.
- **Data availability**: The three benchmark datasets analysed during the current study are publicly available: Adult and South German Credit in the UCI Machine Learning Repository, and COMPAS in the ProPublica *compas-analysis* repository. The synthetic datasets generated during the study, together with the train/test splits they were generated from and the per-run metrics, are deposited at https://doi.org/10.5281/zenodo.22348572; they can equivalently be regenerated from the original datasets using the code, the configurations and the random seeds available in the repository https://github.com/RondinaMR/VAE-Balance-Loss.
- **Materials availability**: Not applicable (see Code availability).
- **Code availability**: The code related to the proposed loss function modifications, along with the materials needed to reproduce the empirical validation study, is archived at https://doi.org/10.5281/zenodo.22348572 and is available at https://github.com/RondinaMR/VAE-Balance-Loss.
- **Author contribution** (CRediT taxonomy): Conceptualization: all authors; Funding acquisition: A. Vetrò, J. C. De Martin; Investigation: M. Rondina; Methodology: M. Rondina, A. Vetrò, L. Gilli; Resources: A. Vetrò, L. Gilli, S. Kurapati, J. C. De Martin; Software: M. Rondina; Supervision: A. Vetrò, L. Gilli, S. Kurapati, J. C. De Martin; Validation: M. Rondina, A. Vetrò, L. Gilli; Visualization: M. Rondina; Writing - original draft: M. Rondina; Writing - review & editing: M. Rondina, A. Vetrò, L. Gilli.

References

[1] Kilkenny, M.F., Robinson, K.M.: Data quality: "Garbage in – garbage out". Health Information Management Journal **47**(3), 103–144 (2018) https://doi.org/10.1177/1833358318774357

[2] European Parliament, Council of the European Union: Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828

(Artificial Intelligence Act) (2024)

[3] Mirowski, P.: The future(s) of open science. Social Studies of Science **48**(2), 171–203 (2018) https://doi.org/10.1177/0306312718772086

[4] European Parliament, European Council: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance) (2016)

[5] European Union Agency for Fundamental Rights: EU Charter of Fundamental Rights - Title II: Freedoms - Article 8 - Protection of Personal Data (2015)

[6] Barocas, S., Selbst, A.D.: Big Data's Disparate Impact. Social Science Research Network, Rochester, NY (2016). https://doi.org/10.2139/ssrn.2477899

[7] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ITCS '12, pp. 214–226. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2090236.2090255

[8] Lu, Y., Wang, H., Wei, W.: Machine Learning for Synthetic Data Generation: A Review. arXiv (2023). https://doi.org/10.48550/arXiv.2302.04062

[9] De Cristofaro, E.: What Is Synthetic Data? The Good, The Bad, and The Ugly. arXiv (2023). https://doi.org/10.48550/arXiv.2303.01230

[10] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., Weller, A.: Synthetic Data – What, Why and How? arXiv (2022). https://doi.org/10.48550/arXiv.2205.03257

[11] Man, K., Chahl, J.: A Review of Synthetic Image Data and Its Use in Computer Vision. Journal of Imaging **8**(11), 310 (2022) https://doi.org/10.3390/jimaging8110310

[12] Abdullin, Y., Molla-Aliod, D., Ofoghi, B., Yearwood, J., Li, Q.: Synthetic Dialogue Dataset Generation Using LLM Agents. arXiv (2024). https://doi.org/10.48550/arXiv.2401.17461

[13] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. arXiv (2022). https://doi.org/10.48550/arXiv.1312.6114

[14] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems, vol. 27, pp. 1–9. Curran Associates, Inc., Red Hook, NY, USA (2014)

[15] Endres, M., Mannarapotta Venugopal, A., Tran, T.S.: Synthetic Data Generation: A Comparative Study. In: Proceedings of the 26th International Database Engineered Applications Symposium. IDEAS '22, pp. 94–102. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3548785.3548793

[16] Koubeissy, H., Amine, A., Kamradt, M., Makhoul, A.: Survey on Tabular Data Privacy and Synthetic Data Generation in Industry 4.0. Applied Intelligence **55**(13), 935 (2025) https://doi.org/10.1007/s10489-025-06823-5

[17] Sohn, K., Lee, H., Yan, X.: Learning Structured Output Representation using Deep Conditional Generative Models. In: Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc., Red Hook, NY, USA (2015)

[18] Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational Autoencoder for Deep Learning of Images, Labels and Captions. In: Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc., Red Hook, NY, USA (2016)

[19] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 7335–7345. Curran Associates Inc., Red Hook, NY, USA (2019)

[20] Rybkin, O., Daniilidis, K., Levine, S.: Simple and Effective VAE Training with Calibrated Decoders. In: Proceedings of the 38th International Conference on Machine Learning, pp. 9179–9189. PMLR, virtual (2021)

[21] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

[22] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE, Hong Kong, China (2008). https://doi.org/10.1109/IJCNN.2008.4633969

[23] Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: Fairness-aware Generative Adversarial Networks. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 570–575. IEEE, Seattle, WA, USA (2018). https://doi.org/10.1109/BigData.2018.8622525

[24] Zhao, Z., Kunar, A., Birke, R., Chen, L.Y.: CTAB-GAN: Effective Table Data Synthesizing. In: Proceedings of The 13th Asian Conference on Machine Learning, pp. 97–112. PMLR, virtual (2021)

[25] Rajabi, A., Garibay, O.O.: TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks. Machine Learning and Knowledge Extraction **4**(2), 488–501 (2022) https://doi.org/10.3390/make4020022

[26] van Breugel, B., Kyono, T., Berrevoets, J., van der Schaar, M.: DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. In: Advances in Neural Information Processing Systems, vol. 34, pp. 22221–22233. Curran Associates, Inc., Red Hook, NY, USA (2021)

[27] Gujar, S., Shah, T., Honawale, D., Bhosale, V., Khan, F., Verma, D., Ranjan, R.: GenEthos: A Synthetic Data Generation System With Bias Detection And Mitigation. In: 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), pp. 1–6. IEEE, Kochi, India (2022). https://doi.org/10.1109/IC3SIS54991.2022.9885653

[28] Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The Variational Fair Autoencoder. arXiv (2017). https://doi.org/10.48550/arXiv.1511.00830

[29] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A Kernel Method for the Two-Sample-Problem. In: Advances in Neural Information Processing Systems, vol. 19. MIT Press, Cambridge, MA, USA (2006)

[30] Grari, V., Lamprier, S., Detyniecki, M.: Fairness without the sensitive attribute via Causal Variational Autoencoder. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pp. 696–702. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria (2022). https://doi.org/10.24963/ijcai.2022/98

[31] Wan, Z., Zhang, Y., He, H.: Variational autoencoder based synthetic data generation for imbalanced learning. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–7. IEEE, Honolulu, HI, USA (2017). https://doi.org/10.1109/SSCI.2017.8285168

[32] Huang, K., Wang, X.: ADA-INCVAE: Improved data generation using variational autoencoder for imbalanced classification. Applied Intelligence **52**(3), 2838–2853 (2022) https://doi.org/10.1007/s10489-021-02566-1

[33] Hong, S., An, S., Jeon, J.-J.: Improving SMOTE via fusing conditional VAE for data-adaptive noise filtering. Applied Intelligence **55**(12), 841 (2025) https://doi.org/10.1007/s10489-025-06692-y

[34] Güldal, S., Gümüş, İ.H.: Balancing the dataset by generating new synthetic data based on heinz mean in medical data. Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi **22**(3), 570–576 (2022) https://doi.org/10.35414/akufemubid.1011058

[35] Chadebec, C., Allassonnière, S.: Data Augmentation with Variational Autoencoders and Manifold Sampling. In: Engelhardt, S., Oksuz, I., Zhu, D., Yuan, Y., Mukhopadhyay, A., Heller, N., Huang, S.X., Nguyen, H., Sznitman, R., Xue, Y. (eds.) Deep Generative Models, and Data Augmentation, Labelling, and Imperfections. Lecture Notes in Computer Science, pp. 184–192. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88210-5_17

[36] European Union Agency For Fundamental Rights: EU Charter of Fundamental Rights - Title III: Quality - Article 21 - Non-Discrimination (2015)

[37] Rini van Solingen, Basili, V., Caldiera, G., Rombach, H.D.: Goal Question Metric (GQM) Approach. In: J.J. Marciniak (ed.) Encyclopedia of Software Engineering. John Wiley & Sons, Ltd, USA (2002). https://doi.org/10.1002/0471028959.sof142

[38] Mecati, M., Vetrò, A., Torchiano, M.: Detecting Risk of Biased Output with Balance Measures. Journal of Data and Information Quality **14**(4), 25–1257 (2022) https://doi.org/10.1145/3530787

[39] Mecati, M., Torchiano, M., Vetrò, A., Martin, J.C.: Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications. IEEE Access **11**, 26996–27011 (2023) https://doi.org/10.1109/ACCESS.2023.3252370

[40] Esteban, C., Hyland, S.L., Rätsch, G.: Real-Valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv (2017). https://doi.org/10.48550/arXiv.1706.02633

[41] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785

[42] Brunelli, D., Kurapati, S., Gilli, L.: SURE: A New Privacy and Utility Assessment Library for Synthetic Data. In: 2024 IEEE International Conference on Blockchain (Blockchain), pp. 643–648. IEEE, Copenhagen, Denmark (2024). https://doi.org/10.1109/Blockchain62396.2024.00094

[43] Gower, J.C.: A General Coefficient of Similarity and Some of Its Properties. Biometrics **27**(4), 857–871 (1971) https://doi.org/10.2307/2528823 2528823

[44] Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). https://doi.org/10.24432/C5XW20

[45] South German Credit. UCI Machine Learning Repository (2020). https://doi.org/10.24432/C5QG88

[46] Propublica/Compas-Analysis. ProPublica (2026)

[47] van der Steen, F., Vink, F., Kaya, H.: Privacy constrained fairness estimation for decision trees. Applied Intelligence **55**(5), 308 (2025) https://doi.org/10.1007/s10489-024-05953-6

[48] Feldt, R., Magazinius, A.: Validity threats in empirical software engineering research-an initial survey. In: Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering, pp. 374–379. Knowledge Systems Institute Graduate School, Redwood City, San Francisco Bay, CA, USA (2010)