This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this contribution is published in "Applied Intelligence 56, 548 (2026)", and is available online at https://doi.org/10.1007/s10489-026-07532-3

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 context 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 elements of a Variational AutoEncoder (VAE) training loss function that penalise deviations from a balanced distribution of sensitive feature classes. These elements are an additive balance component ($\boldsymbol{\mathcal{BL}_{term}}$) and a Simpson-index-based reweighting component ($\boldsymbol{\mathcal{BL}_{weight}}$). 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 $\boldsymbol{\mathcal{BL}_{term}}$ reduces sensitive feature imbalance and improves statistical parity, while preserving utility and privacy at levels comparable to the baseline. The $\boldsymbol{\mathcal{BL}_{weight}}$ 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, by definition depends on the utility of training data [1]. The principle of transparency, which is required by norms [2], 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, privacy and improve 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 originated 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

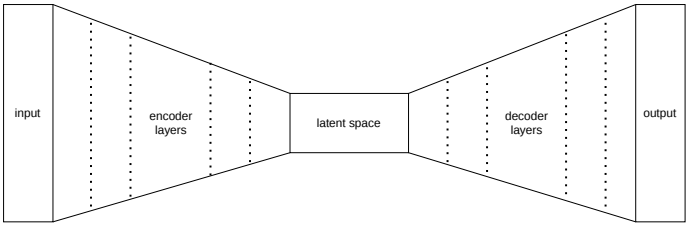


Figure 1: Variational AutoEncoder architecture

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 describe the methodology used for the empirical evaluation. In Section 6 and 7 we present the obtained results and discuss the relative implications. Finally, Section 8 expose the limitations of our findings, Section 9 shows hypothesis of future improvements and in Section 10 are write down 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

When we talk about synthetic data, we are referring to the process in which the original dataset is used as input for the generative model to produce the synthetic dataset. 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 Knigma and Welling [13] is a neural network architecture based on a probabilistic latent space. Figure 1 shows a schematic representation of the architecture of a VAE. 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 translate 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]. Due to the more complex relationship between categorical features and pixels in images, 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 at 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.

Our work differs from all the above in that we do not alter the architecture of the VAE: instead, we propose two modifications of the loss function that directly penalise deviations from the uniform distribution of sensitive feature classes, providing a lightweight and tunable mechanism to steer the generation towards balance.

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[1].

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 standard scaling to continuous features 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 over numerical features, where the squared errors for each feature are normalized by the variance: this allows weighting the contribution of individual continuous features independently of their scale after preprocessing. The term $\mathcal{L}_{\text{cat}}$ operates on categorical features represented as one-hot encodings: it is likewise an MSE between the softmax probabilities produced by the decoder and the original one-hot vectors, normalized 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, which limits overfitting to the training distribution. The term $\mathcal{L}_{\text{KLD}} = -\frac{\alpha}{2}\,\mathbb{E}\left[1 + \log\sigma_z^2 - \mu_z^2 - \sigma_z^2\right]$ 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 $\alpha = 1$; 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 focuses 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. The ideal distribution of a sensitive feature has been identified in the uniform distribution, i.e. where all class frequencies are equal as shown in Eq. 4, where m is the number of classes.

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

[1] https://anonymous.4open.science/r/VAE-Balance-Loss-5730/

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 distance of 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 in 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 $\mathcal{BM}$, function 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 cross-entropy loss term (not affected by our modification) and the balance loss.

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 equilibrium) and 1 (minimum equilibrium).

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

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

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

where $\mathcal{M}_{exp}$ represent an adjustable exponential magnitude term. Following fine-tuning, we set $\mathcal{M}_{exp} = 10.0$ in our experiment. The balance magnitude, denoted by $\mathcal{BM}$, function as the knob employed to adjust the intensity of the balance correction. The Simpson index is computed as detailed in (9).

4 Research Question

In this section we provide a description of the goal that drives this research. We follow the Goal-Question-Metric approach [36]: 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 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** improve 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 affect the **utility**?
RQ3 Does the proposed loss function modification affect the **privacy**?

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

$H1_0$ The proposed loss function modification doesn't improve the **balance** and **fairness** of sensitive features;
$H2_0$ The proposed loss function modification doesn't affect the **utility**;
$H3_0$ The proposed loss function modification doesn't affect 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-Validation

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 2 presents the experiment design.

4.1.1 Balance

The chosen balance measure is a diversity measure used in ecology: the Simpson index [37, 38]. 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 [37].

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

We used the Simpson index to obtain a *Simpson Imbalance Loss of the Dataset Index* (SILDI), a unique 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 based on common social norms. 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 [39]. 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 [40] 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) [41], based on the Gower's distance [42]. 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:

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

It can be concluded that when $\text{DCR}(s) = 0$, it indicates that s is a replication of 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 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 discriminate if a specific row of the training dataset T or of the validation dataset V was used to train the model.

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 2 summarizes the experimental design. The different RQs test different parts of the process:

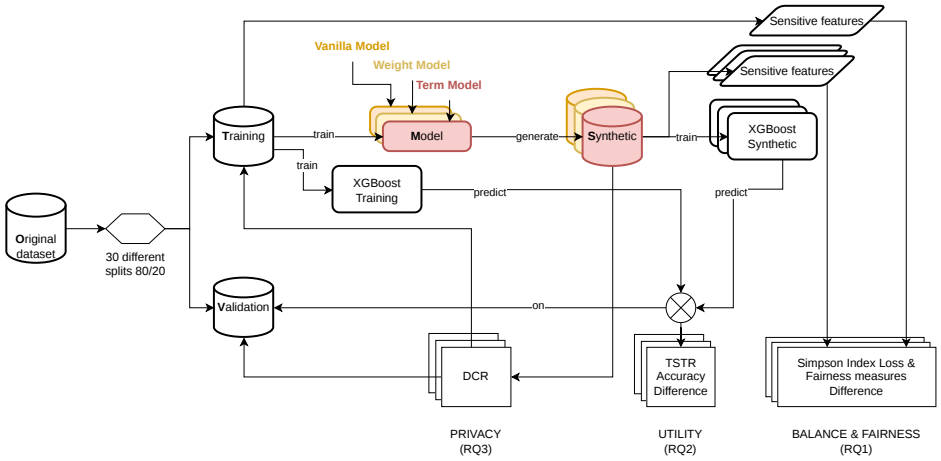


Figure 2: Experiment design.

RQ1 compare the sensitive features of the Training dataset T and of the Synthetic dataset S;

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. Different $\mathcal{BM}$ allow us to evaluate the impact of the balance loss on the generated synthetic dataset and to identify the best configuration for the proposed modifications. 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) is calculated batch by batch, this solution provides a more precise evaluation of the balance loss, since all classes of the entire dataset are taken into account. The downside of this solution is that column by column, all rows must be loaded into memory and this may be unfeasible for some datasets without appropriate hardware. This is why we also include a batch size equal to 512: this size makes it possible to limit the batch size while still ensuring a sufficiently large size to measure an adequate 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: Privileged and unprivileged groups for each sensitive feature, per dataset, used in fairness evaluation.

Dataset	Sensitive Feature	Privileged Group	Unprivileged Group
Adult	*sex*	Male	Female
Adult	*race*	White	All other races
Adult	*native_country*	United States	All other countries
S. German Credit	*famges* (personal status/sex)	Male, married/widowed	All other groups
S. German Credit	*gastarb* (foreign worker)	Non-foreign worker	Foreign worker
COMPAS	*sex*	Male	Female
COMPAS	*race*	Caucasian	All other races
COMPAS	*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 [43], South German Credit [44] and COMPAS [45]. 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 grouping the low and medium risk classes together and considering them both as the positive class, while considering the high risk class as the negative class. We selected the privileged and unprivileged groups for each sensitive feature according to common social norms. This was necessary in order to compute the statistical parity difference (see equation 11). Table 1 summarises the privileged and 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. 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).

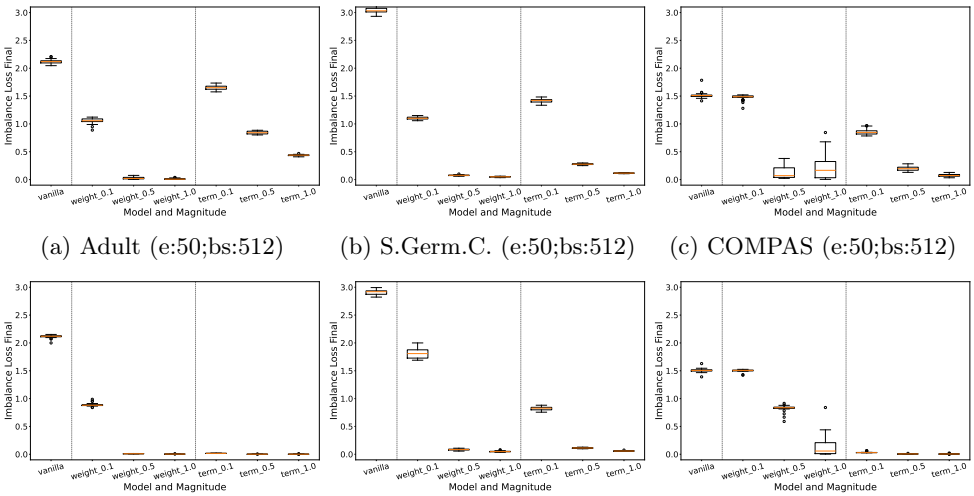
6 Experimental Results

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 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 allow us to visually compare the performance of the different configurations and to identify any significant differences in terms of balance, utility and privacy. The main goal of the figures with balance, utility and privacy metrics 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 unique value for the entire dataset as the other metrics.

6.1 RQ1: Balance and Fairness

The results for the balance measurement are shown in Figure 3. The proposed modifications of the loss function show a significant improvement in terms of balance, as measured by the *Simpson Imbalance Loss of the Dataset Index* (SILDI), for all the three datasets under investigation. In general, the $\mathcal{BL}_{weight}$ configuration appears to have a weaker impact on balance compared to the $\mathcal{BL}_{term}$configuration, which shows a more consistent improvement across all the tested magnitudes. In the mono-batch configurations, the $\mathcal{BL}_{term}$configuration is able to zero 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 similar result. The analysis of results for magnitudes equals to 0.01 and 0.05 (not shown in the figure) present the gradual improvement of the metrics.

By observing the **fairness** improvement for the Adult dataset (Figure 4), we can see that the $\mathcal{BL}_{term}$configuration shows a significant improvement in terms of statistical parity difference for all the sensitive features, while the $\mathcal{BL}_{weight}$ configuration appear to perform similarly to the vanilla model. In the case of *sex* sensitive feature, the $\mathcal{BL}_{term}$configuration shows also a gradual improvement with increasing magnitudes.



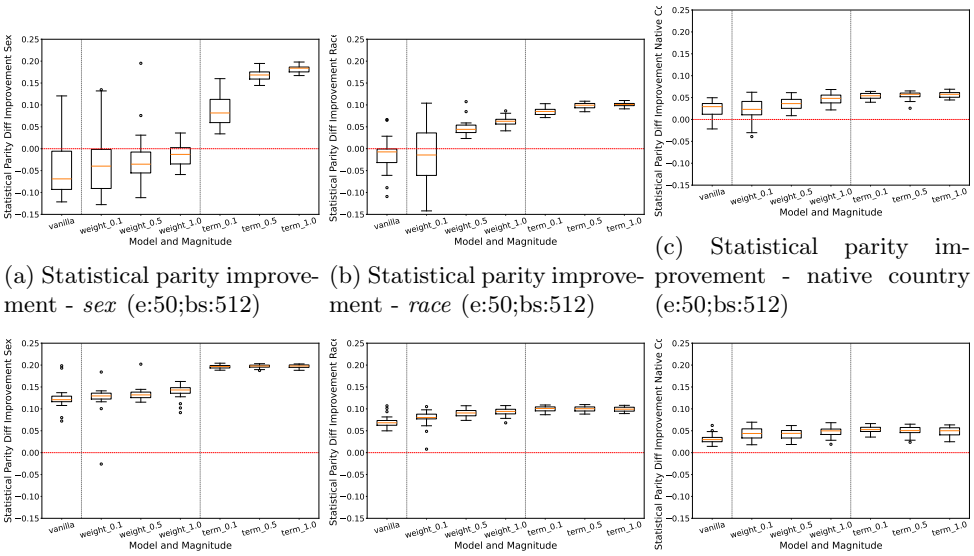
(a) Adult (e:50;bs:512) (b) S.Germ.C. (e:50;bs:512) (c) COMPAS (e:50;bs:512)

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

Figure 3: 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 unique value for a given synthetic dataset. The lower the SILDI, the better the balance of the sensitive features in the generated synthetic dataset.

The results for the South German Credit dataset are shown in Figure 5. 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 can observe a significative improvment 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.

The results for the COMPAS dataset are shown in Figure 6. The sensitive features under investigation are *sex*, *race* and *age* category. The $\mathcal{BL}_{weight}$ configurations present a pitfall for the *sex* sensitive feature, in the case of magnitude equal to 1.0, showing a significant 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 metter of the average results but of the distribution of the results, which is more spread out and includes more negative values compared to the vanilla



(a) Statistical parity improvement - *sex* (e:50;bs:512)

(b) Statistical parity improvement - *race* (e:50;bs:512)

(c) Statistical parity improvement - native country (e:50;bs:512)

(d) Statistical parity improvement - *sex* (e:1250;bs:mono-batch)

(e) Statistical parity improvement - *race* (e:1250;bs:mono-batch)

(f) Statistical parity improvement - native country (e:1250;bs:mono-batch)

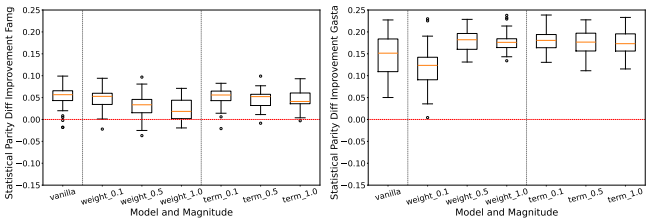
Figure 4: Fairness measurements improvement for the Adult dataset.

model. The $\mathcal{BL}_{term}$configurations, on the other hand, show a significant improvement in terms of statistical parity difference for all the sensitive features, with a gradual improvement for increasing magnitudes.

Answer to RQ1. The proposed modifications of the loss function show a significant improvement in terms of balance, as measured by the *Simpson Imbalance Loss of the Dataset Index* (SILDI) and the statistical parity difference, for all the three datasets under investigation. In general, the $\mathcal{BL}_{weight}$ configuration appears to have a weaker impact on balance compared to the $\mathcal{BL}_{term}$configuration, which shows a more consistent improvement across all the tested magnitudes.

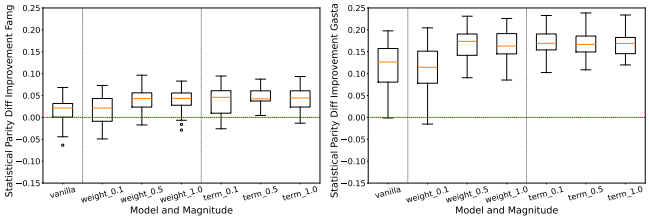
6.2 RQ2: Utility

The results for the utility measurement are shown in Figure 7. In general, the measure used to evaluate utility, namely the TSTR (Train on Synthetic, Test on Real) accuracy difference between the synthetic dataset and the original training dataset, is structured around the value of 0, which represents the threshold at which the synthetic dataset achieves the same classification accuracy as the original training dataset. During the 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.



(a) Statistical parity improvement - famges (e:50;bs:512)

(b) Statistical parity improvement - gastarb (e:50;bs:512)



(c) Statistical parity improvement - famges (e:1250;bs:mono-batch)

(d) Statistical parity improvement - gastarb (e:1250;bs:mono-batch)

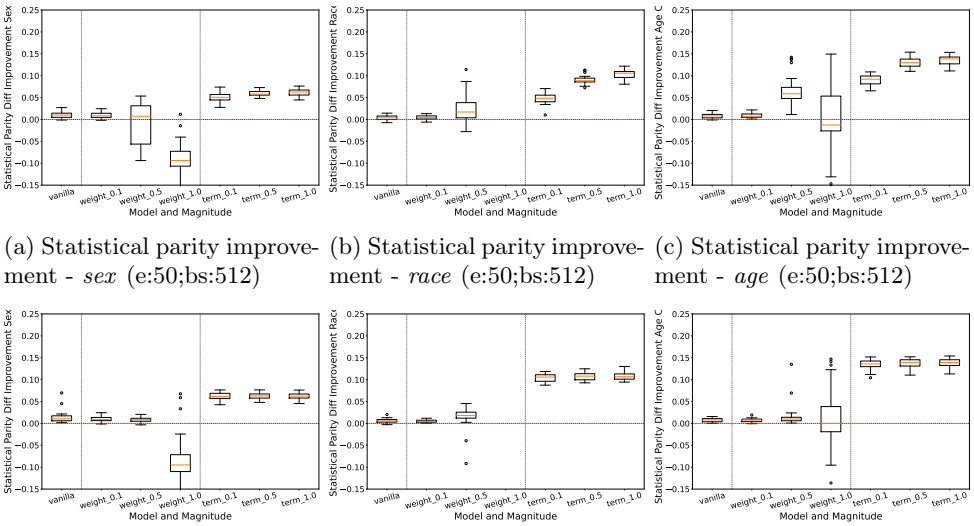
Figure 5: Fairness measurements improvement for the South German Credit dataset.

In fact, what we can see is that the results for all configurations of the loss function, and for all datasets, remain comparable to those of the vanilla model, that is, the model that does not include modifications to the loss function aimed at rebalancing. In the configuration with mono-batch and 1,250 epochs, the South German Credit dataset shows a slight decrease in utility for the $\mathcal{BL}_{term}$configuration, which is more evident for a magnitude of 0.5. In the case of the COMPAS dataset, the worse performance is observed for the $\mathcal{BL}_{weight}$ configuration with a magnitude of 1.0, which shows a significant decrease in utility compared to the vanilla model. However, the difference between the different configurations and the vanilla model appears to be not significant, since the distributions of the results are largely overlapping and the median values are close to each other.

Answer to RQ2. The proposed modifications of the loss function do not appear to have a significant impact on utility, as measured by the TSTR accuracy difference, for all the three datasets under investigation.

6.3 RQ3: Privacy

The results for the privacy measurement are shown in Figure 8. By observing the DCR S-H metric (distance to closest record between the synthetic and holdout datasets), we can see that the configuration with a batch size of 512 and 50 epochs produces comparable results to the vanilla model for the Adult and South German Credit datasets. For the COMPAS dataset, however, the $\mathcal{BL}_{weight}$ model with a magnitude



(a) Statistical parity improvement - *sex* (e:50;bs:512)

(b) Statistical parity improvement - *race* (e:50;bs:512)

(c) Statistical parity improvement - *age* (e:50;bs:512)

(d) Statistical parity improvement - *sex* (e:1250;bs:mono-batch)

(e) Statistical parity improvement - *race* (e:1250;bs:mono-batch)

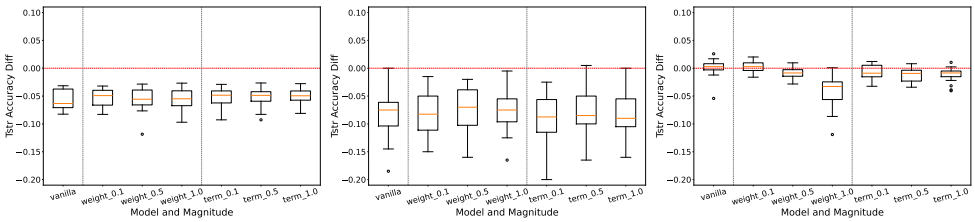
(f) Statistical parity improvement - *age* (e:1250;bs:mono-batch)

Figure 6: Fairness measurements improvement for the COMPAS dataset.

of 1.0 and the $\mathcal{BL}_{term}$model with all magnitudes show a slight decrease in privacy. Although the mean values for the various loss function configurations are similar, the distributions of the results are more varied for configurations with higher balance magnitudes.

In the mono-batch configuration with 1,250 epochs, the results for the Adult dataset are practically identical to those of the vanilla model, with minimal variation between configurations of the loss function. In this batch-epoch configuration, the South German Credit dataset is characterised by a significant deviation of the vanilla model's results from the 50% privacy risk threshold. In this case, both the $\mathcal{BL}_{weight}$ and $\mathcal{BL}_{term}$ models of the proposed loss function modification show a slight improvement in privacy, approaching the 50% threshold, especially for configurations with higher balance magnitudes. For the COMPAS dataset, the $\mathcal{BL}_{weight}$ configuration with a magnitude of 1.0 and the $\mathcal{BL}_{term}$ configuration with all magnitudes show a slight decrease in privacy compared to the vanilla model. However, the mean values still represent a relatively low privacy risk, decreasing from around 50% to around 45%.

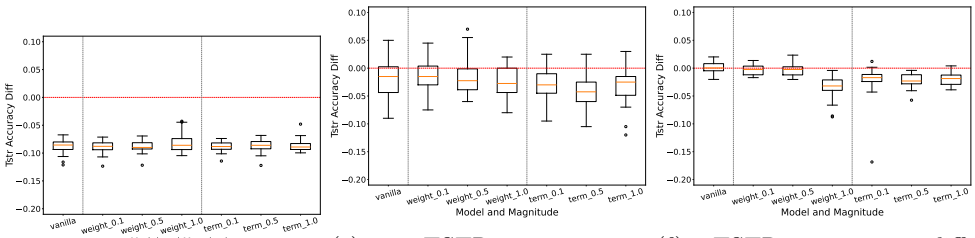
Answer to RQ3. In general, the proposed modifications of the loss function presents comparable results in terms of privacy, as measured by the DCR S-H metric, for all the three datasets under investigation.



(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)

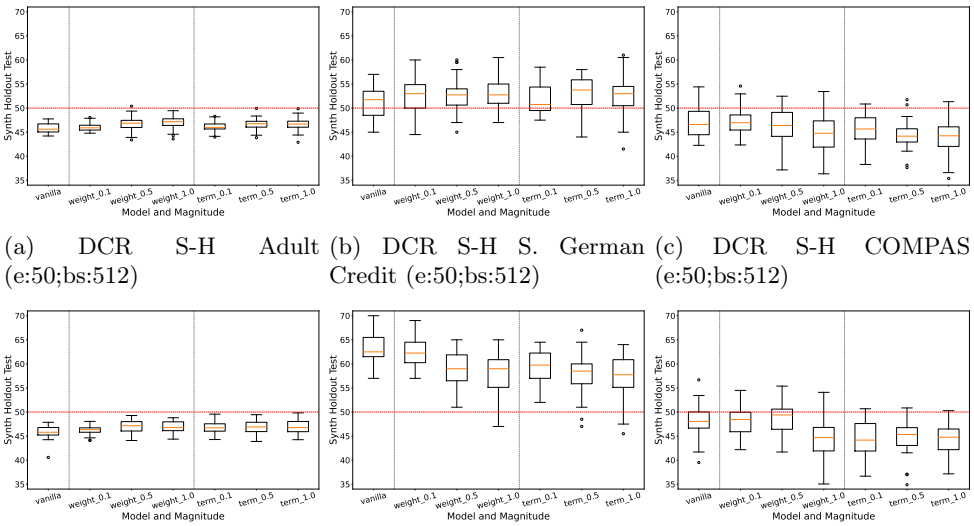
(f) TSTR accuracy diff. COMPAS (e:1250;bs:mono-batch)

Figure 7: 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.

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**, we can reject the null hypothesis $H1_0$: the proposed modifications of the loss function produce a statistically significant improvement in balance, as measured by the *Simpson Imbalance Loss of the Dataset Index* (SILDI), and a consistent improvement in fairness, as measured by the Statistical Parity Difference, across all three datasets under investigation. With respect to **RQ2** and **RQ3**, the research questions ask whether the proposed modifications *affect* utility and privacy rather than whether they improve them. In this context, failing to reject the corresponding null hypotheses $H2_0$ and $H3_0$ does not imply the absence of any effect: it means that no statistically significant degradation was observed within the scope of



(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 8: 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.

this study. The results show that the balance-oriented modifications do not meaningfully degrade utility or privacy relative to the vanilla baseline. The distributions of the TSTR accuracy difference and the DCR S-H metric for the proposed configurations largely overlap with those of the vanilla model, and the median values remain close across all datasets and configurations. Isolated exceptions (such as the decrease in utility for *weight* at magnitude 1.0 on COMPAS, or the marginal privacy reduction for *term* configurations in the mono-batch regime) are confined to specific high-magnitude and high-capacity training settings and can be mitigated through appropriate selection of the balance magnitude hyperparameter.

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 and significant impact on balance and fairness improvement, while the $\mathcal{BL}_{weight}$ configuration appears to have a weaker impact on these metrics. The additive nature of the $\mathcal{BL}_{term}$ model may provide a more direct and effective mechanism for encouraging the model to learn balanced representations, while the multiplicative nature of the $\mathcal{BL}_{weight}$ model may result in a more diffuse influence on the learning process, especially at lower

magnitudes. For this reason, the $\mathcal{BL}_{term}$ model may be a more stable choice for practitioners looking to improve balance without compromising utility or privacy.

Although the results of RQ2 were positive, the analysis of the trade-off with respect to the utility metric, as measured by the difference in TSTR accuracy, presents an interesting exception. The $\mathcal{BL}_{weight}$ model with the COMPAS dataset exhibits more pronounced decreases for the configuration with the maximum magnitude. This suggests, at high magnitudes, the $\mathcal{BL}_{weight}$ model may deviate from the original data distribution, which could negatively impact the utility of the generated data.

Analysis of the trade-off with respect to the privacy metric, as measured by the DCR-S-H, shows that the proposed modifications to the loss function do not significantly impact privacy. In some cases, we observe a slight decrease in the privacy preservation. The small magnitude of this decrease highlights the favourable trade-off of the proposed solutions. In other cases, we can observe a slight improvement. 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. Overall, the differences in these results suggest that the impact of the proposed loss function component on privacy is negligible. The inherent tension between fairness and privacy has been explored in other ML contexts [46]; our results confirm that a similar trade-off can be managed effectively in VAE-based synthetic data generation.

These findings suggest that improved balance and fairness can be achieved in the generation of synthetic data without incurring significant utility or privacy costs by modifying the loss function of a VAE-based generative model. 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 and privacy. Overall, the proposed configuration of the loss function *term* appears to be a promising approach for practitioners seeking to generate balanced synthetic data while maintaining acceptable levels of utility and privacy.

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 [47].

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 [48]. 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 were made based on sociological and demographic considerations that are consistent with the literature on fairness. For example, in the Adult dataset, male was designated as privileged for the *sex* feature and white for the *race* feature. While these choices align with established fairness research, they reflect specific contextual assumptions 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. The proposed loss function modifications have not yet been tested with other architectures. Consequently, our findings cannot be generalised to other types of generative model 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.

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 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 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 significantly reduce the SILDI across all datasets and configurations, and consistently improve statistical parity for the sensitive features under investigation (**RQ1**). The $\mathcal{BL}_{term}$ configuration proved more robust and effective, zeroing the SILDI already at low magnitudes in the mono-batch regime, and producing consistent fairness improvements across all datasets. The $\mathcal{BL}_{weight}$ configuration, while effective in reducing imbalance, showed a weaker and less consistent impact on fairness, and exhibited an instability at high magnitudes (magnitude 1.0) on the COMPAS dataset. Moreover, the proposed modifications do not produce a statistically significant degradation of utility (**RQ2**) or privacy (**RQ3**) relative to the vanilla baseline. The distributions of the TSTR accuracy difference and the DCR S-H metric for the proposed configurations largely overlap with those of the vanilla model across all datasets, with isolated exceptions at high magnitudes that can be controlled through hyperparameter selection.

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, without incurring meaningful costs in terms of data utility or privacy. The $\mathcal{BL}_{term}$ formulation represents a practical and effective choice for practitioners operating in this 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 (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 (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 (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 (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 (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, 250–576 (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] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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 (2024). https://doi.org/10.1109/Blockchain62396.2024.00094

[42] 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

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

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

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

[46] 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

[47] 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)

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