Fairness and Dependency Preservation in Synthetic Credit Data

Rayan Upadhyay[1]

Received October 18, 2025
Accepted March 29, 2026
Electronic access April 15, 2026

In this study, a baseline ensemble model- trained on the original German Credit dataset without synthetic augmentation- showed the most stable fairness-utility trade-off whereas SMOTE introduced the largest distortions. This study focuses on various techniques for generating synthetic data while preserving feature interdependence during fairness-aware preprocessing of loan approval models. Loan approval models rely on interdependent applicant features. Even without sensitive attributes, correlated features can induce unfair outcomes, making dependency preservation critical. Models were trained on the German Credit dataset without synthetic augmentation and compared against SMOTE, CTGAN, TVAE, and Gaussian Copula on the basis of ROC AUC and standard fairness metrics (comparative analysis only included methods that produced valid synthetic datasets). While the baseline consistently maintained interdependencies and provided stable fairness-utility trade-offs, synthetic methods distorted interdependencies and introduced instabilities, with no single method outperforming the baseline.

Introduction

The activity of preparing datasets for machine learning and related tasks is called data preprocessing. The process is usually managed through imputing missing values and bias audit (checking whether features differ meaningfully across protected groups), along with their class balances with SMOTE or reweighting samples if they are too low/high[1,2]. Organizations are increasingly using synthetic data in preprocessing to resemble original data, but not interfere with existing records. By employing this approach, an improvement can be manufactured in predictive performance, fairness, accuracy, and precision[3]. In important decision fields like loan approval, correlation among features plays a crucial role. Factors like income, age, working status and personal status are often highly correlated. Although these relationships are organic; these can indirectly transmit prejudice. Even when sensitive attributes such as gender or nationality are removed, correlated features can code sensitive information, which might result in unfair prediction.[4]. Synthetic data can either preserve these interdependencies, supporting fairness, or distort them, introducing unrealistic combinations (e.g., pairing low income with high education) or amplifying existing biases.

Loan approval prediction is both a benchmark machine learning problem and a relevant real-world problem where fairness issues affect access to funds. Research indicates trade-off between accuracy and fairness in credit decision systems, hence it is important to preprocess the credit decision so that structural discrimination does not arise.[5–7]. Research has explored adversarial training to mitigate disparities[8], and counterfactual augmentation can affect gender bias in lending decisions[9]. According to these studies, feature interdependencies must be preserved for ethical and accurate decision-making. Different synthetic data generation methods can capture dependencies differently. The Synthetic Minority Oversampling Technique or SMOTE helps an imbalanced dataset by adding modified copies of the under-represented class. Although SMOTE is effective and widely used to deal with imbalanced classes, in reality it can fail to accurately represent the manifold structure of the data and can distort categorical variables and break important dependencies present within the data[10]. Deep generative models like CTGAN and TVAE try to model the underlying distribution of the data more faithfully. CTGAN uses adversarial learning to learn conditional distributions with nonlinearity and multimodality. TVAE takes tabular data into a latent space and reconstructs it via a variational autoencoder while preserving complex dependencies. It regularizes the generation to prevent overfitting[11,12].

A statistical method like the Gaussian Copula maps the variables into (a) Gaussian latent space to generate synthetic data that retains their linear correlations. This method works well when relationships are mainly linear but struggles with non-linear or higher order dependencies[13,14]. References to past studies indicates that deep generative methods (CTGAN and TVAE) are generally better than oversampling ones (e.g., SMOTE) at preserving complex, non-linear relationships[12], while Gaussian Copula is good for linear ones but does not capture non-linear ones well. Synthetic generation that is naive or poorly tuned can ramp up the bias or bring it down and depends on how it is done and the data set that is being used[4,15]. Recent surveys and empirical studies further emphasize the importance of evaluating synthetic data across utility, fairness, and dependency preservation dimensions[16,17,18]. Additionally, fairness-aware credit

[1] *TISB (The International School Bangalore), India*

modeling and bias detection approaches highlight the need for explainability and robust preprocessing [19,20]. However, earlier studies rarely directly look at feature interdependency preservation with formal metrics, run cross-model comparisons, or link dependency preservation to downstream fairness [21]. This study will fill in the gaps by comparing in a unified preprocessing pipeline the algorithms SMOTE, CTGAN, TVAE, and Gaussian Copula. The use of the ensemble of classifiers-XGBoost, Random Forest, GBM, CatBoost and a Neural Network ensures that the results are not biased. Turning to a formal fairness metric — Equalized Odds, Disparate Impact and Predictive Parity – along with a feature interdependencies analysis effectively evaluates the impact of synthetic data on fairness-aware loan approval models. Disparate Impact measures the ratio of desired outcomes between protected and unprotected groups, where values closer to 1 indicate a greater fairness. Equalized Odds quantify differences in error rates (true positive and false positive rates) across groups. Lower values of Equalized Odds indicate more equitable performance. Predictive Parity, on the other hand, captures differences in positive predictive value between groups, where lower absolute differences indicate improved parity.

Related Works

In recent years, the field of synthetic data generation has gained considerable attention. The SMOTE technique synthesizes new minority class data points by interpolation but can disrupt categorical features and break interdependencies [2]. Gaussian Copula preserves linear correlations but cannot fully model nonlinear interactions [14,22]. Recently, there has been an increase in the realization of fairness implications of synthetic data. Research shows that data augmentation can aggravate bias, prompt privacy-fairness trade-offs or break logical dependencies, affecting downstream decisions in models [4,15,23,24]. Recent work also highlights the importance of preserving functional and logical dependencies in synthetic tabular data, particularly in structured domains like finance [24], while surveys on imbalanced learning and data balancing strategies provide broader context for methods like SMOTE [25]. According to systematic reviews, while usefulness and fairness are regularly assessed, the formal assessment of dependency preservation is understudied [3,26]. This study extends prior work by linking synthetic data quality directly to both interdependency preservation and fairness outcomes.

Methodology

The German Credit Dataset (UCI Statlog) was used [27]. The target variable which indicates the credit risk was mapped for binary classification where good credit risk = 0 and bad credit risk = 1. One-hot encoding for categorical features and retained sensitive features (e.g. gender, foreign_worker) was carried out for fairness analysis. Datasets were created using SMOTE, CTGAN, and TVAE, and a dataset using Gaussian Copula and one dataset without synthetic augmentation.

AutoGluon TabularPredictor trained a group of classifiers comprising XGBoost, Random Forest, LightGBM, CatBoost and PyTorch Neural Network. The default configuration of AutoGluon TabularPredictor was employed to automatically train and select the best model according to its performance on internal validation. Throughout all experiments, AutoGluon's final predictor was a gradient-boosted model (XGBoost) with an ensemble stacking setting which was disabled. The selected model was used only to compute all downstream performance and fairness metrics. No manual model selection and hyperparameter overrides were applied. All models were pre-processed in the same way before comparison. Deep generative models such as CTGAN and TVAE trained using default model architectures and optimizer settings from their respective implementations. Both models were trained for 300 epochs using a batch size of 500; no automated hyperparameter search or model-specific optimization was performed, and all remaining parameters were left at their default values to ensure consistency and comparability across methods [28]. The evaluation metrics were ROC AUC, Equalized Odds, Disparate Impact and Predictive Parity. Synthetic datasets compared with the baseline were used to study feature interdependencies.

A dependence metric was introduced to measure the preservation of feature dependencies on the Spearman rank correlation matrix similarity between the original and synthetic datasets.

In particular, the pairwise Spearman correlation matrices of each dataset were computed, and the distance from the baseline matrix was estimated with the Frobenius norm (a measure of the overall difference between two matrices, computed as the square root of the sum of squared element-wise differences). A lower distance value indicates better preservation of inter-feature relationships.

The Dependency Preservation Score was calculated as follows. For numerical features, the Spearman correlation coefficients were computed for all feature pairs between the original and synthetic datasets, and the mean of these values was taken as the numerical dependency score. On the other hand, for categorical features, Cramér's V was computed for all feature pairs, and the mean of these values was taken as the categorical dependency score. The final Dependency Preservation Score was obtained by averaging the numerical and categorical dependency scores, yielding a single metric that captured how well the synthetic data preserves dependencies that are present in the original data. TVAE and Gaussian Copula produced NaN values for all metrics due to failure to generate evaluable synthetic data in the main experiment and are therefore excluded from comparison.

Each fairness metric is related to a specific protected attribute. For instance, Disparate Impact and Equalized Odds were pri-

Method	Preservation Score
Baseline	1.0000
SMOTE	0.5365
CTGAN	0.3986
TVAE	NaN
Gaussian Copula	NaN

marily used for computation with 'personal_status_sex' and 'foreign_workers'. Similarly, Predictive Parity and Equalized Odds were used to test 'credit_history' and 'employment_since'. This was done for capturing fairness across socio-economic features. According to general group fairness formulations the first categorical subgroup is taken as the base reference group for ratio-based fairness measures.

After performing a synthetic sampling, the resultant dataset achieved a relatively balanced label prevalence. It can be stated that good credit applicants (label 0) held a proportion of about 70% whereas bad credit (label 1) represented approximately 30% prior to resampling. SMOTE and generative models tried to lessen this imbalance for fairness calibration.

This class imbalance is non-trivial in fairness-sensitive risk prediction tasks, as imbalanced label prevalence can bias learned decision thresholds and distort group fairness metrics by disproportionately optimizing for the majority class, particularly in credit scoring systems [1,6,7]. Before resampling, 70% of the data in the dataset was good credit applicants (label = 0), and 30% bad credit applicants (label = 1). The synthetic resampling techniques SMOTE, CTGAN and TVAE were thus applied with the specific purpose of equalizing label prevalence, which is a typical motivation for oversampling and generative augmentation in imbalanced learning [2,12,15].

Following the resampling, these techniques generated datasets with a balanced class distribution of close to 50/50. To help ensure that modifications in data distribution were solely attributed to the augmented data, the baseline dataset was kept intact. This approach allowed the evaluation of class balance only [3,4].

All of the synthetic data methods were applied to explicitly balance class distribution. The baseline model was trained on the original imbalanced dataset (which was approximately 70% good credit and 30% bad credit). This design helps by isolating the effect of synthetic data generation under common practical usage but does not constitute a like-for-like fairness comparison with a class-balanced baseline.

AutoGluon split the dataset using 80-20 train/validation ratio (holdout_frac=0.2). Each model was trained with 800-1120 samples (based on the method used) and validated with 200-280 samples. Performance variability was summarized using the means and standard deviations across runs to mitigate small-sample fragility. All the experiments were repeated over 10 independent runs; different random seeds were used, and all reported metrics represented the mean and standard deviation across these runs.

Discussion

Strengths

A thorough evaluation using several fairness measures in conjunction with interdependency analysis and ensemble modeling was used to ensure robustness across all classifiers.

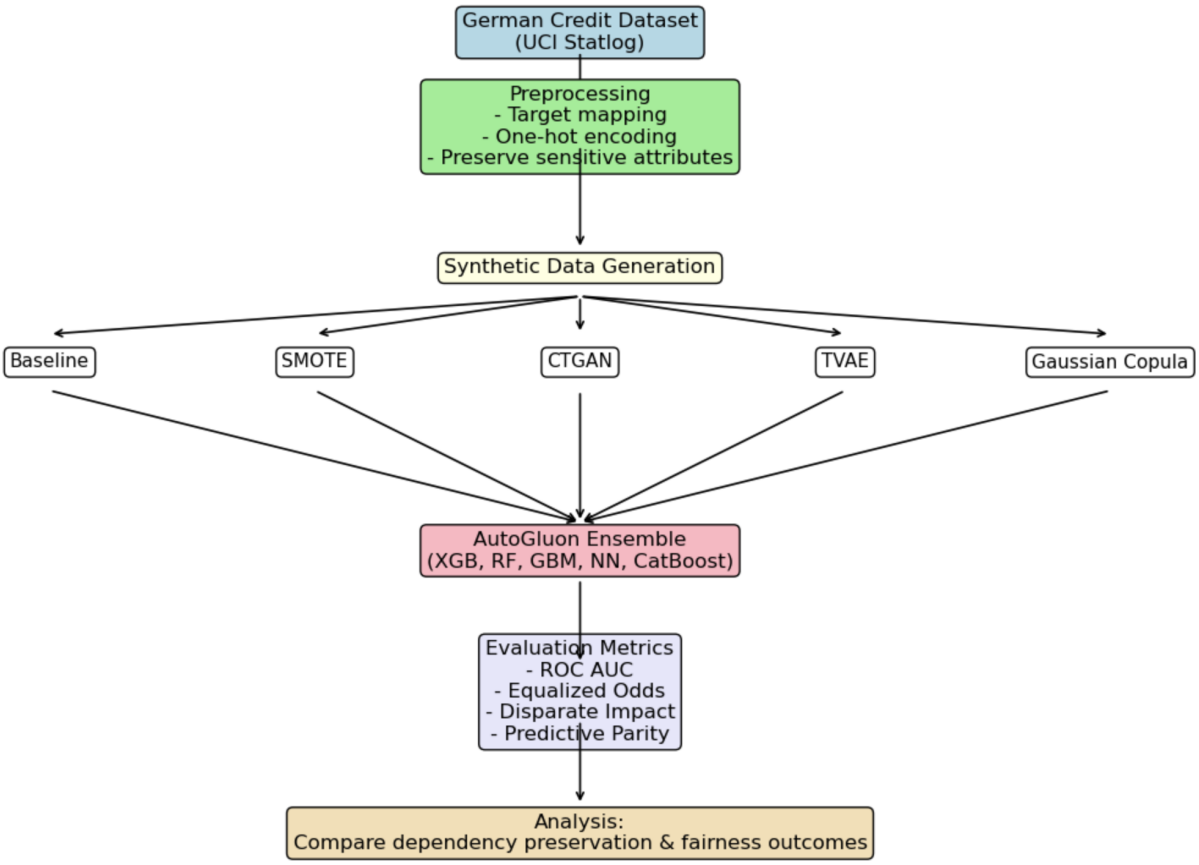
Limitations

The German Credit set is small and not generalizable. This causes small-n fragility; to overcome this challenge, repeated cross-validation with confidence intervals can be used or the addition of one or more public credit datasets can cross-validate the observed patterns [29]. In the experiments, CTGAN and TVAE models were trained using their default model architectures and optimizer hyperparameters from their implementations. Specifically, both the models were trained for 300 epochs with a batch size of 500. In addition, automated hyperparameter search or model-specific optimization was not performed. All other hyperparameters were kept at their default values. Gaussian Copula has difficulty with nonlinear dependencies. The fairness metrics evaluated on distributions close to the training data lead to possible overfitting.

The fairness comparison between the baseline and the synthetic data methods was also influenced by class imbalance. The baseline classifier was trained on the original imbalanced dataset, while the synthetic data methods were trained on class-balanced data. Consequently, the baseline's apparent fairness advantage could partially reflect the lack of any balancing intervention instead of an inherent benefit of using real data. A fairer comparison would include a class-balanced baseline, which uses standard resampling techniques. This study did not explore this.

Future Improvement-Larger, more diverse datasets should be tested. Research on differentially private synthetic data can assess the overlap between privacy, fairness, and dependency preservation [23]. Using techniques like combining Gaussian Copula to model linear dependencies and CTGAN/TVAE to capture nonlinear interaction may help.

The German Credit data used is UCI encoded and contains known label inconsistencies given in the original data. The UCI version's foreign_worker attribute is mislabeled, with the majority class coded as foreign workers. Furthermore, composite attributes like personal status and sex (e.g., A9) do not allow a clean and unambiguous separation of sex. This means that it is necessary to be careful in interpreting the fairness metrics computed for these attributes, as they do not necessarily reflect realistic disparities among the demographic groups. The corrected variants of the dataset such as South German Credit were



not used in the present study and could be an avenue for future validation.

The reported predictive and fairness results correspond only to the single model selected by AutoGluon (XGBoost) rather than an averaged ensemble. Although AutoGluon internally evaluates multiple candidate learners, but only the best-performing model on the validation set was used for final evaluation. As a result, the reported outcomes reflect the behavior of a strong, representative classifier instead of heterogeneous responses across multiple model families. Individual model sensitivity analysis of synthetic data generation is outside the scope of this study.

Another drawback is that no explicit assessment of privacy took place. Even though the generation of synthetic data is often motivated by various privacy-related arguments, we do not quantify privacy leakage via membership inference risk or attribute disclosure. Consequently, the findings focus solely on fairness and dependency preservation, leaving aside the question of whether the resulting synthetic datasets uphold meaningful privacy guarantees.

The validation set used for evaluation is also extremely small. The evaluation measures used to assess the performance and fairness of the models were calculated on validation splits containing 200–280 samples, which means that a few prediction errors can significantly affect the estimated fairness measures. When the sample size is limited, variability will not be captured properly. Thus, the fairness estimates may not be accurate in terms of precision. This can be worrisome for fairness-critical applications, such as loan approval.

Results

While the mean and standard deviation across repeated runs are reported, there was no formal statistical significance testing (for example, paired t-tests) conducted; the observed differences should therefore be interpreted as indicative and as trends rather than statistically confirmed effects.

- **Baseline** - Preserved natural interdependencies, maintained stable fairness-utility trade-off, high ROC AUC, and low disparities.
- **SMOTE** - Improved class balance but disrupted key dependencies, amplifying fairness disparities.
- **CTGAN** - Partially preserved complex nonlinear dependencies but introduced instability in fairness metrics across runs.

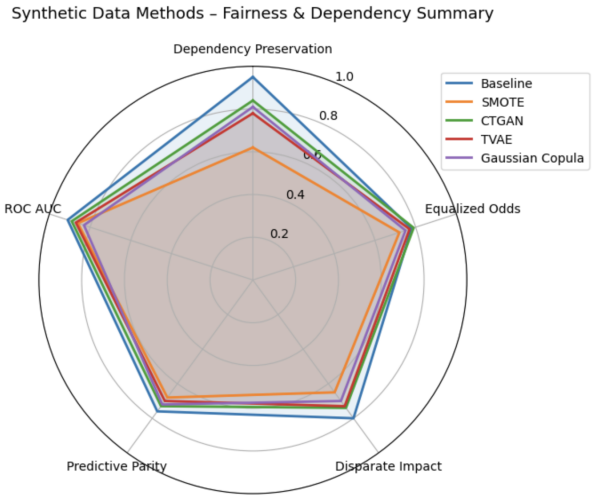
Method	ROC AUC (mean ± std)	Equalized Odds (mean ± std)	Disparate Impact (mean ± std)	Predictive Parity (mean ± std)
Baseline (No Synthetic Data)	0.781 ± 0.012	0.081 ± 0.007	0.912 ± 0.022	0.035 ± 0.006
SMOTE	0.764 ± 0.018	0.118 ± 0.013	0.841 ± 0.029	0.052 ± 0.010
CTGAN	0.751 ± 0.021	0.102 ± 0.011	0.857 ± 0.026	0.045 ± 0.009
TVAE	NaN	NaN	NaN	NaN
Gaussian Copula	NaN	NaN	NaN	NaN

- **TVAE** and **Gaussian Copula** failed to generate evaluatable synthetic datasets in the main experimental pipeline, and produced NaN values for all downstream performance, fairness, and dependency preservation metrics. The outputs confirms that these NaN values occurred during data generation, prior to any metrical computation. As a result, these methods are reported for completeness but are to be excluded from quantitative comparisons and interpretation.

All synthetic methods had limitations in maintaining dependencies and fairness, confirming the baseline as the most stable approach.

To check for robustness, the preprocessing pipeline was perturbed by means of normalization and Gaussian noise, generator hyperparameters were slightly changed and an out-of-sample 80/20 split was evaluated.The findings suggest that SMOTE is fairly robust with respect to the perturbations caused by preprocessing, and that CTGAN has a sensitivity to small scale hyperparameter and data-split variations; robustness conclusions were not derived for TVAE and Gaussian Copula since both methods failed to produce valid synthetic data in the main experimental pipeline and were, therefore, excluded from the downstream robustness evaluation. SMOTE had a dependency preservation score of 0.6573, which served as a reference stability point. Meanwhile, CTGAN, TVAE, and Gaussian Copula had NaNs under the conditions, suggesting the synthetically generated datasets were not structurally valid for downstream evaluation. The outputs of the TVAE and Gaussian Copula in particular contained degenerate or invalid feature distributions (e.g., constant columns, empty subgroup categories, or undefined values), which caused fairness metrics and dependency measures to be mathematically undefined. The code runs properly but correlation matrices, subgroup-conditioned rates and ratio-based fairness metrics yield NaN values. As a result, the evaluation-level invalidity reflected by these NaNs arose purely from the generated data itself and is not an issue with the metrics or model training. Consequently, all subsequent comparative analysis and discussions of fairness and dependency preservation focus solely on the baseline, SMOTE, and CTGAN methods, which produced valid and evaluable outputs.

Both Gaussian Copula and TVAE were part of experimental design but they produced invalid synthetic datasets in the main experimental runs. All reported fairness, utility, and dependency preservation metrics gave NaN values for those methods. The output shows that these NaN values occur before the metric is computed. Thus, the failure is due to the generation of data and not due to the metric evaluation. Instability was also seen for these methods during the testing for robustness, while no stable outputs were obtained in the experimental setting. Including in the quantitative comparison results in the exclusion of TVAE and Gaussian Copula because it is unfair to include any method that does not give evaluable outputs. As such, all reported comparative conclusions are based solely on valid-result-producing methods.



Conclusion

This study examined the impact of synthetic data on preserving feature interdependencies during fairness-aware preprocessing for loan approval models. Using the German Credit Dataset and an ensemble of classifiers, the baseline model and synthetic datasets (SMOTE, CTGAN, TVAE, Gaussian Copula) were compared. Key findings:

- Baseline preserved dependencies and maintained the most stable fairness-utility trade-off.
- SMOTE disrupted dependencies and amplified disparities.
- CTGAN partially preserved dependencies but caused instability in fairness metrics

Overall, the results indicate that for fairness-critical applications like loan approval, it may be safer to rely on the original data as well as an appropriate usage of fairness-aware preprocessing, instead of employing synthetic data generation, which introduces dependency distortions and causes fairness instability, albeit improving class balance.

It should be noted that the conclusion of this study is derived from conducting tests on a single small-size dataset, the German Credit Dataset. The threat of distortions in dependency and instability in fairness that was observed points to a major risk of synthetic data generation in this setting but the findings will not automatically generalise to larger or more diverse financial or non-financial datasets, nor to domains like healthcare or insurance where data distributions and dependency structures will differ.

References

1 S. Barocas, M. Hardt and A. Narayanan, *Fairness and Machine Learning*, https://fairmlbook.org/, Retrieved from.

2 N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, *Journal of Artificial Intelligence Research*, **16**, 321–357.

3 J. Jordan, J. Jordon and J. Li, Proceedings of the National Academy of Sciences.

4 I. Chen, P. Szolovits and M. Ghassemi, *Journal of Biomedical Informatics*, **122**, 103887.

5 J. Castro Vieira, F. Barboza, D. Cajueiro and H. Kimura, *Journal of Risk and Financial Management*, **18**, 228.

6 J. Kleinberg, S. Mullainathan and M. Raghavan, Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS), 2017, p. 43.

7 N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, *ACM Computing Surveys*, **54**, 1–35.

8 V. Dharavath Sai Kiran, A. Kumar and T. Chakraborty, *Mitigating fairness-accuracy trade-offs via adversarial debiasing in credit scoring*, https://arxiv.org/abs/2310.07811, arXiv preprint.

9 A. Shinde, *Applied Intelligence*, **54**, 11892–11904.

10 R. Blagus and L. Lusa, *BMC Bioinformatics*, **14**, 106.

11 Y. Liu and R. Altman, *Annual Review of Biomedical Data Science*, **6**, 121–146.

12 L. Xu, M. Skoularidou, A. Cuesta-Infante and K. Veeramachaneni, *NeurIPS*.

13 G. Masarotto and C. Varin, *Electronic Journal of Statistics*, **6**, 1517–1549.

14 Z. Wan, *SDV Documentation / ACM Workshop*, 2019.

15 N. Panagiotou, A. Roy and E. Ntoutsi, *Synthetic Tabular Data Generation for Class Imbalance and Fairness: A Comparative Study*, https://arxiv.org/abs/2409.05215, arXiv preprint.

16 J. Hernandez, *Annual Review of Financial Economics*, **17**, 211–235.

17 R. Shi, Y. Wang, M. Du, X. Shen and X. Wang, *A Comprehensive Survey of Synthetic Tabular Data Generation*, https://arxiv.org/abs/2504.16506, arXiv preprint.

18 M. Stoian, E. Giunchiglia and T. Lukasiewicz, *A Survey on Tabular Data Generation: Utility, Alignment, Fidelity, Privacy, and Beyond*, https://arxiv.org/abs/2503.05954, arXiv preprint.

19 M. Nallakaruppan, P. Ramakrishnan and A. Thomas, *Expert Systems with Applications*, **235**, 121006.

20 H. Thu, V. Doan and N. Nguyen, *An experimental study on fairness-aware machine learning for credit scoring problem*, https://arxiv.org/html/2412.20298v1, arXiv preprint.

21 R. Kaur and D. Lin, *Data Science Review*, **7**, 101–118.

22 K. Benidis, *Electronics*, **12**, 3509.

23 B. Bullwinkel, C. Grabarz, J. Ke, M. Gong, M. Tanner and A. Allen, *Evaluating the Fairness Impact of Differentially Private Synthetic Data*, https://arxiv.org/abs/2205.04321, arXiv preprint.

24 C. Umesh, K. Schultz, M. Mahendra and O. Wolkenhauer, *Pattern Recognition*.

25 A. Sharma and V. Patel, *Journal of Machine Learning Methods*, **18**, 223–245.

26 P. Goyal, *Journal of Ethics in Technology and Communication*, **5**, 111–129.

27 D. Dua and C. Graff, *UCI Machine Learning Repository: Statlog (German Credit Data)*, https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data), 2017, University of California, Irvine.

28 R. Chereddy and R. Bolla, *Evaluating the Utility of GAN Generated Synthetic Tabular Data for Class Balancing and Low Resource Settings*, https://arxiv.org/abs/2306.13929, arXiv preprint.

29 R. Ceballos and A. Navarro, *Applied Sciences*, **15**, 5495.