

Comparative Evaluation of Synthetic Data Generation Methods for Binary Classification Problems with Class Imbalance

Jhonatan Esquivel , Christian Humpiri , Jose Vega , Nemias Saboya 
Escuela Profesional de Ingeniería de Sistemas-Facultad de Ingeniería y Arquitectura,
Universidad Peruana Unión, Lima, Perú

Abstract—Class imbalance is a critical challenge in the classification of tabular data, since it affects the diagnostic capacity of models in domains such as health and finance. This research compares four synthetic data generation paradigms: traditional interpolation (SMOTE-NC), deep generative models (CTGAN and TVAE), and a hybrid scheme (SMOTE-NC-CTGAN) in which SMOTE-NC only densifies the minority data used to train the generator. Five UC Irvine datasets with imbalance ratios between 2.55:1 and 29.99:1 were used under a dual evaluation that combines multidimensional statistical fidelity (Wasserstein distance, correlation and categorical differences, and diversity) and predictive utility (AUC-PR, Recall, Precision, F1-Score, Specificity and G-mean), with a cost-sensitive baseline as reference. Each experiment was repeated in 10 runs that regenerate the synthetic data, and the differences were tested with paired tests, Holm correction, and effect sizes. SMOTE-NC obtains the highest fidelity in all datasets, while the CTGAN-based methods generate the most dispersed records under extreme scarcity. Synthesis does not improve the global ordering of predictions: of 40 comparisons, AUC-PR decreases significantly in 19 and increases only in 2, both in the most imbalanced dataset with XGBoost (HYBRID: 0.5252 versus 0.4872), an effect not observed with CatBoost. In contrast, Recall increases significantly in 27 of the 40 comparisons (e.g., from 0.1617 to 0.5835 with CTGAN in DS1), at the cost of Precision, and a class-weighted baseline produces a similar displacement. Data synthesis mainly shifts the decision boundary towards the minority class, which is favorable where false negatives dominate the cost.

Keywords—Class imbalance; synthetic data; SMOTE-NC; CTGAN; hybrid models; machine learning; tabular classification

I. INTRODUCTION

The imbalance of classes in datasets represents a very common challenge in machine learning tasks, as well as in the real world, where events of interest occur with low frequency. Consequently, classification algorithms often produce bias by favoring the majority class based on the level of imbalance. This problem has been addressed in various domains such as computer security, medical diagnosis, and detection of financial fraud, where incorrect detection can have critical consequences [1], [2], [3].

To address this problem, various approaches have been proposed, among which the oversampling of the minority class has been widely used. The methods derived from the Synthetic Minority Oversampling Technique (SMOTE) are effective and less computationally expensive with respect to other models.

Likewise, they highlight how versatile they can be [2],[4]. However, these approaches have inherent limitations, since the generation of examples by means of linear interpolation can give rise to less diverse examples and does not adequately capture the complex relationships between variables.

Therefore, in recent years, different deep generative models have been created as a promising alternative for the generation of more realistic synthetic data [5]. In particular, Generative Adversarial Networks (GANs) have demonstrated a remarkable ability to learn complex data distributions and generate new samples that preserve the statistical characteristics of the original dataset [6]. These models have been applied in various domains and with different variations according to the context, such as CTGAN [7], CopulaGAN [8], CGAN [6], or AC-GAN [6].

Likewise, there are Variational Autoencoders (VAE), which seek to minimize the difference between the input and the reconstructed output, thus learning a compact representation of the data [8]. It presents multiple variants, and among them the most applied in the TVAE is a variation aligned to tabular data, which captures that, despite being highly sensitive to its hyperparameters, is a robust model whose performance varies according to the characteristics of the dataset [9]. However, the use of synthetic data introduces new challenges related to the quality and fidelity of the data generated, since some studies have indicated that the predictive capacity of models trained with synthetic data is not sufficient to evaluate their usefulness, as the statistical distributions and relationships between variables present in the original dataset must also be analyzed [10], [11].

More recently, some works have proposed hybrid approaches that combine traditional oversampling techniques with generative models. These strategies seek to take advantage of both approaches. On the one hand, the initial balancing facilitates the learning of the generative model. In contrast, the generative process allows the creation of more diverse and representative samples of the original distribution of the data [12], [13].

In this context, the present study aims to evaluate comparatively different synthetic data generation techniques for imbalanced binary classification problems. For this, traditional oversampling methods and generative models designed for tabular data are analyzed, including SMOTE-NC, CTGAN, and TVAE, as well as a hybrid SMOTE-NC-CTGAN approach. The

experiments are performed on multiple real datasets with unique characteristics, which allows evaluation of the behavior of these methods under different levels of imbalance.

The contributions of this work are: 1) a controlled comparison of interpolation, adversarial, variational, and hybrid generation under a common pipeline, with the same real records and classifiers; 2) a hybrid SMOTE-NC-CTGAN scheme for mixed-type data in which SMOTE-NC only densifies the minority data used to train the generator; 3) a dual evaluation of multidimensional fidelity and predictive utility, including a cost-sensitive baseline; and 4) a statistical validation over repeated runs that regenerate the synthetic data.

II. LITERATURE REVIEW

The problem of class imbalance has been widely studied since it is very relevant to machine learning models because this unequal distribution can produce bias or unjustified benefit towards the majority class, reducing its veracity. In the same way, recent literature has responded to this problem with various forms of oversampling, such as traditional techniques, generative models, variants, or hybrid approaches. This framework is responsible for precisely this: being able to compare and evaluate their fidelity against binary classification datasets with different levels of imbalance.

A. Class Imbalance in Tabular Data Sets

The literature agrees that class imbalance negatively affects models since many learn with a certain bias; consequently, they end up favoring the majority class. The study [2] summarizes that this problem is particularly critical in health, where the minority class usually corresponds to the most important clinical event. In a complementary way, the review [4] shows that both variants of SMOTE and approaches based on GAN have arisen precisely to correct this limitation in imbalanced classification tasks. Along the same lines, the work [3] points out that, in Android malware detection, the imbalance reduces the ability to generalize to the minority class. At the same time, [14] shows that even deep networks can remain biased if the training objective does not adequately control the unequal distribution between classes.

Therefore, generative models become relevant, and this is demonstrated in the study [15], where it is explicitly stated that imbalanced data sets hinder the development of robust models for predictions of obesity levels and that synthetic generation can efficiently support when samples or certain key variables are missing. However, the study [11] mentions that these methods must be analyzed quite carefully, because they can generate differences in distributions, correlations, or even privacy-related risks, especially when it is directly assumed that synthetic data will always be useful. In addition, in another context, the study [16] explains that in imbalanced regression problems, it is not enough to only increase the number of rare values, but it is also important to preserve the relationship between the variables and the continuous objectives. This shows that generating new data alone does not guarantee good results if the existing logic between the original data is not maintained.

B. Traditional Oversampling Techniques

Within traditional techniques, SMOTE continues to be one of the most used methods to treat the imbalance problem. Its operation is based on creating new samples of the minority class through interpolation between close neighbors, simply avoiding copying existing instances. The work [17], applied to the area of phishing detection, demonstrates that the combination of SMOTE-NC together with XGBoost achieved 98.0% accuracy and 98.5% recall, surpassing the model trained without re-sampling. This shows that data balancing can significantly improve the ability to detect the minority class.

Similarly, the study [18] concludes that the combination of SVM + SMOTE-NC presents better results compared to SVM without balancing techniques, especially in data sets containing nominal and continuous variables. In these cases, improvements in both accuracy and F1-score were observed. On the other hand, some applied works also support the usefulness of these techniques. In the case of credit card fraud detection, [1] reports a fairly extreme scenario where fraud only accounts for 0.172% of transactions made. In that study, the combination of Random Oversampling with XGBoost obtained the best performance, reaching an F1-score of 92.43%, accuracy of 97.383%, and recall of 88.418%. Likewise, the study [19] shows that after applying SMOTE to correct an imbalance ratio of 60:40 in zygosity prediction, XGBoost was again positioned as the model with the best results, achieving 85.57% accuracy and 86.72% F1-score.

In general, the studies analyzed suggest that traditional oversampling techniques still maintain a fairly competitive performance, especially when combined with robust classifiers. In addition, they allow for improving the detection of the minority class without increasing the complexity of the pipeline too much. However, studies [4] and [16] also point out that these methods can generate overly generalized or unrealistic samples when the feature space has complex structures. Due to this, their ability to represent heterogeneous tabular distributions may be limited in some scenarios.

C. Generative Models for Tabular Data

The advancement of generative models has opened the possibility of going beyond SMOTE's own local interpolation. The review [6] highlights that GANs have gained relevance as tools to alleviate imbalance in tabular datasets when modeling complex distributions. In this line, CTGAN has become one of the most used models because it was designed to work with mixed and imbalanced tabular data. The study [10] evaluated CTGAN using the Cleveland heart disease dataset and found that the model manages to reproduce the general distribution and some important correlations, especially in variables such as age, thalach, and oldpeak. However, there are still differences in certain categorical attributes. In addition, [10] mentions that the predictive usefulness of synthetic data is moderate, since although the data generated captures relevant patterns, they fail to replicate the original labels fully.

Similarly, the work [20] compared CTGAN and TVAE in tabular datasets of different sizes and concluded that both models can retain statistical properties. However, their performance depends quite a bit on the size of the data and the configuration used during training. On the other hand, [21] presents an improved conditional GAN architecture for cardiovascular data and reports a mean KS of 0.3900, close to TVAE (0.3858) and higher than CTGAN (0.4803). This shows that the behavior of generative models can vary depending on the characteristic evaluated. In fact, the study points out that some variables are better represented by TVAE and others by the proposed model, so there is no single dominant model. This idea coincides with [11], where TVAE is described as a fairly robust and stable method even in complex datasets with many categorical variables. However, it is also mentioned that it may present greater privacy risks, as some synthetic data tends to look too much like actual records.

D. Hybrid Approaches to Synthetic Generation

On the other hand, a particularly relevant line for this research is that of hybrid methods, which try to combine the simplicity of classic oversampling with the ability of generative models to learn more complex distributions. Article [12] is central at this point, since it proposes SMOTified-GAN, where the samples generated by SMOTE replace random noise as input to the GAN. The work [12] reports that this approach improved the F1-score by up to 9% with respect to the next-best method and that it surpassed the other algorithms in 10 out of 11 datasets, showing especially clear advantages in sets with very small minority classes.

Subsequent evidence goes in the same direction. The study [22] focused on financial fraud and compared SMOTE, GAN, SMOTified-GAN, SMOTE+GAN, and GANified-SMOTE, concluding that hybrid approaches are more consistent than individual methods and that performance also depends on the proportion of synthetic samples generated. In turn, [23] compares traditional oversampling techniques with CGAN and CTGAN in medical data with imbalance ratios of 6.8, 9, and 19, finding that GAN-BASED variants achieve better performance than several classic techniques in terms of AUC and F1-score. Finally, [24] shows that even the proportion between synthetic samples from SMOTE and GAN can be optimized by genetic algorithms, which indicates that the success of hybrids does not depend only on the generative model but also on the balance between diversity, realism, and the number of artificial samples. This line of evidence methodologically supports the SMOTE-NC-CTGAN approach evaluated in this study, which is inspired by [12] but differs from it in three aspects.

First, SMOTified-GAN replaces the random noise of the generator with SMOTE samples, whereas here SMOTE-NC only expands the minority class up to 50% of the majority class, and CTGAN is then trained on this expanded set. Second, the interpolated records are discarded afterwards, so the final set contains only real and CTGAN records. Third, SMOTE-NC and the conditional generator of CTGAN [37] handle mixed numerical and categorical variables natively, whereas [12] and [22] used standard GANs on mainly continuous features.

E. Classification Models in Imbalanced Contexts

In addition to the generation method, the literature shows that the final performance depends on the classifier used. The article [1] shows that, in credit card fraud, ensemble and boosting models especially benefit from balancing, and that XGBoost can outperform other classifiers when combined with appropriate oversampling techniques. Similarly, [19] reports that XGBoost, Logistic Regression, and Gradient Boosting had comparable behaviors in zygoty prediction, although XGBoost was ultimately selected for its slight advantage and better subsequent interpretability by SHAP. In cybersecurity, [17] also shows that XGBoost combined with re-sampling techniques achieves a very high balance between sensitivity and specificity, supporting its choice in imbalanced tabular classification contexts.

On the other hand, [25] demonstrates that balancing can also enhance explainable ensemble models in cardiovascular prediction, while [1] and [17] reinforce the idea that tree-based and boosting models are usually more robust than simple linear alternatives when the dataset is heterogeneous or highly imbalanced. Taken together, this evidence justifies the use of XGBoost and CatBoost as the main classifiers in the experimental phase of this study, leaving Logistic Regression as an auxiliary tool for the selection of hyperparameters during synthetic generation.

F. Metrics to Evaluate and Verify Synthetic Data

An important aspect of recent research is that the evaluation of synthetic data should not focus only on the final performance of the classifier. The study [26] explains that, in imbalanced problems, metrics such as precision, recall, and F1-score are usually more useful than accuracy to analyze behavior about the minority class. In addition, mentions that the metric used should depend on the objective of each problem.

Similarly, the work [8] compares CTGAN, CopulaGAN, VAE, and SMOTE in anonymous traffic classification. It points out that it is not enough to only measure predictive performance, but it is also necessary to evaluate the quality of the data generated. In that study, SMOTE managed to increase the F1-score values by about 7.5% compared to the control group, while LightGBM reached an F1-score of 90.3% over eight types of applications. This shows that the final performance changes depending on the generative method and the classifier used subsequently.

In addition, the study [10] mentions that the evaluation should include other aspects such as numerical and categorical fidelity, the preservation of correlations, and the similarity between real and synthetic data. On the other hand, [27] reports that AE-CTGAN obtained better results than an autoencoder combined with a conventional GAN, reaching an average accuracy of 0.9929 for BMR in bladder cancer and 0.9748 for menopausal status in breast cancer. It also significantly decreased the distance between real and synthetic samples.

Finally, [28] demonstrates that synthetic data can also improve generalizability in medical vision tasks. Through the use of conditional StyleGAN, the mAP increased from 0.86 to 0.93, while the recall in flat or depressed lesions increased from 0.72 to 0.87. These results suggest that the evaluation of synthetic data should not focus only on statistical fidelity but also on the real utility and generalization capacity they bring to the model.

G. Research Gap

Despite the advances made in the treatment of class imbalance, there are still some important limitations within the current literature that this study seeks to address:

1) *Homogeneity of the approaches compared*: Many investigations tend to compare methods belonging to the same family of algorithms, such as variants of SMOTE or different GAN architectures. This makes it more difficult to analyze how different generative approaches respond to the same scenario.

2) *Lack of a dual evaluation*: There are still few studies that jointly analyze the statistical fidelity of synthetic data and its predictive utility, although both dimensions do not always exhibit the same behavior.

3) *Little evidence in extreme scenarios*: There is still little evidence on the performance of these techniques when the imbalance exceeds critical ratios greater than 20:1, a context where several interpolation methods tend to lose effectiveness.

In this context, the present research seeks to cover these limitations through a direct comparison between four different approaches: SMOTE-NC as a traditional interpolation method, CTGAN based on deep adversarial learning, TVAE through variational modeling, and a hybrid proposal called SMOTE-NC-CTGAN. In this way, the behavior of these models is evaluated using health and finance datasets, especially in scenarios with high levels of data scarcity.

III. METHODOLOGY

This research follows a quasi-experimental and empirical experimental design, developed through a six-stage pipeline. The objective is to evaluate how generative techniques influence the performance of gradient boosting classifiers (Fig. 1).

A. Datasets Presentation

For the study, five datasets from the UC Irvine Machine Learning Repository [29] were used, selected to present different levels of class imbalance and different degrees of

complexity. The sets mainly belong to the areas of health and finance, in addition to containing numerical and categorical variables. Likewise, the imbalance ratios range from moderate scenarios to extreme cases close to 29.99:1, allowing the behavior of the techniques to be analyzed under different conditions.

It is important to note that, although all data sets were obtained through the UCI repository, there are differences in their provenance. Datasets 2, 4, and 5 correspond to sets directly integrated into the repository, while datasets 1 and 3 are curated versions of external sources. In particular, dataset 1 originally came from the Behavioral Risk Factor Surveillance System of the Centers for Disease Control and Prevention (CDC), and dataset 3 from ClinicalTrials.gov, both of which were later structured and adapted by UCI for use in machine learning tasks.

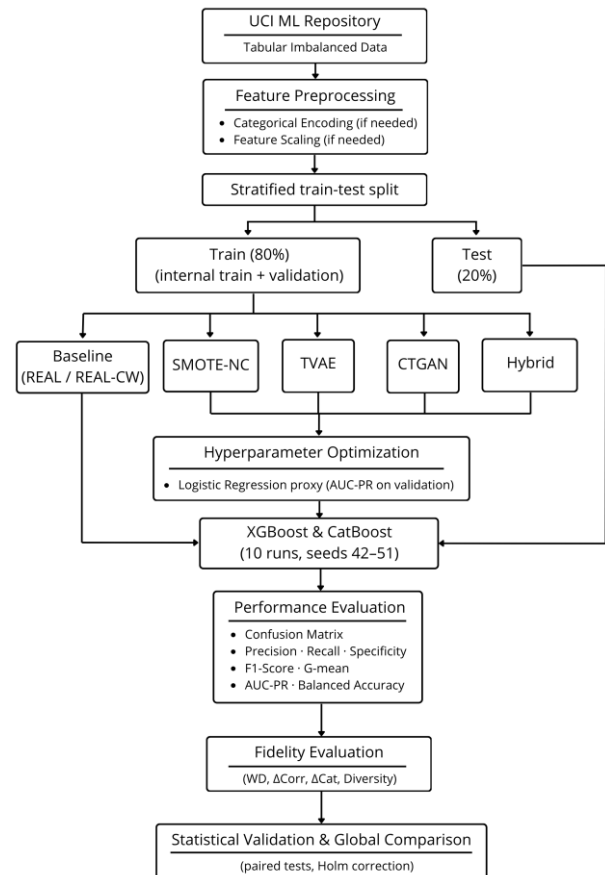


Fig. 1. Methodological flow.

TABLE I. DATASET DISTRIBUTION TABLE

Dataset	Name	Domain	Samples	Features	Numerical	Categorical	Ratio	Minority n (%)	Ref.
DS1	Diabetes Indicators	Healthcare	253,680	21	5	16	6.18	35,346 (13.9%)	[30], [31]
DS2	Thyroid Recurrence	Healthcare	383	16	1	15	2.55	108 (28.2%)	[32]
DS3	AIDS Clinical Trial (175)	Healthcare	2,139	23	9	14	3.1	521 (24.4%)	[33], [34]
DS4	Credit Card Default	Finance	30,000	23	14	9	3.5	6,636 (22.1%)	[35]
DS5	Taiwanese Bankruptcy	Finance	6,819	95	93	2	29.99	220 (3.2%)	[36]

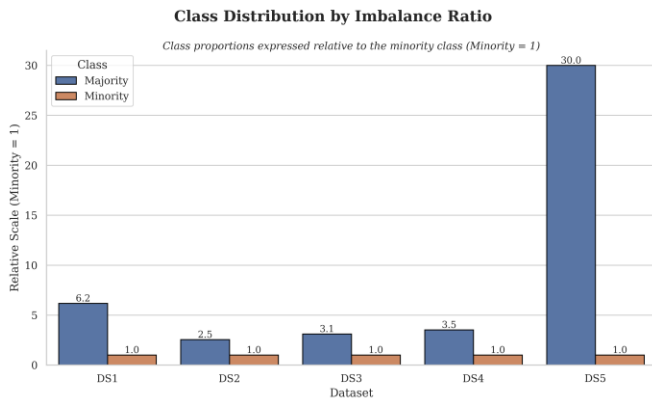


Fig. 2. Distribution of classes by dataset.

It should be noted that the DS5 (Taiwanese Bankruptcy) constitutes the "stress case" of this research, due to its extreme imbalance ratio and its high dimensionality (95 features), factors that impose a critical challenge to the generalizability of generative models.

Imbalance ratio and dataset size are not independent in this benchmark: DS2 and DS5 have the smallest minority classes (Fig. 2). Table I therefore reports the size of the minority class, which is considered in the interpretation (Section V).

B. Data Pre-Processing

Although the datasets come from consolidated sources, a standardized transformation protocol was applied to ensure convergence of the generative models. Both transformations were fitted exclusively on the training partition and then applied, without refitting, to the test partition:

1) *Categorical coding*: Label Encoding was used to transform nominal variables into integer representations, guaranteeing compatibility with the *Gumbel-Softmax* activations of the CTGAN and TVAE architectures.

2) *Numerical standardization*: Continuous variables were normalized using StandardScaler, as this facilitates the learning process in neural networks and improves the behavior of distance-based methods, such as SMOTE-NC. CTGAN and TVAE were trained on the variables in their original scale, since they apply their own mode-specific normalization [37]; their outputs were then transformed with the same scaler before being used by the classifiers.

C. Data Division

Each dataset was divided into 80% for training and 20% for testing by stratified sampling (seed 42). The test partition was isolated before any transformation and used only once, for the final evaluation, which prevents data leakage. The training partition was further divided (80/20, stratified) into an internal training subset, the only real data used by the generators and by all classifiers, including the baseline, and a validation subset used only for hyperparameter selection.

D. Generation of Synthetic Data

Four strategies representing different approaches to treating class imbalance were used in this research: a traditional technique (SMOTE-NC), two modern generative models

(CTGAN and TVAE [37]), and a hybrid proposal called SMOTE-NC-CTGAN. This allowed us to compare methods with different ways of generating synthetic data.

For all methods, only the minority class of the internal training subset was used as the source, and $M - m$ records were generated to reach a 1:1 ratio (M and m : majority and minority sizes). SMOTE-NC interpolates numerical variables between minority neighbours and assigns the most frequent category among them; CTGAN and TVAE were trained only on minority records. The hybrid approach follows Algorithm 1.

Algorithm 1. SMOTE-NC-CTGAN hybrid generation

Input: internal training subset (X, y) with M majority and m minority records; densification floor $\alpha = 0.5$; k selected for SMOTE-NC; CTGAN hyperparameters θ ; scaler S fitted on the training partition.

- 1: Apply SMOTE-NC with k to the scaled data until the minority class reaches $\lceil \alpha \cdot M \rceil$ records (real and interpolated).
- 2: Map the expanded minority set back to the original scale with S^{-1} .
- 3: Train CTGAN(θ) exclusively on the expanded minority set, declaring the categorical variables as discrete columns.
- 4: Sample $M - m$ synthetic records from the trained CTGAN.
- 5: Discard the interpolated SMOTE-NC records; the final set is formed by the original real records and the $M - m$ CTGAN records (1:1 ratio).
- 6: Scale the numerical variables with S before training the classifiers.

Output: balanced training set.

The floor $\alpha = 0.5$ gives CTGAN a larger minority sample while limiting the share of interpolated records; it was fixed a priori and not tuned.

The hyperparameters of the generators were selected per dataset by grid search on the internal train/validation split: $k \in \{3, 5, 7\}$ for SMOTE-NC; (epochs, batch size) $\in \{(300, 500), (500, 500), (500, 1000)\}$ with $\text{pac} = 10$ for CTGAN and the CTGAN stage of the hybrid; and (epochs, batch size, embedding dimension) $\in \{(300, 256, 128), (500, 256, 128), (500, 512, 256)\}$ for TVAE. Each candidate was scored by the AUC-PR on the validation subset of a Logistic Regression proxy, chosen for its low cost and its sensitivity to the training distribution; the selected values are listed in Table VI (Appendix).

E. Classifier Models

Once the balanced datasets were generated, the XGBoost and CatBoost classifiers were trained with a fixed configuration (300 trees, learning rate 0.05, depth 6, log-loss; XGBoost also with 0.8 row and column subsampling) and the default threshold of 0.5, without tuning, so that the differences can be attributed to the training data. Two references were trained on the same real data: REAL, without treatment, and REAL-CW, which weights the minority class by the imbalance ratio (cost-sensitive baseline).

Each experiment was repeated in 10 runs (seeds 42 to 51); each run regenerates all synthetic data and retrain the classifiers, while the train/test partition is fixed. The results are presented as averages accompanied by their Standard Deviation.

F. Fidelity Evaluation and Evaluation Metrics

For this research, an evaluation divided into two parts was proposed, since it is not only important that the synthetic data resemble the originals, but also to check if they really help the performance of the models. In the statistical fidelity part, the Wasserstein Distance was mainly used because it allows comparison of how similar the distributions of the real data are to the synthetic ones. In addition, this metric is based on the theory of optimal transport, so it helps to observe whether the generator model manages to preserve the original structure of the data even when there is high dimensionality.

Likewise, the analysis also considers the Correlation Difference and the Categorical Difference because both help to verify whether the relationships between variables and the proportions of the categories remain similar with respect to the original dataset. Likewise, the Diversity Score allows evaluation of whether the model is capable of generating varied data and not simply repeating very similar records. It is compared with the diversity of the real minority records: values far above it indicate records dispersed beyond the real data.

On the other hand, to measure predictive utility, the Area Under the Precision-Recall Curve (AUC-PR) was taken as the main metric, mainly because it offers a better analysis in scenarios where there is a severe imbalance of classes: its baseline equals the prevalence of the minority class, and it does not depend on a threshold [38]. In addition, this analysis was complemented by Precision and F1-Score, since they allow observing the balance between correctly detecting the minority class and avoiding unnecessary false alarms.

Recall was also considered because it allows us to measure more directly how many cases of the minority class are recovered correctly by the model, and Specificity and the G-mean [39] were added to quantify the cost on the majority class. Meanwhile, the Balanced Accuracy helps maintain a more balanced assessment considering the performance of both classes. Finally, the confusion matrices allowed these results to be interpreted in a more practical way.

Additionally, a hypothesis testing protocol was incorporated to determine if the differences observed between the methods are attributable to the synthesis procedure or to random variability between runs. Since all conditions share runs, seeds, and test partition, the design is paired. For each family of four comparisons (synthetic methods versus REAL, same dataset, classifier and metric), p-values were corrected with Holm-Bonferroni, and the effect size (d_z) and the 95% confidence interval are reported; REAL-CW is tested separately.

Normality of the paired differences was checked with Shapiro-Wilk; when it was rejected ($p < 0.05$), which is frequent for Recall or Precision in small test sets, the Wilcoxon signed-rank test was used instead of the t-test. Independence holds at the level of runs, although all runs share the same test partition,

so the inference refers to the variability of the procedure. Non-significant results indicate insufficient evidence, not absence of effect.

Finally, all comparisons keep the classifier constant; Table III and the figures show CatBoost, and Table VII (Appendix) shows XGBoost.

IV. RESULTS

Below are the main findings obtained after executing the experimental pipeline:

A. Statistical Fidelity Analysis (Quality Metrics)

In this first stage, the ability of generative models to replicate the distribution of the minority class is evaluated. This phase is important because if the synthetic data does not maintain adequate statistical fidelity, the classifier could lose the ability to correctly generalize to real data.

The results are presented in Table II, where the performance of the four paradigms is compared using different quality metrics. Among them is the Wasserstein Distance (WD), used to measure the similarity between real and synthetic data. In addition, the Correlation Difference (Corr) and the Categorical Difference (CAT) were used to verify whether the relationships between variables and the categorical proportions are adequately preserved. Finally, the Diversity Score (DIV) allowed evaluation of whether the model generates varied samples and avoids producing repetitive data that may affect learning.

The analysis of Table II shows different behaviors according to the characteristics of each dataset:

Domain of interpolation in all scenarios: SMOTE-NC obtains the lowest Wasserstein distance in the five datasets. In DS1 (Diabetes Health Indicators), with a ratio of 6.18:1, it reaches 0.0429 ± 0.0005 , compared with 0.1652 for CTGAN and 0.1471 for TVAE, which indicates that interpolation preserves the statistical structure well when there is a sufficient amount of real data.

Response in categorical datasets: An interesting case appears in the DS2 (Thyroid Cancer). Although it has the lowest ratio (2.55:1), the presence of many categorical variables represents a difficulty for several traditional methods. In this scenario, HYBRID achieves a better adaptation to the nominal proportions, reaching the smallest Categorical Difference (0.1269 ± 0.0279), thus surpassing SMOTE-NC (0.2887) and TVAE (0.3208).

Greater diversity in extreme scenarios: In the DS5, with a ratio of 29.99:1, there is a more marked difference between fidelity and diversity. SMOTE-NC keeps a low WD (0.0913) and a diversity (9.24) close to that of the real records (12.11), whereas CTGAN and HYBRID reach 42.55 and 46.20, together with the highest WD of the study (1.1981 and 1.0716). This diversity mostly reflects records dispersed outside the real minority class, consistent with the only 141 minority records available in 95 dimensions.

To complement this analysis, Fig. 3 shows the evaluation of fidelity and diversity corresponding to the base scenario DS1.

TABLE II. GLOBAL SUMMARY OF STATISTICAL FIDELITY METRICS (MEAN $\pm$ SD OVER 10 RUNS)

Dataset	Ratio	Method	WD ($\downarrow$)	Δ Corr* ($\downarrow$)	Δ Cat ($\downarrow$)	Div. ($\approx$ real)
DS1	6.18	SMOTE-NC	0.0429 ($\pm$ 0.0005)	0.0087 ($\pm$ 0.0008)	0.0548 ($\pm$ 0.0006)	2.8363 ($\pm$ 0.0277)
		CTGAN	0.1652 ($\pm$ 0.0449)	0.0726 ($\pm$ 0.0265)	0.1453 ($\pm$ 0.0178)	3.0061 ($\pm$ 0.2121)
		TVAE	0.1471 ($\pm$ 0.0370)	0.0635 ($\pm$ 0.0114)	0.1168 ($\pm$ 0.0168)	2.4460 ($\pm$ 0.1620)
		HYBRID	0.1449 ($\pm$ 0.0309)	0.0635 ($\pm$ 0.0184)	0.1265 ($\pm$ 0.0344)	2.8552 ($\pm$ 0.1382)
		Real minority	—	—	—	2.8756
DS2	2.55	SMOTE-NC	0.1278 ($\pm$ 0.0476)	-	0.2887 ($\pm$ 0.0250)	1.0809 ($\pm$ 0.0565)
		CTGAN	0.5730 ($\pm$ 0.1989)	-	0.1982 ($\pm$ 0.0576)	1.7503 ($\pm$ 0.1011)
		TVAE	0.1347 ($\pm$ 0.0422)	-	0.3208 ($\pm$ 0.0317)	1.0814 ($\pm$ 0.0358)
		HYBRID	0.7813 ($\pm$ 0.2234)	-	0.1269 ($\pm$ 0.0279)	1.2114 ($\pm$ 0.1222)
		Real minority	—	—	—	1.1628
DS3	3.10	SMOTE-NC	0.0958 ($\pm$ 0.0077)	0.0355 ($\pm$ 0.0028)	0.1659 ($\pm$ 0.0080)	3.5990 ($\pm$ 0.0348)
		CTGAN	0.6533 ($\pm$ 0.0976)	0.1335 ($\pm$ 0.0100)	0.1195 ($\pm$ 0.0146)	5.0332 ($\pm$ 0.2063)
		TVAE	0.2305 ($\pm$ 0.0232)	0.0751 ($\pm$ 0.0069)	0.2188 ($\pm$ 0.0199)	3.4832 ($\pm$ 0.0438)
		HYBRID	0.5045 ($\pm$ 0.0975)	0.1311 ($\pm$ 0.0125)	0.1273 ($\pm$ 0.0109)	4.9659 ($\pm$ 0.2123)
		Real minority	—	—	—	4.0682
DS4	3.52	SMOTE-NC	0.0281 ($\pm$ 0.0035)	0.0170 ($\pm$ 0.0027)	0.0621 ($\pm$ 0.0028)	3.6468 ($\pm$ 0.1274)
		CTGAN	0.1358 ($\pm$ 0.0132)	0.1981 ($\pm$ 0.0079)	0.3000 ($\pm$ 0.0282)	4.4143 ($\pm$ 0.1854)
		TVAE	0.0901 ($\pm$ 0.0083)	0.0701 ($\pm$ 0.0024)	0.1052 ($\pm$ 0.0256)	3.5753 ($\pm$ 0.1314)
		HYBRID	0.1159 ($\pm$ 0.0138)	0.0909 ($\pm$ 0.0059)	0.2758 ($\pm$ 0.0401)	3.8830 ($\pm$ 0.1048)
		Real minority	—	—	—	3.8432
DS5	29.99	SMOTE-NC	0.0913 ($\pm$ 0.0021)	0.0303 ($\pm$ 0.0015)	0.0146 ($\pm$ 0.0011)	9.2412 ($\pm$ 0.1190)
		CTGAN	1.1981 ($\pm$ 0.0699)	0.1523 ($\pm$ 0.0004)	0.0937 ($\pm$ 0.0032)	42.5466 ($\pm$ 1.4128)
		TVAE	0.2637 ($\pm$ 0.0037)	0.1177 ($\pm$ 0.0023)	0.0213 ($\pm$ 0.0001)	5.4916 ($\pm$ 0.1263)
		HYBRID	1.0716 ($\pm$ 0.0569)	0.1482 ($\pm$ 0.0058)	0.1144 ($\pm$ 0.0110)	46.2043 ($\pm$ 1.0787)
		Real minority	—	—	—	12.1129

($\downarrow$) indicates lower values are better; ($\approx$ real) indicates that values closest to the diversity of the real minority records are preferable.

WD: Wasserstein Distance; Δ Corr: Correlation Difference; Δ Cat: Categorical Difference; Div: Mean Pairwise Distance; Real minority: diversity of the real minority records of the internal training subset, used as reference.

*Note: Δ Corr is not applicable (-) for DS2 because it contains only one numerical feature (Table I), making pairwise correlation uncalculable.

Fidelity and Diversity Assessment
Dataset 1 | Imbalance Ratio 6.18:1 | Mean $\pm$ SD of 10 runs

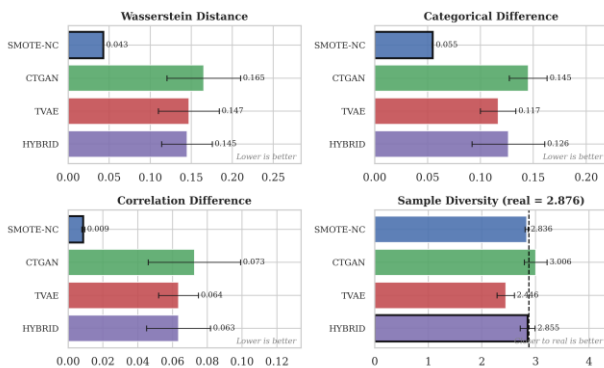


Fig. 3. Dataset fidelity and diversity evaluation 1.

Fig. 3 shows that in the Correlation Difference panel all methods stay at low values, with SMOTE-NC presenting the lowest error and an almost zero variance between runs. The Sample Diversity panel includes the diversity of the real minority records as a reference: in DS1, SMOTE-NC and HYBRID remain closest to it, CTGAN slightly exceeds it, and TVAE generates less varied records.

Distribution Fidelity on Most Predictive Features
Real minority records vs. SMOTE-NC synthetic records (seed 42)
Dataset 1 | Imbalance Ratio 6.18:1

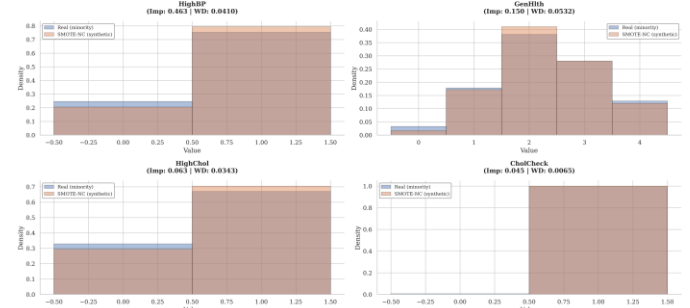


Fig. 4. Distribution fidelity of SMOTE-NC in the most predictive features of Dataset 1 (real minority vs. synthetic records).

Finally, Fig. 4 compares the real minority records of DS1 with the SMOTE-NC synthetic records (seed 42) in the four most important features (XGBoost importance on the internal training subset).

Fig. 4 shows real minority records in blue and synthetic records in orange. All distances are low (WD between 0.0065 and 0.0532), consistent with Table II, and three patterns are identified:

1) *Almost perfect mimicry*: In CholCheck (WD = 0.0065), 99.34% of the real and 99.99% of the synthetic records take the value 1, so the distribution is practically reproduced.

2) *Reinforcement of the dominant category*: In HighBP (Imp: 0.463), the most relevant feature, the share of the value 1 rises from 75.3% to 79.4% (WD = 0.0410), and in HighChol from 67.1% to 70.5%. This occurs because SMOTE-NC assigns to each synthetic record the most frequent category among its neighbours, which slightly inflates the dominant categories.

3) *Ordinal smoothing*: GenHlth shows the largest distance (WD = 0.0532): the central category gains weight (38.1% to 41.0%) and the extreme ones lose it (e.g., 3.3% to 1.7% in category 0), so rare values are under-represented, although the ordinal structure of the variable is preserved.

B. Evaluation of Predictive Performance (Utility Analysis)

After validating the statistical fidelity, we proceed to evaluate the predictive utility of synthetic data, understood as its ability to improve the training of classifiers in scenarios with severe imbalance. This analysis seeks to determine whether the samples generated help to better identify the minority class without significantly affecting the overall performance of the system.

The consolidated results with CatBoost are presented in Table III, and those with XGBoost in Table VII (Appendix).

These values were obtained over 10 runs that regenerate the synthetic data, and they include the cost-sensitive baseline REAL-CW.

1) *Ranking resilience analysis (AUC-PR)*: The analysis of Table III allows us to identify different behaviors depending on the level of imbalance present in each dataset:

a) *Behavior in moderate ratios (DS1 to DS4)*: In scenarios with ratios between 2.55:1 and 6.18:1, no synthesis method improves the AUC-PR of the model trained with real data. In DS2 (Thyroid Cancer), all conditions are practically saturated (REAL: 0.9821; CTGAN: 0.9859), whereas in DS1 and DS4 all synthetic methods reduce AUC-PR significantly, with the largest losses for the deep models (e.g., DS1: 0.4242 with REAL versus 0.3482 with CTGAN and 0.3222 with TVAE).

b) *Performance at the critical limit (DS5)*: In the scenario with a 29.99:1 ratio, the effect depends on the classifier. With XGBoost, HYBRID (0.5252 ± 0.0222) and CTGAN (0.5137) significantly exceed the baseline (0.4872; $p=0.017$ and $p=0.021$ after Holm correction); with CatBoost, no method exceeds the baseline (0.5131), and HYBRID obtains 0.4902. Moreover, the best XGBoost result remains at the level of the CatBoost baseline, so the improvement is relative to a weaker classifier and should not be interpreted as a global superiority of the hybrid method (Section IV-D).

TABLE III. PREDICTIVE PERFORMANCE ON THE TEST PARTITION WITH CATBOOST (MEAN $\pm$ SD OVER 10 RUNS)

Dataset	Ratio	Method	AUC-PR	Recall	Precision	F1-Score	Specificity	G-mean	Bal. Acc.
DS1	6.18	REAL	0.4242 $\pm$ 0.0005	0.1617 $\pm$ 0.0016	0.5581 $\pm$ 0.0030	0.2508 $\pm$ 0.0021	0.9793 $\pm$ 0.0002	0.3980 $\pm$ 0.0020	0.5705 $\pm$ 0.0008
		REAL-CW	0.4237 $\pm$ 0.0003	0.7938 $\pm$ 0.0008	0.3079 $\pm$ 0.0002	0.4437 $\pm$ 0.0002	0.7112 $\pm$ 0.0003	0.7513 $\pm$ 0.0003	0.7525 $\pm$ 0.0003
		SMOTE-NC	0.4182 $\pm$ 0.0003	0.2549 $\pm$ 0.0017	0.5021 $\pm$ 0.0019	0.3382 $\pm$ 0.0017	0.9591 $\pm$ 0.0004	0.4945 $\pm$ 0.0016	0.6070 $\pm$ 0.0008
		CTGAN	0.3482 $\pm$ 0.0079	0.5835 $\pm$ 0.0249	0.3256 $\pm$ 0.0084	0.4176 $\pm$ 0.0057	0.8040 $\pm$ 0.0145	0.6847 $\pm$ 0.0093	0.6937 $\pm$ 0.0065
		TVAE	0.3222 $\pm$ 0.0084	0.4582 $\pm$ 0.0436	0.3257 $\pm$ 0.0090	0.3797 $\pm$ 0.0139	0.8460 $\pm$ 0.0177	0.6216 $\pm$ 0.0243	0.6521 $\pm$ 0.0135
		HYBRID	0.4138 $\pm$ 0.0024	0.1781 $\pm$ 0.0032	0.5320 $\pm$ 0.0049	0.2669 $\pm$ 0.0037	0.9746 $\pm$ 0.0006	0.4166 $\pm$ 0.0037	0.5764 $\pm$ 0.0014
DS2	2.55	REAL	0.9821 $\pm$ 0.0030	0.9091 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9524 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9535 $\pm$ 0.0000	0.9545 $\pm$ 0.0000
		REAL-CW	0.9819 $\pm$ 0.0042	0.9091 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9524 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9535 $\pm$ 0.0000	0.9545 $\pm$ 0.0000
		SMOTE-NC	0.9807 $\pm$ 0.0026	0.9091 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9524 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9535 $\pm$ 0.0000	0.9545 $\pm$ 0.0000
		CTGAN	0.9859 $\pm$ 0.0031	0.9136 $\pm$ 0.0144	0.9861 $\pm$ 0.0309	0.9483 $\pm$ 0.0176	0.9945 $\pm$ 0.0123	0.9532 $\pm$ 0.0102	0.9541 $\pm$ 0.0102
		TVAE	0.9803 $\pm$ 0.0033	0.9091 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9524 $\pm$ 0.0000	1.0000 $\pm$ 0.0000	0.9535 $\pm$ 0.0000	0.9545 $\pm$ 0.0000
		HYBRID	0.9857 $\pm$ 0.0058	0.9136 $\pm$ 0.0258	0.9907 $\pm$ 0.0196	0.9503 $\pm$ 0.0137	0.9964 $\pm$ 0.0077	0.9540 $\pm$ 0.0126	0.9550 $\pm$ 0.0120
DS3	3.10	REAL	0.8717 $\pm$ 0.0025	0.7442 $\pm$ 0.0081	0.8090 $\pm$ 0.0178	0.7752 $\pm$ 0.0110	0.9435 $\pm$ 0.0062	0.8380 $\pm$ 0.0062	0.8439 $\pm$ 0.0060
		REAL-CW	0.8683 $\pm$ 0.0053	0.7913 $\pm$ 0.0091	0.7609 $\pm$ 0.0158	0.7757 $\pm$ 0.0094	0.9201 $\pm$ 0.0070	0.8533 $\pm$ 0.0056	0.8557 $\pm$ 0.0054
		SMOTE-NC	0.8692 $\pm$ 0.0039	0.7885 $\pm$ 0.0111	0.7826 $\pm$ 0.0110	0.7854 $\pm$ 0.0062	0.9296 $\pm$ 0.0050	0.8561 $\pm$ 0.0052	0.8590 $\pm$ 0.0047
		CTGAN	0.8666 $\pm$ 0.0092	0.7644 $\pm$ 0.0130	0.7662 $\pm$ 0.0166	0.7652 $\pm$ 0.0114	0.9250 $\pm$ 0.0071	0.8409 $\pm$ 0.0077	0.8447 $\pm$ 0.0073

		TVAE	0.8652 0.0050	±	0.7856 0.0111	±	0.7950 0.0171	±	0.7902 0.0118	±	0.9349 0.0066	±	0.8570 0.0074	±	0.8602 0.0072	±
		HYBRID	0.8677 0.0070	±	0.7481 0.0174	±	0.7789 0.0095	±	0.7631 0.0100	±	0.9318 0.0040	±	0.8348 0.0091	±	0.8399 0.0082	±
DS4	3.52	REAL	0.5593 0.0021	±	0.3648 0.0025	±	0.6613 0.0025	±	0.4702 0.0024	±	0.9469 0.0006	±	0.5877 0.0020	±	0.6559 0.0012	±
		REAL-CW	0.5598 0.0024	±	0.6130 0.0026	±	0.4825 0.0019	±	0.5400 0.0016	±	0.8133 0.0016	±	0.7061 0.0013	±	0.7131 0.0012	±
		SMOTE-NC	0.5515 0.0018	±	0.4669 0.0050	±	0.5638 0.0054	±	0.5108 0.0039	±	0.8974 0.0024	±	0.6473 0.0032	±	0.6822 0.0024	±
		CTGAN	0.5300 0.0036	±	0.3949 0.0041	±	0.6318 0.0057	±	0.4860 0.0031	±	0.9346 0.0019	±	0.6075 0.0028	±	0.6648 0.0017	±
		TVAE	0.5250 0.0030	±	0.4352 0.0053	±	0.5771 0.0076	±	0.4962 0.0041	±	0.9094 0.0032	±	0.6291 0.0034	±	0.6723 0.0023	±
		HYBRID	0.5328 0.0063	±	0.3958 0.0047	±	0.6226 0.0120	±	0.4839 0.0048	±	0.9318 0.0037	±	0.6073 0.0034	±	0.6638 0.0025	±
DS5	29.99	REAL	0.5131 0.0245	±	0.2409 0.0342	±	0.6157 0.0540	±	0.3456 0.0415	±	0.9950 0.0008	±	0.4885 0.0351	±	0.6180 0.0171	±
		REAL-CW	0.4727 0.0286	±	0.6023 0.0193	±	0.4504 0.0187	±	0.5152 0.0163	±	0.9755 0.0018	±	0.7664 0.0123	±	0.7889 0.0096	±
		SMOTE-NC	0.5045 0.0264	±	0.6273 0.0117	±	0.4168 0.0187	±	0.5007 0.0160	±	0.9707 0.0020	±	0.7803 0.0077	±	0.7990 0.0064	±
		CTGAN	0.5119 0.0268	±	0.2455 0.0383	±	0.6093 0.0575	±	0.3484 0.0404	±	0.9947 0.0012	±	0.4929 0.0366	±	0.6201 0.0189	±
		TVAE	0.4472 0.0220	±	0.3955 0.0469	±	0.5790 0.0433	±	0.4695 0.0461	±	0.9905 0.0010	±	0.6248 0.0376	±	0.6930 0.0236	±
		HYBRID	0.4902 0.0514	±	0.2409 0.0267	±	0.6160 0.0901	±	0.3453 0.0367	±	0.9948 0.0017	±	0.4889 0.0273	±	0.6179 0.0135	±

REAL: real data without treatment; REAL-CW: real data with class weights (cost-sensitive baseline). The results with XGBoost are reported in Table VII (Appendix)

To complement this analysis, Fig. 5 shows the evolution of the performance of the different methods as the level of imbalance increases.

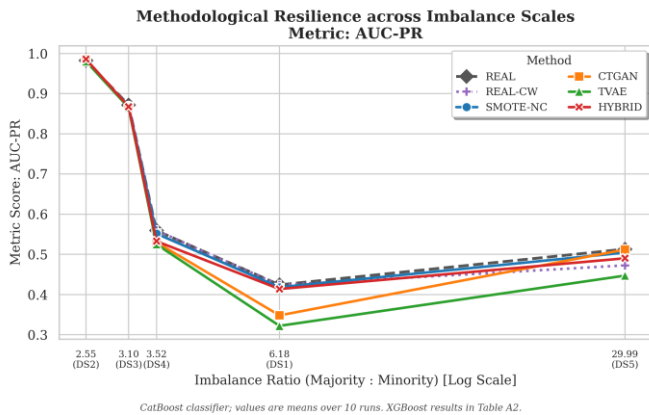


Fig. 5. Methodological resilience through imbalance scales (Metric: AUC-PR).

Fig. 5 (CatBoost) shows that the AUC-PR varies more with the dataset than with the generation method; the only significant improvements appear with XGBoost in DS5 (Table V). Fig. 6 details the differences in DS5 for both classifiers; given the small number of positive records in the test partition of this dataset (44), these differences are sensitive to a few cases.

2) *Stability and fair evaluation (balanced accuracy)*: Because traditional Accuracy can be misleading in imbalanced scenarios, Balanced Accuracy was used to evaluate performance by considering both classes in a more balanced way.

a) *Impact of balancing*: Balanced Accuracy increases significantly in 21 of the 40 comparisons. In DS1, for example,

CTGAN increases it from 0.5705 to 0.6937, and the cost-sensitive baseline REAL-CW reaches 0.7525. This shows that the classifier stops mainly favoring the majority class.

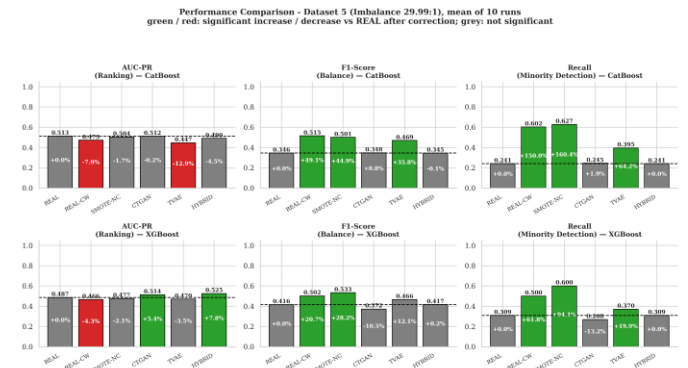


Fig. 6. Relative performance improvements among generative methods for dataset 5.

b) *Behavior in the extreme scenario*: In DS5, SMOTE-NC (0.7990) and REAL-CW (0.7889) obtain the highest Balanced Accuracy, followed by TVAE (0.6930), while CTGAN (0.6201) and HYBRID (0.6179) remain at the level of the baseline (0.6180).

3) *Operational sensitivity and stability (recall and F1-score)*: The most important impact of data synthesis is observed in the model's ability to detect rare events, evaluated by Recall and F1-Score. According to the results of Table III, the use of synthetic data considerably modifies the behavior of the classifier.

a) *Displacement of the decision boundary*: In DS1 (Diabetes), the model trained with real data obtains a Recall of only 0.1617 ± 0.0016 . By incorporating CTGAN, this value

increases to 0.5835 ± 0.0249 , and REAL-CW reaches 0.7938 without any synthetic records. The hybrid approach is the most conservative (0.1781), since its final set contains only CTGAN records generated from a densified minority.

Gain in extreme scenarios (DS5): In the scenario of financial bankruptcy, SMOTE-NC obtains the largest increase in Recall, from 0.2409 to 0.6273 (CatBoost), similar to REAL-CW (0.6023), and both obtain the highest F1-Score (0.5007 and 0.5152 versus 0.3456). This improvement is accompanied by a reduction in precision. CTGAN and HYBRID do not change Recall significantly in this dataset.

As seen in Fig. 7, the F1-Score improves mainly in the scenarios where the baseline detects few positive cases (DS1, DS4 and DS5), and the largest values are obtained by REAL-CW and SMOTE-NC. In DS5, the F1-Score of the CTGAN-based methods also shows the largest variability between runs.

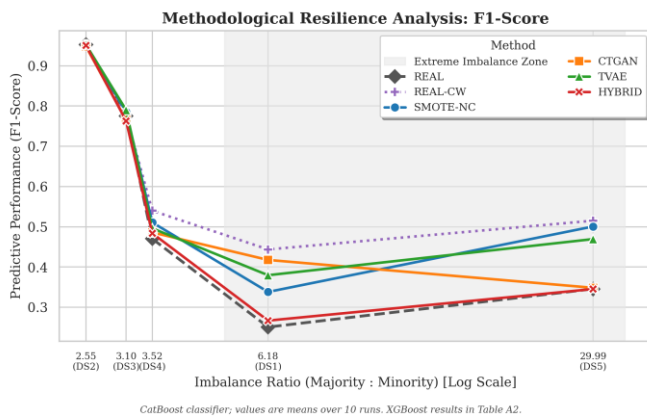


Fig. 7. F1 score across imbalance scales and methodological resilience.

4) Diagnostic quality and trade-off (precision): To complement the analysis, precision was also evaluated, a metric that allows us to measure how reliable the positive predictions made by the model are. In imbalanced scenarios, this metric is important because it helps to observe the cost of increasing sensitivity through synthetic data.

a) Relationship between precision and recall: In DS1 (Diabetes), CTGAN raises Recall to 58.3%, but reduces Precision from 0.5581 (REAL) to 0.3256; that is, about two out of every three positive predictions are false alarms. Across the 40 comparisons, Precision decreases significantly in 25 and Specificity in 27, while the G-mean, which balances Recall and Specificity, increases in 23.

b) Behavior in extreme imbalance (DS5): While SMOTE-NC reduces Precision to 0.4168 to reach a Recall of 0.6273, TVAE (0.5790) and HYBRID (0.6160) maintain a precision close to that of the baseline (0.6157), but with a smaller gain in Recall. The methods therefore move along the precision-recall trade-off rather than improving both metrics at once.

To complement this analysis, Fig. 8 shows the operating point (Recall, Precision) of each condition at the default threshold, together with iso-F1 curves. In all datasets except DS2, the synthetic methods and REAL-CW move the operating point towards higher Recall and lower Precision, and the points

remain close to the same iso-F1 curves, which confirms that the main effect is a displacement of the decision threshold rather than a better separation of the classes.

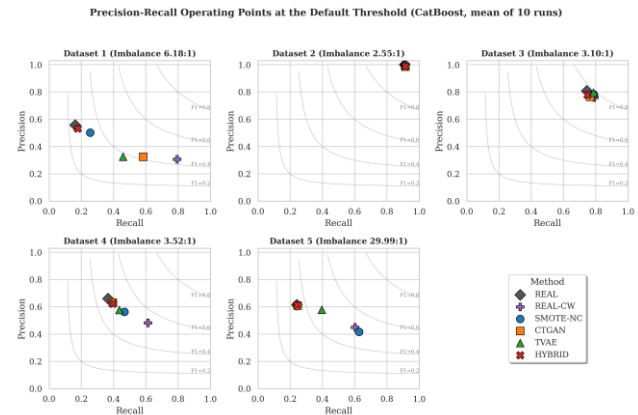


Fig. 8. Precision-Recall operating points at the default threshold (CatBoost).

5) Operational breakdown of positive detections: Fig. 9 to 13 show the mean confusion matrices over the 10 runs (CatBoost) of the baseline and of one representative method per dataset.

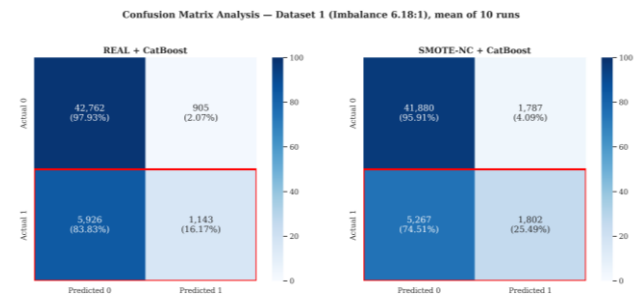


Fig. 9. Confusion matrix analysis: Dataset 1

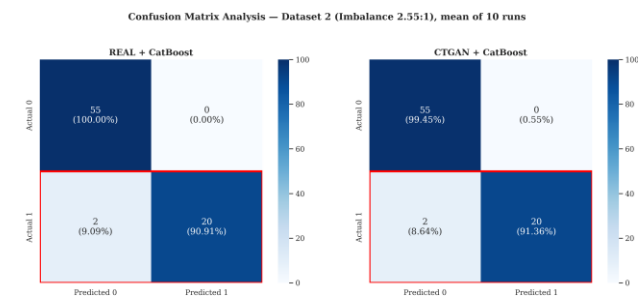


Fig. 10. Confusion matrix analysis: Dataset 2.

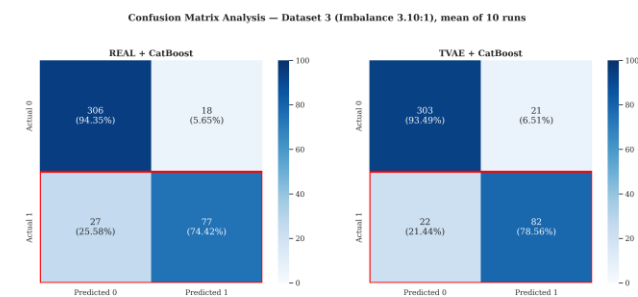


Fig. 11. Confusion matrix analysis: Dataset 3.

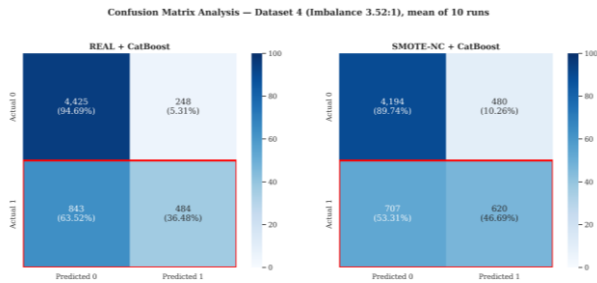


Fig. 12. Confusion matrix analysis: Dataset 4.

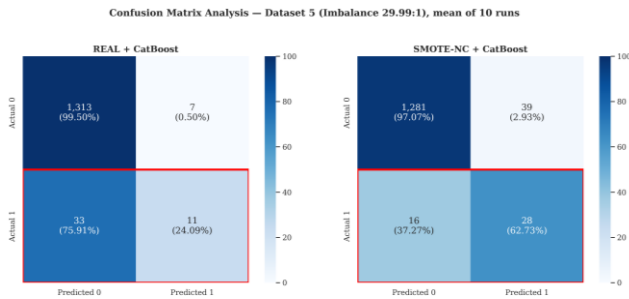


Fig. 13. Confusion matrix analysis: Dataset 5.

Scenario 1: Large-Scale Health Indicators (DS1). With an imbalance ratio of 6.18:1, the base classifier detects, on average, 1,143 of the 7,069 positive cases (16.2%). SMOTE-NC raises this number to 1,802, CTGAN to 4,125, and REAL-CW to 5,611; however, false positives grow from 905 to 8,559 with CTGAN and 12,612 with REAL-CW.

Scenario 2: Moderate Imbalance (DS2, DS3 and DS4). In DS2 (Thyroid Cancer), all conditions detect about 20 of the 22 positive cases, so synthesis does not disturb an already saturated learning. In DS3 (AIDS), SMOTE-NC and TVAE raise the detections from 77.4 to 82.0 and 81.7 of 104 cases. The largest impact of this group occurs in DS4 (Credit Default), where SMOTE-NC increases the detections from 484 to 620 of 1,327 cases, and REAL-CW to 813.

Scenario 3: Behavior at the Operational Limit (DS5). With a ratio of 29.99:1, the base model detects, on average, 10.6 of the 44 companies in bankruptcy. SMOTE-NC (27.6) and REAL-CW (26.5) more than double this number, whereas CTGAN (10.8) and HYBRID (10.6) remain at the level of the baseline. Given the small number of positive cases, each detected company represents 2.3 points of Recall.

C. Trade-Off Analysis: Statistical Fidelity Vs. Predictive Utility

Fig. 14 relates the fidelity error (Wasserstein distance) of each method to its AUC-PR (CatBoost); the dotted lines mark the medians of both axes. SMOTE-NC occupies the low-error side in all datasets, while the CTGAN-based methods move to the right, especially in DS2, DS3, and DS5.

The vertical position depends mainly on the dataset rather than on the method, and the largest fidelity errors (CTGAN and HYBRID in DS5) do not correspond to the largest losses of AUC-PR, which occur in DS1 and DS4 with moderate errors. Therefore, fidelity alone is not a sufficient predictor of utility, which reinforces the need for the dual evaluation.

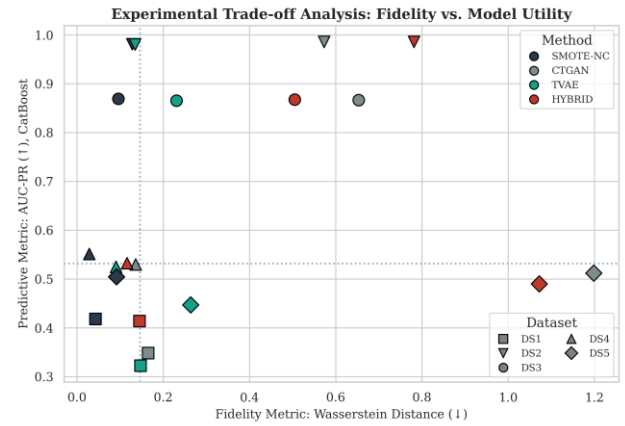


Fig. 14. Trade-off: fidelity vs. model utility.

D. Statistical Validation of Results

Tables II and III present averages and standard deviations over 10 runs, which describe the stability of each method but do not allow establishing whether the differences observed are attributable to the synthesis procedure or to random variability. To resolve this issue, the hypothesis testing protocol described in Section III-F is used. The results are summarized in Table IV.

The analysis of Table IV reveals an asymmetric pattern that substantively qualifies the interpretation of the results:

Few Improvements in AUC-PR. of the 40 Comparisons (five Datasets $\times$ Two Classifiers $\times$ Four Methods): AUC-PR Decreases Significantly in 19, Does Not Change Significantly in 19, and Increases in Only 2: CTGAN and HYBRID in DS5 with XGBoost. The decreases are concentrated in DS1 and DS4, where the four methods reduce AUC-PR with both classifiers.

Robust and Consistent Improvements in Recall. The picture is reversed when evaluating the Detection of the Minority Class. Of the 40 Comparisons, 27 Show Significant Increases in Recall, none show a Significant Decrease, 8 Are Inconclusive, and 5 Correspond to DS2, where the Metric Is Saturated. The counterpart is precision, Which Decreases in 25 Comparisons. The G-mean Increases in 23, and the F1-Score in 18, with No Significant Decrease in Either Metric.

Dependence on the Classifier in the Extreme Scenario. As shown in Table V, the Sign of the Effect in DS5 depends on the Classifier: under XGBoost, HYBRID exceeds the baseline by +0.0380 ($p=0.017$ after Holm correction), while under CatBoost it is below by -0.0229 ($p=0.640$). In addition, the XGBoost baseline (0.4872) is lower than the CatBoost baseline (0.5131), so the improvement obtained with XGBoost brings this classifier to the level of CatBoost rather than above it. A possible explanation of this reversal is discussed in Section V.

On the whole, statistical validation leads to a more accurate reading of the results: data synthesis does not improve the overall ordering of predictions measured by AUC-PR and, in several scenarios, degrades it significantly. Its contribution is concentrated in a reproducible displacement of the decision boundary towards the minority class, which translates into increases in Recall and G-mean of large magnitude, at the cost of Precision and Specificity.

TABLE IV. STATISTICAL VALIDATION: MEAN DIFFERENCE WITH RESPECT TO REAL (Δ), SIGNIFICANCE AND EFFECT SIZE D_Z (N = 10 PAIRED RUNS, SAME CLASSIFIER)

Dataset	Method	AUC-PR		Recall		Precision		G-mean	
		CatBoost	XGBoost	CatBoost	XGBoost	CatBoost	XGBoost	CatBoost	XGBoost
DS1 (6.18)	REAL-CW	-0.0006** (-1.1)	+0.0009** (1.2)	+0.6320*** (337)	+0.6183*** (330)	-0.2502*** (-87)	-0.2414*** (-81)	+0.3534*** (184)	+0.3465*** (200)
	SMOTE-NC	-0.0060*** (-10)	-0.0102*** (-9.3)	+0.0932*** (69)	+0.1702** (68)	-0.0561*** (-15)	-0.0920*** (-18)	+0.0965*** (66)	+0.1576** (79)
	CTGAN	-0.0760*** (-9.6)	-0.0847*** (-12)	+0.4218*** (17)	+0.3921*** (16)	-0.2325*** (-25)	-0.2271*** (-26)	+0.2867*** (31)	+0.2692*** (26)
	TVAE	-0.1020*** (-12)	-0.1029*** (-12)	+0.2964*** (6.9)	+0.2671*** (6.6)	-0.2325*** (-22)	-0.2255*** (-24)	+0.2237*** (9.5)	+0.2040*** (8.6)
	HYBRID	-0.0104** (-3.8)	-0.0076*** (-3.2)	+0.0164*** (4.1)	+0.0118*** (2.5)	-0.0262*** (-5.2)	-0.0221*** (-3.8)	+0.0187*** (4.0)	+0.0133*** (2.5)
DS2 (2.55)	REAL-CW	-0.0002 (-0.0)	-0.0010* (-1.0)	0 (—)	0 (—)	0 (—)	-0.0314* (-1.4)	0 (—)	-0.0063* (-1.4)
	SMOTE-NC	-0.0014 (-0.4)	-0.0003 (-0.2)	0 (—)	0 (—)	0 (—)	-0.0136 (-0.6)	0 (—)	-0.0027 (-0.6)
	CTGAN	+0.0038 (0.8)	-0.0006 (-0.2)	+0.0045 (0.3)	0 (—)	-0.0139 (-0.4)	-0.0265 (-1.2)	-0.0003 (-0.0)	-0.0054 (-1.2)
	TVAE	-0.0018 (-0.4)	-0.0009 (-0.9)	0 (—)	0 (—)	0 (—)	-0.0045 (-0.2)	0 (—)	-0.0009 (-0.2)
	HYBRID	+0.0036 (0.6)	+0.0003 (0.1)	+0.0045 (0.2)	-0.0045 (-0.3)	-0.0093 (-0.5)	-0.0132 (-0.4)	+0.0005 (0.0)	-0.0050 (-0.5)
DS3 (3.10)	REAL-CW	-0.0034 (-0.6)	-0.0030* (-0.8)	+0.0471*** (4.5)	+0.0288*** (2.4)	-0.0481*** (-2.3)	-0.0269*** (-2.7)	+0.0153*** (2.1)	+0.0105** (1.5)
	SMOTE-NC	-0.0025 (-0.5)	-0.0026 (-0.7)	+0.0442*** (2.9)	+0.0250** (1.9)	-0.0264** (-1.6)	-0.0278*** (-2.7)	+0.0181*** (2.3)	+0.0084* (1.2)
	CTGAN	-0.0051 (-0.6)	-0.0089 (-0.8)	+0.0202** (1.6)	+0.0269*** (2.0)	-0.0428*** (-2.2)	-0.0523** (-3.0)	+0.0029 (0.4)	+0.0047 (0.5)
	TVAE	-0.0065** (-1.4)	-0.0114** (-1.6)	+0.0413*** (2.5)	+0.0231* (1.1)	-0.0140 (-0.6)	-0.0365*** (-2.0)	+0.0190** (1.7)	+0.0058 (0.5)
	HYBRID	-0.0040 (-0.5)	-0.0050 (-0.7)	+0.0038 (0.2)	+0.0192* (0.9)	-0.0302** (-1.6)	-0.0349*** (-1.8)	-0.0031 (-0.3)	+0.0041 (0.4)
DS4 (3.52)	REAL-CW	+0.0005 (0.2)	-0.0019 (-0.5)	+0.2482*** (63)	+0.2255*** (42)	-0.1787*** (-53)	-0.1735*** (-40)	+0.1183*** (43)	+0.1079*** (30)
	SMOTE-NC	-0.0078*** (-3.5)	-0.0212*** (-5.4)	+0.1021*** (16)	+0.1039*** (28)	-0.0975*** (-15)	-0.1259*** (-13)	+0.0596*** (13)	+0.0554*** (16)
	CTGAN	-0.0293*** (-6.9)	-0.0127*** (-2.5)	+0.0301*** (6.4)	+0.0170*** (3.2)	-0.0295*** (-4.1)	-0.0208*** (-2.8)	+0.0197*** (5.8)	+0.0110*** (2.5)
	TVAE	-0.0342*** (-8.1)	-0.0264*** (-6.5)	+0.0704*** (12)	+0.0444*** (6.5)	-0.0841*** (-12)	-0.0686*** (-10)	+0.0413*** (11)	+0.0255*** (5.3)
	HYBRID	-0.0265*** (-3.8)	-0.0159*** (-2.1)	+0.0310*** (6.2)	+0.0130*** (2.2)	-0.0387*** (-3.2)	-0.0251*** (-2.0)	+0.0195*** (5.2)	+0.0077*** (1.6)
DS5 (29.99)	REAL-CW	-0.0404* (-0.9)	-0.0210* (-0.7)	+0.3614*** (9.6)	+0.1909*** (3.3)	-0.1652*** (-3.5)	-0.1341*** (-2.2)	+0.2779*** (7.8)	+0.1476*** (3.2)
	SMOTE-NC	-0.0086 (-0.3)	-0.0104 (-0.5)	+0.3864*** (11)	+0.2909*** (6.0)	-0.1988*** (-3.9)	-0.1588*** (-3.1)	+0.2918*** (8.5)	+0.2126*** (5.4)
	CTGAN	-0.0012 (-0.0)	+0.0265* (1.1)	+0.0045 (0.1)	-0.0409 (-0.7)	-0.0063 (-0.1)	-0.0264 (-0.3)	+0.0044 (0.1)	-0.0380 (-0.7)
	TVAE	-0.0659*** (-2.1)	-0.0168 (-0.4)	+0.1545*** (3.3)	+0.0614* (1.0)	-0.0367 (-0.5)	-0.0062 (-0.1)	+0.1364*** (3.5)	+0.0525* (1.0)
	HYBRID	-0.0229 (-0.4)	+0.0380* (1.2)	+0.0000 (0.0)	+0.0000 (0.0)	+0.0003 (0.0)	+0.0060 (0.1)	+0.0004 (0.0)	+0.0004 (0.0)

Cell format: Δ with significance (effect size d_z). *** p<0.001; ** p<0.01; * p<0.05; no mark: not significant. p-values from the paired t-test, or from the Wilcoxon signed-rank test when Shapiro-Wilk rejects the normality of the differences, with Holm correction within each family of four synthetic methods (same dataset, classifier and metric). REAL-CW is tested separately without correction. 0 (—): identical results in all runs (metric saturated). Complete results for all metrics, confidence intervals and uncorrected p-values are available from the corresponding author upon request.

TABLE V. DECOMPOSITION OF THE DS5 SCENARIO (AUC-PR) BY CLASSIFIER (MEAN $\pm$ SD OVER 10 RUNS)

Method	CatBoost	Δ vs REAL	p (Holm)	XGBoost	Δ vs REAL	p (Holm)
REAL	0.5131 $\pm$ 0.0245	—	—	0.4872 $\pm$ 0.0173	—	—
REAL-CW	0.4727 $\pm$ 0.0286	-0.0404	0.017†	0.4662 $\pm$ 0.0176	-0.0210	0.043†
SMOTE-NC	0.5045 $\pm$ 0.0264	-0.0086	0.822	0.4768 $\pm$ 0.0101	-0.0104	0.246
CTGAN	0.5119 $\pm$ 0.0268	-0.0012	0.922	0.5137 $\pm$ 0.0162	+0.0265	0.021
TVAE	0.4472 $\pm$ 0.0220	-0.0659	<0.001	0.4704 $\pm$ 0.0300	-0.0168	0.246
HYBRID	0.4902 $\pm$ 0.0514	-0.0229	0.640	0.5252 $\pm$ 0.0222	+0.0380	0.017

Δ vs REAL: mean difference compared to the baseline trained with the same classifier. The p-values are corrected with Holm within each classifier; † REAL-CW: uncorrected p-value.

Comparison with the Cost-sensitive Baseline. REAL-CW Significantly Increases Recall in All Non-saturated Comparisons (e.g., DS1: from 0.1617 to 0.7938) without the substantial loss of AUC-PR observed with the synthetic methods in DS1 and DS4, and it obtains the highest F1-Score and G-mean in DS1 and DS4; in DS5 it is comparable to SMOTE-NC.

V. DISCUSSION

Together, the results show that data synthesis does not work the same in all cases, but rather depends on the type of dataset, the level of imbalance, and how much information is actually available. In terms of fidelity, SMOTE-NC performs better when the data has a well-defined numerical part, and the imbalance is not as extreme as in DS1. There, it manages to maintain the structure of the data well with a low Wasserstein Distance, which basically shows that interpolation works well when the database is "stable". On the other hand, when the scenario becomes more complicated and there are many categorical variables, as in DS2, the CTGAN-based methods (HYBRID and CTGAN) reproduce the categorical proportions better, which is consistent with the conditional generator of CTGAN, designed to sample the categories of discrete columns in a balanced way [37].

When moving to predictive utility, the statistical validation of Section IV-D forces us to distinguish two planes that initially seemed to be one. In DS5, HYBRID and CTGAN improve the AUC-PR of XGBoost significantly (HYBRID: 0.5252 versus 0.4872) but not that of CatBoost, and they do not exceed the best baseline; the hypothesis that synthesis improves the global ordering under extreme scarcity is therefore only partially supported. The consistent effect is on sensitivity: CTGAN raises Recall in DS1 from 0.1617 to 0.5835, and SMOTE-NC in DS5 from 0.2409 to 0.6273. This suggests that the real contribution of the synthesis is to relocate the decision boundary to regions where the base model lacked information to risk a positive prediction.

This interpretation is reinforced by REAL-CW, which produces a similar or larger displacement without generating records, since oversampling and class reweighting act on the same mechanism: the relative weight of the minority class in the loss.

The dependence on the classifier in DS5 admits two explanations. First, the differences (0.023 to 0.038 in AUC-PR) are of the same order as the variability between runs, and the test partition contains only 44 positive records. Second, CatBoost builds symmetric trees, which limit the isolation of small regions, whereas XGBoost grows asymmetric trees that can exploit the dispersed CTGAN records in 95 dimensions; moreover, the XGBoost baseline is weaker in this dataset. This second explanation is a hypothesis, not a verified mechanism.

Regarding extreme imbalance, the generators learn from 22,622 minority records in DS1 but only 141 in DS5 (about 1.5 per feature) and 69 in DS2. This explains the large fidelity errors of the deep models in DS5 and DS2, and why SMOTE-NC, which only interpolates between existing neighbours, obtains the largest gain in Recall in DS5. The behavior in DS5 therefore reflects extreme imbalance, a very small minority class, and high

dimensionality together, and, being the only dataset above 20:1, it should be confirmed in additional datasets.

In this sense, the hybrid SMOTE-NC-CTGAN approach is the most conservative synthetic method: it keeps precision close to the baseline and loses less AUC-PR than CTGAN and TVAE in DS1, but its gain in Recall is small, and its only significant improvement in AUC-PR appears in DS5 with XGBoost. SMOTE-NC is the most reliable option when the minority class is reasonably represented.

From a practical point of view, this displacement has a cost: in DS1, CTGAN lowers Precision from 0.56 to 0.33, so about two out of three positive predictions are false alarms, which may mean unnecessary clinical tests or the rejection of solvent clients. The choice of method should therefore consider the relative cost of both errors, and reweighting the classes should be considered first, since it obtains a similar displacement at a lower cost.

Of course, the study was not free of limitations. The datasets come from a curated repository (UCI) with few missing values and little noise, so the patterns may differ in messier real-world health and finance data [11]; only one dataset exceeds 20:1; the hyperparameter grid of the generators was small [9], [20]; and a single train/test partition and the default threshold were used. Finally, privacy was not addressed; future work should include multi-class problems and differential privacy.

VI. CONCLUSION

From the results obtained, several conclusions can be drawn about the impact of data synthesis on imbalanced classification problems: Dependence on the Ratio and Behavior under Extreme Imbalance: the Performance of Synthesis Techniques Depends Directly on the Level of Imbalance. SMOTE-NC Preserves the Structure of the Data Best in All Scenarios and Is the Most Reliable Option in Moderate Imbalance. When the Imbalance Is Extreme (29.99:1), CTGAN and the Hybrid Approach Significantly Improve the AUC-PR of XGBoost but Not that of CatBoost, and They Do Not Exceed the Best Baseline; This Behavior Is Therefore Classifier-dependent. This Scenario Is Represented by a Single Dataset with a Very Small Minority Class, Which Limits the Generality of This Conclusion.

Diversity Must Be Interpreted Together with Fidelity: in High-scarcity Scenarios, the GAN-based and Hybrid Models Generate the Most Diverse Records, but This Diversity Coincides with the Largest Fidelity Errors and Translates into a Significant Gain in AUC-PR Only in One Scenario (DS5 with XGBoost). Therefore, Diversity Alone Is Not an Indicator of the Usefulness of Synthetic Data.

Hybrid Approach Balance: the SMOTE-NC-CTGAN Model Is the Most Conservative of the Synthetic Methods: it Keeps Precision Close to that of the Baseline and achieves the Only Significant Improvements in AUC-PR, together with CTGAN, in the Most Critical Scenario with XGBoost (DS5: 0.5252 Versus 0.4872). However, Its Gain in Recall Is Small, and Its Advantage Depends on the Classifier, so Its Merit Lies in This Specific Scenario and Not in a General Superiority over the Baseline.

Improvement in the Detection of the Minority Class: Finally, and as a Main Result of This Research, it Is Concluded that Data Synthesis Improves the Model's Ability to identify the Minority Class in a Statistically Robust Way. Recall Increases Significantly in 27 of 40 Comparisons and Never Decreases, at the Cost of AUC-PR and of Precision (lower in 25); Reweighting the Classes Achieves a Comparable Displacement. This Trade-off Is Favorable Where the Cost of a False Negative Dominates, Such as in clinical diagnosis and Credit Risk, and must be weighed against the Cost of False Alarms.

REFERENCES

- [1] M. G. Wijaya, M. F. Pinarnggi, A. Y. Zakiyyah, and Meiliana, "Comparative analysis of machine learning algorithms and data balancing techniques for credit card fraud detection," *Procedia Computer Science*, vol. 245, pp. 677–688, 2024, <https://doi.org/10.1016/j.procs.2024.10.294>
- [2] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing class imbalance of health data: a systematic literature review on modified synthetic minority oversampling technique strategies," *JOIV: Int. J. Informat. Visualiz.*, vol. 8, no. 3, pp. 1310–1318, 2024, <http://dx.doi.org/10.62527/joiv.8.3.2283>
- [3] A. G. Akintola, "Performance analysis of machine learning methods with class imbalance problem in Android malware detection," *Int. J. Interact. Mob. Technol.*, vol. 16, no. 10, pp. 140–162, May 2022, <https://doi.org/10.3991/ijim.v16i10.29687>
- [4] A. G. Udu, M. T. Salman, M. K. Ghalati, A. Lecchini-Visintini, D. R. Siddle, and H. Dong, "Emerging SMOTE and GAN variants for data augmentation in imbalance machine learning tasks: A review," *IEEE Access*, vol. 13, pp. 113838–113853, 2025, <https://doi.org/10.1109/ACCESS.2025.3584532>
- [5] N. Peppes, T. Alexakis, E. Daskalakis, and E. Adamopoulou, "Evaluating synthetic malicious network traffic generated by GAN and VAE models: a data quality perspective," *Future Internet*, vol. 17, no. 12, p. 561, 2025, <https://doi.org/10.3390/fi17120561>
- [6] R. Sauber-Cole and T. M. Khoshgoftaar, "The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey," *J. Big Data*, vol. 9, p. 98, 2022, <https://doi.org/10.1186/s40537-022-00648-6>
- [7] T. Hasan and S. Tasnim, "Real-time explainable IoT security with machine learning and CTGAN-enhanced detection for resource-constrained devices," *Ad Hoc Netw.*, vol. 178, p. 103937, 2025, <https://doi.org/10.1016/j.adhoc.2025.103937>
- [8] D. Cullen, J. Halladay, N. Briner, R. Basnet, J. Bergen, and T. Doleck, "Evaluation of synthetic data generation techniques in the domain of anonymous traffic classification," *IEEE Access*, vol. 10, pp. 129612–129625, 2022, <https://doi.org/10.1109/ACCESS.2022.3228507>
- [9] D.-H. Won, K.-S. Shin, and S. Youm, "Synthetic data augmentation for imbalanced tabular data: a comparative study of generation methods," *Electronics*, vol. 15, no. 4, p. 883, 2026, <https://doi.org/10.3390/electronics15040883>
- [10] W. A. W. A. Bakar, N. L. N. Josdi, M. Man, and E. A. Kadir, "Evaluating CTGAN-generated synthetic data for heart disease prediction: fidelity, predictive utility, and feature preservation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 12, 2025, <http://dx.doi.org/10.14569/IJACSA.2025.0161296>
- [11] M. Miletic and M. Sariyar, "Challenges of using synthetic data generation methods for tabular microdata," *Appl. Sci.*, vol. 14, no. 14, p. 5975, 2024, <https://doi.org/10.3390/app14145975>
- [12] A. Sharma, P. K. Singh, and R. Chandra, "SMOTified-GAN for class imbalanced pattern classification problems," *IEEE Access*, vol. 10, pp. 30655–30665, 2022, <https://doi.org/10.1109/ACCESS.2022.3158977>
- [13] D.-K. Kim and Y. K. Chung, "A GAN-based hybrid approach for addressing class imbalance in machine learning," *Comput. Sci. Inf. Syst.*, vol. 22, no. 3, pp. 673–691, 2025, <https://doi.org/10.2298/CSIS231113014K>
- [14] I. B. Mustapha, S. Hasan, H. S. Y. Nabbus, M. M. A. Montaser, S. O. Olatunji, and S. M. Shamsuddin, "Investigating group distributionally robust optimization for deep imbalanced learning: a case study of binary tabular data classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, 2023, <https://doi.org/10.14569/ijacsa.2023.0140286>
- [15] H. A. Eren, H. İ. Emek, and S. B. Keser, "Comparison of synthetic data generation techniques for obesity level prediction based on dietary habits and physical status," *Artif. Intell. Health*, vol. 2, no. 4, pp. 47–74, 2025, <https://doi.org/10.36922/AIH025140027>
- [16] S. B. Belhaouari, A. Islam, K. Kassoul, A. Al-Fuqaha, A. Bouzerdoum, "Oversampling techniques for imbalanced data in regression," *Expert Syst. Appl.*, vol. 252, part B, p. 124118, 2024, <https://doi.org/10.1016/j.eswa.2024.124118>
- [17] K. Omari, C. Taoussi, and A. Oukhatar, "Comparative analysis of undersampling, oversampling, and SMOTE techniques for addressing class imbalance in phishing website detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, 2025, <http://dx.doi.org/10.14569/IJACSA.2025.0160276>
- [18] D. Utari, "Integration of SVM and SMOTE-NC for classification of heart failure patients," *Barekeng: J. Ilmu Mat. Ter.*, vol. 17, no. 4, pp. 2263–2272, 2023, <https://doi.org/10.30598/barekengvol17iss4pp2263-2272>
- [19] H. Abtahi, M. Edalatfard, M. Gholamzadeh, and E. Khakvatan, "Predicting zygosity based on physical similarity of twin pairs with the aid of machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 26, no. 1, p. 32, 2025, <https://doi.org/10.1186/s12911-025-03334-8>
- [20] P. Yadav, M. Gaur, R. K. Madhukar, G. Verma, P. Kumar, N. Fatima, S. Sarwar, and Y. R. Dwivedi, "Rigorous experimental analysis of tabular data generated using TVAE and CTGAN," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, 2024, <http://dx.doi.org/10.14569/IJACSA.2024.01504125>
- [21] M. Alqulaity and P. Yang, "Enhanced conditional GAN for high-quality synthetic tabular data generation in mobile-based cardiovascular healthcare," *Sensors*, vol. 24, no. 23, p. 7673, 2024, <https://doi.org/10.3390/s24237673>
- [22] P. C. Y. Cheah, Y. Yang, and B. G. Lee, "Enhancing financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques," *Int. J. Financial Stud.*, vol. 11, no. 3, p. 110, 2023, <https://doi.org/10.3390/ijfs11030110>
- [23] G. Eom and H. Byeon, "Searching for optimal oversampling to process imbalanced data: generative adversarial networks and synthetic minority over-sampling technique," *Mathematics*, vol. 11, no. 16, p. 3605, 2023, <https://doi.org/10.3390/math11163605>
- [24] H.-Y. Cho and Y.-H. Kim, "A genetic algorithm to optimize SMOTE and GAN ratios in class imbalanced datasets," in *Proc. 2020 Genet. Evol. Comput. Conf. Companion (GECCO '20)*, 2020, pp. 33–34, <https://doi.org/10.1145/3377929.3398153>
- [25] N. Uddin, P. Mahmud, M. Ahammed, et al., "Explainable ensemble learning model for cardiovascular disease prediction with feature optimization and data balancing," *Discov. Comput.*, vol. 29, p. 44, 2026, <https://doi.org/10.1007/s10791-026-09930-0>
- [26] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, 2023, <http://dx.doi.org/10.14569/IJACSA.2023.01406116>
- [27] I. Al-Hurani, S. H. ElFar, A. Alkhateeb, and S. Ikki, "AE-CTGAN: autoencoder-conditional tabular GAN for multi-omics imbalanced class handling and cancer outcome prediction," *Algorithms*, vol. 19, no. 2, p. 95, 2026, <https://doi.org/10.3390/a19020095>
- [28] Y. Lin, C. Huang, H. Tian, et al., "Improving generalization of polyp detection via conditional StyleGAN augmented training," *npj Digit. Med.*, vol. 9, p. 113, 2026, <https://doi.org/10.1038/s41746-025-02293-8>
- [29] M. Kelly, R. Longjohn, and K. Nottingham, "The UCI Machine Learning Repository." [Online]. Available: <https://archive.ics.uci.edu>
- [30] Centers for Disease Control and Prevention (CDC), "Behavioral Risk Factor Surveillance System (BRFSS) 2014 Data and Documentation," 2014. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2014.html
- [31] "CDC Diabetes Health Indicators," UCI Machine Learning Repository, 2017. [Online]. Available: <https://doi.org/10.24432/C53919>

- [32] S. Borzooei and A. Tarokhian, "Differentiated Thyroid Cancer Recurrence," UCI Machine Learning Repository, 2023. [Online]. Available: <https://doi.org/10.24432/C5632J>.
- [33] National Institute of Allergy and Infectious Diseases (NIAID), "A Randomized, Double-Blind Phase II/III Trial of Monotherapy vs. Combination Therapy With Nucleoside Analogs in HIV-Infected Persons With CD4 Cells of 200–500/mm³," ClinicalTrials.gov, NCT00000625, 2021.
- [34] S. Hammer et al., "AIDS Clinical Trials Group Study 175," UCI Machine Learning Repository, 2023. [Online]. Available: <https://doi.org/10.24432/C5ZG8F>.
- [35] I. Yeh, "Default of Credit Card Clients," UCI Machine Learning Repository, 2009. [Online]. Available: <https://doi.org/10.24432/C55S3H>.
- [36] J. J. S. Huang, "Taiwanese Bankruptcy Prediction," UCI Machine Learning Repository, 2020. [Online]. Available: <https://doi.org/10.24432/C5004D>.
- [37] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Advances in Neural Information Processing Systems, vol. 32, 2019.
- [38] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, p. e0118432, 2015, <https://doi.org/10.1371/journal.pone.0118432>
- [39] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in Proc. 14th Int. Conf. Machine Learning (ICML), 1997, pp. 179–186.

APPENDIX

TABLE VI. HYPERPARAMETERS SELECTED FOR EACH DATASET (GRID SEARCH, LOGISTIC REGRESSION PROXY, AUC-PR ON VALIDATION)

Dataset	Minority records (train-in)	SMOTE-NC k	CTGAN (epochs, batch, pac)	TVAE (epochs, batch, emb. dim.)	HYBRID CTGAN stage (epochs, batch, pac)	α (HYBRID)
DS1	22,622	3	300, 500, 10	500, 256, 128	500, 1000, 10	0.5
DS2	69	7	500, 1000, 10	500, 512, 256	300, 500, 10	0.5
DS3	333	7	500, 500, 10	300, 256, 128	500, 1000, 10	0.5
DS4	4,247	3	300, 500, 10	500, 256, 128	500, 1000, 10	0.5
DS5	141	3	300, 500, 10	500, 512, 256	300, 500, 10	0.5

TABLE VII. PREDICTIVE PERFORMANCE ON THE TEST PARTITION WITH XGBOOST (MEAN $\pm$ SD OVER 10 RUNS)

Dataset	Ratio	Method	AUC-PR	Recall	Precision	F1-Score	Specificity	G-mean	Bal. Acc.
DS1	6.18	REAL	0.4219 $\pm$ 0.0005	0.1670 $\pm$ 0.0015	0.5518 $\pm$ 0.0031	0.2564 $\pm$ 0.0021	0.9780 $\pm$ 0.0002	0.4041 $\pm$ 0.0019	0.5725 $\pm$ 0.0008
		REAL-CW	0.4228 $\pm$ 0.0006	0.7852 $\pm$ 0.0009	0.3104 $\pm$ 0.0005	0.4449 $\pm$ 0.0006	0.7176 $\pm$ 0.0007	0.7507 $\pm$ 0.0005	0.7514 $\pm$ 0.0005
		SMOTE-NC	0.4117 $\pm$ 0.0009	0.3372 $\pm$ 0.0029	0.4599 $\pm$ 0.0029	0.3891 $\pm$ 0.0015	0.9359 $\pm$ 0.0012	0.5618 $\pm$ 0.0021	0.6365 $\pm$ 0.0010
		CTGAN	0.3372 $\pm$ 0.0073	0.5591 $\pm$ 0.0236	0.3247 $\pm$ 0.0086	0.4105 $\pm$ 0.0060	0.8114 $\pm$ 0.0139	0.6733 $\pm$ 0.0093	0.6853 $\pm$ 0.0063
		TVAE	0.3189 $\pm$ 0.0082	0.4341 $\pm$ 0.0403	0.3263 $\pm$ 0.0092	0.3715 $\pm$ 0.0127	0.8545 $\pm$ 0.0173	0.6082 $\pm$ 0.0231	0.6443 $\pm$ 0.0119
		HYBRID	0.4143 $\pm$ 0.0022	0.1788 $\pm$ 0.0040	0.5297 $\pm$ 0.0061	0.2673 $\pm$ 0.0042	0.9743 $\pm$ 0.0010	0.4174 $\pm$ 0.0045	0.5766 $\pm$ 0.0016
DS2	2.55	REAL	0.9939 $\pm$ 0.0008	0.9545 $\pm$ 0.0000	0.9818 $\pm$ 0.0235	0.9679 $\pm$ 0.0115	0.9927 $\pm$ 0.0094	0.9734 $\pm$ 0.0046	0.9736 $\pm$ 0.0047
		REAL-CW	0.9929 $\pm$ 0.0011	0.9545 $\pm$ 0.0000	0.9504 $\pm$ 0.0131	0.9524 $\pm$ 0.0067	0.9800 $\pm$ 0.0057	0.9672 $\pm$ 0.0028	0.9673 $\pm$ 0.0029
		SMOTE-NC	0.9937 $\pm$ 0.0016	0.9545 $\pm$ 0.0000	0.9682 $\pm$ 0.0220	0.9612 $\pm$ 0.0107	0.9873 $\pm$ 0.0088	0.9708 $\pm$ 0.0043	0.9709 $\pm$ 0.0044
		CTGAN	0.9934 $\pm$ 0.0023	0.9545 $\pm$ 0.0000	0.9553 $\pm$ 0.0290	0.9547 $\pm$ 0.0145	0.9818 $\pm$ 0.0121	0.9681 $\pm$ 0.0060	0.9682 $\pm$ 0.0061
		TVAE	0.9930 $\pm$ 0.0012	0.9545 $\pm$ 0.0000	0.9773 $\pm$ 0.0240	0.9656 $\pm$ 0.0117	0.9909 $\pm$ 0.0096	0.9725 $\pm$ 0.0047	0.9727 $\pm$ 0.0048
		HYBRID	0.9942 $\pm$ 0.0032	0.9500 $\pm$ 0.0144	0.9686 $\pm$ 0.0298	0.9589 $\pm$ 0.0139	0.9873 $\pm$ 0.0123	0.9684 $\pm$ 0.0077	0.9686 $\pm$ 0.0076
DS3	3.10	REAL	0.8637 $\pm$ 0.0034	0.7337 $\pm$ 0.0129	0.8075 $\pm$ 0.0066	0.7687 $\pm$ 0.0070	0.9438 $\pm$ 0.0028	0.8321 $\pm$ 0.0067	0.8387 $\pm$ 0.0057
		REAL-CW	0.8607 $\pm$ 0.0026	0.7625 $\pm$ 0.0129	0.7806 $\pm$ 0.0093	0.7714 $\pm$ 0.0081	0.9312 $\pm$ 0.0039	0.8426 $\pm$ 0.0068	0.8468 $\pm$ 0.0062
		SMOTE-NC	0.8612 $\pm$ 0.0034	0.7587 $\pm$ 0.0071	0.7797 $\pm$ 0.0088	0.7690 $\pm$ 0.0066	0.9312 $\pm$ 0.0033	0.8405 $\pm$ 0.0045	0.8449 $\pm$ 0.0042
		CTGAN	0.8549 $\pm$ 0.0118	0.7606 $\pm$ 0.0132	0.7552 $\pm$ 0.0206	0.7577 $\pm$ 0.0131	0.9207 $\pm$ 0.0086	0.8368 $\pm$ 0.0083	0.8406 $\pm$ 0.0079
		TVAE	0.8523 $\pm$ 0.0055	0.7567 $\pm$ 0.0144	0.7710 $\pm$ 0.0152	0.7637 $\pm$ 0.0125	0.9278 $\pm$ 0.0059	0.8379 $\pm$ 0.0089	0.8423 $\pm$ 0.0084
		HYBRID	0.8588 $\pm$ 0.0047	0.7529 $\pm$ 0.0111	0.7726 $\pm$ 0.0167	0.7625 $\pm$ 0.0087	0.9287 $\pm$ 0.0072	0.8362 $\pm$ 0.0058	0.8408 $\pm$ 0.0053
DS4	3.52	REAL	0.5498 $\pm$ 0.0021	0.3656 $\pm$ 0.0037	0.6547 $\pm$ 0.0053	0.4692 $\pm$ 0.0042	0.9453 $\pm$ 0.0009	0.5878 $\pm$ 0.0031	0.6554 $\pm$ 0.0021

DS5	29.99	REAL-CW	0.5479 0.0023	±	0.5910 0.0024	±	0.4812 0.0042	±	0.5305 0.0029	±	0.8190 0.0030	±	0.6958 0.0018	±	0.7050 0.0019	±
		SMOTE-NC	0.5286 0.0026	±	0.4695 0.0041	±	0.5288 0.0077	±	0.4974 0.0048	±	0.8812 0.0034	±	0.6432 0.0032	±	0.6753 0.0028	±
		CTGAN	0.5371 0.0048	±	0.3826 0.0046	±	0.6339 0.0055	±	0.4772 0.0048	±	0.9373 0.0011	±	0.5988 0.0037	±	0.6599 0.0025	±
		TVAE	0.5234 0.0033	±	0.4099 0.0072	±	0.5862 0.0031	±	0.4824 0.0050	±	0.9178 0.0019	±	0.6134 0.0049	±	0.6639 0.0029	±
		HYBRID	0.5339 0.0059	±	0.3786 0.0063	±	0.6297 0.0099	±	0.4728 0.0053	±	0.9367 0.0030	±	0.5955 0.0046	±	0.6577 0.0028	±
	29.99	REAL	0.4872 0.0173	±	0.3091 0.0389	±	0.6388 0.0489	±	0.4158 0.0420	±	0.9942 0.0010	±	0.5534 0.0341	±	0.6516 0.0195	±
		REAL-CW	0.4662 0.0176	±	0.5000 0.0321	±	0.5046 0.0216	±	0.5020 0.0246	±	0.9836 0.0011	±	0.7010 0.0225	±	0.7418 0.0160	±
		SMOTE-NC	0.4768 0.0101	±	0.6000 0.0287	±	0.4800 0.0113	±	0.5331 0.0160	±	0.9783 0.0010	±	0.7660 0.0182	±	0.7892 0.0141	±
		CTGAN	0.5137 0.0162	±	0.2682 0.0352	±	0.6123 0.0618	±	0.3722 0.0420	±	0.9943 0.0011	±	0.5154 0.0345	±	0.6312 0.0177	±
		TVAE	0.4704 0.0300	±	0.3705 0.0340	±	0.6326 0.0555	±	0.4661 0.0342	±	0.9927 0.0017	±	0.6059 0.0276	±	0.6816 0.0169	±
		HYBRID	0.5252 0.0222	±	0.3091 0.0287	±	0.6448 0.0484	±	0.4167 0.0275	±	0.9942 0.0013	±	0.5538 0.0254	±	0.6517 0.0141	±

Same protocol and notation as Table III.