Close to Reality: Interpretable and Feasible Data Augmentation for Imbalanced Learning

Anonymous authors
Paper under double-blind review

Abstract

Many machine learning classification tasks involve imbalanced datasets, which are often subject to over-sampling techniques aimed at improving model performance. However, these over-sampling methods are prone to generating unrealistic or infeasible samples. Furthermore, they often function as black boxes, lacking interpretability in their procedures. This opacity makes it difficult to track their effectiveness and provide necessary adjustments, and they may ultimately fail to yield significant performance improvements. To bridge this gap, we introduce the Decision Predicate Graphs for Data Augmentation (DPG-da), a framework that extracts interpretable decision predicates from trained models to capture domain rules and enforce them during sample generation. This design ensures that over-sampled data remain diverse, constraint-satisfying, and interpretable. In experiments on synthetic and real-world benchmark datasets, DPG-da consistently improves classification performance over traditional over-sampling methods, while guaranteeing logical validity and offering clear, interpretable explanations of the over-sampled data.

1 Introduction

A recurring challenge in supervised classification is the presence of imbalanced datasets, where one or more classes are represented by significantly fewer samples than others. This imbalance can significantly create biases in the learning algorithms toward the majority class, leading to poor generalization on minority classes. As a result, classifiers may achieve high overall accuracy while systematically misclassifying rare but critical events, a limitation that is particularly problematic in high-stakes domains such as fraud detection (Makki et al., 2019; Rubaidi et al., 2022; Silva et al., 2023) and clinical diagnosis (Roy et al., 2024; Ejiyi et al., 2025), where correct classification of the minority class is often the central goal.

Existing approaches to class imbalance, focusing on tabular data, fall into three categories: problem definition, algorithmic modification, and data-level processing (He & Ma, 2013; Weiss, 2013; Haixiang et al., 2017). The first reformulates evaluation metrics or tasks to emphasize minority performance but does not alter the data or model. The second, often referred to as cost-sensitive or cost-adjusted learning, integrates class-specific misclassification penalties directly into the loss function (Cao et al., 2013; Hu et al., 2017), but its effectiveness hinges on defining appropriate costs and often results in more complex models. The third, and most widely adopted, modifies the data distribution directly, either by under-sampling the majority class or over-sampling the minority class. Under-sampling, as in Random Under-Sampling (RUS)(Weiss, 2013), risks discarding informative instances(He & Ma, 2013), while over-sampling techniques like Random Over-Sampling (ROS)(Chawla et al., 2002), which simply duplicate existing minority instances, may induce overfitting(Dixit et al., 2023).

More sophisticated over-sampling methods, such as the interpolation-based SMOTE (Chawla et al., 2002) and generative approaches like GANs (Goodfellow et al., 2020) and VAEs (Xu et al., 2019), aim to generate diverse synthetic samples. However, these methods face three key limitations: (i) interpolation-based techniques can generate infeasible instances that violate domain constraints (Matharaarachchi et al., 2024; Yang et al., 2023), (ii) generative models may suffer from instability and provide no guarantees of logical validity (Di Liello et al.,

2020; Tarawneh et al., 2022), and (iii) both approaches offer little transparency into how synthetic samples are produced, limiting their usefulness in sensitive domains.

These issues are particularly acute in high-stakes applications, where synthetic data must not only improve classifier performance but also remain valid, auditable, and interpretable. Interpretability, in contrast to post-hoc explainability, refers to methods whose internal logic and outputs can be directly understood without approximation layers (Ross & Doshi-Velez, 2018; Rudin, 2019). In the context of data augmentation, the process of generating synthetic samples from a query sample to expand or balance existing datasets, interpretability requires not only that the generated records respect domain-specific constraints, but also that users gain insight into how these synthetic samples are produced and in what ways they differ from the original data they are derived from.

To address these challenges, we propose *DPG-da*, a constraint-aware augmentation method guided by Decision Predicate Graphs (DPGs) (Arrighi et al., 2024). DPGs provide a global, interpretable representation of decision logic, from which class-specific predicates are extracted and enforced during generation. In our framework, interpretability arises in two complementary ways: (i) constraints define a transparent and auditable valid region for minority augmentation, offering users an explicit understanding of the problem space; and (ii) the evolutionary transformation process from query samples to synthetic samples makes it possible to trace which features changed most during augmentation, thereby explaining *how* new instances were created.

The contributions of this work are:

- A novel augmentation framework, DPG-da, that integrates constraint discovery and enforcement directly into sample generation, ensuring validity and interpretability of synthetic data.
- A systematic violation analysis across three complementary groups of datasets (handcrafted rule-based, high-stakes real-world, and synthetic overlapping) and over 10 baseline over-samplers, showing that existing interpolation- and generative-based methods frequently produce infeasible samples, whereas DPG-da guarantees strict constraint adherence.
- A comprehensive performance evaluation on 27 benchmark datasets, demonstrating that DPG-da consistently outperforms interpolation and generative over-sampling methods in classification performance while avoiding violations and providing interpretability into the augmentation process.

Experiments confirm that DPG-da outperforms state-of-the-art over-sampling techniques in both accuracy and reliability, while eliminating infeasible samples. Although this comes with higher computational cost, the trade-off is justified in domains where validity and interpretability of augmented data are as critical as predictive accuracy.

The remainder of this paper is organized as follows: Section 2 provides the necessary background, Section 3 reviews related work, Section 4 presents the DPG-da framework, Section 5 details the experimental design, Section 6 reports results and discussion, and Section 7 concludes with contributions and future directions.

2 Background

2.1 Violation Space

Let $\mathcal{X}$ denote the problem space, i.e., all possible combinations of x feature values (e.g., $\mathcal{X} \subseteq \mathbb{R}^d$ for d features). Not every point in $\mathcal{X}$ represents a feasible or meaningful instance in reality.

The *violation space* $V \subseteq \mathcal{X}$ is the subset of points that break one or more domain-specific constraints, such as physical laws, regulatory requirements, or common-sense rules. For example, a record describing a patient with negative blood pressure, or a loan applicant under 18 years old, belongs to V. These samples are numerically valid but semantically invalid.

Formally, let $\mathcal{C}$ be the set of Boolean-valued constraints over $\mathcal{X}$. Then:

$$V = \{x \in \mathcal{X} \mid \exists c \in \mathcal{C} : c(x) = \text{False}\}.$$

Its complement, the *feasible region*, consists of instances that satisfy all constraints. The literature sometimes refers to V as the *infeasible region* (Peres et al., 2018; Alsouly et al., 2022) or the *constraint violation region* (Bernardo et al., 2011).

2.2 Decision Predicate Graphs

DPGs (Arrighi et al., 2024) are graph-based structures that capture decision logic in an interpretable and compact form. A predicate $p = (f, o, t)$ encodes a simple test over a feature f with operator $o \in \{<, \leq, >, \geq\}$ and threshold $t \in \mathbb{R}$. Nodes correspond to predicates or leaf outcomes; edges connect predicates that co-occur along decision paths, with weights indicating their frequency. Figure 1 shows an illustrative binary-class example.

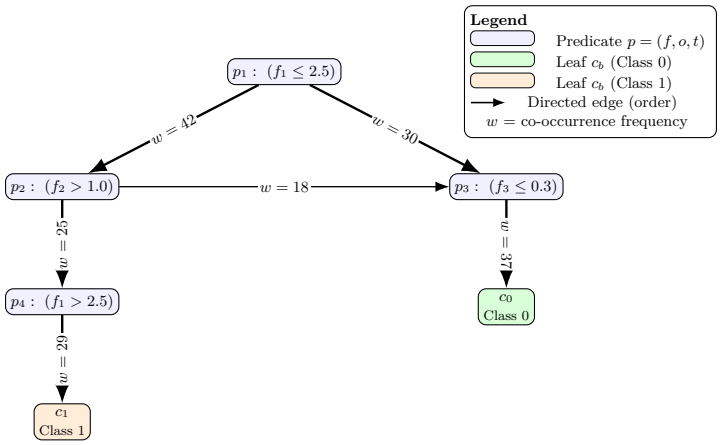

Figure 1: DPG structure showing predicates as nodes and class leaves.

DPGs are typically extracted from surrogate ensemble models trained to approximate a base classifier. Formally, a DPG for model M_n is a directed weighted graph $\mathrm{DPG}(M_n) = (P, E)$, where P is the set of predicates and E the edges weighted by co-occurrence frequency. Leaf nodes $c_b \in P$ denote final class labels.

From a DPG, one can derive *class-bound predicates*, i.e., conjunctions of rules that consistently lead to a given class. These yield explicit descriptions of feasible regions, such as $18 \leq \texttt{Age} \leq 65$ (Loan requirement for approval) or $\texttt{Glucose} < 140$ (Diabetes diagnosis). In our setting, these constraints serve two roles: they restrict augmentation to the feasible region, preventing violations, and they provide interpretable insights into the structure of class-specific domains.

3 Related Works

Data augmentation for tabular data has been widely studied, primarily to mitigate class imbalance. Early over-sampling methods such as SMOTE and its variants (Han et al., 2005; Kovács, 2019) interpolate between minority instances, while generative models such as GANs and VAEs (Ding et al., 2024; Ai et al., 2023; Tian et al., 2024; Stocksieker et al., 2024; Yadav et al., 2025; Berti et al., 2025) attempt to learn full data distributions. Although effective at enlarging datasets, both families often yield implausible or noisy samples, as they lack mechanisms to enforce domain-specific constraints and operate largely as black-box processes (Thekumparampil et al., 2018; Sakho et al., 2024).

To improve trust and transparency, researchers have paired augmentation with XAI. Prior work combined SMOTE with post-hoc explanation tools such as SHAP, LIME, or LRP in domains including fraud detection, diabetes prediction, and student performance forecasting (Patil et al., 2020; Nayan et al., 2023; Sahlaoui et al., 2024). While these studies demonstrate the value of combining over-sampling with interpretability, augmentation itself remains unconstrained and prone to producing unrealistic samples. More recent domain-specific frameworks embed interpretability directly into augmentation pipelines, for example

using model explanations to guide low-resource NLP augmentation (Mersha et al., 2025) or steering over-sampling with feature-importance signals in healthcare (Ejiyi et al., 2025). These methods, however, remain tied to particular data modalities.

A related line of work employs counterfactual explanations to generate locally plausible alternatives (Mohammed et al., 2022; Temraz & Keane, 2022; Panagiotou et al., 2024). Counterfactual augmentation improves minority coverage and interpretability, but is inherently instance-specific, sensitive to learned decision boundaries, and not designed for large-scale rebalancing. Manifold-based approaches such as TabMDA (Margeloiu et al., 2024) address augmentation at the dataset level by leveraging pre-trained tabular models, but they lack transparency and do not enforce domain-specific constraints, limiting their applicability in sensitive domains.

In summary, existing methods trade off between diversity (e.g., generative, manifold) and plausibility or interpretability (e.g., counterfactuals, XAI-assisted over-sampling). Our framework, DPG-da, bridges this gap by embedding model-derived constraints directly into the augmentation process, ensuring that generated samples are both semantically valid and interpretable.

4 Proposed Approach

DPG-da consists of two key components: (i) extraction of class-specific constraints using Decision Predicate Graphs (DPGs) and (ii) augmentation of minority samples via a heuristic optimization process of a query sample. The framework ensures that generated data points are both diverse and strictly confined to feasible regions defined by the DPG, addressing limitations of SMOTE-based and generative approaches that may produce infeasible or overlapping samples (Figure 2).

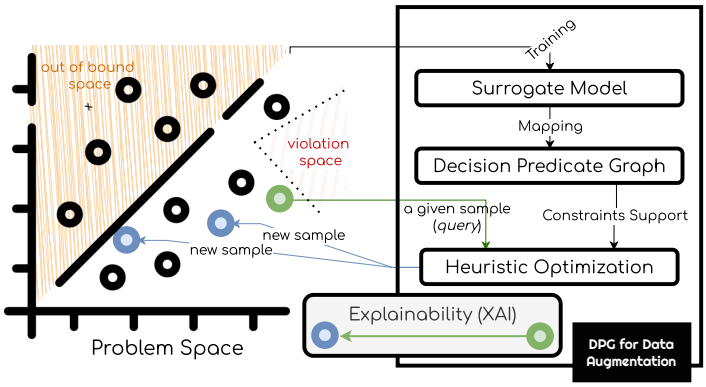

Figure 2: Workflow of the proposed DPG-da. Constraints extracted from DPGs guide the optimization for valid and diverse synthetic samples.

4.1 Surrogate Model

The original imbalanced dataset is first used to train a Random Forest, which can be substituted with any ensemble-based algorithm. This Random Forest ensemble acts as a surrogate model to approximate the decision boundaries of the base classifier. The surrogate then enables the systematic extraction of decision predicates, which are subsequently used to construct the DPG. These predicates define class-specific constraints that delimit feasible regions for synthetic sample generation, thereby ensuring interpretability through explicit, data-driven rules.

4.2 Heuristic Optimization

To generate augmented samples confined to the valid class-bound regions (*constraints*) identified by the DPG, we employ a heuristic optimization strategy based on Genetic Algorithms (GAs). Each augmentation begins from a query or seed sample x^q drawn from the minority class, which serves as a reference for generating new candidates. We selected GAs for their simplicity and robustness: although faster or more sophisticated optimization methods exist, our focus is on constraint-guided augmentation rather than optimizing the search process itself.

Each individual in the GA represents a candidate augmented sample x^a, initialized by uniform sampling within the DPG-defined bounds for the target class. Candidates are evaluated using a composite fitness function that captures multiple objectives:

- **Classification validity** $V(x^a, x^q)$: A binary indicator that evaluates whether the candidate sample x^a and the query sample x^q are assigned to the same target class, thereby validating semantic alignment with the intended class label.
- **DPG-constraint adherence** $A(x^a, P)$: Rewards candidate samples that satisfy class-specific predicates, remaining strictly within the feasible region defined by the DPG class bounds. This prevents over-generalization and preserves the structure of the original data manifold.
- **Distance from the query sample** $D(x^a, x^q)$: Promotes diversity among the augmented samples by maximizing the distance between each new augmented sample and the reference query sample. Distance is measured using Euclidean distance for continuous features.
- **Augmentation sparsity** $S(x^a, x^q)$: Encourages minimal and interpretable modifications by penalizing candidates that differ significantly from the query sample. Sparsity is quantified using the ℓ_0 norm, which counts the number of feature changes, thereby promoting localized, meaningful transformations.

The overall fitness of a candidate x^a, with respect to the query sample x^q and the class-bound predicates P (obtained from the DPG), is computed by weighting the contributions of multiple objectives. Specifically, the fitness is defined as:

$$\text{fitness}(x^a, x^q, P) = V(x^a, x^q) \times \big(w_1 \cdot A(x^a, P) + w_2 \cdot D(x^a, x^q) + w_3 \cdot (1 - S(x^a, x^q))\big) \tag{1}$$

where w_1, w_2, w_3 control the relative importance of each objective. Higher fitness scores guide the GA toward candidates that best satisfy all objectives.

The GA iterates until a stopping criterion is met: either the best fitness plateaus for a fixed number of generations or a maximum number of generations is reached. This ensures efficient exploration of feasible regions while generating diverse and interpretable samples without excessive computational overhead.

4.3 Interpretability of Augmentation

Our proposed DPG-da method introduces a novel form of intrinsic interpretability by tracking the entire augmentation process. It leverages interpretable the evolutionary trajectory of each augmented sample produced via a GA. To facilitate analysis and interpretation, we generate feature-level analysis that visualize how candidate solutions evolve over successive generations. Various XAI methods have been developed, with many producing visual outputs (e.g., heatmap) to highlight important features in image classification tasks (Richard et al., 2024; Sawada & Toyoda, 2019). Inspired by these approaches, the heatmap and barplot represent the magnitude and direction of changes in each feature dimension for the best-performing individuals at each step, exposing the dynamics of the optimization process.

For example, consider a binary loan approval classification problem with four numerical features: x_1 (Age in years), x_2 (Annual Income in thousands of USD), x_3 (Credit Score ranging from 300 to 850), and x_4 (Number

of Children). Let $\mathbf{x}^{(0)} = [45,\ 50,\ 620,\ 2]$ be a query instance from the minority class (i.e., *approved loans*). The DPG derived from the surrogate model defines the feasible region for synthetic instances through the following class-specific constraints: $x_1 \geq 30$, $x_2 \geq 45$, $x_3 \geq 600$, and $x_4 \leq 3$. The goal of DPG-da is to generate synthetic instances $\mathbf{x}^{(g)}$ at each generation g of the GA, following the constraint adherence.

In this example, the evolutionary process considers $\mathbf{x}^{(0)} = [45,\ 50,\ 620,\ 2]$ and produces the best individuals over three generations as follows: $\mathbf{x}^{(1)} = [47,\ 52,\ 610,\ 2]$, $\mathbf{x}^{(2)} = [50,\ 55,\ 605,\ 3]$, and $\mathbf{x}^{(3)} = [52,\ 60,\ 590,\ 3]$, with the last instance being invalid due to violation of constraint C_3. The algorithm adjusts x_1 and x_2 to explore the feasible region and promote diversity, while the violation of constraint C_3 in generation 3 (where $x_3 = 590 < 600$) results in penalization and exclusion of the candidate, highlighting the interpretable and traceable nature of the evolutionary trajectory for post-hoc validation of synthetic data feasibility. The change in feature values between successive generations as visible in Table1.

Table 1: Feature-wise changes (Δx_i) between successive generations of the GA in the DPG-da augmentation process. A violation of constraint C_3 occurs in generation 3 (signed by $\times$).

Feature	$\mathbf{\Delta^{0\to1}}$	$\mathbf{\Delta^{1\to2}}$	$\mathbf{\Delta^{2\to3}}$
x_1 (Age)	$+2$	$+3$	$+2$
x_2 (Income)	$+2$	$+3$	$+5$
x_3 (Credit Score)	-10	-5	-15 ($\times$)
x_4 (Number of Children)	0	$+1$	0

5 Materials and Methods

To evaluate the proposed DPG-da framework, we designed a comprehensive experimental protocol combining violation-prone, synthetic imbalanced, and benchmark datasets. Our methodology aims to test (i) data realism through adherence to domain constraints, (ii) classification performance under varying augmentation levels, and (iii) interpretability of synthetic samples. This section details the DPG-da implementation, baseline over-sampling algorithms and datasets.

5.1 DPG-da Implementation

The implementation of DPG-da is publicly available[1]. The codebase provides a reproducible workflow for generating realistic, constraint-compliant synthetic samples in imbalanced tabular datasets. To avoid data leakage, all augmentation procedures were applied exclusively to the training set; the surrogate model was trained only on training data, and augmented samples were added back to this set before classifier training. The test set remained untouched to ensure fair evaluation.

The stages of the implementation are as follows:

1. **Surrogate Model Training:** A tree-based ensemble (Random Forest) is trained to approximate the classification behaviour of the dataset, providing interpretable decision predicates.
2. **DPG Construction:** Decision predicates are extracted from the surrogate model to delineate class-specific feasible regions, supported by the DPG library [2].
3. **GA-Based Augmentation:** Synthetic samples are generated by evolving a population of candidate solutions through a GA.
4. **Fitness Function and Optimization:** The GA is guided by a composite fitness function promoting (i) strict adherence to DPG-derived constraints, (ii) diversity among synthetic samples, and (iv) sparsity of modifications relative to a query sample. In the current implementation, these objectives are combined via a weighted sum with $(w_1, w_2, w_3) = (2.0, 1.0, 3.0)$ (see Appendix C).

[1] Anonymized
[2] `https://github.com/LeonardoArrighi/DPG`

5.2 Baseline Algorithms

We compared it against 12 established over-sampling techniques: classical interpolation-based methods (SMOTE, SMOTE-ENN, SMOTE-SVM), advanced variants from the *smote-variants* library (DE, SMOTE-POLYNOM, MSMOTE, SMOTE-LVQ), generative approaches for tabular data (CTGAN, TVAE, COPULA-GAN) and ROS. Baselines were run with standard hyperparameters to ensure fair comparison. A summary is provided in Table 2 in the Appendix.

5.3 Violation-Prone Datasets

To evaluate DPG-da under constraint-sensitive scenarios, we curated three types of violation-prone datasets:

Synthetic violation-prone datasets: We synthesized a set of datasets covering domains such as healthcare, finance, manufacturing, fraud detection, energy grid, and education. Features were sampled from plausible distributions, and domain-specific rules were encoded to evaluate violation-space adherence.

Real-world imbalanced datasets: Tabular datasets from the KEEL repository (Derrac et al., 2015) with imbalance ratios between 1.5 and 9 (e.g., Wisconsin and Pima) (Fernández et al., 2008). These allow testing on naturally occurring imbalances.

Synthetic imbalanced datasets with noise and borderline examples: From Napierała et al. (2010), including three minority-class shapes (*subclus*, *clover*, *paw*) to simulate challenging scenarios with noise and non-linear decision boundaries.

These datasets allow us to evaluate the over-sampling techniques ability to ensure generated data realism, by verifying that synthetic samples respect domain-specific constraints and remain semantically valid. A complete summary of the data and their valid feature ranges is provided in Appendix D.1.

5.4 Benchmarking Datasets

Furthermore, to evaluate the classification performance on augmented data from these over-samplers, we also selected 27 benchmark datasets from (Ding, 2011) and the *imbalanced-learn* library, spanning diverse domains, imbalance ratios, sample sizes, and feature dimensionalities. These datasets allow comprehensive assessment of method performance under realistic, challenging conditions. A summary is provided in Appendix D.2.

6 Results and Discussions

Following the experimental protocol described in Section 5, we compare our method against established over-sampling baselines, spanning both classical interpolation-based techniques and recent generative models. The evaluation focuses on three complementary aspects: (i) the realism of augmented data, i.e., adherence to domain constraints; (ii) classification performance across multiple augmentation levels; and (iii) interpretability of the generated instances. This approach enables a holistic assessment of predictive performance alongside the plausibility and transparency of synthetic samples.

6.1 Ensuring generated data realism

We evaluate DPG-da alongside baseline over-sampling methods using three widely adopted classifiers: Logistic Regression, k-Nearest Neighbors (kNN), and Decision Trees. These classifiers represent a spectrum of model complexities and inductive biases: Logistic Regression assumes linear separability, kNN relies on local neighborhoods and is sensitive to sample distribution, and Decision Trees capture non-linear relationships but can overfit to noisy or artificially augmented instances.

Each method is applied at three augmentation levels (15%, 30%, and 50% minority class proportion relative to the total dataset), and classifiers are trained on the resulting augmented datasets. To account for variability introduced by the random nature of over-sampling and model initialization, each experiment is repeated ten times per dataset and classifier combination.

To assess data realism, we check whether augmented samples adhere to the inherent constraints of each dataset. A violation is defined as any synthetic instance with feature values that are unrealistic or semantically inconsistent with the dataset domain. The number of violations serves as a quantitative measure of how well each method respects dataset constraints.

Figure 3 presents a heatmap of the average violation number across the over-sampling methods (Appendix 2) and representative datasets (Table 3). Interpolation-based methods are the primary source of constraint breaches. SMOTE-LVQ shows the highest number of violations, with 98.6 in Wisconsin, 53.3 in Pima, 10.4 in Fraud Detection, and roughly 5 in Quality Control, Finance, Clover, and Iris. DE also generates violations, though generally at lower rates (e.g., 18.1 in Wisconsin, 26.9 in Pima, 26.5 in Fraud Detection, and smaller counts elsewhere). SMOTE-SVM exhibits moderate violations, particularly in Wisconsin (14.9) and Fraud Detection (16.3), while SMOTE-POLYNOM is mainly problematic for Pima (40 violations). All remaining datasets and methods show no violations.

To illustrate these violations in detail, Appendix E.1 shows a pair-wise plot of the Wisconsin dataset augmented with SMOTE-LVQ. Several synthetic samples lie outside valid ranges, particularly in features such as *Mitoses*, *NormalNucleoli*, and *MarginalAdhesion*, highlighting the practical impact of constraint violations.

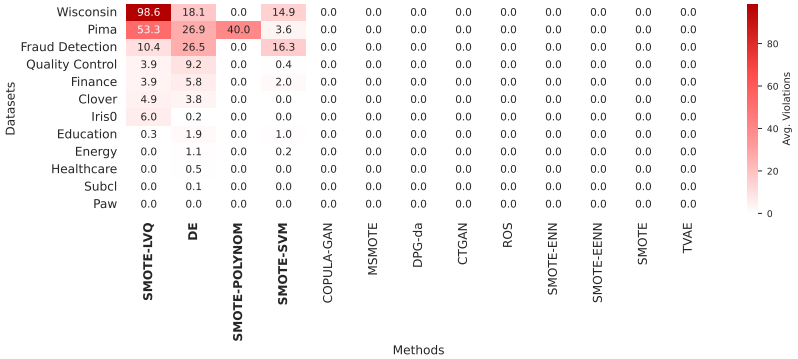

Figure 3: Heatmap of violations per over-sampling method and dataset.

These findings demonstrate that some widely used over-sampling techniques can produce semantically invalid samples, especially in datasets with complex feature interdependencies. In contrast, DPG-da, which incorporates constraint-awareness into the augmentation process, avoids such violations, underscoring the importance of enforcing data realism when generating synthetic minority-class instances in sensitive domains.

6.2 Performance Analysis

6.2.1 Classification improvement

We assess classification performance using the macro F1-score and runtime, reporting the average across 10 repetitions. The F1-score is particularly suitable for imbalanced classification as it balances precision and recall, offering a more informative metric than accuracy when class distributions are skewed.

Across 27 imbalanced datasets, evaluated at three augmentation levels, Figure 4 summarizes the results. Most methods showed a gradual decline in F1-score as augmentation increased, indicating that over-sampling primarily introduced non-descriptive or noisy instances.

Among the evaluated methods, DPG-da, SMOTE-SVM, and TVAE consistently maintained high F1-scores across all augmentation levels, while also outperforming the baseline performance of 71%. This demonstrates their robustness to increasing amounts of synthetic data and highlights their suitability for imbalanced classification tasks under varying augmentation conditions. In contrast, MSMOTE exhibited moderate sensitivity to augmentation, while SMOTE and SMOTE-ENN showed notable performance degradation at higher sampling percentages.

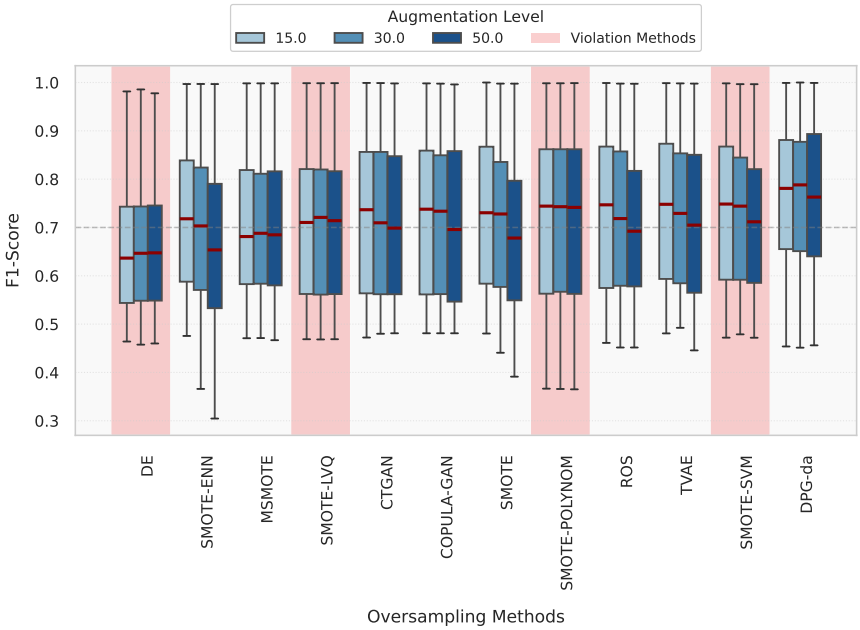

Figure 4: Mean classification performance across augmentation methods and sampling percentages. Red background areas indicate methods that violated constraints (DE, SMOTE-LVQ, SMOTE-POLYNOM, and SMOTE-SVM). The dashed line indicates the average performance of the datasets before any augmentation (71%).

To determine whether the differences in terms of predictive improvements were statistically significant, we conducted a Friedman test ($\alpha = 0.05$) followed by a Nemenyi post-hoc test (Demsar, 2006). The Friedman test confirmed significant differences among the eight populations ($p < 0.001$), providing formal evidence that performance varies across augmentation methods. Appendix Table 7 reports each method's median (MD), median absolute deviation (MAD), and mean rank (MR).

We considered only methods that did not produce violations (described in Section 6.1). The critical difference (CD) diagram (Figure 5) shows the post-hoc analysis (CD = 1.167). DPG-da (MR = 2.531) achieves the highest overall rank and is statistically superior to all other methods. TVAE (MR = 3.704) and ROS (MR = 3.975) form a high-performing cluster with SMOTE (MR = 4.506), where differences within the group are not statistically significant. MSMOTE (MR = 5.494), SMOTE-ENN (MR = 5.494), CTGAN (MR = 5.272), and COPULA-GAN (MR = 5.025) form a lower tier, with no significant differences observed within this cluster. These results reinforce the earlier observation that DPG-da consistently outperforms competing over-sampling techniques, demonstrating that constraint-aware augmentation produces both realistic and effective synthetic samples for imbalanced classification.

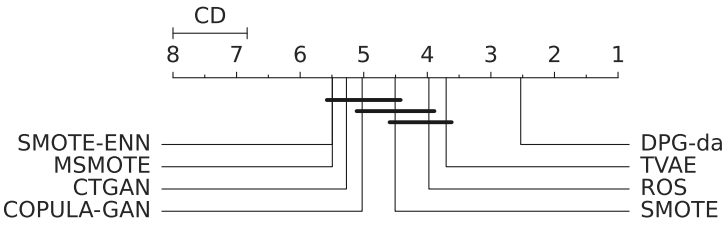

Figure 5: Critical difference diagram for F1-score performance. Methods not connected by a bar differ significantly (Nemenyi test, $\alpha = 0.05$, CD = 1.167).

6.2.2 Runtime

Figure 6 shows the log-scaled average runtime for generating synthetic samples across all evaluated oversampling methods.

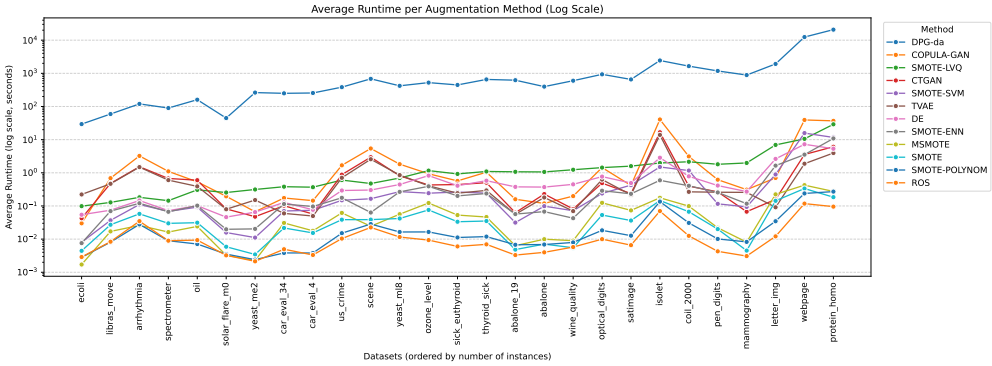

Figure 6: Log scaled Time performance of the augmentation methods.

The interpolation-based methods, ROS, SMOTE, and SMOTE-POLYNOM, are extremely fast, with runtimes below 0.060 seconds. MSMOTE (0.072 seconds), SMOTE-ENN (0.740 seconds), and SMOTE-LVQ (2.520 seconds) require slightly more time due to additional operations such as noise filtering or SVM-based boundary computation.

Generative approaches, CTGAN (1.410 seconds), TVAE (1.120 seconds) and COPULA-GAN (5.240 seconds), introduce moderate overhead from model fitting and sample generation, but remain practical for typical tabular datasets. DE and SMOTE-SVM similarly incur small additional costs (0.950 and 1.250 seconds, respectively).

By contrast, DPG-da is substantially slower (1,809.45 seconds) due to its genetic algorithm optimization and the need to enforce DPG-derived constraints. This clearly illustrates the trade-off: DPG-da delivers highly realistic and interpretable synthetic data, but at the cost of considerably longer runtime, whereas all other methods remain extremely fast and scalable. We discuss the limitations of our proposal in the Appendix A.

6.3 Interpreting augmentation

The final part of our analysis examines the interpretability of the proposed DPG-da augmentation method, which we define as the traceability and interpretation of the synthetic data generation process. Since DPG-da is grounded in the DPG framework, it inherits a transparent mechanism that reveals how new samples are produced.

To illustrate this property, we first present a heatmap (Figure 7) showing the average change (Δ) in feature values between consecutive generations for the best individual in each generation of the evolutionary process on the Abalone dataset. Each cell reflects the magnitude of change for a specific feature, highlighting which attributes were most actively modified during the search.

The query sample, represented by the vector:

$$\mathbf{x}^q = [5.00,\ 1.00,\ 1.00,\ 4.00,\ 2.00,\ 1.00,\ 3.00,\ 1.00,\ 1.00] \tag{2}$$

Was augmented into a synthetic sample:

$$\mathbf{x}^a = [5.98,\ 2.56,\ 2.58,\ 3.49,\ 1.65,\ 2.92,\ 2.30,\ 3.15,\ 4.05] \tag{3}$$

corresponding to the features: Clump Thickness, Cell Size, Cell Shape, Marginal Adhesion, Epithelial Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses.

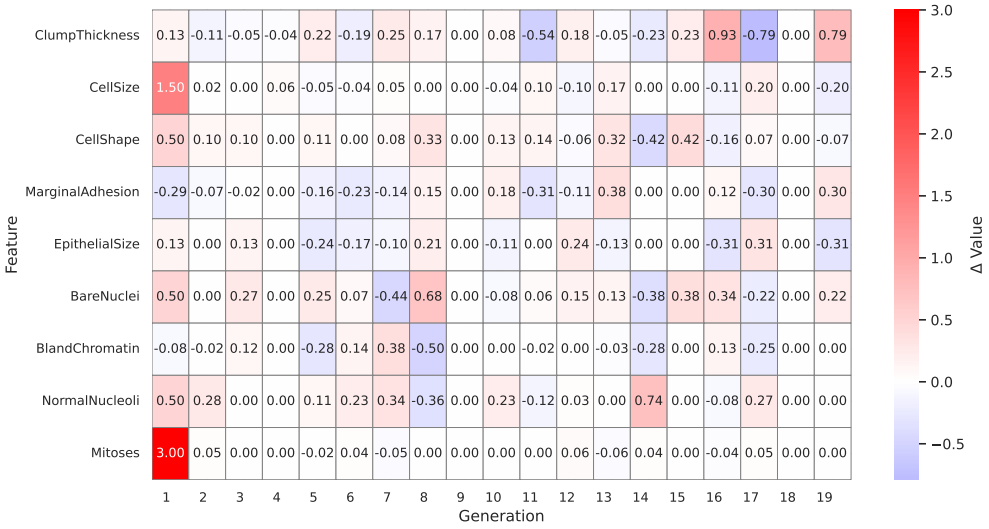

Figure 7: Heatmap of the evolutive process in the Wisconsin dataset. Showcasing the change (Δ) in feature values between generations in the GA.

All features were potentially adjusted by the GA, with the largest changes occurring in the most discriminative attributes for the target class. Clump Thickness, Cell Size, and Cell Shape are moderately modified to remain consistent with the DPG constraints. Continuous features such as Bare Nuclei, Normal Nucleoli, and Mitoses experience larger adjustments, reflecting targeted exploration of the feasible minority-class region. This pattern demonstrates how DPG-da selectively modified key features in this run while preserving the overall structure of the query sample, ensuring realistic and interpretable synthetic data.

We now turn to examples that trace how a single query sample is augmented into a synthetic counterpart. Figure 8 shows this process for three datasets: Pima, Ecoli, and Wisconsin. Each subplot visualizes the cumulative change applied to each feature across generations of the GA.

Pima. The Pima Indian Diabetes dataset describes patient medical measurements used for diagnosing diabetes. In our example, the query sample is already from the minority (diabetes) class, and the augmented instance remains almost identical, differing only in the *Glucose* feature. This adjustment is clinically plausible: blood glucose is one of the most discriminative variables for diabetes diagnosis, and a shift still yields a sample consistent with the minority class. Importantly, other attributes such as *Age* and (number of) *Pregnancies* remain unchanged, showing that DPG-da preserves the original patient profile while exploring variability within the minority region.

Ecoli. The Ecoli dataset concerns protein localization sites. Features such as *LIP* (signal peptide), *GVH* (Von Heijne signal sequence), and *MCG* (McGeoch method for signal sequence recognition) describe biochemical properties relevant to protein sorting. Here, the minority-class query is augmented with only small perturbations to *LIP* and *MCG*, while all other features remain unchanged. These modifications are subtle but biologically meaningful: since signal peptide properties are highly influential for localization, adjusting them slightly generates a new synthetic protein record that stays within the minority localization class but introduces plausible diversity.

Wisconsin. The Wisconsin Breast Cancer dataset encodes cytological measurements of cell nuclei obtained from fine needle aspirates. In this case, the minority-class query is augmented with broader changes. Features such as *Clump Thickness*, *Cell Size*, and *Cell Shape* are moderately adjusted, while continuous variables like *Bare Nuclei*, *Normal Nucleoli*, and *Mitoses* undergo larger shifts. These modifications are biologically interpretable: nuclei with irregular shapes, prominent nucleoli, and increased mitotic activity are characteristic of malignant tumors, the minority class in this dataset. The augmented sample thus represents a realis-

tic new patient record that preserves minority-class pathology while exploring alternative manifestations of malignancy.

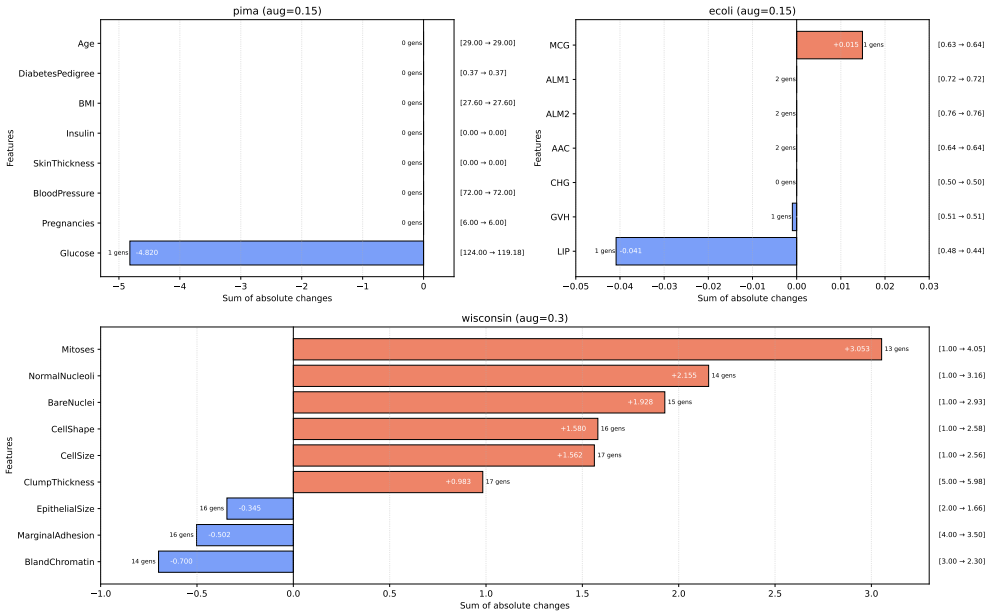

Figure 8: Interpretable feature changes during the genetic algorithm evolution. The top row shows Pima (left) and Ecoli (right), while the bottom row shows Wisconsin. Each horizontal bar represents the absolute difference of a feature from the first to the last generation. The white numbers inside the bars indicate the magnitude of the change, the black numbers outside the bars indicate the number of generations with non-zero change, and the text to the right shows the query-to-augmented transformation for each feature.

Together, these three examples highlight how DPG-da adapts its augmentation strategy to the context of each dataset, only small feature changes are introduced. When larger structural differences are needed (Wisconsin), the method applies broader but still interpretable modifications. In all cases, the augmentation remains transparent and traceable, producing synthetic data that not only looks realistic but also aligns with the underlying semantics of the domain.

7 Conclusion

We introduced DPG-da, a novel data augmentation method for imbalanced classification that integrates predictive performance with inherent interpretability. By leveraging decision pattern graphs to define rule-based boundaries, DPG-da generates instances that are semantically valid and transparent, avoiding the unrealistic samples often produced by traditional over-sampling techniques. Empirical results across benchmark and violation-prone datasets demonstrate that DPG-da consistently outperforms widely used over-sampling methods, achieving higher F1 scores. Future work may explore scalability to larger datasets, integration with deep learning pipelines, and incorporation of domain-specific constraints for high-stakes applications.

References

Qingzhong Ai, Pengyun Wang, Lirong He, Liangjian Wen, Lujia Pan, and Zenglin Xu. Generative oversampling for imbalanced data via majority-guided vae. In *International Conference on Artificial Intelligence and Statistics*, pp. 3315–3330. PMLR, 2023.

Hanan Alsouly, Michael Kirley, and Mario Andrés Muñoz. An instance space analysis of constrained multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 27(5):1427–1439, 2022.

Leonardo Arrighi, Luca Pennella, Gabriel Marques Tavares, and Sylvio Barbon Junior. Decision predicate graphs: Enhancing interpretability in tree ensembles. In *World Conference on Explainable Artificial Intelligence*, pp. 311–332. Springer, 2024.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

J Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Optimization under unknown constraints. *Bayesian Statistics*, 9(9):229, 2011.

Michele Berti, Matheus Camilo da Silva, Sebastiano Saccani, and Sylvio Barbon Junior. Meta-learning approach for variational autoencoder hyperparameter tuning. *Journal of Universal Computer Science*, 31 (7):668, 2025.

Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13*, pp. 475–482. Springer, 2009.

Peng Cao, Dazhe Zhao, and Osmar Zaiane. An optimized cost-sensitive svm for imbalanced data learning. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 280–292. Springer, 2013.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

L. Chen, Z. Cai, L. Chen, and Q. Gu. A novel differential evolution-clustering hybrid resampling algorithm on imbalanced datasets. In *Third International Conference on Knowledge Discovery and Data Mining*, pp. 81–85, 2010.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

J Derrac, S Garcia, L Sanchez, and F Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Logic Soft Comput*, 17:255–287, 2015.

Luca Di Liello, Pierfrancesco Ardino, Jacopo Gobbi, Paolo Morettin, Stefano Teso, and Andrea Passerini. Efficient generation of structured objects with constrained adversarial networks. *Advances in neural information processing systems*, 33:14663–14674, 2020.

Hongwei Ding, Nana Huang, and Xiaohui Cui. Leveraging gans data augmentation for imbalanced medical image classification. *Applied Soft Computing*, 165:112050, 2024.

Zejin Ding. *Diversified ensemble classifiers for highly imbalanced data learning and its application in bioinformatics*. PhD thesis, Georgia State University, 2011.

Krishna Kant Dixit, Upendra Singh Aswal, V Saravanan, Manishn Sararswat, N Shalini, and Amit Srivastava. Data augmentation with generative adversarial networks for deep learning in healthcare. In *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, volume 1, pp. 1–6. IEEE, 2023.

Chukwuebuka Joseph Ejiyi, Dongsheng Cai, Francis Ofoma Eze, Makuachukwu Bennedith Ejiyi, Jennifer Ene Idoko, Sarpong Kwadwo Asere, and Thomas Ugochukwu Ejiyi. Polynomial-shap as a smote alternative in conglomerate neural networks for realistic data augmentation in cardiovascular and breast cancer diagnosis. *Journal of Big Data*, 12(1):97, 2025.

Alberto Fernández, Salvador García, María José del Jesus, and Francisco Herrera. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, 159(18):2378–2398, 2008.

S. Gazzah and N. Essoukri Ben Amara. New oversampling approaches based on polynomial fitting for imbalanced data sets. In *Proc. of the Eighth IAPR Int. Workshop on Document Analysis Systems*, pp. 677–684, 2008.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer, 2005.

Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013.

Junjie Hu, Haiqin Yang, Michael R Lyu, Irwin King, and Anthony Man-Cho So. Online nonlinear auc maximization for imbalanced data sets. *IEEE transactions on neural networks and learning systems*, 29 (4):882–895, 2017.

Shengguo Hu, Yanfeng Liang, Lintao Ma, and Ying He. Msmote: Improving classification performance when training data is imbalanced. In *2009 second international workshop on computer science and engineering*, volume 2, pp. 13–17. IEEE, 2009.

György Kovács. Smote-variants: A python implementation of 85 minority oversampling techniques. *Neurocomputing*, 366:352–354, 2019.

Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Said Hacid, and Hassan Zeineddine. An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE access*, 7:93010–93022, 2019.

Andrei Margeloiu, Adrián Bazaga, Nikola Simidjievski, Pietro Liò, and Mateja Jamnik. Tabmda: Tabular manifold data augmentation for any classifier using transformers with in-context subsetting. In *ICML 1st Workshop on In-Context Learning*, 2024.

Surani Matharaarachchi, Mike Domaratzki, and Saman Muthukumarana. Enhancing smote for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, 18:100597, 2024.

Melkamu Abay Mersha, Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Hassan Shakil, Samer Iskander, Olga Kolesnikova, and Jugal Kalita. Explainable ai: Xai-guided context-aware data augmentation. *Expert Systems with Applications*, 289:128364, 2025.

Azhar Mohammed, Dang Nguyen, Bao Duong, and Thin Nguyen. Efficient classification with counterfactual reasoning and active learning. In *Asian Conference on Intelligent Information and Database Systems*, pp. 27–38. Springer, 2022.

Munehiro Nakamura, Yusuke Kajiwara, Atsushi Otsuka, and Haruhiko Kimura. Lvq-smote–learning vector quantization based synthetic minority over–sampling technique for biomedical data. *BioData mining*, 6: 1–10, 2013.

Krystyna Napierała, Jerzy Stefanowski, and Szymon Wilk. Learning from imbalanced data in presence of noisy and borderline examples. In *International conference on rough sets and current trends in computing*, pp. 158–167. Springer, 2010.

Nasim Mahmud Nayan, Ashraful Islam, Muhammad Usama Islam, Eshtiak Ahmed, Mohammad Mobarak Hossain, and Md Zahangir Alam. Smote oversampling and near miss undersampling based diabetes diagnosis from imbalanced dataset with xai visualization. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6. IEEE, 2023.

Emmanouil Panagiotou, Manuel Heurich, Tim Landgraf, and Eirini Ntoutsi. Tabcf: Counterfactual explanations for tabular data using a transformer-based vae. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 274–282, 2024.

Aum Patil, Aman Framewala, and Faruk Kazi. Explainability of smote based oversampling for imbalanced dataset problems. In *2020 3rd international conference on information and computer technologies (ICICT)*, pp. 41–45. IEEE, 2020.

Louise Manha Peres, Sylvio Barbon Jr, Estefânia Mayumi Fuzyi, Ana Paula AC Barbon, Douglas Fernandes Barbin, Priscila Tiemi Maeda Saito, Nayara Andreo, and Ana Maria Bridi. Fuzzy approach for classification of pork into quality grades: coping with unclassifiable samples. *Computers and Electronics in Agriculture*, 150:455–464, 2018.

Guy-Junior Richard, Jérôme Habonneau, Didier Guériot, and Jean-Marc Le Caillec. Ai explainability and acceptance: A case study for underwater mine hunting. *ACM Journal of Data and Information Quality*, 16(1):1–20, 2024.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Debashis Roy, Anandarup Roy, and Utpal Roy. Learning from imbalanced data in healthcare: State-of-the-art and research challenges. *Computational Intelligence in Healthcare Informatics*, pp. 19–32, 2024.

Zainab Saad Rubaidi, Boulbaba Ben Ammar, and Mohamed Ben Aouicha. Fraud detection using large-scale imbalance dataset. *International Journal on Artificial Intelligence Tools*, 31(08):2250037, 2022.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Hayat Sahlaoui, El Arbi Abdellaoui Alaoui, Said Agoujil, and Anand Nayyar. An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models. *Education and Information Technologies*, 29(5):5447–5483, 2024.

Abdoulaye Sakho, Emmanuel Malherbe, and Erwan Scornet. Do we need rebalancing strategies? a theoretical and empirical study around smote and its variants. *arXiv preprint arXiv:2402.03819*, 2024.

Shoko Sawada and Masashi Toyoda. Model-agnostic visual explanation of machine learning models based on heat map. In *EuroVis (Posters)*, pp. 37–39, 2019.

Matheus Camilo da Silva, Gabriel Marques Tavares, Marcos Cesar Gritti, Paolo Ceravolo, and Sylvio Barbon Junior. Using process mining to reduce fraud in digital onboarding. *FinTech*, 2(1):120–137, 2023.

Samuel Stocksieker, Denys Pommeret, and Arthur Charpentier. Data augmentation with variational autoencoder for imbalanced dataset. In *International Conference on Neural Information Processing*, pp. 354–370. Springer, 2024.

Ahmad S Tarawneh, Ahmad B Hassanat, Ghada Awad Altarawneh, and Abdullah Almuhaimeed. Stop oversampling for class imbalance learning: A review. *IEEe Access*, 10:47643–47660, 2022.

Mohammed Temraz and Mark T Keane. Solving the class imbalance problem using a counterfactual method for data augmentation. *Machine Learning with Applications*, 9:100375, 2022.

Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. *Advances in neural information processing systems*, 31, 2018.

Wuxin Tian, Yanping Shen, Na Guo, Jing Yuan, and Yanqing Yang. Vae-wacgan: an improved data augmentation method based on vaegan for intrusion detection. *Sensors (Basel, Switzerland)*, 24(18):6035, 2024.

Gary M Weiss. Foundations of imbalanced learning. *Imbalanced learning: Foundations, algorithms, and applications*, pp. 13–41, 2013.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Pankaj Yadav, Gulshan Sihag, and Vivek Vijay. Rebalancing the scales: A systematic mapping study of generative adversarial networks (gans) in addressing data imbalance. *arXiv preprint arXiv:2502.16535*, 2025.

Sujin Yang, Kee-Deog Kim, Eiichiro Ariji, Natsuho Takata, and Yoshitaka Kise. Evaluating the performance of generative adversarial network-synthesized periapical images in classifying c-shaped root canals. *Scientific reports*, 13(1):18038, 2023.

A Limitations

Although DPG-da demonstrates notable improvements over conventional over-sampling approaches, certain limitations should be acknowledged.

First, the method introduces a higher computational burden. The extraction of decision predicates from a surrogate model, followed by the use of a GA for optimization, increases runtime compared to classical interpolation-based methods. This effect is particularly evident in large-scale or high-dimensional datasets, which may limit the applicability of the method in time-constrained scenarios. While we used a vanilla GA to provide a straightforward understanding of our approach, this step could be replaced with a more lightweight optimization method to reduce computational costs.

Second, the effectiveness of the extracted constraints depends directly on the quality of the surrogate model. If the surrogate does not approximate the original decision boundaries with sufficient accuracy, the derived constraints may either be too restrictive or misrepresent the structure of the data. This dependence is particularly critical in problems with highly non-linear or complex decision boundaries, potentially limiting the validity of the generated samples.

Third, the procedure introduces an additional level of methodological complexity. The framework ensures that generated samples adhere to identified constraints, but the implementation requires more effort and expertise than simpler alternatives. This complexity is further compounded by the design of the fitness function used in the GA. While our current approach provides a straightforward mechanism to balance multiple objectives, it relies on manually chosen weights, which may affect the quality of the generated data. A more principled multi-objective optimization method (e.g., NSGA-II) could address this limitation, but would introduce additional complexity and computational cost.

Despite these limitations, several aspects of DPG-da mitigate their impact. The higher computational cost may be acceptable in application domains where the validity and consistency of augmented data are critical. Surrogate-derived constraints ensure that new samples remain coherent with the learned decision logic, reducing the risk of unrealistic or infeasible instances. Additionally, the structured nature of the augmentation process facilitates transparency and auditability, which are increasingly important in contexts requiring reliability and compliance. Overall, DPG-da represents a trade-off between computational efficiency and the assurance of constraint-adherent, consistent data augmentation.

B Baseline Over-sampling Methods

The baseline over-sampling methods considered in this work, along with the proposed DPG-da, are summarized in Table 2, highlighting their main features and objectives.

Table 2: Summary of over-sampling methods compared in this study. Methods range from classical interpolation-based techniques to recent generative approaches.

Method	**Description**
ROS	Random over-sampling; simple, replicates existing minority instances.
SMOTE	Interpolation-based over-sampling (Chawla et al., 2002); improves class balance.
SMOTE-ENN	SMOTE followed by Edited Nearest Neighbors (Batista et al., 2004); reduces noise.
SMOTE-SVM	Focuses augmentation near decision boundaries (Bunkhumpornpat et al., 2009).
MSMOTE	Targets difficult minority regions for improved discrimination (Hu et al., 2009).
SMOTE-POLYNOM	Captures non-linear feature relationships (Gazzah & Essoukri Ben Amara, 2008).
DE	Differential evolution-based over-sampling (Chen et al., 2010); enhances diversity.
SMOTE-LVQ	Selects informative minority prototypes for over-sampling (Nakamura et al., 2013).
CTGAN	GAN-based model for tabular data; learns conditional distributions (Xu et al., 2019).
TVAE	Variational Autoencoder for tabular data; handles mixed feature types (Xu et al., 2019).
COPULA-GAN	Combines copula modeling with GANs; balances flexibility and statistical fidelity (Xu et al., 2019).
DPG-da	Ensures domain constraint adherence, interpretable, avoids over-generalization.

C Ablation Study

We conducted a systematic ablation study by evaluating all $3 \times 3 \times 3 = 27$ possible combinations of the three weighting factors from 1.0 to 3.0: distance (w_1), sparsity (w_2), and constraints (w_3). Figure 9 reports the top-performing configurations ranked by macro F1-score. While some configurations with higher values of w_1 (distance factor) consistently appear near the top, the overall differences among the ten best settings are relatively small. This indicates that performance is not overly sensitive to any single factor, and instead depends on a balanced interaction between distance, sparsity, and constraints.

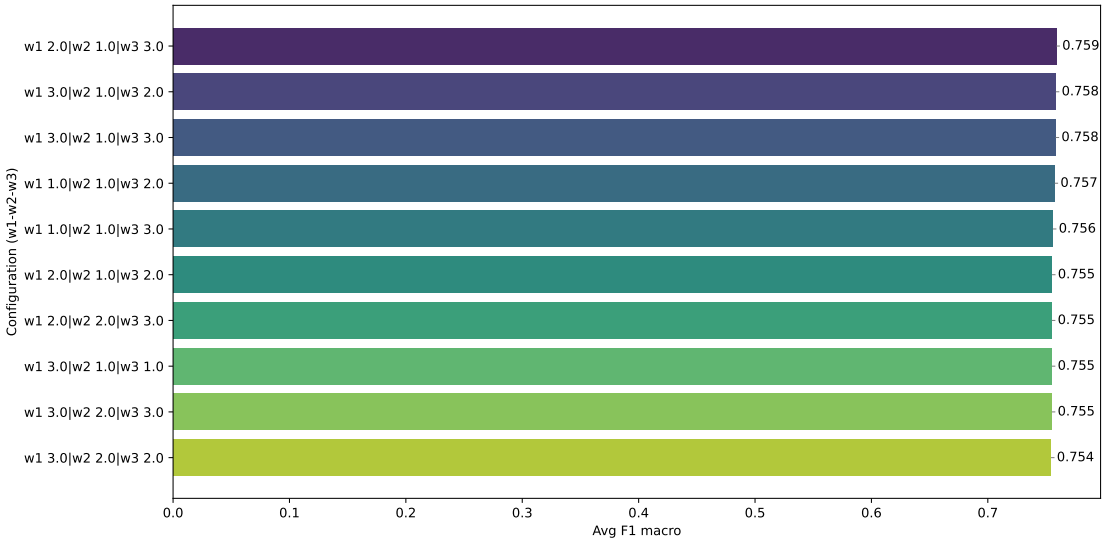

Figure 9: Top-performing ablation weight configurations for the optimization function, ranked by macro F1-score. Each configuration is denoted by its $(w1, w2, w3)$ tuple corresponding to distance, sparsity, and constraints factors. The results indicate which combinations yield the highest predictive performance across all datasets.

D Datasets

D.1 Violation Benchmark

Table 3 provides an overview of the violation-prone datasets used in this study. Each dataset includes clearly defined feature ranges and domain-specific constraints, allowing systematic evaluation of whether synthetic samples generated by augmentation techniques respect valid and feasible values.

D.2 Classification Benchmark

Table 3: Overview of datasets curated for the benchmark of violations

Name	**Ratio**	**#Samples**	**#Features**
Healthcare	3:1	1,000	4
Finance	1:1	1,000	4
Quality Control	2:1	1,000	4
Fraud Detection	3:1	1,000	4
Energy	1:1	1,000	3
Education	1:1	1,000	3
Wisconsin	2:1	683	9
Pima	2:1	768	8
Iris0	2:1	150	4
Subclus	5:1	600	2
Clover	5:1	600	2
Paw	5:1	600	2

Table 4: Summary of the UCI imbalanced benchmarking datasets used in the experiments on Classifier performance. Ratio refers to the imbalance level.

Name	**Ratio**	**#Samples**	**#Features**
ecoli	8.6:1	336	7
optical digits	9.1:1	5,620	64
satimage	9.3:1	6,435	36
pen digits	9.4:1	10,992	16
abalone	9.7:1	4,177	8
sick euthyroid	9.8:1	3,162	25
spectrometer	11:1	531	93
car eval_34	12:1	1,728	21
isolet	12:1	7,797	617
us crime	12:1	1,994	100
yeast ml8	13:1	2,417	103
scene	13:1	2,407	294
libras move	14:1	360	90
thyroid sick	15:1	3,772	52
coil 2000	16:1	9,822	85
arrhythmia	17:1	452	278
solar flare m0	19:1	1,389	32
oil	22:1	937	49
car eval 4	26:1	1,728	21
wine quality	26:1	4,898	11
letter img	26:1	20,000	16
yeast me2	28:1	1,484	8
webpage	33:1	34,780	300
ozone level	34:1	2,536	72
mammography	42:1	11,183	6
protein homo	111:1	145,751	74
abalone 19	130:1	4,177	10

E Constraint Extraction and Guided Augmentation

DPG-da automatically derives constraints from the classifier decision boundaries, which serve as the foundation for generating valid, semantically plausible synthetic samples. Constraints are represented as inequalities over the input features (e.g., bounds on feature combinations, threshold conditions) and are stored in a structured format (JSON) for programmatic evaluation.

As displayed in 5, for each dataset, we analyse the extracted constraints and report the values and types of constraints identified, illustrating how the feasible region for augmentation in DPG-da is defined. These constraints prevent generation of samples violating domain-specific or model-derived rules, ensuring that synthetic instances remain realistic while enriching minority-class coverage.

Table 5: Model-derived constraints extracted by DPG-da for the violation set of datasets.

Dataset	**Class**	**Feature Ranges / Constraints**
Education	negative	Attendance: [50.68, 99.96]; Grades: [52.51, 99.56]; StudyHours: [0.47, 19.72]
	positive	Attendance: [50.68, 99.96]; Grades: [52.51, 99.56]; StudyHours: [0.47, 19.72]
Energy	negative	Usage: [0.52, 2.47]; Baseline: [103.32, 495.35]; Voltage: [190.39, 249.85]
	positive	Usage: [0.52, 2.44]; Baseline: [103.32, 498.23]; Voltage: [191.02, 249.85]
Finance	negative	CreditScore: [308.5, 845.0]; Income: [21340.5, 149246.5]; NumChildren: [0.5, 4.5]
	positive	CreditScore: [309.0, 845.0]; Income: [21755.5, 149246.5]; NumChildren: [0.5, 4.5]
Fraud Detection	negative	TransactionAmount: [134.14, 9884.44]; TransactionTime: [0.89, 29.88]
	positive	TransactionAmount: [134.14, 9955.39]; TransactionTime: [0.3, 29.88]
Healthcare	negative	BMI: [18.09, 39.81]; Cholesterol: [101.0, 296.0]; Age: [20.5, 88.5]
	positive	BMI: [18.14, 39.81]; Cholesterol: [101.0, 294.0]; Age: [20.5, 88.5]
Quality Control	negative	Temperature: [80.25, 118.42]; Pressure: [5.17, 14.87]; Speed: [100.5, 298.5]; Vibration: [1.24, 9.96]
	positive	Temperature: [80.25, 119.13]; Pressure: [5.17, 14.88]; Speed: [106.5, 298.5]; Vibration: [1.24, 9.96]
Wisconsin	negative	ClumpThickness: [3.5, 8.0]; CellSize: [2.5, 4.5]; CellShape: [1.5, 7.5]; MarginalAdhesion: [1.5, 7.5]; EpithelialSize: [1.5, 5.0]; BareNuclei: [1.5, 8.5]; BlandChromatin: [1.5, 4.5]; NormalNucleoli: [1.5, 8.5]; Mitoses: [4.0, 5.0]
	positive	ClumpThickness: [3.5, 8.0]; CellSize: [2.0, 4.5]; CellShape: [2.5, 7.5]; MarginalAdhesion: [1.5, 7.5]; EpithelialSize: [2.5, 6.5]; BareNuclei: [1.5, 8.5]; BlandChromatin: [1.5, 4.0]; NormalNucleoli: [1.5, 7.5]; Mitoses: [1.5, 5.0]
Pima	negative	Preg: [0.5, 13.5]; Plas: [42.5, 180.5]; Pres: [25.0, 99.0]; Skin: [3.5, 41.0]; Insu: [34.0, 471.0]; Mass: [9.65, 47.55]; Pedi: [0.14, 1.31]; Age: [22.5, 66.5]
	positive	Preg: [0.5, 13.5]; Plas: [42.5, 178.0]; Pres: [36.0, 93.0]; Skin: [6.0, 40.0]; Insu: [34.0, 235.0]; Mass: [9.65, 47.55]; Pedi: [0.14, 1.31]; Age: [22.5, 66.5]
Iris0	negative	SepalLength: [4.9, 7.9]; SepalWidth: [2.0, 3.8]; PetalLength: [2.5, 3.5]; PetalWidth: [0.8, 1.8]
	positive	SepalLength: [4.3, 5.8]; SepalWidth: [2.3, 4.4]; PetalLength: [1.7, 2.7]; PetalWidth: [-0.2, 0.8]
Subclus	negative	dim1: [52.5, 360.0]; dim2: [-118.0, 1312.5]
	positive	dim1: [52.5, 360.0]; dim2: [-118.0, 1312.5]
Clover	negative	dim1: [-335.5, 377.5]; dim2: [-269.0, 402.0]
	positive	dim1: [-335.5, 377.5]; dim2: [-367.5, 402.0]
Paw	negative	dim1: [139.5, 616.5]; dim2: [232.0, 854.0]
	positive	dim1: [139.5, 616.5]; dim2: [232.0, 854.0]

E.1 Case Study: Violations

Figure 10 provides a detailed example of violations generated by an over-sampling method on the Wisconsin dataset. Many augmented samples fall outside the valid feature space, entering regions that are semantically invalid. This is particularly evident for features such as *Mitoses*, *NormalNucleoli*, and *MarginalAdhesion*, which take on negative values, an outcome that is unrealistic and inconsistent with domain knowledge.

Such violations are not only theoretically problematic but can also negatively impact downstream model performance. Classifiers trained on datasets containing semantically invalid samples may learn spurious patterns or overfit to unrealistic feature combinations, potentially reducing generalization to real-world data. This case study highlights the practical importance of constraint-aware augmentation methods like DPG-da, which systematically avoid these issues by ensuring that all generated samples remain within valid feature ranges.

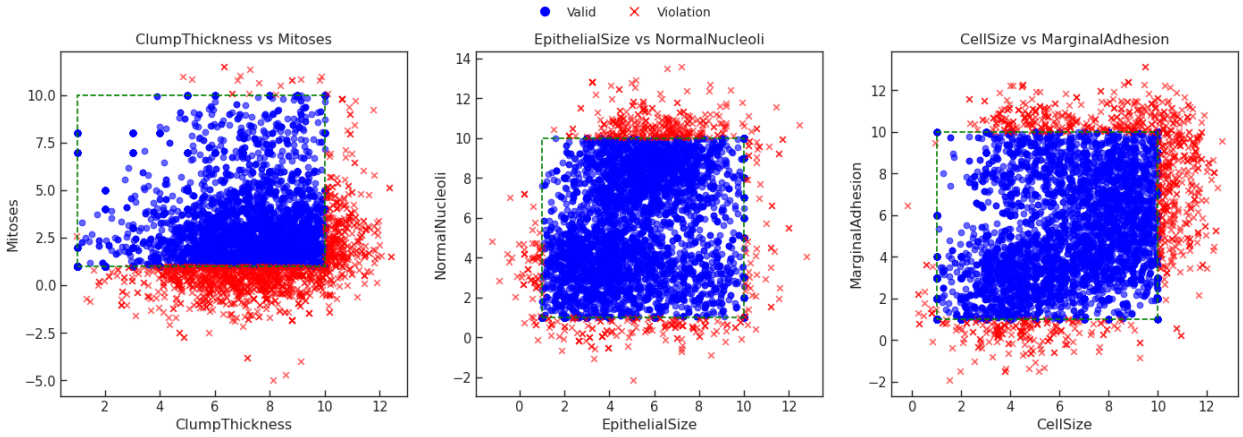

Figure 10: Pairwise scatter plots of SMOTE-LVQ augmented samples from the Wisconsin dataset. Blue circles indicate valid samples within feature constraints, red crosses indicate samples violating at least one feature range. Green rectangles show the valid region for each feature pair.

F Traceability across generations

To illustrate the traceability of our augmentation process, we track how candidate samples evolve during the GA. For each dataset, we record the feature-wise differences (Δ) between successive generations relative to the initial query sample. This provides an interpretable view of how synthetic samples are formed: which features are modified, when changes occur, and whether evolution converges to stable solutions.

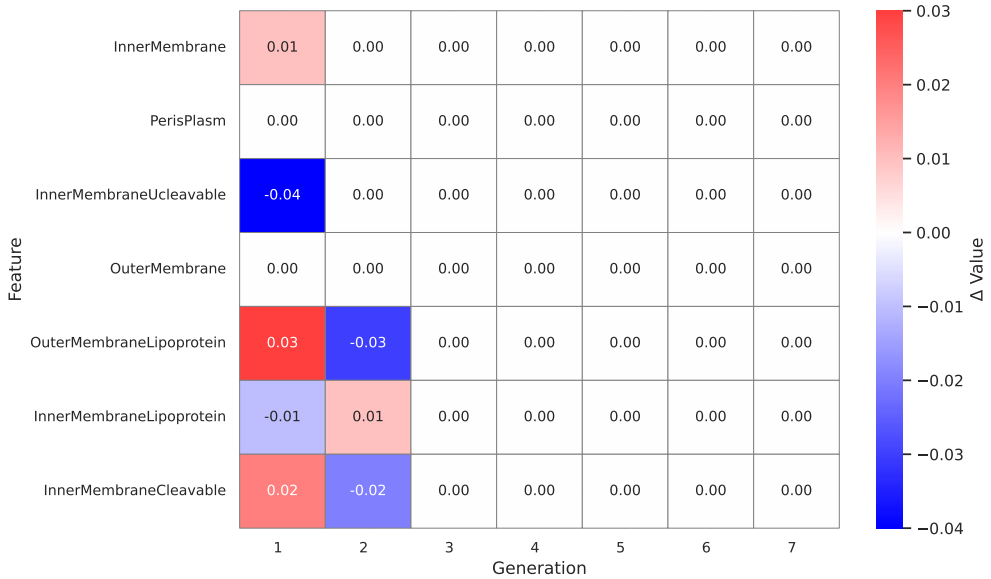

Figure 11: Heatmap of feature-wise changes (Δ) during the GA on the **Ecoli** dataset. Early generations explore a broad range of features, with substantial variation across most attributes. As evolution proceeds, the process converges, reflecting stabilization of the candidate samples within the feasible region.

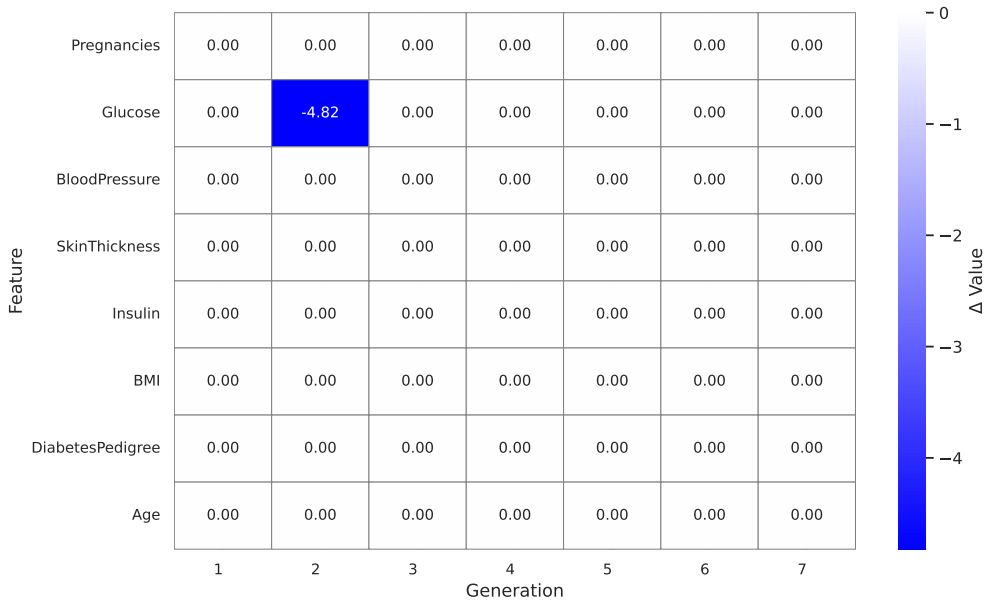

Figure 12: Heatmap of feature-wise changes (Δ) during the GA on the **Pima** dataset. Unlike Ecoli, the augmentation converges rapidly by modifying exclusively the *glucose* feature. This highlights a highly localized and interpretable transformation path, where minority samples are generated through a single, domain-relevant attribute shift.

G Performance Across Datasets

Table 6 reports the F1-Score for each dataset and over-sampling method, averaged over multiple augmentation levels. This table provides a detailed comparison of classifier performance under different augmentation strategies, highlighting both the effectiveness and limitations of each approach. Notably, it allows readers to observe which methods consistently improve classification across datasets and which are less robust, complementing the violation-focused analyses presented in the main text.

Table 6: F1-Score by dataset and over-sampling method, averaged over augmentation levels. The highest performance per dataset is in bold.

Dataset	DE	SMOTE-ENN	MSMOTE	SMOTE-LVQ	COPULA-GAN	CTGAN	SMOTE-POLYNOM	SMOTE	ROS	TVAE	SMOTE-SVM	DPG-da
abalone	0.600	0.602	**0.614**	0.584	0.588	0.599	0.601	0.594	0.593	0.598	0.608	0.566
abalone_19	0.487	0.505	0.535	0.492	0.497	0.525	0.511	0.474	0.509	0.505	0.536	**0.627**
arrhythmia	0.639	0.659	0.590	0.617	0.629	0.672	0.651	0.716	0.617	0.654	0.663	**0.860**
car_eval_34	0.821	0.829	0.823	0.866	0.874	0.889	0.872	0.870	0.881	**0.921**	0.882	0.862
car_eval_4	0.853	0.866	0.897	0.885	0.902	0.900	0.895	0.917	0.896	**0.934**	0.907	0.903
coil_2000	0.523	0.522	0.530	0.529	0.520	0.521	0.530	0.537	0.533	0.526	0.534	**0.619**
ecoli	0.656	0.762	0.733	0.755	0.781	0.751	0.760	0.751	0.722	0.782	0.749	**0.842**
isolet	0.729	0.810	0.826	0.862	0.872	0.876	0.841	0.870	0.861	0.868	0.849	**0.884**
letter_img	0.866	0.913	0.899	0.907	0.883	0.897	0.917	**0.931**	0.915	0.918	0.899	0.908
libras_move	0.667	0.826	0.833	0.805	0.831	0.862	0.855	0.834	0.840	**0.884**	0.867	0.857
mammography	0.683	0.740	0.759	0.752	0.721	0.745	0.750	0.732	0.772	0.762	**0.794**	0.554
oil	0.601	0.602	0.627	0.610	0.600	0.594	0.615	0.624	0.622	0.598	0.638	**0.713**
optical_digits	0.755	0.927	0.921	0.935	0.927	0.930	0.927	0.936	0.928	0.933	0.922	**0.937**
ozone_level	0.546	0.560	0.554	0.552	0.530	0.517	0.562	0.561	0.564	0.580	0.570	**0.682**
pen_digits	0.754	0.946	0.939	0.942	0.908	0.915	0.946	0.903	**0.947**	0.935	0.931	0.923
protein_homo	0.637	0.701	0.752	0.749	0.850	**0.852**	0.711	0.846	0.776	0.730	0.789	0.835
satimage	0.674	0.683	0.679	0.695	0.676	0.676	0.695	0.685	0.697	0.693	0.694	**0.739**
scene	0.572	0.538	0.568	0.573	0.553	0.559	0.561	0.588	0.580	0.574	0.589	**0.657**
sick_euthyroid	0.715	0.733	0.706	0.726	0.743	0.731	0.754	0.742	0.754	0.763	**0.768**	0.760
solar_flare_m0	0.505	0.552	0.536	0.524	**0.561**	0.553	0.538	0.552	0.540	0.534	0.551	0.520
spectrometer	0.839	0.869	0.886	0.837	0.858	0.876	0.884	0.861	0.884	0.872	0.881	**0.900**
thyroid_sick	0.722	0.748	0.709	0.730	0.763	0.735	0.769	0.737	0.771	0.781	0.778	**0.813**
us_crime	0.678	0.681	0.696	0.702	0.704	0.689	0.697	0.701	0.699	0.697	0.720	**0.731**
webpage	0.731	0.692	0.731	0.808	0.781	0.780	0.746	0.807	0.783	0.753	0.769	**0.836**
wine_quality	0.558	0.582	0.587	0.563	0.601	0.577	0.591	0.583	0.606	0.580	0.615	**0.627**
yeast_me2	0.553	0.597	0.609	0.591	0.583	0.524	0.602	0.567	0.602	0.612	0.634	**0.690**
yeast_ml8	0.495	0.430	0.477	0.500	0.499	0.493	0.481	0.446	0.504	0.504	0.499	**0.676**

Table 7: Detailed statistical results (MD, MAD, MR) complementing the CD diagram in Figure 5

Method	MD ± MAD	MR
DPG-da	0.757 ± 0.109	2.531
TVAE	0.719 ± 0.144	3.704
ROS	0.728 ± 0.131	3.975
SMOTE	0.708 ± 0.130	4.506
CTGAN	0.722 ± 0.135	5.272
COPULA-GAN	0.691 ± 0.158	5.025
SMOTE-ENN	0.693 ± 0.115	5.494
MSMOTE	0.705 ± 0.120	5.494