Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants

Abdoulaye Sakho[A,B], **Emmanuel Malherbe**[A] **and Erwan Scornet**[B]
[A]Artefact Research Center, France
[B]Sorbonne Université, Université Paris Cité, CNRS, LPSM, France

Synthetic Minority Oversampling Technique **(SMOTE) is a common rebalancing strategy for handling imbalanced tabular data sets. However, few works analyse SMOTE theoretically. In this paper, we derive several non-asymptotic upper bound on SMOTE density. From these results, we prove that SMOTE (with default parameter) tends to copy the original minority samples asymptotically. We confirm and illustrate this first theoretical behavior numerically. Furthermore, we prove that SMOTE density vanishes near the boundary of the support of the minority class distribution. We then adapt SMOTE based on our theoretical findings to introduce two new variants. These strategies are compared on** 13 **tabular data sets with** 10 **state-of-the-art rebalancing procedures, including deep generative and diffusion models. First, for most data sets, applying no rebalancing strategy is competitive in terms of predictive performances, would it be with LightGBM, tuned random forests or logistic regression. Second, when the imbalance ratio is artificially augmented, one of our two modifications of SMOTE leads to promising predictive performances compared to SMOTE and other strategies.**

Keywords: *Imbalanced Learning, SMOTE, Oversampling, Tabular Data*

1. INTRODUCTION

Imbalanced data sets for binary classification are encountered in various fields such as fraud detection (Hassan and Abraham, 2016), medical diagnosis (Khalilia et al., 2011) and churn detection (Nguyen and Duong, 2021). In our study, we focus on imbalanced data in the context of binary classification on *tabular data* (thus excluding images and text data sets), for which most machine learning algorithms have a tendency to predict the majority class. This leads to biased predictions, so that several rebalancing strategies have been developed in order to handle this issue, as explained by Krawczyk (2016) and Ramyachitra and Manikandan (2014). These procedures can be divided into model-level and data-level categories.

Model-level approaches modify existing classifiers in order to prevent predicting only the majority class. Among such techniques, Class-Weight (CW) works by assigning higher weights to minority samples. Another related procedure proposed by Zhu et al. (2018) assigns data-driven weights to each tree of a random forest, in order to improve aggregated metrics such as F1 score or ROC AUC. Another model-level technique is to modify the loss function of the classifier. For instance, Cao et al. (2019), Lin et al. (2017) and Ren et al. (2018) respectively introduce LDAM, Focal and L2RW losses, in order to produce neural network classifiers that better handle imbalanced data sets. However, model-level approaches are not model agnostic, and thus cannot be applied to a wide variety of machine learning algorithms. Consequently, we focus in this paper on data-level approaches.

Data-level approaches can be divided into two groups: synthetic and non-synthetic procedures. Non-synthetic procedures works by removing or copying original data points. Mani and Zhang (2003) explain that Random Under Sampling (RUS) is one of the most used resampling strategy and design new adaptive versions called Nearmiss. RUS produces the prespecified balance between classes by dropping uniformly at random majority class samples. The Nearmiss1 strategy (Mani and Zhang, 2003) includes a distinction between ma-

Published at the International Conference on Artificial Intelligence and Statistics (AISTATS) 2026
Artefact Research Center, Paris, France

jority samples by ranking them with their mean distance to their nearest neighbor from the minority class. Then, low-ranked majority samples are dropped until a given balancing ratio is reached. In contrast, Random Over Sampling (ROS) duplicates original minority samples. The main limitation of all these sampling strategies is the fact that they either remove information from the data or do not add new information.

On the contrary, synthetic procedures generate new synthetic samples in the minority class. For instance, Random Over Sampling Examples (see Menardi and Torelli, 2014) is a variant of ROS that produces duplicated samples and then add a noise in order to get these samples slightly different from the original ones. This leads to the generation of new samples on the neighborhood of original minority samples. One of the most famous synthetic strategies is *Synthetic Minority Oversampling Technique* (SMOTE, see Chawla et al., 2002)[1]. In SMOTE, new minority samples are generated via linear interpolation between an original minority sample and one of its nearest neighbor in the minority class. Other approaches are based on Generative Adversarial Networks (GAN Xu et al., 2019; Islam and Zhang, 2020) or diffusion (Jolicoeur-Martineau et al., 2024), which are computationally expensive and mostly designed for specific data structures, such as images. Furthermore, some recent works (2023) have shown that SMOTE remains competitive even compared with recent diffusion models (see Table 4 and Table 5 in Kotelnikov et al., 2023).

Contributions We place ourselves in the setting of imbalanced classification on tabular data, which is very common in real-world applications (see Shwartz-Ziv and Armon, 2022, for details). In this paper:

- We prove that, without tuning the hyperparameter K (usually set to 5), SMOTE asymptotically copies the original minority samples, therefore lacking the intrinsic variability required in any synthetic generative procedure. We provide numerical illustrations of this limitation (Section 3). We also establish that SMOTE density vanishes near the boundary of the support of the minority distribution.
- Our theoretical analysis naturally leads us to introduce two SMOTE alternatives, SMOTE K-tuned and Multivariate Gaussian SMOTE (MGS). In Section 4, we evaluate our new strategies and state-of-the-art rebalancing strategies on several real-world data sets using random forests, logistic regression and LightGBM. We find that, for most data sets, applying no strategy is competitive in terms of predictive performances. When the imbalance ratio is dramatically increased, our proposed SMOTE modification MGS is among the best strategies (Section 5).

[1]More than 25.000 papers found in GoogleScholar with a title including "SMOTE" over the last decade.

2. RELATED WORKS

In this section, we focus on the literature that is the most relevant to our work: long-tail learning, SMOTE variants and studies of rebalancing strategies.

Long-tailed learning (see, e.g., Zhang et al., 2023) is a relatively new field, originally designed to handle image classification with numerous output classes such as for ImageNet-LT (Liu et al., 2019) or iNaturalist (Van Horn et al., 2018) data sets. Most techniques in long-tailed learning are based on neural networks or use the large number of classes to build or adapt aggregated predictors. However, in most tabular classification data sets, the number of classes to predict is relatively small, usually equal to two (Chawla et al., 2004; He and Garcia, 2009; Grinsztajn et al., 2022). Therefore, long-tailed learning methods are not intended for our setting as (i) we only have two output classes and (ii) state-of-the-art models for tabular data are not neural networks but tree-based methods, such as random forests or gradient boosting (see Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022). However, we decide to include in our study several model-level rebalancing strategies from this framework, such as LDAM, Focal and L2RW losses. Finally, we highlight that SMOTE-related techniques are widely used for image data sets. For example, Mix-up methodology (Zhang et al., 2018) performs a linear interpolation between two images and is commonly used by prac-

titioners.

SMOTE has seen many variants proposed in the literature. Several of them focus on generating synthetic samples near the boundary of the minority class support, such as ADASYN (He et al., 2008), SVM-SMOTE (Nguyen et al., 2011) or Borderline SMOTE (Han et al., 2005). Many other variants exist such as SMOTEBoost (Chawla et al., 2003), Adaptive-SMOTE (Pan et al., 2020), GDO (Xie et al., 2020) or DBSMOTE (Bunkhumpornpat et al., 2012). From a computational perspective, several synthetic methods (SMOTE, ADASYN, Borderline SMOTE and SVM-SMOTE) are available in the open-source package *imb-learn* (see Lemaître et al., 2017). Several papers study experimentally some specificities of the sampling strategies. For example, Kamalov et al. (2022) study the optimal sampling ratio for imbalanced data sets when using synthetic approaches. Aguiar et al. (2023) give an overview of imbalance data sets in the context of online learning and propose a standardized framework in order to compare rebalancing strategies in this context. Furthermore, Wongvorachan et al. (2023) aim at analysing the data-level approaches (ROS, RUS and SMOTE) on educational data in order to predict probability of pursuing higher education or the risk of leaving high school before graduation.

Rebalancing strategies are studied in several works. Xu et al. (2020) consider the weighted risk of plug-in classifiers, for arbitrary weights. They establish rates of convergence and derive a new robust risk that may in turn improve classification performance in imbalanced scenarios. Based on this work, Aghbalou et al. (2024) derive a sharp error bound on the balanced risk for binary classification with severe class imbalance. Using extreme value theory, Chaudhuri et al. (2023) show that applying Random Under Sampling in binary classification improve the worst-group error when learning from imbalanced classes with tails. Wallace and Dahabreh (2014) study the class probability estimates for several rebalancing strategies before introducing a generic methodology in order to improve all these estimates. Dal Pozzolo et al. (2015) focus on the effect of RUS on the posterior probability of the selected classifier. They show that RUS affect the accuracy and the probability calibration of the model. To the best of our knowledge, there are only few theoretical works dissecting the intrinsic machinery in SMOTE algorithm, with the notable exception of Elreedy and Atiya (2019) and Elreedy et al. (2023) who established the density of synthetic observations generated by SMOTE, the associated expectation and covariance matrix. Elreedy and Atiya (2019) also highlights the effects of the number of minority samples and the input dimension on SMOTE procedure. Indeed, results on simulated and real-world data sets show that the predictive performance of the classifier applied after SMOTE increases with the number of minority samples. Increasing the number of input variables decreases SMOTE ability of regenerating the minority distribution, in terms of Total Variance Difference and Kullback–Leibler divergence.

3. A STUDY OF SMOTE

Notations We denote by $\mathcal{U}([a,b])$ the uniform distribution over $[a,b]$. We denote by $\mathcal{N}(\mu,\Sigma)$ the multivariate normal distribution of mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. For any set A, we denote by $Vol(A)$, the Lebesgue measure of A. For any $z \in \mathbb{R}^d$ and $r > 0$, let $B(z,r)$ be the ball for the L_2 norm, centered at z of radius r. We note $c_d = Vol(B(0,1))$ the volume of the unit ball in $\mathbb{R}^d$. For any $p,q \in \mathbb{N}$, and any $z \in [0,1]$, we denote by $\mathcal{B}(p,q;z) = \int_{t=0}^{z} t^{p-1}(1-t)^{q-1}\mathrm{d}t$ the incomplete beta function.

3.1. SMOTE algorithm

We assume to be given a training sample $\mathcal{D}_N$ composed of N pairs (X_i,Y_i), independent and identically distributed as (X,Y), where X and Y are random variables that take values respectively in $\mathcal{X} \subset \mathbb{R}^d$ and $\{0,1\}$. We consider a class imbalance problem, in which the class $Y=1$ is underrepresented, compared to the class $Y=0$, and thus called the minority class. We assume that we have n minority samples in our training set. We define the imbalance ratio as n/N. In this paper, we consider continuous input variables only, as SMOTE was originally designed for.

Algorithm 1 Single SMOTE Sample Generation

Require: Minority class samples $\{X_1, \dots, X_n\} \subset \mathbb{R}^d$, number of neighbors K
Ensure: Synthetic sample Z
1: Sample a central point X_c uniformly from $\{X_1, \dots, X_n\}$
2: Compute the set I of the K nearest neighbors of X_c (Euclidean distance)
3: Sample a neighbor X_k uniformly from I
4: Sample $w \sim \mathcal{U}(0, 1)$
5: Generate $Z \leftarrow X_c + w(X_k - X_c)$
6: **return** Z

SMOTE procedure generates synthetic data through linear interpolations between two pairs of original samples of the minority class. SMOTE algorithm has a single hyperparameter, K, by default set to 5, which stands for the number of nearest neighbors considered when interpolating. We denote by $Z_{K,n}$ the distribution of samples generated by SMOTE. A single SMOTE iteration is detailed in Algorithm 1. In a machine learning pipeline, SMOTE procedure is repeated in order to obtain a prespecified ratio between the two classes, before training a classifier.

A classical statistical framework is to let the total number of samples grow to infinity, i.e. $N \to \infty$. In this setting, we assume that the imbalance ratio is asymptotically constant, that is

$$\lim_{N\to\infty} \frac{n}{N} = \mathbb{P}(Y = 1). \tag{1}$$

In particular, the number of minority samples tends to infinity. Thus, throughout the following section, we study the theoretical behavior of SMOTE, by letting $n \to \infty$ or by providing finite-sample upper bounds.

3.2. Theoretical results on SMOTE

SMOTE has been shown to exhibit good performances when combined to standard classification algorithms (see, e.g., Mohammed et al., 2020). However, there exist only few works that aim at understanding theoretically SMOTE behavior. In this section, we assume that $X_1, \dots, X_n$ are i.i.d samples from the minority class (that is, $Y_i = 1$ for all $i \in [n]$), with a common density f_X with bounded support, denoted by $\mathcal{X}$.

Lemma 3.1 (Convexity). *Given f_X the distribution density of the minority class, with support $\mathcal{X}$, for all K, n, the associated SMOTE density $f_{Z_{K,n}}$ satisfies*

$$Supp(f_{Z_{K,n}}) \subseteq Conv(\mathcal{X}). \tag{2}$$

By construction, synthetic observations generated by SMOTE cannot fall outside the convex hull of $\mathcal{X}$. Equation (2) is not an equality, as SMOTE samples are the convex combination of only two original samples. For example, in dimension two, if $\mathcal{X}$ is concentrated near the vertices of a triangle, then SMOTE samples are distributed near the triangle edges, whereas $\mathrm{Conv}(\mathcal{X})$ is the surface delimited by the triangle.

SMOTE algorithm has only one hyperparameter K, which is the number of nearest neighbors taken into account for building the linear interpolation. By default, this parameter is set to 5. The following theorem describes the behavior of SMOTE distribution asymptotically, as $K/n \to 0$.

Theorem 3.2. *For all Borel sets $B \subset \mathbb{R}^d$, if $K/n \to 0$, as n tends to infinity, we have*

$$\lim_{n\to\infty} \mathbb{P}[Z_{K,n} \in B] = \mathbb{P}[X \in B]. \tag{3}$$

The proof of Theorem 3.2 can be found in Section D.2. Theorem 3.2 proves that the random variables $Z_{K,n}$ generated by SMOTE converges in distribution to the original random variable X, provided that K/n tends to zero. From a practical point of view, Theorem 3.2 guarantees asymptotically the ability of SMOTE to regenerate the distribution of the minority class. This highlights a good behavior of the default setting of SMOTE ($K = 5$), as it can create more data points, different from the original samples while preserving the original data distribution. Note that Theorem 3.2 is very generic, as it makes no assumptions on the distribution of X.

SMOTE distribution has been derived in Theorem 1 and Lemma 1 in Elreedy et al. (2023) using geometrical arguments. We recall the expression of this density in Lemma C.1 in Appendix, with a proof based on random variables (see Section D.3). When no confusion is possible, we simply write f_Z instead of $f_{Z_{K,n}}$. A close inspection of the SMOTE density allows us to derive more

precise bounds about the behavior of SMOTE, as established in Theorem 3.4.

Assumption 3.3. There exists $R > 0$ such that $\mathcal{X} \subset B(0,R)$. Besides, there exist $0 < C_2 < \infty$ such that for all $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x\in\mathcal{X}}$.

Theorem 3.4. *Grant Assumption 3.3. Let $x_c \in \mathcal{X}$ and $\alpha \in (0, 2R)$. For all $K \leq (n-1)\mu_X(B(x_c,\alpha))$, we have*

$$\mathbb{P}(\|Z_{K,n} - X_c\|_2 \geq \alpha | X_c = x_c) \leq \eta_{\alpha,R,d} \times \exp\left(-2(n-1)\left(\mu_X(B(x_c,\alpha)) - \frac{K}{n-1}\right)^2\right), \tag{4}$$

$$\eta_{\alpha,R,d} = C_2 c_d R^d \times \begin{cases} \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1, \\ \frac{1}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{if } d > 1. \end{cases}$$

Consequently, if $\lim_{n\to\infty} K/n = 0$, we have, for all $x_c \in \mathcal{X}$, $Z_{K,n}|X_c = x_c \to x_c$ in probability.

The proof of Theorem 3.4 can be found in Section D.4. Theorem 3.4 establishes an upper bound on the distance between an observation generated by SMOTE and its central point. Asymptotically, when K/n tends to zero, the new synthetic observation concentrates around the central point. Recall that, by default, $K = 5$ in SMOTE algorithm. Therefore, Theorem 3.2 and Theorem 3.4 prove that, with the default settings, SMOTE asymptotically targets the original density of the minority class and generates new observations very close to the original ones. Besides, the assumption for all $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x\in\mathcal{X}}$ can be removed at the cost of replacing the factor $(2R/\alpha)^{d-1}$ by $(2R/\alpha)^d$ when $d > 1$, or $\ln\left(\frac{2R}{\alpha}\right)$ by $(2R/\alpha)^{-1}\ln\left(\frac{2R}{\alpha}\right)$ when $d = 1$ (see eq (87)). These resulting factors are less sharp than those obtained under the boundedness assumption on C_2, which enables a more precise control of the concentration of SMOTE observations around the original data points. The following result establishes the characteristic distance between SMOTE observations and their central points.

Corollary 3.5. *Grant Assumption 3.3. For all $d \geq 2$, for all $\gamma \in (0, 1/d)$, we have*

$$\mathbb{P}\left[\|Z_{K,n} - X_c\|_2 > 12R(K/n)^\gamma\right] \leq \left(\frac{K}{n}\right)^{2/d-2\gamma}. \tag{5}$$

The proof of Corollary 3.5 can be found in Section D.5. The characteristic distance between a SMOTE observation and the associated central point is of order $(K/n)^{1/d}$. As expected from the curse of dimensionality, this distance increases with the dimension d. Choosing K that increases with n leads to larger characteristic distances: SMOTE observations are more distant from their central points. Corollary 3.5 leads us to choose K such that K/n does not tend too fast to zero, so that SMOTE observations are not too close to the original minority samples. However, choosing such a K can be problematic, especially near the boundary of the support, as shown in the following theorem.

Theorem 3.6. *There exists $R > 0$ such that $\mathcal{X} = B(0,R)$. Besides, there exist $0 < C_1 < C_2 < \infty$ such that for all $x \in \mathbb{R}^d$, $C_1 \mathbb{1}_{x\in\mathcal{X}} \leq f_X(x) \leq C_2 \mathbb{1}_{x\in\mathcal{X}}$. Let $\varepsilon \in (0,R)$ such that $\left(\frac{\varepsilon}{R}\right)^{1/2} \leq \frac{c_d}{\sqrt{2}dC_2}$. Then, for all $1 \leq K < n$, and all $z \in B(0,R)\backslash B(0,R-\varepsilon)$, and for all $d > 1$, we have*

$$f_{Z_{K,n}}(z) \leq C_2^{3/2}\left(\frac{2^{d+2}c_d^{1/2}}{d^{1/2}}\right)\left(\frac{n-1}{K}\right)\left(\frac{\varepsilon}{R}\right)^{1/4}. \tag{6}$$

The proof of Theorem 3.6 can be found in Section D.6. Theorem 3.6 establishes an upper bound of SMOTE density at points distant from less than ε from the boundary of the minority class support. More precisely, Theorem 3.6 shows that SMOTE density vanishes as $\varepsilon^{1/4}$ near the boundary of the support. Choosing $\varepsilon/R = o((K/n)^4)$ leads to a vanishing upper bound, which proves that SMOTE density is unable to reproduce the original density $f_X \geq C_1$ in the peripheral area $B(0,R)\backslash B(0,R-\varepsilon)$. Such a behavior was expected since the boundary bias of local averaging methods (kernels, nearest neighbors) has been extensively studied (see, e.g. Jones, 1993; Arya et al., 1995; Arlot and Genuer, 2014; Mourtada et al., 2020).

For default settings of SMOTE (i.e., $K = 5$), and large sample size, this area is relatively small ($\varepsilon = o(n^{-4})$). Still, Theorem 3.6 provides a theoretical ground for understanding the behavior of SMOTE near the boundary, a phenomenon that has led to introduce variants of SMOTE to circumvent this issue (see Borderline SMOTE in Han et al., 2005). While increasing K leads to more

diverse generated observations (Theorem 3.4), it increases the boundary bias of SMOTE. Indeed, choosing $K = n^{3/4}$ implies a boundary effect in the peripheral area $B(0,R)\backslash B(0,R-\varepsilon)$ for $\varepsilon = o(1/n)$, which may not be negligible. Finally, note that constants in the upper bounds are of reasonable size. Letting $d = 3$, $K = 5$, $X \sim \mathcal{U}(B_d(0,1))$, the upper bound turns into $0.89n\varepsilon^{1/4}$.

3.3. Numerical illustrations

We illustrate Theorem 3.4 and Corollary 3.5 with simulated data. In Figure 1, we define a diversity measure $d(\mathbf{Z},\mathbf{X})$ for the synthetic samples (y-axis, detailed below) and plot it as a function of the number of minority samples n for different values of K.

Simulated data In order to measure the similarity between any generated data set $\mathbf{Z} = \{Z_1,\dots,Z_m\}$ and the original data set $\mathbf{X} = \{X_1,\dots,X_n\}$, we compute $C(\mathbf{Z},\mathbf{X}) = \frac{1}{m}\sum_{i=1}^{m}\|Z_i - X_{(1)}(Z_i)\|_2$, where $X_{(1)}(Z_i)$ is the nearest neighbor of Z_i among $X_1,\dots,X_n$. Intuitively, this quantity measures how far the generated data set is from the original observations: if the new data are copies of the original ones, this measure equals zero. We apply the following protocol: for each value of n,

1. Generate $\mathbf{X}$ composed of n i.i.d samples distributed as $\mathcal{U}([-3,3]^2)$.
2. Generate $\mathbf{Z}$ composed of $m = 1000$ new i.i.d observations by applying SMOTE procedure on the original data set $\mathbf{X}$, with different values of K. Compute $C(\mathbf{Z},\mathbf{X})$.
3. Generate $\tilde{\mathbf{X}}$ composed of m i.i.d new samples distributed as $\mathcal{U}([-3,3]^2)$. Compute $C(\tilde{\mathbf{X}},\mathbf{X})$, which is a reference value in the ideal case of new points sampled from the same distribution.

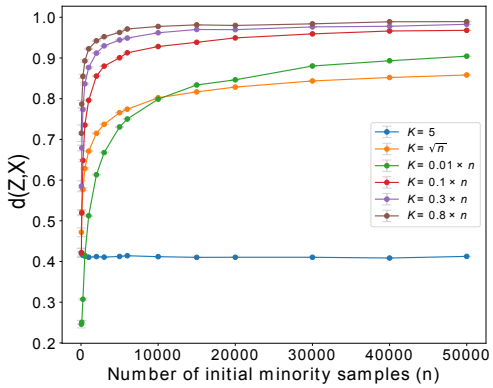


Figure 1 | Diversity measure with $\mathcal{U}([-3,3]^2)$

Steps 1-3 are repeated 75 times. The average of $C(\mathbf{Z},\mathbf{X})$ (resp. $C(\tilde{\mathbf{X}},\mathbf{X})$) over these repetitions is computed and denoted by $\bar{C}(\mathbf{Z},\mathbf{X})$ (resp. $\bar{C}(\tilde{\mathbf{X}},\mathbf{X})$). We consider the diversity measure $d(\mathbf{Z},\mathbf{X}) = \bar{C}(\mathbf{Z},\mathbf{X})/\bar{C}(\tilde{\mathbf{X}},\mathbf{X})$, depicted in Figure 1 (see also Figure 3 in Section B for $\bar{C}(\mathbf{Z},\mathbf{X})$).

The diversity measure close to one corresponds to synthetic SMOTE samples that are independent from the initial samples, and generated as the original samples. On the contrary, a diversity measure close to zero corresponds to synthetic SMOTE samples that are close from their central points, regenerating the original distribution with less diversity than the true underlying distribution (uniform distribution).

We note that the diversity measure increases with n for all settings (different choices of K), except for the constant choice $K = 5$. Thus, default SMOTE ($K = 5$) generates samples that are less diverse, than those generated with other choices for K. We extended our simulations to real-world data sets and obtained the same conclusions (see Figure 2 in Appendix B).

4. PREDICTIVE EVALUATION ON INITIAL REAL-WORLD DATA SETS

In this section, we propose two minor modifications of SMOTE, in order to circumvent the theoretical limitations of SMOTE identified in the previous section. Then we analyse and compare their associated predictive performances to the state-of-the-art rebalancing strategies. We start by presenting existing strategies to deal with imbalanced data.

4.1. Existing strategies

Class-weight (CW) Class weighting assigns the same weight (chosen as hyperparameter) to each minority samples. The default setting for this

strategy is to choose a weight ρ such that $\rho n = N - n$, where n and $N - n$ are respectively the number of minority and majority samples in the data set.

Random Over/Under Sampling Random Under Sampling (RUS) acts on the majority class by selecting uniformly without replacement several samples in order to obtain a prespecified size for the majority class. Similarly, Random Over Sampling (ROS) acts on the minority class by selecting uniformly with replacement several samples to be copied.

NearMissOne NearMissOne is an undersampling procedure. For each sample X_i in the majority class, the averaged distance of X_i to its K nearest neighbors in the minority class is computed. Then, the samples X_i are ordered according to this averaged distance. Finally, iteratively, the first X_i is dropped (smallest mean first) until the given ratio is reached.

Borderline SMOTE 1 and 2 Borderline SMOTE 1 (Han et al., 2005) procedure works as follows. For each individual X_i in the minority class, let $m_-(X_i)$ be the number of samples of the majority class among the m nearest neighbors of X_i, where m is a hyperparameter. For all X_i in the minority class such that $m/2 \leq m_-(X_i) < m$, generate successive samples $Z = WX_i + (1 - W)X_k$ where $W \sim \mathcal{U}([0, 1])$ and X_k is selected among the K nearest-neighbors of X_i in the minority class. In Borderline SMOTE 2 (Han et al., 2005), the selected neighbor X_k is chosen from the neighbors of both positive and negative classes, and Z is sampled with $W \sim \mathcal{U}([0, 0.5])$.

CTGAN and ForestDiff CTGAN (Conditional Tabular GAN) (Xu et al., 2019) is a conditional generative adversarial network that is specially designed for generating new synthetic samples for tabular data sets. CTGAN uses conditional generation to sample specific categories during training and applies a mode-specific normalization. ForestDiffusion (Jolicoeur-Martineau et al., 2024) is a score-based diffusion and a conditional flow matching models which substitute the score function learner, traditionally a neural networks, by gradient-boosted decision trees classifier (XGBoost). This substitution capitalizes on XGBoost proficiency in handling tabular data.

4.2. Variants of SMOTE

Now, we introduce two slight modifications of SMOTE in order to alleviate the theoretical limitations (tendency to copy the minority samples for fixed K and artifact boundaries) demonstrated in Section 3.

***K*-tuned SMOTE** The analysis of Theorem 3.4 and Corollary 3.5 suggests that the parameter K of SMOTE should not be fixed to a default value. Through several numerical experiments, we did not identify a clear way to set K as a function of n that systemically induces better predictive performances. SMOTE K-tuned finds the best hyperparameter K among a prespecified grid via a 5-fold cross-validation procedure. The grid is composed of the set $\{1, 2, \ldots, 15\}$ extended with the values $\lfloor 0.01 n_{train} \rfloor$, $\lfloor 0.1 n_{train} \rfloor$, $\lfloor 0.5 n_{train} \rfloor$, $\lfloor 0.7 n_{train} \rfloor$ and $\lfloor \sqrt{n_{train}} \rfloor$, where n_{train} is the number of minority samples in the training set. SMOTE K-tuned is one of the simplest ideas to solve some SMOTE limitations highlighted theoretically in Section 3. To the best of our knowledge, there is no previous work that recommends tuning the hyperparameter K of SMOTE, neither recommend values of K that depends on n.

Multivariate Gaussian SMOTE (MGS) Lemma 3.1 leads us to modify SMOTE algorithm such that the generated samples are no longer contained in the convex hull of the original minority samples. Furthermore, we expect from Theorem 3.4 and Theorem 3.6 that adopting a different distribution function to generate new samples, given the central point and its neighbors, would limit the copying effect and the boundary artifact of SMOTE. Thus, we slightly modify SMOTE to introduce a new oversampling strategy in which new samples are generated from several *multivariate* Gaussian distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$, where the empirical mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ are estimated using the K neighbors and the central point (see Algorithm 2 for details). Note that MGS generated samples follow a mixture of Gaussian distributions. By default, we choose $K = d + 1$, so that estimated

covariance matrices can be of full rank. MGS produces more diverse synthetic observations than SMOTE as they are spread in all directions around the central point (when the covariance matrix is full rank).

4.3. Protocol and results on initial data sets

Protocol We compare the different rebalancing strategies on the 13 initial data sets (described below). We employ a 5-fold stratified cross-validation, and apply each rebalancing strategy on four training folds, in order to obtain the same number of minority/majority samples. Then, we train a Random Forest classifier (showing good predictive performance on tabular data sets, see Grinsztajn et al., 2022) on the same folds, and evaluate its performance on the remaining fold, via the PR AUC. Results are averaged over the five test folds and over 20 repetitions of the cross-validation. We use the `RandomForestClassifier` module in *scikit-learn* (Pedregosa et al., 2011) and tune the tree depth (when desired) via nested cross-validation Cawley and Talbot (2010). We use the implementation of *imb-learn* (Lemaître et al., 2017) for state-of-the-art rebalancing strategies (see Appendix A for details). The procedures implementation, and code for experiments reproducibility are available at the following GitHub repository: `https://github.com/artefactory/mgs-grf`.

Metrics Saito and Rehmsmeier (2015) show that the ROC AUC might be biased for evaluating classifiers on imbalanced data sets, so that we choose to use the PR AUC. While the balanced accuracy (Brodersen et al., 2010) is a widespread alternative to the usual accuracy in the imbalance context, we prefer the PR AUC which is independent of the threshold on classifier output probabilities. In addition, using the balanced accuracy is equivalent to applying class weights to the test set, as suggested by Lemma 4.1 below, which is not recommended in machine learning as the test set must reflect the true data distribution. Thus, we do not advocate for using the balanced accuracy as a final metric.

Algorithm 2 Single MGS Sample Generation.

Require: Minority class samples $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$, number of neighbors K
Ensure: Synthetic sample Z
1: Select uniformly X_c among $X_1, \ldots, X_n$.
2: Compute the set I of the K nearest neighbors of X_c (Euclidean distance)
3: $\hat{\mu} \leftarrow \frac{1}{K+1} \sum_{x \in I} x$
4: $\hat{\Sigma} \leftarrow \frac{1}{K+1} \sum_{x \in I} (x - \hat{\mu})^T (x - \hat{\mu})$
5: Generate synthetic sample $Z \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$
6: **return** Z

Lemma 4.1. *Let $\hat{f} : \mathcal{X} \to \{0, 1\}$ be a trained classifier. Let $\mathcal{D}_{test} = \{(X_i, Y_i)\}_{i=1}^N$ be a test set for $\hat{f}$, with $n \in \mathbb{N}$ and $\sum_{i=1}^N Y_i = n$. The balanced accuracy of $\hat{f}$, denoted by $Acc_{Balanced}(\hat{f})$, is defined as the average of the recall and the specificity of $\hat{f}$ on $\mathcal{D}_{test}$. Let us consider the weighted accuracy of $\hat{f}$, denoted by $Acc_{Weighted}(\hat{f})$, which is nothing but the standard accuracy where samples in $\mathcal{D}_{test}$ are weighted by $\frac{N}{2(N-n)}$ if they belong to the class $Y = 0$ and by $\frac{N}{2n}$ if they belong to the class $Y = 1$. Then, we have:*

$$Acc_{Balanced}(\hat{f}) = Acc_{Weighted}(\hat{f}). \tag{7}$$

The proof of Lemma 4.1 is available in Section D.7.

Initial data sets We employ 13 tabular data sets from Grinsztajn et al. (2022), UCI Irvine (Dua and Graff, 2017) and other public data sources (Phoneme (Alinat, 1993) and CreditCard (Dal Pozzolo et al., 2015)). All data sets are described in Table 3 in Section A and we call them *initial data sets*. As we want to compare several rebalancing methods including SMOTE, originally designed to handle continuous variables only, we have included data sets with continuous input only.

First results: None is competitive for slightly imbalanced data sets For 11 initial data sets out of 13, applying no strategy (named None in all tables) is the best, probably highlighting that the imbalance ratio is not low enough or the learning task not difficult enough to require a tailored rebalancing strategy. Therefore, considering only continuous input variables, and measuring the

predictive performance with PR AUC, we observe that dedicated rebalancing strategies are not required for most data sets. While applying no rebalancing strategy was already perceived as competitive in the literature (see, e.g., Han et al., 2005; He et al., 2008; Elor and Averbuch-Elor, 2022; Carriero et al., 2025), we believe that our analysis advocates for its broad use in practice, at least as a default method. Note that for these 11 data sets, qualified as slightly imbalanced, applying no rebalancing strategy is on par with the CW strategy, one of the most common rebalancing strategies (regardless of tree depth tuning, see Table 2 and Table 7).

5. EXTREMELY IMBALANCED DATA SETS

Strengthening the imbalance To analyse what could happen for data sets with lower imbalance ratio, we subsample the minority class for each one of the initial data sets mentioned above, so that the resulting imbalance ratio is set to 20%, 10% or 1% (when possible, taking into account dimension d). By doing so, we reproduce the extreme imbalance that is often encountered in practice (see He and Garcia, 2009). We apply our subsampling strategy once for each data set and each imbalance ratio in a nested fashion, so that the minority samples of the 1% data set are included in the minority samples of the 10% data set. The 18 new data sets thus obtained are called *subsampled data sets* and presented in Table 4 in Section F.

5.1. Results

For the sake of brevity, we display in Table 1 the initial and subsampled (with the new final imbalance ratio) data sets for which the None strategy is not the best, up to its standard deviation. The results in terms of PR AUC are displayed in Table 2 (see Table 5 in Appendix F for the other data sets). Thus, we focus our discussion on the performances of rebalancing methods in Table 1. We remark that the included data sets, referred as *extremely imbalanced data sets*, correspond to the most imbalanced subsampling, or simply the initial data set in case of extreme initial imbalance.

Whilst in the vast majority of experiments, applying no rebalancing is among the best approaches to deal with imbalanced data (see Table 6), it seems to be outperformed by dedicated rebalancing strategies for extremely imbalanced data sets (Table 2). Surprisingly, most rebalancing strategies do not benefit drastically from tree depth tuning, with the exception of None and CW strategies (see the differences between Table 2 and Table 7).

SMOTE and SMOTE K-tuned Default SMOTE (with $K = 5$) has a tendency to duplicate original observations, as shown by Theorem 3.4. This behavior is illustrated through our experiments when the tree depth is fixed. In this context, SMOTE ($K = 5$) has the same behavior as ROS, a method that copies original samples (see Table 9). When the tree depth is tuned, SMOTE may exhibit better performances compared to reweighting methods (ROS, RUS, CW), probably due to a higher tree depth. Indeed, even if synthetic data are close to the original samples, they are distinct and thus allow for more splits in the tree structure. However, SMOTE K-tuned performances are not systemically higher than those of default SMOTE (see Table 2). This may come from the fact that SMOTE K-tuned does not include a way of limiting the boundary artifacts phenomenon described in Theorem 3.6.

MGS Our second new available strategy exhibits good predictive performances (best performance in 4 out of 6 data sets in Table 2 and

Table 1 | Extremely imbalanced data sets, among the 13 initial data sets, for which None strategy is not the best (up to its standard deviation). N_{new} is the total number of samples after strengthening the imbalance.

	d	N	n/N	n/N_{new}
CreditCard	29	284 315	0.2%	0.2%
Abalone	8	4 177	1%	1%
Phoneme	5	5 404	29%	1%
Haberman	3	306	26%	10%
MagicTel	10	13 376	50%	20%
California	8	20 634	50%	1%

best average). This could be explained by the multivariate Gaussian sampling of synthetic observations that allows generating data points outside the convex hull of the minority class, thus limiting the border phenomenon, established in Theorem 3.6. Furthermore, MGS is associated to better performances for 5 data sets out of 6 compare to BS1 and BS2. Thus, the boundary artifact phenomenon of SMOTE is better circumvented by MGS. Note that with MGS, there is no need of tuning the tree depth: predictive performances of default RF are on par with tuned RF. All in all MGS is a promising new strategy compared to deep generative models and diffusion.

5.2. Supplementary results

CTGAN and ForestDiffusion We note in Table 2 that CTGAN performances are significantly lower than those of None strategy on 3 data sets and ForestDiffusion is among the best strategies for most data sets. Then, note that both CTGAN and ForestDiffusion computation time are among the longest. Furthermore, we remark that SMOTE and all the others SMOTE-related performances are not outperformed by GAN or diffusion based strategies in Table 2. This observation is corroborated by recent works showing that SMOTE-related methodologies remain competitive even compared to recent diffusion models (see Table 4 and Table 5 in Kotelnikov et al., 2023).

LightGBM and Logistic Regression When replacing random forests with LightGBM (a second-order boosting algorithm, see Ke et al., 2017) or logistic regression in the above protocol, we still do not observe strong benefits of using a rebalancing strategy for most data sets. Thus we keep only extremely imbalanced data sets for those classifiers in Table 10 and Table 11. As shown in Table 10, LightGBM reduces the performance gap between the different rebalancing methodologies, while MGS remains competitive. Additionally, Table 11 shows that the None strategy can still be competitive, even on some extremely imbalanced data sets, when used in conjunction with logistic regression. We compared in Table 12 the LDAM, Focal and L2RW losses intended for long-tailed learning, using PyTorch. Table 12 shows that LDAM loss performances are on par with the None strategy ones, while the performances of Focal are significantly higher for 3 data sets. We also observe that L2RW introduces the best PR AUC improvement for Abalone data set but does not seem to be competitive on the majority of the data sets. Methods from long-tailed learning do not seem promising for binary classification on tabular data, for which they were not initially intended. In Table 10, we note that our introduced strategy MGS, display again good predictive results.

Multiclass We extend our strategies to the multiclass label with 3 data sets that were originally multiclass : Wine, Yeast and Ecoli (described in Section A). For each data sets, the minority classes are augmented until the equilibrium is reached with the majority class. We applied our experimental protocol with LightGBM as classifier and show the results in Table 13. First we remark that for Ecoli data set, None is among the best strategy, strengthening our statement that rebalancing strategies are not always necessary. Borderline SMOTE 1 induces the best predictive performances for Wine while we note that MGS is the best strategy for the last data set.

6. CONCLUSION AND PERSPECTIVES

In this paper, we analysed the impact of rebalancing strategies on predictive performance for binary classification tasks on tabular data. First, we prove that default SMOTE tends to copy the original minority samples asymptotically, and exhibits boundary artifacts. From a computational perspective, we show that applying no rebalancing is competitive for most datasets, when used in conjunction with a tuned random forest/Logistic regression/LightGBM considering the PR AUC. For extremely imbalanced data sets, rebalancing strategies lead to improved predictive performances, with or without tuning the maximum tree depth. Based on our theoretical findings, we propose two slight modifications of SMOTE. Tuning the hyperparameter K of SMOTE appears to be necessary in theory (Theorem 3.4) and in simulations (Section 3.3), but does not improve the

Table 2 | Random Forests PR AUC (tree depth tuned on PR AUC) on extremely imbalanced data sets. Only data sets whose PR AUC of at least one rebalancing strategy is larger than that of None strategy plus its standard deviation are displayed. Undersampled data sets are in italics. Std are displayed in Table 5. Rebalancing and training times for *Phoneme* (1%) data set of one protocol iteration are displayed.

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS $(d+1)$	CT GAN	Forest Diff
CreditCard (0.2%)	0.776	0.783	0.751	**0.787**	0.586	0.786	0.784	0.773	0.771	0.747	0.757	0.752
Abalone (1%)	0.048	0.055	0.052	0.052	0.024	0.044	0.039	0.046	0.051	**0.056**	0.050	**0.056**
Phoneme (1%)	0.198	0.211	0.086	0.200	0.054	0.222	0.211	0.224	0.240	**0.269**	0.130	0.249
Haberman (10%)	0.252	0.268	0.289	0.295	0.268	0.308	0.294	0.306	0.307	**0.311**	0.280	0.306
MagicTel (20%)	0.754	0.759	0.741	**0.760**	0.338	0.739	0.742	0.757	0.757	0.750	0.693	0.756
California (1%)	0.298	0.283	0.207	0.277	0.019	0.234	0.226	0.242	0.258	**0.322**	0.152	0.294
Average	0.388	0.393	0.354	0.395	0.214	0.389	0.382	0.391	0.397	**0.409**	0.344	0.402
Time (s)	6.0	5.1	1.7	5.9	1.7	5.5	5.5	6.1	120.2	7.2	32.8	306.3

predictive performance on the real-world data sets we considered. MGS, the other variant we propose, appears promising, with good predictive performances regardless of the tuning of random forests. Thus, a simple modification based on theoretical findings can improve SMOTE performances.

SMOTE was originally designed for continuous features, since it is based on interpolation, which is irrelevant for qualitative variables. That is our primary reason for studying SMOTE for continuous features only. Extensions of SMOTE to generate categorical features have been proposed, based on univariate nearest neighbors vote. However, nearest neighbors struggle in high dimensions. We plan to extend MGS to categorical features, replacing nearest neighbor vote by multi-output decision tree predictions, potentially resulting in more plausible generated samples (see Sakho et al., 2025, [2]).

Acknowledgements

The authors would like to express their gratitude to anonymous reviewers for their constructive feedback.

[2]A version of this future work was published before the present article.

References

Anass Aghbalou, Anne Sabourin, and François Portier. Sharp error bounds for imbalanced classification: how many examples in the minority class? In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR, 2024.

Gabriel Aguiar, Bartosz Krawczyk, and Alberto Cano. A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine learning*, pages 1–79, 2023.

Pierre Alinat. Periodic progress report 4, roars project esprit ii-number 5516. *Technical Thomson Report TS ASM 93/S/EGS/NC*, 79, 1993.

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

Sunil Arya, David M Mount, and Onuttom Narayan. Accounting for boundary effects in nearest neighbor searching. In *Proceedings of the eleventh annual symposium on computational geometry*, pages 336–344, 1995.

Thomas Benjamin Berrett. Modern k-nearest neighbour methods in entropy estimation, independence testing and classification. *Apollo - University of Cambridge Repository*, 2017. doi: 10.17863/CAM.13756. URL https://www.repository.cam.ac.uk/handle/1810/267832.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Dbsmote: density-based synthetic minority oversampling technique. *Applied Intelligence*, 36: 664–684, 2012.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Alex Carriero, Kim Luijken, Anne de Hond, Karel GM Moons, Ben van Calster, and Maarten van Smeden. The harms of class imbalance corrections for machine learning based prediction models: a simulation study. *Statistics in Medicine*, 44(3-4):e10320, 2025.

Gavin C Cawley and Nicola LC Talbot. On overfitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Nilotpal Chakravarti. Isotonic median regression: a linear programming approach. *Mathematics of operations research*, 14(2):303–308, 1989.

Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? In *International Conference on Machine Learning*, pages 4144–4188. PMLR, 2023.

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.

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7*, pages 107–119. Springer, 2003.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced

classification. In *2015 IEEE symposium series on computational intelligence*, pages 159–166. IEEE, 2015.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Yotam Elor and Hadar Averbuch-Elor. To smote, or not to smote? *arXiv preprint arXiv:2201.08528*, 2022.

Dina Elreedy and Amir F Atiya. A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, 505:32–64, 2019.

Dina Elreedy, Amir F. Atiya, and Firuz Kamalov. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, January 2023. ISSN 1573-0565. doi: 10.1007/s10994-022-06296-4. URL https://doi.org/10.1007/s10994-022-06296-4.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

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*, pages 878–887. Springer, 2005.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Amira Kamil Ibrahim Hassan and Ajith Abraham. Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pages 117–127. Springer, 2016.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

Jyoti Islam and Yanqing Zhang. Gan-based synthetic brain pet image generation. *Brain informatics*, 7:1–12, 2020.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International Conference on Artificial Intelligence and Statistics*, pages 1288–1296. PMLR, 2024.

M Chris Jones. Simple boundary correction for kernel density estimation. *Statistics and computing*, 3:135–146, 1993.

Firuz Kamalov, Amir F Atiya, and Dina Elreedy. Partial resampling of imbalanced data. *arXiv preprint arXiv:2207.04631*, 2022.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11:1–13, 2011.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL http://jmlr.org/papers/v18/16-365.html.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML, 2003.

Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28:92–122, 2014.

Ahmed Jameel Mohammed, Masoud Muhammed Hassan, and Dler Hussein Kadir. Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(3):3161–3172, 2020.

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for mondrian trees and forests. *Annals of Statistics*, 48(4): 2253–2276, 2020.

Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, 2011.

Nam N Nguyen and Anh T Duong. Comparison of two main approaches for handling imbalanced data in churn prediction problem. *Journal of advances in information technology*, 12(1), 2021.

Tingting Pan, Junhong Zhao, Wei Wu, and Jie Yang. Learning imbalanced datasets based on smote and gaussian distribution. *Information Sciences*, 512:1214–1233, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Duraisamy Ramyachitra and Parasuraman Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5 (4):1–29, 2014.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3): e0118432, 2015.

Abdoulaye Sakho, Emmanuel Malherbe, Carl-Erik Gauthier, and Erwan Scornet. Harnessing mixed features for imbalance data oversampling: Application to bank customers scoring. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 247–264. Springer, 2025.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

George Proctor Wadsworth, Joseph G Bryan, and A Cemal Eringen. Introduction to probability and random variables. *Journal of Applied Mechanics*, 28(2):319, 1961.

Byron C Wallace and Issa J Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and information systems*, 41(1):33–52, 2014.

Tarid Wongvorachan, Surina He, and Okan Bulut. A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information*, 14(1):54, 2023.

Yuxi Xie, Min Qiu, Haibo Zhang, Lizhi Peng, and Zhenxiang Chen. Gaussian distribution based oversampling for imbalanced data classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):667–679, 2020.

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

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, pages 10544–10554. PMLR, 2020.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Min Zhu, Jing Xia, Xiaoqing Jin, Molei Yan, Guolong Cai, Jing Yan, and Gangmin Ning. Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access*, 6: 4641–4652, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes, we present computation times.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes, we provide a Github repository.**
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes.**
   (b) Complete proofs of all theoretical results. **Yes.**
   (c) Clear explanations of any assumptions. **Yes.**
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes.**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes.**
   (b) The license information of the assets, if applicable. **Yes.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes.**
   (d) Information about consent from data providers/curators. **Yes.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

Part

Supplementary Materials to “Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants"

Table of Contents

A PREDICTIVE EVALUATION REAL-WORLD DATA SETS: REMARKS **18**

B NUMERICAL ILLUSTRATIONS OF THEOREMS: REMARKS **20**

B.1 Simulated data 20

B.2 Real-world data sets 21

C ADDITIONAL THEORETICAL RESULTS **22**

C.1 Lemma C.1 22

D MAIN PROOFS **22**

D.1 Proof of Lemma 3.1 22

D.2 Proof of Theorem 3.2 22

D.3 Proof of Lemma C.1 24

D.4 Proof of Theorem 3.4 27

D.5 Proof of Corollary 3.5 29

D.6 Proof of Theorem 3.6 31

D.7 Proof of Lemma 4.1 35

E TECHNICAL LEMMAS **35**

E.1 Cumulative distribution function of a binomial distribution 35

E.2 Bounds for the incomplete beta function 36

E.3 Upper bound for binomial coefficient 37

E.4 Inequality $x \ln\left(\frac{1}{x}\right) \leq \sqrt{x}$ 37

F TABLES **38**

A. PREDICTIVE EVALUATION REAL-WORLD DATA SETS: REMARKS

Table 3 | Initial data sets.

	N	n/N	d
Haberman	306	26%	3
Ionosphere	351	36%	32
Breast cancer	630	36%	9
Pima	768	35%	8
Vehicle	846	23%	18
Yeast	1 462	11%	8
Abalone	4 177	1%	8
Wine	4 974	4%	11
Phoneme	5 404	29%	5
MagicTel	13 376	50%	10
California	20 634	50%	8
House_16H	22 784	30%	16
CreditCard	284 315	0.2%	29

General comment on the protocol for real-world data sets The state-of-the-art rebalancing strategies(see Lemaître et al., 2017) are used with their default hyperparameter values.

The subsampled data sets (see Table 4) can be obtained through the repository (the functions and the seeds are given in a jupyter notebook). For the CreditCard data set, a Time Series split is performed instead of a Stratified 5−fold, because of the temporality of the data. Furthermore, a group out is applied on the different scope time value.

For MagicTel and California data sets, the initial data sets are already balanced, leading to no opportunity for applying a rebalancing strategy. This is the reason why we do not include these original data sets in our study but only their subsampled associated data sets.

The max_depth hyperparameter is tuned using GridSearch function from scikit-learn. The grids minimum is 5 and the grid maximum is the mean depth of the given strategy for the given data set (when random forest is used without tuning depth hyperparameter). Then, using numpy (Harris et al., 2020), a list of integer of size 8 between the minimum value and the maximum is value is built. Finally, the "None" value is added to this list.

Mean standard deviation For each protocol run, we computed the mean of the PR AUC over the 5-fold. Then, all of these 20 means are averaged in order to get what we call in some of our tables the mean standard deviations.

Multiclass data sets The Ecoli data set is composed of 327 samples with 7 features. The target is formed of 5 categories. The multiclass Wine data set is composed of 11 features with 6492 samples. The target is formed of 6 categories. The multiclass Yeast data set is composed of 8 features with 1394 samples. The target is formed of 6 categories.

More results about None strategy from seminal papers It appears in the results of two seminal papers that the None strategy is competitive in terms of predictive performances when applied with

Table 4 | Subsampled data sets.

	N	n/N	d
Haberman (10%)	250	10%	3
Ionosphere (20%)	281	20%	32
Ionosphere (10%)	250	10%	32
Breast cancer (20%)	500	20%	9
Breast cancer (10%)	444	10%	9
Pima (20%)	625	20%	8
Vehicle (10%)	718	10%	18
Yeast (1%)	1 334	1%	8
Phoneme (20%)	4 772	20%	5
Phoneme (10%)	4 242	10%	5
Phoneme (1%)	3 856	1%	5
MagicTel (20%)	8 360	20%	10
House_16H (20%)	20 050	20%	16
House_16H (10%)	17 822	10%	16
House_16H (1%)	16 202	1%	16
California (20%)	12 896	20%	8
California (10%)	11 463	10%	8
California (1%)	10 421	1%	8

weak classifiers, a contrario of our work. Furthermore, only He et al. (2008) highlight this phenomenon in section III.C. He et al. (2008) compare the None strategy, ADASYN and SMOTE, followed by a decision tree classifier on 5 data sets (including Vehicle, Pima, Ionosphere and Abalone). In terms of Precision and F1 score, the None strategy is on par with the two other rebalancing methods. Han et al. (2005) study the impact of Borderline SMOTE and others SMOTE variant on 4 data sets (including Pima and Haberman) when using C4.5 classifier. The None strategy is competitive (in terms of F1 score) on two of these data sets.

Elor and Averbuch-Elor (2022) performs an empirical comparison of oversampling strategies combined with various classifiers. The purpose is to find in which setting SMOTE, and others oversampling strategies, can improve predictive performances. While informative, it suffers from a key limitation: categorical variables are ordinally encoded, making linear interpolation via SMOTE producing not "plausible" samples. Moreover, their evaluation relies on F1 score, tuned only for the "no rebalancing" case (in Table 3 from Elor and Averbuch-Elor (2022)), and ROC AUC, which is not relevant in imbalanced settings as stated above and by Saito and Rehmsmeier (2015). The F1-score is tuned for all classifiers in Figure 2 from Elor and Averbuch-Elor (2022) but we observe that oversampling improves LightGBM and XGBoost performances. Finally, the authors of Elor and Averbuch-Elor (2022) claim that applying no rebalancing strategy performs well but their analysis excludes both undersampling and deep-learning-based oversampling methods (such as CTGAN or ForestDiffusion Xu et al., 2019; Jolicoeur-Martineau et al., 2024).

Carriero et al. (2025) investigates classifier calibration post-rebalancing on both synthetic and MIMIC-III data sets. Carriero et al. (2025) state that when applying no rebalancing strategy, the quality of the calibration is better. While useful, its conclusions are based on a single real data set and focus solely on calibration quality, not on predictive performance. A model can be poorly calibrated yet perform well on classification tasks with respect to a classification metric. Well-calibrated models are required for use cases where output probabilities are combined with other sources of

information for decision making. Furthermore, post-processing calibration methods (such as isotonic regression Chakravarti, 1989) exist and were not considered in their study to circumvent drawbacks of rebalancing strategies.

Hardware For all the numerical experiments, we use the following processor : AMD Ryzen Threadripper PRO 5955WX: 16 cores, 4.0 GHz, 64 MB cache, PCIe 4.0. We also add access to 250GB of RAM.

Computation time The computation times reported in Table 2 were obtained using the Phoneme data set.

B. NUMERICAL ILLUSTRATIONS OF THEOREMS: REMARKS

Through Section 3, we highlighted that SMOTE asymptotically regenerates the distribution of the minority class, by tending to copy the minority samples. The purpose of this section is to numerically illustrate the theoretical limitations of SMOTE, typically with the default value $K = 5$.

B.1. Simulated data

The quantity $\bar{C}(Z, X)$ measures how far SMOTE observations Z_i are from the original observations X_i. The quantity $\bar{C}(\tilde{X}, X)$ measures how far data $\tilde{X}_i$, distributed as the original observations X_i, are from these observations (X_i). The metric $\bar{C}(Z, X)/\bar{C}(\tilde{X}, X)$ quantify the ability of the synthetic samples to fill the space in the same way as samples generated with the true distribution. When $\bar{C}(Z, X)/\bar{C}(\tilde{X}, X)$ is close to one, SMOTE observations are comparable to observations $\tilde{X}$, in terms of the metric $\bar{C}$. In this case, data generated via SMOTE behave in the same way as the original data, with respect to the quantity $\bar{C}$. Similarly, when the quantity $\bar{C}(Z, X)/\bar{C}(\tilde{X}, X)$ is close to zero, observations generated via SMOTE are much closer to their central point than observations generated with the true distribution. In this case, this highlights that SMOTE has a tendency to generate data close to the original observations. Conversely, when $\bar{C}(Z, X)/\bar{C}(\tilde{X}, X)$ is larger than one, data generated via SMOTE are further away from their central point than observations generated with the true distribution. In this case, this highlight that diversity of newly created samples is too high.

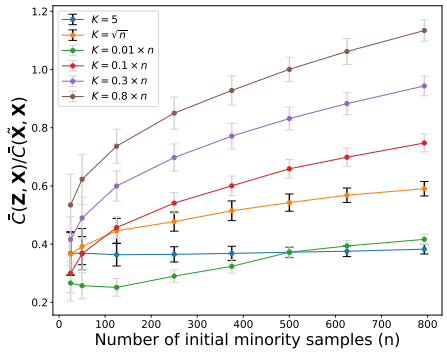


Figure 2 | $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$with Phoneme data.

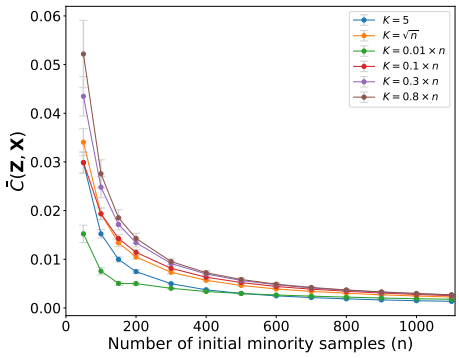


Figure 3 | Average distance $\bar{C}(\mathbf{Z}, \mathbf{X})$.

Results. Figure 1 shows the renormalized quantity $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$ as a function of n. We notice that the asymptotic for $K = 5$ is different since it is the only one where the distance between SMOTE data points and original data points does not vary with n. Besides, this distance is smaller than the other ones, thus stressing out that the SMOTE data points are very close to the original distribution for $K = 5$. Note that, for the other asymptotics in K, the diversity of SMOTE observations increases with n, meaning $\bar{C}(\mathbf{Z}, \mathbf{X})$ gets closer from $\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$. This behavior in terms of average distance is ideal, since $\tilde{\mathbf{X}}$ is drawn from the same theoretical distribution as $\mathbf{X}$. On the contrary, $K = 5$ keeps a lower average distance, showing a lack of diversity of generated points. Besides, this diversity is asymptotically more important for $K = 0.1n$ and $K = 0.01n$. This corroborates our theoretical findings (Theorem 3.2) as these asymptotics do not satisfy $K/n \to 0$. Indeed, when K is set to a fraction of n, the SMOTE distribution does not converge to the original distribution anymore, therefore generating data points that are not simple copies of the original uniform samples. By construction, SMOTE data points are close to central points, which may explain why the quantity of interest in Figure 1 is smaller than 1.

B.2. Real-world data sets

Extension to real-world data sets We extended our protocol to a real-world data set by splitting the data into two sets of equal size $\mathbf{X}$ and $\tilde{\mathbf{X}}$. The first one is used for applying SMOTE strategies to sample $\mathbf{Z}$ and the other set is used to compute the normalization factor $\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$.

Results We apply the adapted protocol to Phoneme data set, described in Table 3. Figure 2 displays the quantity $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$ as a function of the size n of the minority class. As above, we observe in Figure 2 that the average normalized distance $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$ increases for all strategies but the one with $K = 5$. The strategies using a value of hyperparameter K such that $K/n \to 0$ seem to converge to a value smaller than all the strategies with K such that $K/n \not\to 0$.

Results with $\bar{C}(\mathbf{Z}, \mathbf{X})$ Figure 3 depicts the quantity $\bar{C}(\mathbf{Z}, \mathbf{X})$ as a function of the size of the minority class, for different values of K. The metric $\bar{C}(\mathbf{Z}, \mathbf{X})$ is consistently smaller for $K = 5$ than for other values of K, therefore highlighting that data generated by SMOTE with $K = 5$ are closer to the original data sample. This phenomenon is strengthened as n increases. This is an artifact of the simulation setting as the original data samples fill the input space as n increases.

More details on the numerical illustrations protocol applied to real-world data sets We apply SMOTE on real-world data and compare the distribution of the generated data points to the original distribution, using the metric $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$.

For each value of n, we subsample n data points from the minority class. Then,

1. We uniformly split the data set into $X_1, \ldots, X_{n/2}$ (denoted by $\mathbf{X}$) and $\tilde{X}_1, \ldots, \tilde{X}_{n/2}$ (denoted by $\tilde{\mathbf{X}}$).
2. We generate a data set $\mathbf{Z}$ composed of $m = n/2$ i.i.d new observations $Z_1, \ldots, Z_m$ by applying SMOTE procedure on the original data set $\mathbf{X}$, with different values of K. We compute $C(\mathbf{Z}, \mathbf{X})$.
3. We use $\tilde{\mathbf{X}}$ in order to compute $C(\tilde{\mathbf{X}}, \mathbf{X})$.

This procedure is repeated $B = 100$ times to compute averages values as in Section 3.3.

C. ADDITIONAL THEORETICAL RESULTS

C.1. Lemma C.1

Lemma C.1. *Let X_c be the central point chosen in a SMOTE iteration. Then, for all $x_c \in \mathcal{X}$, the random variable $Z_{K,n}$ generated by SMOTE has a conditional density $f_{Z_{K,n}}(.|X_c = x_c)$ which satisfies*

$$f_{Z_{K,n}}(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \mathcal{B}\left(n-K-1, K; 1-\beta_{x_c,z,w}\right) \mathrm{d}w, \tag{8}$$

where $\beta_{x_c,z,w} = \mu_X\left(B\left(x_c, ||z-x_c||/w\right)\right)$ and μ_X is the probability measure associated to f_X. Using the following substitution $w = \|z-x_c\|/r$, we have,

$$f_{Z_{K,n}}(z|X_c = x_c) = (n-K-1) \quad \times \binom{n-1}{K}\int_{r=\|z-x_c\|}^{\infty} f_X\left(x_c + \frac{(z-x_c)r}{\|z-x_c\|}\right) \times \frac{r^{d-2}\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, r\right)\right)\right)}{\|z-x_c\|^{d-1}}\mathrm{d}r. \tag{9}$$

The proof of Lemma C.1 is available in Section D.3.

D. MAIN PROOFS

This section contains the main proof of our theoretical results. The technicals lemmas used by several proofs are available on Section E.

D.1. Proof of Lemma 3.1

Proof of Lemma 3.1. Let $\mathcal{X}$ be the support of P_X. SMOTE generates new points by linear interpolation of the original minority sample. This means that for all x, y in the minority samples or generated by SMOTE procedure, we have $(1-t)x+ty \in Conv(\mathcal{X})$ by definition of $Conv(\mathcal{X})$. This leads to the fact that precisely, all the new SMOTE samples are contained in $Conv(\mathcal{X})$. This implies $Supp(P_Z) \subseteq Conv(\mathcal{X})$.

□

D.2. Proof of Theorem 3.2

Proof of Theorem 3.2. For any event A, B, we have

$$1 - \mathbb{P}[A \cap B] = \mathbb{P}[A^c \cup B^c] \leq \mathbb{P}[A^c] + \mathbb{P}[B^c], \tag{10}$$

which leads to

$$\mathbb{P}[A \cap B] \geq 1 - \mathbb{P}[A^c] - \mathbb{P}[B^c] \tag{11}$$

$$= \mathbb{P}[A] - \mathbb{P}[B^c]. \tag{12}$$

By construction,

$$\|X_c - Z\| \leq \|X_c - X_{(K)}(X_c)\|. \tag{13}$$

Let $x \in \mathcal{X}$ and $\eta > 0$. Let $\alpha, \varepsilon > 0$. We have,

$$\begin{aligned}
&\mathbb{P}[X_c \in B(x, \alpha - \varepsilon)] - \mathbb{P}[\|X_c - X_{(K)}(X_c)\| > \varepsilon] && (14)\\
&\leq \mathbb{P}[X_c \in B(x, \alpha - \varepsilon), \|X_c - X_{(K)}(X_c)\| \leq \varepsilon] && (15)\\
&\leq \mathbb{P}[X_c \in B(x, \alpha - \varepsilon), \|X_c - Z\| \leq \varepsilon] && (16)\\
&\leq \mathbb{P}[Z \in B(x, \alpha)]. && (17)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{P}[Z \in B(x, \alpha)] - \mathbb{P}[\|X_c - X_{(K)}(X_c)\| > \varepsilon] && (18)\\
&\leq \mathbb{P}[Z \in B(x, \alpha), \|X_c - X_{(K)}(X_c)\| \leq \varepsilon] && (19)\\
&\leq \mathbb{P}[Z \in B(x, \alpha), \|X_c - Z\| \leq \varepsilon] && (20)\\
&\leq \mathbb{P}[X_c \in B(x, \alpha + \varepsilon)]. && (21)
\end{aligned}$$

Since X_c admits a density, for all $\varepsilon > 0$ small enough

$$\mathbb{P}[X_c \in B(x, \alpha + \varepsilon)] \leq \mathbb{P}[X_c \in B(x, \alpha)] + \eta, \tag{22}$$

and

$$\mathbb{P}[X_c \in B(x, \alpha)] - \eta \leq \mathbb{P}[X_c \in B(x, \alpha - \varepsilon)]. \tag{23}$$

Let ε such that (22) and (23) are verified. According to Lemma 2.3 in Biau and Devroye (2015), since $X_1, \ldots, X_n$ are i.i.d., if K/n tends to zero as $n \to \infty$, we have

$$\mathbb{P}[\|X_c - X_{(K)}(X_c)\| > \varepsilon] \to 0. \tag{24}$$

Thus, for all n large enough,

$$\mathbb{P}[X_c \in B(x, \alpha)] - 2\eta \leq \mathbb{P}[Z \in B(x, \alpha)] \tag{25}$$

and

$$\mathbb{P}[Z \in B(x, \alpha)] \leq 2\eta + \mathbb{P}[X_c \in B(x, \alpha)]. \tag{26}$$

Finally, for all $\eta > 0$, for all n large enough, we obtain

$$\mathbb{P}[X_c \in B(x, \alpha)] - 2\eta \leq \mathbb{P}[Z \in B(x, \alpha)] \leq 2\eta + \mathbb{P}[X_c \in B(x, \alpha)], \tag{27}$$

which proves that

$$\mathbb{P}[Z \in B(x, \alpha)] \to \mathbb{P}[X_c \in B(x, \alpha)]. \tag{28}$$

Therefore, by the Monotone convergence theorem, for all Borel sets $B \subset \mathbb{R}^d$,

$$\mathbb{P}[Z \in B] \to \mathbb{P}[X_c \in B]. \tag{29}$$

□

D.3. Proof of Lemma C.1

Proof of Lemma C.1. We consider a single SMOTE iteration. Recall that the central point X_c (see Algorithm 1) is fixed, and thus denoted by x_c.

The random variables $X_{(1)}(x_c), \dots, X_{(n-1)}(x_c)$ denote a reordering of the initial observations $X_1, X_2, \dots, X_n$ such that

$$||X_{(1)}(x_c) - x_c|| \leq ||X_{(2)}(x_c) - x_c|| \leq \dots \leq ||X_{(n-1)}(x_c) - x_c||.$$

For clarity, we remove the explicit dependence on x_c. Recall that SMOTE builds a linear interpolation between x_c and one of its K nearest neighbors chosen uniformly. Then the newly generated point Z satisfies

$$Z = (1 - W)x_c + W \sum_{k=1}^{K} X_{(k)} \mathbb{1}_{\{I=k\}}, \tag{30}$$

where W is a uniform random variable over $[0, 1]$, independent of $I, X_1, \dots, X_n$, with I distributed as $\mathcal{U}(\{1, \dots, K\})$.

From now, consider that the k-th nearest neighbor of x_c, $X_{(k)}(x_c)$, has been chosen (that is $I = k$). Then Z satisfies

$$\begin{aligned} Z &= (1 - W)x_c + WX_{(k)} && (31)\\ &= x_c - Wx_c + WX_{(k)}, && (32) \end{aligned}$$

which implies

$$Z - x_c = W(X_{(k)} - x_c). \tag{33}$$

Let f_{Z-x_c}, f_W and $f_{X_{(k)}-x_c}$ be respectively the density functions of $Z - x_c$, W and $X_{(k)} - x_c$. Let $z, z_1, z_2 \in \mathbb{R}^d$. Recall that $z \leq z_1$ means that each component of z is lower than the corresponding component of z_1. Since W and $X_{(k)} - x_c$ are independent, we have,

$$\begin{aligned} \mathbb{P}(z_1 \leq Z - x_c \leq z_2) &= \int_{w\in\mathbb{R}} \int_{x\in\mathbb{R}^d} f_{W,X_{(k)}-x_c}(w, x) \mathbb{1}_{\{z_1 \leq wx \leq z_2\}} \mathrm{d}w\mathrm{d}x && (34)\\ &= \int_{w\in\mathbb{R}} \int_{x\in\mathbb{R}^d} f_W(w) f_{X_{(k)}-x_c}(x) \mathbb{1}_{\{z_1 \leq wx \leq z_2\}} \mathrm{d}w\mathrm{d}x && (35)\\ &= \int_{w\in\mathbb{R}} f_W(w) \left(\int_{x\in\mathbb{R}^d} f_{X_{(k)}-x_c}(x) \mathbb{1}_{\{z_1 \leq wx \leq z_2\}} \mathrm{d}x \right) \mathrm{d}w. && (36) \end{aligned}$$

Besides, let $u = wx$. Then $x = (\frac{u_1}{w}, \dots, \frac{u_d}{w})^T$.The Jacobian of such transformation equals:

$$\begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \dots & \frac{\partial x_1}{\partial u_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial u_1} & \dots & \frac{\partial x_d}{\partial u_d} \end{vmatrix} = \begin{vmatrix} \frac{1}{w} & & 0 \\ & \ddots & \\ 0 & \dots & \frac{1}{w} \end{vmatrix} = \frac{1}{w^d} \tag{37}$$

Therefore, we have $x = u/w$ and $\mathrm{d}x = \mathrm{d}u/w^d$, which leads to

$$\begin{aligned} &\mathbb{P}(z_1 \leq Z - x_c \leq z_2) && (38)\\ &= \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) \left(\int_{u\in\mathbb{R}^d} f_{X_{(k)}-x_c}\left(\frac{u}{w}\right) \mathbb{1}_{\{z_1 \leq u \leq z_2\}} \mathrm{d}u \right) \mathrm{d}w. && (39) \end{aligned}$$

Note that a random variable Z' with density function

$$f_{Z'}(z') = \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) f_{X_{(k)}-x_c}\left(\frac{z'}{w}\right) \mathrm{d}w \tag{40}$$

satisfies, for all $z_1, z_2 \in \mathbb{R}^d$,

$$\mathbb{P}(z_1 \leq Z - x_c \leq z_2) = \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) \left(\int_{u\in\mathbb{R}^d} f_{X_{(k)}-x_c}\left(\frac{u}{w}\right) \mathbb{1}_{\{z_1\leq u\leq z_2\}} \mathrm{d}u \right) \mathrm{d}w. \tag{41}$$

Therefore, the variable $Z - x_c$ admits the following density

$$f_{Z-x_c}(z'|X_c = x_c, I = k) = \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) f_{X_{(k)}-x_c}\left(\frac{z'}{w}\right) \mathrm{d}w. \tag{42}$$

Since W follows a uniform distribution on $[0, 1]$, we have

$$f_{Z-x_c}(z'|X_c = x_c, I = k) = \int_0^1 \frac{1}{w^d} f_{X_{(k)}-x_c}\left(\frac{z'}{w}\right) \mathrm{d}w. \tag{43}$$

The density $f_{X_{(k)}-x_c}$ of the k-th nearest neighbor of x_c can be computed exactly (see, Lemma 6.1 in Berrett, 2017), that is

$$\begin{aligned} f_{X_{(k)}-x_c}(u) = (n-1)\binom{n-2}{k-1} f_X(x_c+u) \left[\mu_X\left(B(x_c, ||u||)\right)\right]^{k-1} \\ \times \left[1 - \mu_X\left(B(x_c, ||u||)\right)\right]^{n-k-1}, \end{aligned} \tag{44}$$

where

$$\mu_X\left(B(x_c, ||u||)\right) = \int_{B(x_c,\|u\|)} f_X(x)\mathrm{d}x. \tag{45}$$

We recall that $B(x_c, \|u\|)$ is the ball centered on x_c and of radius $\|u\|$. Hence we have

$$f_{X_{(k)}-x_c}(u) = (n-1)\binom{n-2}{k-1} f_X(x_c+u) \mu_X\left(B(x_c, ||u||)\right)^{k-1} \left[1 - \mu_X\left(B(x_c, ||u||)\right)\right]^{n-k-1}. \tag{46}$$

Since $Z - x_c$ is a translation of the random variable Z, we have

$$f_Z(z|X_c = x_c, I = k) = f_{Z-x_c}(z - x_c|X_c = x_c, I = k). \tag{47}$$

Injecting Equation (46) in Equation (43), we obtain

$$f_Z(z|X_c = x_c, I = k) \tag{48}$$

$$= f_{Z-x_c}(z - x_c|X_c = x_c, I = k) \tag{49}$$

$$= \int_0^1 \frac{1}{w^d} f_{X_{(k)}-x_c}\left(\frac{z-x_c}{w}\right) \mathrm{d}w \tag{50}$$

$$= (n-1)\binom{n-2}{k-1} \int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k-1} \tag{51}$$

$$\times \left[1 - \mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-1} \mathrm{d}w \tag{52}$$

Recall that in SMOTE, k is chosen at random in $\{1, \dots, K\}$ through the uniform random variable I. So far, we have considered I fixed. Taking the expectation with respect to I, we have

$$
\begin{aligned}
& f_Z(z|X_c = x_c) && (53)\\
&= \sum_{k=1}^{K} f_Z(z|X_c = x_c, I = k)\mathbb{P}[I = k] && (54)\\
&= \frac{1}{K}\sum_{k=1}^{K}\int_0^1 \frac{1}{w^d} f_{X_{(k)}-x_c}\left(\frac{z-x_c}{w}\right)\mathrm{d}w && (55)\\
&= \frac{1}{K}\sum_{k=1}^{K}(n-1)\binom{n-2}{k-1}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k-1} && (56)\\
&\quad \times [1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)]^{n-k-1}\mathrm{d}w && (57)\\
&= \frac{(n-1)}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\sum_{k=1}^{K}\binom{n-2}{k-1}\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k-1} && (58)\\
&\quad \times [1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)]^{n-k-1}\mathrm{d}w && (59)\\
&= \frac{(n-1)}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\sum_{k=0}^{K-1}\binom{n-2}{k}\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k} && (60)\\
&\quad \times \left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-2}\mathrm{d}w. && (61)
\end{aligned}
$$

Note that the sum can be expressed as the cumulative distribution function of a Binomial distribution parameterized by $n-2$ and $\mu_X(B(x_c, \|z-x_c\|/w))$, so that

$$
\begin{aligned}
& \sum_{k=0}^{K-1}\binom{n-2}{k}\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k}\left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-2} && (62)\\
=&(n-K-1)\binom{n-2}{K-1}\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right), && (63)
\end{aligned}
$$

(see Technical Lemma E.1 for details). We inject Equation (63) in Equation (53)

$$
\begin{aligned}
f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\\
\times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w. \qquad (64)
\end{aligned}
$$

We know that

$$
f_Z(z) = \int_{x_c \in \mathcal{X}} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c.
$$

Combining this remark with the result of Equation (64) we get

$$
\begin{aligned}
f_Z(z) = (n-K-1)\binom{n-1}{K}\int_{x_c\in\mathcal{X}}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\\
\times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) f_X(x_c)\mathrm{d}w\mathrm{d}x_c. \qquad (65)
\end{aligned}
$$

Link with Elreedy's formula According to the Elreedy formula

$$f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_{r=\|z-x_c\|}^{\infty} f_X\left(x_c + \frac{(z-x_c)r}{\|z-x_c\|}\right)\frac{r^{d-2}}{\|z-x_c\|^{d-1}} \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, r\right)\right)\right) \mathrm{d}r. \tag{66}$$

Now, let $r = \|z - x_c\|/w$ so that $\mathrm{d}r = -\|z - x_c\|\mathrm{d}w/w^2$. Thus,

$$\begin{aligned} & f_Z(z|X_c = x_c) \\ & = (n-K-1)\binom{n-1}{K}\int_0^1 f_X\left(x_c + \frac{z-x_c}{w}\right)\frac{1}{w^{d-2}}\frac{1}{\|z-x_c\|} && (67) \\ & \quad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{z-x_c}{w}\right)\right)\right)\frac{\|z-x_c\|}{w^2}\mathrm{d}w && (68) \\ & = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \\ & \quad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{z-x_c}{w}\right)\right)\right)\mathrm{d}w. && (69) \end{aligned}$$

□

D.4. Proof of Theorem 3.4

Proof of Theorem 3.4. Let $x_c \in \mathcal{X}$ be a central point in a SMOTE iteration. From Lemma C.1, we have,

$$\begin{aligned} & f_Z(z|X_c = x_c) \\ & = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \\ & \quad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w && (70) \\ & = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathbb{1}_{\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\}} \\ & \quad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w. && (71) \end{aligned}$$

Let $R \in \mathbb{R}$ such that $\mathcal{X} \subset \mathcal{B}(0, R)$. For all $u = x_c + \frac{z-x_c}{w}$, we have

$$w = \frac{||z-x_c||}{||u-x_c||}. \tag{72}$$

If $u \in \mathcal{X}$, then $u \in \mathcal{B}(0, R)$. Besides, since $x_c \in \mathcal{X} \subset B(0, R)$, we have $||u - x_c|| < 2R$ and

$$w > \frac{||z-x_c||}{2R}. \tag{73}$$

Consequently,

$$\mathbb{1}_{\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\}} \leq \mathbb{1}_{\left\{w>\frac{||z-x_c||}{2R}\right\}}. \tag{74}$$

So finally

$$\mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}} = \mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}}\mathbb{1}_{\left\{w>\frac{||z-x_c||}{2R}\right\}}. \tag{75}$$

Hence,

$$\begin{aligned}
&f_Z(z|X_c=x_c)\\
&=(n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c+\frac{z-x_c}{w}\right)\mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}}\mathbb{1}_{\left\{w>\frac{||z-x_c||}{2R}\right\}} && (76)\\
&\quad\times\mathcal{B}\left(n-K-1,K;1-\mu_X\left(B\left(x_c,\frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w && (77)\\
&=(n-K-1)\binom{n-1}{K}\int_{\frac{||z-x_c||}{2R}}^1 \frac{1}{w^d} f_X\left(x_c+\frac{z-x_c}{w}\right)\\
&\quad\times\mathcal{B}\left(n-K-1,K;1-\mu_X\left(B\left(x_c,\frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w. && (78)
\end{aligned}$$

Now, let $0<\alpha\leq 2R$ and $z\in\mathbb{R}^d$ such that $||z-x_c||>\alpha$. In such a case, $w>\frac{\alpha}{2R}$ and:

$$\begin{aligned}
&f_Z(z|X_c=x_c) && (79)\\
&=(n-K-1)\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d} f_X\left(x_c+\frac{z-x_c}{w}\right)\\
&\quad\times\mathcal{B}\left(n-K-1,K;1-\mu_X\left(B\left(x_c,\frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w && (80)\\
&\leq(n-K-1)\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d} f_X\left(x_c+\frac{z-x_c}{w}\right)\\
&\quad\times\mathcal{B}\left(n-K-1,K;1-\mu_X\left(B\left(x_c,\alpha\right)\right)\right)\mathrm{d}w. && (81)
\end{aligned}$$

Let $\mu\in[0,1]$ and S_n be a binomial random variable of parameters $(n-1,\mu)$. For all K,

$$\mathbb{P}[S_n\leq K]=(n-K-1)\binom{n-1}{K}\mathcal{B}\left(n-K-1,K;1-\mu\right). \tag{82}$$

According to Hoeffding's inequality, we have, for all $K\leq(n-1)\mu$,

$$\mathbb{P}[S_n\leq K]\leq\exp\left(-2(n-1)\left(\mu-\frac{K}{n-1}\right)^2\right). \tag{83}$$

Thus, for all $z\notin B(x_c,\alpha)$, for all $K\leq(n-1)\mu_X\left(B\left(x_c,\alpha\right)\right)$,

$$\begin{aligned}
&f_Z(z|X_c=x_c) && (84)\\
&\leq\exp\left(-2(n-1)\left(\mu_X\left(B\left(x_c,\alpha\right)\right)-\frac{K}{n-1}\right)^2\right)\int_{\frac{\alpha}{2R}}^1\frac{1}{w^d}f_X\left(x_c+\frac{z-x_c}{w}\right)\mathrm{d}w && (85)\\
&\leq C_2\exp\left(-2(n-1)\left(\mu_X\left(B\left(x_c,\alpha\right)\right)-\frac{K}{n-1}\right)^2\right)\int_{\frac{\alpha}{2R}}^1\frac{1}{w^d}\mathrm{d}w && (86)\\
&\leq C_2\eta(\alpha,R)\exp\left(-2(n-1)\left(\mu_X\left(B\left(x_c,\alpha\right)\right)-\frac{K}{n-1}\right)^2\right), && (87)
\end{aligned}$$

with

$$\eta(\alpha, R) = \begin{cases} \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1 \\ \frac{1}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{otherwise} \end{cases}.$$

Letting

$$\epsilon(n, \alpha, K, x_c) = C_2 \eta(\alpha, R) \exp\left(-2(n-1)\left(\mu_X\left(B\left(x_c, \alpha\right)\right) - \frac{K}{n-1}\right)^2\right), \tag{88}$$

we have, for all $\alpha \in (0, 2R)$, for all $K \leq (n-1)\mu_X\left(B\left(x_c, \alpha\right)\right)$,

$$\begin{aligned} \mathbb{P}\left(|Z - X_c| \geq \alpha | X_c = x_c\right) &= \int_{z \notin \mathcal{B}(x_c, \alpha), z \in \mathcal{X}} f_Z(z|X_c = x_c) \mathrm{d}z & (89) \\ &\leq \int_{z \notin \mathcal{B}(x_c, \alpha), z \in \mathcal{X}} \varepsilon(n, \alpha, K, x_c) \mathrm{d}z & (90) \\ &= \varepsilon(n, \alpha, K, x_c) \int_{z \notin \mathcal{B}(x_c, \alpha), z \in \mathcal{X}} \mathrm{d}z & (91) \\ &\leq c_d R^d \varepsilon(n, \alpha, K, x_c), & (92) \end{aligned}$$

as $\mathcal{X} \subset B(0, R)$. Since $x_c \in \mathcal{X}$, by definition of the support, we know that for all $\rho > 0$, $\mu_X(B(x_c, \rho)) > 0$. Thus, $\mu_X\left(B\left(x_c, \alpha\right)\right) > 0$. Consequently, $\varepsilon(n, \alpha, K, x_c)$ tends to zero, as K/n tends to zero. □

D.5. Proof of Corollary 3.5

We adapt the proof of Theorem 2.1 and Theorem 2.4 in Biau and Devroye (2015) to the case where X belongs to $B(0, R)$. We prove the following result.

Lemma D.1. *Let X takes values in $B(0, R)$. For all $d \geq 2$,*

$$\mathbb{E}[\|X_{(1)}(X) - X\|_2^2] \leq 36R^2\left(\frac{k}{n+1}\right)^{2/d}, \tag{93}$$

where $X_{(1)}(X)$ is the nearest neighbor of X among $X_1, \ldots, X_n$.

Proof of Lemma D.1. Let us denote by $X_{(i,1)}$ the nearest neighbor of X_i among $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}$. By symmetry, we have

$$\mathbb{E}[\|X_{(1)}(X) - X\|_2^2] = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{E}\|X_{(i,1)} - X_i\|_2^2. \tag{94}$$

Let $R_i = \|X_{(i,1)} - X_i\|_2$ and $B_i = \{x \in \mathbb{R}^d : \|x - X_i\| < R_i/2\}$. By construction, B_i are disjoint. Since $R_i \leq 2R$, we have

$$\cup_{i=1}^{n+1} B_i \subset B(0, 3R), \tag{95}$$

which implies,

$$\mu\left(\cup_{i=1}^{n+1} B_i\right) \leq (3R)^d c_d. \tag{96}$$

Thus, we have

$$\sum_{i=1}^{n+1} c_d \left(\frac{R_i}{2}\right)^d \leq (3R)^d c_d. \tag{97}$$

Besides, for all $d \geq 2$, we have

$$\left(\frac{1}{n+1}\sum_{i=1}^{n+1} R_i^2\right)^{d/2} \leq \frac{1}{n+1}\sum_{i=1}^{n+1} R_i^d, \tag{98}$$

which leads to

$$\begin{aligned}
\mathbb{E}[\|X_{(1)}(X) - X\|_2^2] &= \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{E}\|X_{(i,1)} - X_i\|_2^2 && (99)\\
&= \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} R_i^2\right] && (100)\\
&\leq \left(\frac{(6R)^d}{n+1}\right)^{2/d} && (101)\\
&\leq 36R^2\left(\frac{1}{n+1}\right)^{2/d}. && (102)
\end{aligned}$$

□

Lemma D.2. *Let X takes values in $B(0,R)$. For all $d \geq 2$,*

$$\mathbb{E}[\|X_{(k)}(X) - X\|_2^2] \leq (2^{1+2/d})36R^2\left(\frac{k}{n}\right)^{2/d}, \tag{103}$$

where $X_{(k)}(X)$ is the nearest neighbor of X among $X_1, \ldots, X_n$.

Proof of Lemma D.2. Set $d \geq 2$. Recall that $\mathbb{E}[\|X_{(k)}(X) - X\|_2^2] \leq 4R^2$. Besides, for all $k > n/2$, we have

$$\begin{aligned}
(2^{1+2/d})36R^2\left(\frac{k}{n}\right)^{2/d} &> (2^{1+2/d})36R^2\left(\frac{1}{2}\right)^{2/d} && (104)\\
&> 72R^2 && (105)\\
&> \mathbb{E}[\|X_{(k)}(X) - X\|_2^2]. && (106)
\end{aligned}$$

Thus, the result is trivial for $k > n/2$. Set $k \leq n/2$. Now, following the argument of Theorem 2.4 in Biau and Devroye (2015), let us partition the set $\{X_1, \ldots, X_n\}$ into $2k$ sets of sizes $n_1, \ldots, n_{2k}$ with

$$\sum_{j=1}^{2k} n_j = n \quad \text{and} \quad \left\lfloor \frac{n}{2k} \right\rfloor \leq n_j \leq \left\lfloor \frac{n}{2k} \right\rfloor + 1. \tag{107}$$

Let $X_{(1)}^\star(j)$ be the nearest neighbor of X among all X_i in the jth group. Note that

$$\|X_{(k)}(X) - X\|^2 \leq \frac{1}{k}\sum_{j=1}^{2k} \|X_{(1)}^\star(j) - X\|^2, \tag{108}$$

since at least k of these nearest neighbors have values larger than $\|X_{(k)}(X) - X\|^2$. By Lemma D.1, we have

$$\|X_{(k)}(X) - X\|^2 \leq \frac{1}{k}\sum_{j=1}^{2k} 36R^2\left(\frac{1}{n_j+1}\right)^{2/d} \tag{109}$$

$$\leq \frac{1}{k}\sum_{j=1}^{2k} 36R^2\left(\frac{2k}{n}\right)^{2/d} \tag{110}$$

$$\leq 2^{1+2/d} \times 36R^2\left(\frac{k}{n}\right)^{2/d}. \tag{111}$$

□

Proof of Corollary 3.5. Let $d \geq 2$. By Markov's inequality, for all $\varepsilon > 0$, we have

$$\mathbb{P}\left[\|X_{(k)}(X) - X\|_2 > \varepsilon\right] \leq \frac{\mathbb{E}[\|X_{(k)}(X) - X\|_2^2]}{\varepsilon^2}. \tag{112}$$

Let $\gamma \in (0, 1/d)$ and $\varepsilon = 12R(k/n)^\gamma$, we have

$$\mathbb{P}\left[\|X_{(k)}(X) - X\|_2 > 12R(k/n)^\gamma\right] \leq \left(\frac{k}{n}\right)^{2/d-2\gamma}. \tag{113}$$

Noticing that, by construction of a SMOTE observation $Z_{K,n}$, we have

$$\|Z_{K,n} - X\|_2 \leq \|X_{(K)}(X) - X\|_2. \tag{114}$$

Thus,

$$\mathbb{P}\left[\|Z_{K,n} - X\|_2 > 12R(k/n)^\gamma\right] \leq \mathbb{P}\left[\|X_{(K)}(X) - X\|_2^2 > 12R(k/n)^{1/d}\right] \tag{115}$$

$$\leq \left(\frac{k}{n}\right)^{2/d-2\gamma}. \tag{116}$$

□

D.6. Proof of Theorem 3.6

Proof of Theorem 3.6. Let $\varepsilon > 0$ and $z \in B(0,R)$ such that $\|z\| \geq R-\varepsilon$. Let $A_\varepsilon = \{x \in B(0,R), \langle x-z, z\rangle \leq 0\}$. Let $0 < \alpha < 2R$ and $\tilde{A}_{\alpha,\varepsilon} = A_\varepsilon \cap \{x, \|z-x\| \geq \alpha\}$. An illustration is displayed in Figure 4.

We have

$$f_Z(z) = \int_{x_c \in \tilde{A}_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c) \mathrm{d}x_c + \int_{x_c \in \tilde{A}^c_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c) \mathrm{d}x_c \tag{117}$$

First term Let $x_c \in \tilde{A}_{\alpha,\varepsilon}$. In order to have $x_c + \frac{z-x_c}{w} = z + \left(-1 + \frac{1}{w}\right)(z - x_c) \in B(0,R)$, it is necessary that

$$\left(-1 + \frac{1}{w}\right)\|z - x_c\| \leq \sqrt{2\varepsilon R} \tag{118}$$

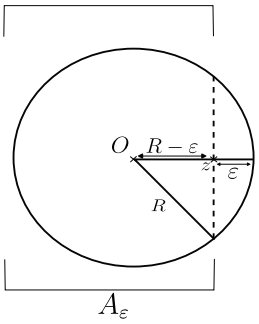


Figure 4 | Illustration of Theorem 3.6.

which leads to

$$w \geq \frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\|z-x_c\|}} \tag{119}$$

Since $x_c \in \tilde{A}_{\alpha,\varepsilon}$, we have $\|x_c - z\| \geq \alpha$. Thus, according to inequality (119), $x_c + \frac{z-x_c}{w} \in B(0,R)$ implies

$$w \geq \frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\alpha}}. \tag{120}$$

Recall that $x_c + \frac{z-x_c}{w} \in \mathcal{X}$. Consequently, according to Lemma C.1, for all $x_c \in \tilde{A}_{\alpha,\varepsilon}$,

$$f_Z(z|X_c = x_c) \tag{121}$$

$$= (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \mathrm{d}w \tag{122}$$

$$\leq C_2(n-K-1)\binom{n-1}{K}\int_{\frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\alpha}}}^1 \frac{1}{w^d} \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \mathrm{d}w. \tag{123}$$

Besides,

$$(n-K-1)\binom{n-1}{K}\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \tag{124}$$

$$= \left(\frac{n-1}{K}\right)(n-K-1)\binom{n-2}{K-1}\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \tag{125}$$

$$\leq \frac{n-1}{K}, \tag{126}$$

according to Lemma E.1. Thus,

$$f_Z(z|X_c = x_c) \leq C_2\left(\frac{n-1}{K}\right)\int_{\frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\alpha}}}^1 \frac{1}{w^d}\mathrm{d}w \tag{127}$$

$$\leq C_2\left(\frac{n-1}{K}\right)\eta(\alpha, R), \tag{128}$$

with

$$\eta(\alpha, R) = \begin{cases} \ln\left(1 + \frac{\sqrt{2\varepsilon R}}{\alpha}\right) & \text{if } d = 1 \\ \frac{1}{d-1}\left(\left(1 + \frac{\sqrt{2\varepsilon R}}{\alpha}\right)^{d-1} - 1\right) & \text{otherwise} \end{cases}.$$

Second term According to Lemma C.1, we have

$$f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \mathrm{d}w \tag{129}$$

$$\leq \left(\frac{n-1}{K}\right)\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \mathrm{d}w \tag{130}$$

Since $\mathcal{X} \subset B(0, R)$, all points $x, z \in \mathcal{X}$ satisfy $\|x - z\| \leq 2R$.

Consequently, if $\|z - x_c\|/w > 2R$,

$$x_c + \frac{\|z-x_c\|}{w} \notin \mathcal{X}. \tag{131}$$

Hence, for all $w \leq \|z - x_c\|/2R$,

$$f_X\left(x_c + \frac{z-x_c}{w}\right) = 0. \tag{132}$$

Plugging this equality into (130), we have

$$f_Z(z|X_c = x_c) \tag{133}$$

$$\leq \left(\frac{n-1}{K}\right)\int_{\|z-x_c\|/2R}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \mathrm{d}w \tag{134}$$

$$\leq C_2\left(\frac{n-1}{K}\right)\int_{\|z-x_c\|/2R}^1 \frac{1}{w^d} \mathrm{d}w \tag{135}$$

$$\leq C_2\left(\frac{n-1}{K}\right)\left[-\frac{1}{d-1} w^{-d+1}\right]_{\|z-x_c\|/2R}^1 \tag{136}$$

$$\leq C_2\left(\frac{n-1}{K}\right)\frac{(2R)^{d-1}}{d-1}\frac{1}{\|z-x_c\|^{d-1}}. \tag{137}$$

Besides, note that, for all $\alpha > 0$, we have

$$\int_{B(z,\alpha)} \frac{1}{\|z-x_c\|^{d-1}} f_X(x_c) \mathrm{d}x_c \tag{138}$$

$$\leq C_2 \int_{B(0,\alpha)} \frac{1}{r^{d-1}} r^{d-1} \sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2)\ldots\sin(\varphi_{d-2})\mathrm{d}r\mathrm{d}\varphi_1\ldots\mathrm{d}\varphi_{d-2}, \tag{139}$$

where $r, \varphi_1, \ldots, \varphi_{d-2}$ are the spherical coordinates. A direct calculation leads to

$$\int_{B(z,\alpha)} \frac{1}{\|z-x_c\|^{d-1}} f_X(x_c) \mathrm{d}x_c$$

$$\leq C_2 \int_0^\alpha \mathrm{d}r \int_{S(0,\alpha)} \sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2)\ldots\sin(\varphi_{d-2})\mathrm{d}\varphi_1\ldots\mathrm{d}\varphi_{d-2} \tag{140}$$

$$\leq \frac{2C_2\pi^{d/2}}{\Gamma(d/2)}\alpha, \tag{141}$$

as

$$\int_{S(0,\alpha)} \sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2)\ldots\sin(\varphi_{d-2})\mathrm{d}\varphi_1\ldots\mathrm{d}\varphi_{d-2} \tag{142}$$

is the surface of the S^{d-1} sphere. Finally, for all $z \in \mathcal{X}$, for all $\alpha > 0$, and for all K, N such that $1 \leq K \leq N$, we have

$$\int_{B(z,\alpha)} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c \leq \frac{2C_2^2(2R)^{d-1}\pi^{d/2}}{(d-1)\Gamma(d/2)}\left(\frac{n-1}{K}\right)\alpha. \tag{143}$$

Final result Using Figure 4 and Pythagore's Theorem, we have $a^2 \leq \sqrt{2\varepsilon R}$. Let $d > 1$ and $\epsilon > 0$. Then we have for all α such that $\alpha > a$.

$$\begin{aligned}
&f_Z(z) && (144)\\
&= \int_{x_c \in \tilde{A}_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c + \int_{x_c \in \tilde{A}^c_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c && (145)\\
&\leq \frac{C_2}{d-1}\left(\left(1+\frac{\sqrt{2\varepsilon R}}{\alpha}\right)^{d-1} - 1\right)\left(\frac{n-1}{K}\right) + \frac{2C_2^2(2R)^{d-1}\pi^{d/2}}{(d-1)\Gamma(d/2)}\left(\frac{n-1}{K}\right)\alpha && (146)\\
&= \frac{C_2}{d-1}\left(\frac{n-1}{K}\right)\left[\left(\left(1+\frac{\sqrt{2\varepsilon R}}{\alpha}\right)^{d-1} - 1\right) + \frac{2C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)}\alpha\right], && (147)
\end{aligned}$$

But this inequality is true if $\alpha \geq a$. We know that $(1+x)^{d-1} \leq (2^{d-1}-1)x + 1$ for $x \in [0,1]$ and $d - 1 \geq 0$. Then, for α such that $\frac{\sqrt{2\varepsilon R}}{\alpha} \leq 1$,

$$\begin{aligned}
&f_Z(z) && (148)\\
&\leq \frac{C_2}{d-1}\left(\frac{n-1}{K}\right)\left[\left(\left((2^{d-1}-1)\frac{\sqrt{2\varepsilon R}}{\alpha} + 1\right) - 1\right) + \frac{2C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)}\alpha\right] && (149)\\
&\leq \frac{C_2}{d-1}\left(\frac{n-1}{K}\right)\left[\left((2^{d-1}-1)\frac{\sqrt{2\varepsilon R}}{\alpha}\right) + \frac{2C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)}\alpha\right]. && (150)
\end{aligned}$$

Since $\frac{\sqrt{2\varepsilon R}}{\alpha} \leq 1$, then $\alpha \geq \sqrt{2\varepsilon R} \geq a$. So our initial condition on α to get the upper bound of the second term is still true. Now, we choose α such that,

$$(2^{d-1}-1)\frac{\sqrt{2\varepsilon R}}{\alpha} \leq \frac{2C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)}\alpha, \tag{151}$$

which leads to the following condition

$$\alpha \geq \left(\frac{\Gamma(d/2)(2^{d-1}-1)\sqrt{2\varepsilon R}}{2C_2(2R)^{d-1}\pi^{d/2}}\right)^{1/2}, \tag{152}$$

assuming that

$$\left(\frac{\varepsilon}{R}\right)^{1/2} \leq \frac{1}{\sqrt{2}dC_2} Vol(B_d(0,1)). \tag{153}$$

Finally, for

$$\alpha = \left(\frac{\Gamma(d/2)(2^{d-1}-1)\sqrt{2\varepsilon R}}{2C_2(2R)^{d-1}\pi^{d/2}} \right)^{1/2}, \tag{154}$$

we have,

$$f_Z(z) \leq \frac{C_2}{d-1} \left(\frac{n-1}{K} \right) \left[\frac{4C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)} \alpha \right] \tag{155}$$

$$\leq \frac{C_2}{d-1} \left(\frac{n-1}{K} \right) \left[\frac{4C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)} \left(\frac{\Gamma(d/2)(2^{d-1}-1)\sqrt{2\varepsilon R}}{2C_2(2R)^{d-1}\pi^{d/2}} \right)^{1/2} \right] \tag{156}$$

$$= 2^{d+2} \left(\frac{n-1}{K} \right) \left(\frac{C_2^3 Vol(B_d(0,1))}{d} \right)^{1/2} \left(\frac{\varepsilon}{R} \right)^{1/4}. \tag{157}$$

□

D.7. Proof of Lemma 4.1

Proof of Lemma 4.1. Let $\hat{f}$ be a trained classifier. Let $\mathcal{D}_{test} = \{(X_i, Y_i)\}_{i=1}^N$ be a test set for $\hat{f}$, with $n \in \mathbb{N}$ and $\sum_{i=1}^N Y_i = n$. We have $Y \in \{0, 1\}$.

The balanced accuracy of $\hat{f}$, denoted by $\text{Acc}_{Balanced}(\hat{f})$, is defined as the average of the recall and the specificity of $\hat{f}$ on $\mathcal{D}_{test}$. Let us consider the weighted accuracy of $\hat{f}$, denoted by $Acc_{Weighted}(\hat{f})$, which is nothing but the standard accuracy where samples in $\mathcal{D}_{test}$ are weighted by $\frac{N}{2(N-n)}$ if they belong to the class $Y = 0$ and by $\frac{N}{2n}$ if they belong to the class $Y = 1$.

We denote respectively by TP,TN,FP,FN the number of true positive samples, the number of true negative, the number of false positive and the number of false negative from $\hat{f}$ predictions on $\mathcal{D}_{test}$. Thus we have $N = TP + TN + FP + FN$ and $n = TP + FN$.

We have:

$$\text{Acc}_{Weighted}(\hat{f}) = \frac{TP\frac{1}{2}\frac{N}{n} + TN\frac{1}{2}\frac{N}{N-n}}{N} \tag{158}$$

$$= \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{159}$$

$$= \frac{1}{2} \left(Recall + Specificity \right) \tag{160}$$

$$= \text{Acc}_{Balanced}(\hat{f}). \tag{161}$$

□

E. TECHNICAL LEMMAS

E.1. Cumulative distribution function of a binomial distribution

Lemma E.1 (Cumulative distribution function of a binomial distribution)**.** *Let X be a random variable following a binomial law of parameter $n \in \mathbf{N}$ and $p \in [0, 1]$. The cumulative distribution function F of X can be expressed as Wadsworth et al. (1961):*

(i)

$$F(k;n,p) = \mathbb{P}(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i},$$

(ii)

$$\begin{aligned} F(k;n,p) &= (n-k)\binom{n}{k} \int_0^{1-p} t^{n-k-1}(1-t)^k \mathrm{d}t \\ &= (n-k)\binom{n}{k} \mathcal{B}(n-k, k+1; 1-p), \end{aligned}$$

with $\mathcal{B}(a,b;x) = \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t$*, the incomplete beta function.*

Proof. see Wadsworth et al. (1961). □

E.2. Bounds for the incomplete beta function

Lemma E.2. *Let* $B(a,b;x) = \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t$*, be the incomplete beta function. Then we have*

$$\frac{x^a}{a} \leq B(a,b;x) \leq x^{a-1}\left(\frac{1-(1-x)^b}{b}\right),$$

for $a > 0$.

Proof. We have

$$\begin{aligned} B(a,b;x) &= \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &\leq \int_{t=0}^{x} x^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &= x^{a-1}\int_{t=0}^{x} (1-t)^{b-1}\mathrm{d}t \\ &= x^{a-1}\left[(-1)\frac{(1-t)^b}{b}\right]_0^x \\ &= x^{a-1}\left[-\frac{(1-x)^b}{b} + \frac{1}{b}\right] \\ &= x^{a-1}\frac{1-(1-x)^b}{b}. \end{aligned}$$

On the other hand,

$$\begin{aligned} B(a,b;x) &= \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &\geq \int_{t=0}^{t} x^{a-1}\mathrm{d}t \\ &= \left[\frac{t^a}{a}\right]_0^x \\ &= \frac{x^a}{a} - \frac{0^a}{a} \\ &= \frac{x^a}{a}. \end{aligned}$$

□

E.3. Upper bound for binomial coefficient

Lemma E.3. *For $k, n \in \mathbb{N}$ such that $k < n$, we have*

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k. \tag{162}$$

Proof. We have,

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k!} \leq \frac{n^k}{k!}. \tag{163}$$

Besides,

$$e^k = \sum_{i=0}^{+\infty} \frac{k^i}{i!} \Longrightarrow e^k \geq \frac{k^k}{k!} \Longrightarrow \frac{e^k}{k^k} \geq \frac{1}{k!}. \tag{164}$$

Hence,

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k!} \leq \frac{n^k}{k!} \leq \left(\frac{en}{k}\right)^k. \tag{165}$$

□

E.4. Inequality $x \ln\left(\frac{1}{x}\right) \leq \sqrt{x}$

Lemma E.4. *For $x \in]0, +\infty[$,*

$$x \ln\left(\frac{1}{x}\right) \leq \sqrt{x}. \tag{166}$$

Proof. Let,

$$f(x) = \sqrt{x} - x \ln\left(\frac{1}{x}\right) \tag{167}$$

$$= \sqrt{x} + x \ln(x). \tag{168}$$

Then,

$$f^{'}(x) = \frac{1}{2\sqrt{x}} + \ln x + 1. \tag{169}$$

And,

$$f^{''}(x) = \frac{1}{x} - \frac{1}{4x^{3/2}}. \tag{170}$$

We have,

$$\begin{aligned} f''(x) \geq 0 &\Longrightarrow \frac{1}{x} - \frac{1}{4x^{3/2}} \geq 0 \\ &\Longrightarrow \frac{1}{x} \geq \frac{1}{4x^{3/2}} \end{aligned} \tag{171}$$

Since $x \in]0, +\infty[$,

$$\textit{Equation (171)} \Longrightarrow \frac{x^{3/2}}{x} \geq \frac{1}{4} \tag{172}$$

$$\Longrightarrow \sqrt{x} \geq \frac{1}{4} \tag{173}$$

$$\Longrightarrow x \geq \frac{1}{16}. \tag{174}$$

This result leads to,

x	0		$\frac{1}{16}$		$+\infty$
f''		$-$	0	$+$	
f'	$\searrow$		$2 + \ln(\frac{1}{16}) + 1$		$\nearrow$

(175)

We have $2 + \ln(\frac{1}{16}) + 1 > 0$. So $f'(x) > 0$ for all $x \in]0, \infty[$. Furthermore $\lim_{x \to 0^+} f(x) = 0$, hence $f(x) > 0$ for all $x \in]0, \infty[$, therefore $\sqrt{x} > x \ln\left(\frac{1}{x}\right)$ for all $x \in]0, \infty[$.

□

F. TABLES

In this section, the additional tables from the experimental on real world data sets are displayed.

Tables The tables in appendix can be divided into 3 categories. First, we have the tables related to random forests. Then the tables related to logistic regression. Finally, we have the tables of LightGBM classifiers. Here are some details fore each group:

- **Random Forest:** In Table 6, PR AUC of data sets not presented in Table 2 are displayed. In Table 7 and Table 8 PR AUC of default random forests are displayed for all the data sets. In Table 9 is displayed default random forests PR AUC with a max tree depth fixed to the value of RUS depth.
- **LightGBM:** The PR AUC of the extremely imbalanced data sets when using a LightGBM, with tuned maximum tree depth,are displayed in Table 10. Table 13 shows the PR AUC of LightGBM (not tuned) on three multiclass data set.

Table 5 | Table 2 with standard deviations. Random Forest (tree depth tuned with PR AUC) PR AUC for different rebalancing strategies and different data sets. The other data sets are displayed in Table 6.

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
CreditCard (0.2%)	0.776	0.783	0.751	**0.787**	0.586	0.786	0.784	0.773	0.771	0.747	0.757	0.752
std	0.0	0.0	0.008	0.003	0.0	0.003	0.003	0.002	0.006	0.01	0.012	0.0
Abalone (1%)	0.048	0.055	0.052	0.052	0.024	0.044	0.039	0.046	0.051	**0.056**	0.050	**0.056**
std	0.006	0.009	0.017	0.011	0.007	0.008	0.006	0.011	0.012	0.014	0.016	0.006
Phoneme (1%)	0.198	0.211	0.086	0.200	0.054	0.222	0.211	0.224	0.240	**0.269**	0.130	0.249
std	0.026	0.026	0.031	0.037	0.014	0.03	0.028	0.035	0.035	0.032	0.042	0.047
Haberman (10%)	0.252	0.268	0.289	0.295	0.268	0.308	0.294	0.306	0.307	**0.311**	0.280	0.306
std	0.033	0.035	0.045	0.038	0.04	0.036	0.039	0.034	0.036	0.032	0.024	0.042
MagicTel (20%)	0.754	0.759	0.741	**0.760**	0.338	0.739	0.742	0.757	0.757	0.750	0.693	0.756
std	0.002	0.003	0.004	0.004	0.009	0.004	0.004	0.004	0.003	0.003	0.012	0.003
California (1%)	0.298	0.283	0.207	0.277	0.019	0.234	0.226	0.242	0.258	**0.322**	0.152	0.294
std	0.015	0.02	0.023	0.017	0.001	0.012	0.015	0.02	0.019	0.024	0.023	0.013

- **Logistic Regression:** Table 11 dispalys PR AUC of of extremely imbalanced data sets when using Logistic regression. Table 12 shows PR AUC of None, LDAM and Focal loss strategies when using a logistic regression reimplemented using PyTorch.

Table 6 | Remaining data sets (without those of Table 2). Random Forest (tree depth tuned with PR AUC) PR AUC for different rebalancing strategies and different data sets. Only datasets such that the None strategy is on par with the best strategies are displayed. Other data sets are presented in Table 6. Mean standard deviations are computed. The best strategy is highlighted in bold for each data set.

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
Phoneme	**0.920**	0.918	0.883	0.919	0.844	0.912	0.911	0.916	0.915	0.911	0.910	0.912
std	0.0	0.0	0.004	0.002	0.003	0.001	0.001	0.001	0.002	0.0	0.005	0.001
Phoneme (20%)	**0.864**	0.859	0.781	0.861	0.577	0.844	0.847	0.857	0.857	0.854	0.824	0.003
std	0.002	0.004	0.012	0.005	0.013	0.005	0.004	0.004	0.001	0.005	0.008	0.004
Phoneme (10%)	**0.714**	0.700	0.540	0.705	0.278	0.676	0.690	0.696	0.691	0.705	0.609	0.686
std	0.009	0.008	0.015	0.014	0.015	0.014	0.013	0.006	0.014	0.007	0.015	0.003
Yeast	0.830	**0.846**	0.839	0.845	0.732	0.787	0.778	0.831	0.824	0.814	0.823	0.703
std	0.015	0.01	0.012	0.007	0.014	0.013	0.015	0.012	0.011	0.005	0.019	0.016
Yeast (1%)	**0.380**	0.323	0.254	0.298	0.149	0.333	0.320	0.333	0.335	0.295	0.252	0.243
std	0.053	0.054	0.069	0.06	0.051	0.058	0.041	0.055	0.06	0.063	0.016	0.031
Wine (4%)	0.601	0.598	0.387	**0.606**	0.149	0.578	0.585	0.589	0.589	0.545	0.424	0.559
std	0.024	0.029	0.024	0.026	0.014	0.021	0.025	0.025	0.027	0.023	0.031	0.022
Pima	**0.713**	0.705	0.701	0.712	0.697	0.681	0.684	0.698	0.706	0.701	0.691	0.703
std	0.012	0.016	0.007	0.005	0.008	0.004	0.006	0.01	0.002	0.009	0.004	0.001
Pima (20%)	**0.527**	0.513	0.517	0.519	0.484	0.502	0.503	0.504	0.507	0.520	0.495	0.481
std	0.018	0.018	0.015	0.018	0.017	0.017	0.013	0.016	0.013	0.015	0.019	0.017
Haberman	0.461	**0.477**	0.441	0.457	0.471	0.457	0.455	0.462	0.440	0.468	0.433	0.455
std	0.01	0.013	0.031	0.025	0.023	0.005	0.015	0.03	0.001	0.028	0.011	0.006
California (20%)	**0.889**	0.885	0.872	0.884	0.674	0.875	0.878	0.882	0.882	0.880	0.854	0.852
std	0.001	0.0	0.002	0.001	0.0	0.000	0.001	0.001	0.0	0.0	0.005	0.002
California (10%)	**0.803**	0.796	0.757	0.797	0.411	0.773	0.778	0.787	0.784	0.791	0.718	0.728
std	0.000	0.002	0.001	0.004	0.007	0.002	0.005	0.004	0.003	0.002	0.007	0.002
House_16H (20%)	**0.858**	0.846	0.834	0.849	0.587	0.829	0.832	0.840	0.838	0.835	0.846	0.814
std	0.001	0.001	0.001	0.000	0.006	0.002	0.001	0.001	0.002	0.001	0.002	0.000
House_16H (10%)	**0.756**	0.729	0.693	0.733	0.291	0.701	0.706	0.714	0.712	0.692	0.723	0.648
std	0.001	0.001	0.002	0.002	0.001	0.000	0.003	0.003	0.005	0.004	0.008	0.001
House_16H (1%)	**0.317**	0.244	0.144	0.220	0.034	0.210	0.227	0.194	0.199	0.178	0.231	0.187
std	0.011	0.021	0.018	0.004	0.003	0.009	0.003	0.004	0.000	0.001	0.007	0.002
Vehicle	0.978	0.979	0.961	0.980	0.962	0.976	0.978	0.978	0.978	**0.981**	0.980	0.980
std	0.003	0.001	0.002	0.001	0.005	0.003	0.004	0.001	0.001	0.004	0.0	0.001
Vehicle (10%)	**0.958**	0.950	0.872	0.935	0.710	0.927	0.928	0.953	0.944	0.956	0.939	0.954
std	0.003	0.003	0.011	0.015	0.02	0.003	0.003	0.005	0.009	0.003	0.011	0.001
Ionosphere	**0.975**	0.973	0.963	0.971	0.933	0.974	0.973	0.971	0.971	0.962	0.969	0.969
std	0.003	0.001	0.009	0.004	0.002	0.003	0.001	0.001	0.000	0.002	0.001	0.000
Ionosphere (20%)	**0.972**	0.963	0.943	0.966	0.688	0.928	0.949	0.937	0.929	0.945	0.958	0.923
std	0.002	0.005	0.002	0.000	0.013	0.025	0.008	0.013	0.015	0.007	0.001	0.014
Ionosphere (10%)	**0.950**	0.929	0.885	0.925	0.522	0.840	0.868	0.815	0.848	0.826	0.916	0.854
std	0.001	0.005	0.006	0.006	0.056	0.014	0.011	0.029	0.002	0.023	0.013	0.027
Breast Cancer	0.985	0.984	0.986	0.987	0.987	0.984	0.979	0.985	**0.988**	0.987	0.985	0.984
std	0.004	0.004	0.002	0.000	0.001	0.003	0.000	0.003	0.001	0.000	0.001	0.002
Breast Cancer (20%)	0.982	0.976	0.975	0.980	0.984	0.972	0.960	0.982	0.982	0.981	**0.985**	0.976
std	0.002	0.008	0.001	0.001	0.002	0.001	0.007	0.001	0.003	0.003	0.001	0.007
Breast Cancer (10%)	0.967	0.952	0.903	0.958	0.976	0.920	0.915	0.947	0.934	0.966	**0.979**	0.948
std	0.006	0.004	0.002	0.002	0.008	0.010	0.002	0.004	0.011	0.009	0.005	0.004

Table 7 | Extremely imbalanced data sets. Random Forest (tree depth=∞) PR AUC for different rebalancing strategies and different data sets. Data sets artificially undersampled for minority class are in italics. Other data sets are presented in Table 8. Mean standard deviations are computed.

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
CreditCard (0.2%)	0.772	0.776	0.751	0.780	0.524	**0.784**	0.780	0.772	0.770	0.731	0.754	0.743
std	0.004	0.003	0.010	0.003	0.024	0.003	0.003	0.004	0.004	0.005	0.008	0.005
Abalone (1%)	0.046	0.040	0.046	0.041	0.024	0.047	0.039	0.040	0.051	**0.056**	0.048	0.047
std	0.009	0.008	0.014	0.008	0.006	0.008	0.005	0.007	0.009	0.010	0.007	0.010
Phoneme (1%)	0.210	0.228	0.082	0.227	0.052	0.237	0.220	0.241	0.259	**0.269**	0.133	0.255
std	0.027	0.030	0.030	0.030	0.013	0.034	0.027	0.037	0.034	0.026	0.036	0.036
Haberman (10%)	0.216	0.223	0.284	0.228	0.266	0.278	0.255	0.275	0.289	**0.294**	0.277	0.265
std	0.022	0.019	0.043	0.021	0.034	0.029	0.030	0.027	0.033	0.035	0.041	0.035
MagicTel (20%)	0.751	**0.758**	0.737	**0.758**	0.340	0.736	0.739	0.754	0.753	0.746	0.695	0.754
std	0.003	0.003	0.005	0.003	0.007	0.003	0.004	0.003	0.003	0.003	0.007	0.003
California (1%)	0.294	0.292	0.197	0.282	0.019	0.234	0.230	0.249	0.265	**0.322**	0.195	0.294
std	0.012	0.016	0.023	0.016	0.003	0.015	0.016	0.018	0.017	0.024	0.025	0.018

Table 8 | Remaining data sets (without those of Table 2). Random Forest (tree depth=∞) PR AUC for different rebalancing strategies and different data sets. Only datasets such that the None strategy is on par with the best strategies are displayed. Other data sets are presented in Table 7. Mean standard deviations are computed. The best strategy is highlighted in bold for each data set.

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
Phoneme	**0.919**	0.916	0.882	0.919	0.845	0.910	0.913	0.916	0.914	0.913	0.905	0.911
std	0.003	0.003	0.004	0.003	0.006	0.004	0.003	0.003	0.003	0.003	0.003	0.003
Phoneme(20%)	**0.862**	0.858	0.774	0.861	0.572	0.843	0.848	0.854	0.854	0.854	0.824	0.849
std	0.004	0.003	0.011	0.004	0.023	0.005	0.005	0.005	0.003	0.004	0.007	0.004
Phoneme(10%)	**0.724**	0.707	0.538	0.713	0.271	0.675	0.685	0.695	0.693	0.701	0.620	0.685
std	0.011	0.012	0.020	0.013	0.023	0.015	0.011	0.012	0.011	0.009	0.021	0.014
Pima	**0.700**	0.699	0.688	0.695	0.691	0.674	0.674	0.685	0.680	0.684	0.699	0.689
std	0.011	0.010	0.012	0.009	0.009	0.009	0.013	0.011	0.016	0.011	0.014	0.013
Pima (20%)	**0.506**	0.501	**0.506**	0.497	0.481	0.480	0.482	0.479	0.485	0.484	0.488	0.472
std	0.024	0.019	0.019	0.020	0.017	0.017	0.020	0.019	0.021	0.018	0.023	0.020
Yeast	0.836	**0.839**	0.823	0.830	0.716	0.790	0.788	0.824	0.822	0.820	0.830	0.814
std	0.009	0.010	0.010	0.013	0.013	0.014	0.011	0.010	0.009	0.015	0.010	0.013
Yeast (1%)	0.314	0.308	0.242	0.267	0.112	0.340	0.312	**0.353**	0.351	0.288	0.300	0.269
std	0.056	0.057	0.066	0.046	0.027	0.059	0.044	0.057	0.047	0.050	0.073	0.063
Wine (4%)	0.602	0.599	0.385	**0.604**	0.150	0.579	0.587	0.587	0.585	0.543	0.460	0.567
std	0.024	0.027	0.032	0.028	0.014	0.022	0.025	0.026	0.025	0.020	0.030	0.025
Haberman	0.423	0.423	0.436	0.410	**0.466**	0.414	0.412	0.421	0.430	0.434	0.427	0.419
std	0.023	0.027	0.024	0.022	0.019	0.022	0.023	0.021	0.022	0.023	0.031	0.024
California (20%)	**0.887**	0.884	0.868	0.885	0.672	0.873	0.878	0.881	0.880	0.880	0.855	0.875
std	0.001	0.001	0.002	0.002	0.004	0.001	0.002	0.002	0.002	0.002	0.005	0.001
California (10%)	**0.799**	0.795	0.756	0.796	0.382	0.772	0.777	0.786	0.784	0.791	0.698	0.782
std	0.003	0.003	0.006	0.003	0.012	0.004	0.004	0.003	0.003	0.003	0.016	0.003
House_16H (20%)	**0.854**	0.847	0.830	0.847	0.577	0.826	0.829	0.836	0.836	0.831	0.847	0.821
std	0.001	0.001	0.002	0.000	0.005	0.001	0.000	0.000	0.001	0.002	0.001	0.000
House_16H (10%)	**0.752**	0.730	0.686	0.728	0.240	0.702	0.704	0.708	0.705	0.690	0.725	0.671
std	0.002	0.002	0.002	0.002	0.010	0.001	0.002	0.002	0.003	0.002	0.006	0.003
House_16H (1%)	**0.303**	0.244	0.157	0.241	0.032	0.212	0.235	0.205	0.212	0.193	0.186	0.187
std	0.010	0.014	0.018	0.013	0.006	0.013	0.010	0.011	0.012	0.011	0.022	0.011
Vehicle	**0.982**	0.980	0.964	0.981	0.958	0.980	0.982	0.978	0.979	0.983	0.979	0.979
std	0.002	0.003	0.006	0.003	0.008	0.003	0.003	0.002	0.002	0.003	0.003	0.003
Vehicle (10%)	0.952	0.941	0.866	0.939	0.717	0.932	0.925	0.943	0.945	**0.954**	0.940	0.951
std	0.007	0.011	0.029	0.010	0.025	0.012	0.014	0.008	0.008	0.008	0.011	0.006
Ionosphere	**0.971**	0.971	0.964	0.970	0.933	0.969	0.971	0.968	0.968	0.964	0.970	0.970
std	0.003	0.003	0.005	0.003	0.008	0.003	0.003	0.003	0.004	0.004	0.004	0.003
Ionosphere (20%)	**0.968**	0.960	0.942	0.962	0.701	0.911	0.942	0.926	0.917	0.956	0.964	0.941
std	0.001	0.000	0.004	0.004	0.012	0.018	0.007	0.005	0.009	0.001	0.009	0.006
Ionosphere (10%)	**0.932**	0.922	0.860	0.925	0.642	0.819	0.852	0.792	0.804	0.880	0.928	0.904
std	0.007	0.000	0.014	0.001	0.076	0.020	0.013	0.008	0.037	0.025	0.009	0.003
Breast Cancer	**0.987**	0.986	0.984	0.986	0.986	0.982	0.983	0.985	0.986	0.987	0.986	0.987
std	0.002	0.003	0.004	0.003	0.002	0.003	0.004	0.003	0.003	0.002	0.003	0.003
Breast Cancer (20%)	**0.984**	0.981	0.970	0.978	0.982	0.972	0.971	0.978	0.977	0.981	0.981	0.979
std	0.006	0.006	0.008	0.007	0.007	0.007	0.007	0.006	0.007	0.006	0.007	0.005
Breast Cancer (10%)	**0.978**	0.966	0.933	0.965	0.976	0.940	0.938	0.964	0.959	0.967	0.969	0.967
std	0.008	0.009	0.020	0.009	0.010	0.014	0.015	0.009	0.013	0.012	0.014	0.010

Table 9 | Extremely imbalanced data sets. PR AUC Random Forest with tree depth=RUS. On the last column, the value of maximal depth when using Random forest (max_depth=∞) with RUS strategy for each data set. For each data set, the best strategy is highlighted in bold and the mean of the standard deviation is computed (and rounded to 10^{-3}).

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CTGAN	Forest Diff	depth
CreditCard (±0.005)	0.773	**0.781**	0.751	0.774	0.508	**0.781**	0.777	0.752	0.773	0.760	0.714	0.773	10
Abalone (1%)(±0.01)	0.048	0.055	0.049	0.057	0.023	0.044	0.041	0.054	0.054	**0.058**	0.048	0.057	7
Phoneme (1%)(±0.025)	0.162	0.102	0.087	0.104	0.049	0.166	0.142	0.124	**0.255**	0.146	0.084	0.120	6
Haberman (10%)(±0.033)	0.242	0.260	0.292	0.261	0.268	0.301	0.280	0.291	0.294	**0.321**	0.278	0.305	7
MagicTel (20%) (±0.004)	0.755	**0.759**	0.741	0.760	0.339	0.737	0.740	0.758	0.752	0.751	0.697	0.758	20
California (1%)(±0.017)	**0.295**	0.232	0.200	0.204	0.018	0.192	0.190	0.223	0.265	0.301	0.158	0.288	10

Table 10 | Extremely imbalanced data sets. LightGBM (max_depth=tuned on PR AUC) PR AUC.

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
CreditCard (0.2%)	0.431	0.767	0.727	**0.772**	0.306	0.730	0.734	0.742	0.749	0.754	0.755	0.770
std	0.03	0.0	0.015	0.006	0.0	0.006	0.012	0.007	0.011	0.001	0.013	0.005
Abalone (1%)	0.051	0.048	0.039	0.051	0.031	0.041	0.040	0.046	**0.053**	**0.053**	0.048	0.050
std	0.013	0.01	0.013	0.014	0.009	0.005	0.01	0.007	0.008	0.013	0.013	0.007
Phoneme (1%)	0.181	0.177	0.052	0.174	0.039	0.257	0.248	0.237	0.245	0.239	0.106	**0.268**
std	0.03	0.029	0.013	0.039	0.005	0.031	0.039	0.039	0.037	0.027	0.026	0.05
Haberman (10%)	0.297	0.316	0.125	0.300	0.134	0.294	0.263	0.302	0.322	**0.323**	0.305	0.282
std	0.047	0.038	0.023	0.041	0.027	0.039	0.041	0.039	0.055	0.037	0.047	0.033
MagicTel (20%)	0.762	**0.764**	0.730	0.763	0.281	0.729	0.729	0.760	0.761	0.746	0.702	0.763
std	0.004	0.003	0.005	0.004	0.008	0.005	0.006	0.004	0.004	0.003	0.009	0.003
California (1%)	0.358	0.342	0.229	0.330	0.037	0.308	0.292	0.342	0.352	**0.401**	0.209	0.352
std	0.031	0.024	0.031	0.028	0.009	0.017	0.021	0.022	0.015	0.023	0.033	0.012

Table 11 | Extremely imbalanced data sets.. Logistic Regression PR AUC. For each data set, the best strategy is highlighted in bold and the mean of the standard deviation is computed (and rounded to 10^{-3}).

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
CreditCard (0.2%)(±0.007)	0.618	0.726	0.456	0.733	0.008	0.635	0.559	0.707	0.715	0.677	0.714	**0.753**
Abalone (1%)(±0.012)	0.086	0.080	0.078	0.079	0.058	0.077	0.084	0.077	0.077	0.076	0.051	**0.088**
Phoneme (1%)(±0.015)	**0.091**	0.060	0.066	0.063	0.052	0.043	0.044	0.047	0.044	0.043	0.040	0.068
Haberman (10%)(±0.029)	0.348	0.365	0.326	0.360	0.350	0.359	0.357	0.358	**0.366**	0.353	0.230	0.365
MagicTel (20%) (±0.004)	**0.558**	0.525	0.524	0.524	0.178	0.464	0.462	0.521	0.521	0.518	0.423	0.525
California (1%)(±0.013)	**0.316**	0.278	0.211	0.278	0.049	0.313	0.305	0.293	0.290	0.309	0.103	0.291

Table 12 | Extremely imbalanced data sets PR AUC. Logistic regression reimplemented in PyTorch using the implementation of Cao et al. (2019).

Strategy	None	LDAM loss	Focal loss	L2RW
CreditCard	0.677 ± 0.041	0.677 ± 0.041	**0.776** ± 0.007	0.714 ± 0.033
Abalone	0.048 ± 0.0143	0.048 ± 0.014	0.085 ± 0.013	**0.091** ± 0.010
Phoneme (1%)	**0.078** ± 0.029	**0.078** ± 0.029	0.055 ± 0.016	0.033 ± 0.003
Haberman (10%)	0.273 ± 0.273	0.273 ± 0.273	**0.349** ± 0.026	0.154 ± 0.015
MagicTel (20%)	0.521 ± 0.031	0.521 ± 0.031	**0.555** ± 0.002	0.404 ± 0.012
California (1%)	**0.311** ± 0.012	**0.311** ± 0.012	0.250 ± 0.010	0.238 ± 0.018

Table 13 | LightGBM (not tuned) PR AUC computed on three **multiclass** data sets (average of one-vs-rest PR AUC).

Strategy	None	CW	RUS	ROS	NM1	BS1	BS2	SMOTE	SMOTE K-tuned	MGS ($d+1$)	CT GAN	Forest Diff
Wine	0.496	0.496	0.418	0.491	0.412	0.496	**0.502**	0.490	0.462	0.436	0.430	0.454
std	0.001	0.006	0.006	0.003	0.004	0.001	0.000	0.001	0.003	0.006	0.003	0.003
Yeast	0.673	0.671	0.647	0.680	0.619	0.671	0.677	0.673	0.683	**0.689**	0.686	0.684
std	0.002	0.000	0.006	0.006	0.002	0.002	0.000	0.002	0.003	0.000	0.008	0.007
Ecoli	**0.880**	0.873	0.848	0.873	0.833	0.866	0.872	0.871	0.871	0.867	0.868	0.868
std	0.002	0.001	0.008	0.012	0.009	0.010	0.002	0.003	0.015	0.004	0.005	0.006