Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240563

Evaluating Synthetic Data Augmentation to Correct for Data Imbalance in Realistic Clinical Prediction Settings

Nina WAHLER [a,1], Bayrem KAABACHI [a], Bogdan KULYNYCH [a], Jérémie DESPRAZ [a], Christian SIMON [a] and Jean Louis RAISARO [a]
[a] *Lausanne University Hospital (CHUV), Switzerland*
ORCiD ID: Nina Wahler: https://orcid.org/0009-0003-5024-4492, Bayrem Kaabachi https://orcid.org/0009-0002-7534-8493, Bogdan Kulynych https://orcid.org/0000-0001-5923-3931, Jérémie Despraz https://orcid.org/0000-0002-2435-4079, Jean Louis Raisaro https://orcid.org/0000-0003-2052-6133

Abstract. Predictive modeling holds a large potential in clinical decision-making, yet its effectiveness can be hindered by inherent data imbalances in clinical datasets. This study investigates the utility of synthetic data for improving the performance of predictive modeling on realistic small imbalanced clinical datasets. We compared various synthetic data generation methods including Generative Adversarial Networks, Normalizing Flows, and Variational Autoencoders to the standard baselines for correcting for class underrepresentation on four clinical datasets. Although results show improvement in F1 scores in some cases, even over multiple repetitions, we do not obtain statistically significant evidence that synthetic data generation outperforms standard baselines for correcting for class imbalance. This study challenges common beliefs about the efficacy of synthetic data for data augmentation and highlights the importance of evaluating new complex methods against simple baselines.

Keywords. Synthetic Data, Imbalanced Data, Minority Oversampling

1. Introduction

Predictive modeling holds a large potential to help clinical decision making and improve patient outcomes [1]. Yet, the reliability of predictive models depends on the quality and representativeness of the data used for training them. In clinical datasets, predictive modeling can be challenging. A common issue is that the number of records in the data with a given value prediction target (e.g., disease) are significantly outnumbered by other values (e.g., healthy patients). Such class imbalance is natural due to either relatively small prevalence rates of diseases within populations or sampling bias in data collection. This can lead to models that are poorly calibrated and are biased towards the majority class, which critically undermines their utility in clinical settings.

Although there exist many standard methods for correcting for class imbalance, such as minority class oversampling [2] and SMOTE [3], they may fail when it comes to

[1] Corresponding Author: Nina Wahler; E-mail: nina.wahler@chuv.ch.

capturing the complexity of the minority class distribution. Recent advances in the generation of synthetic data offer a potential technical solution to this [4], as synthetic data could generate realistic synthetic instances of the underrepresented minority class [5]. In this work, we investigate the effectiveness of synthetic data augmentation for improving performance of predictive models in the context of realistic, small, imbalanced clinical datasets, aiming to answer the question whether this new line of methods can lead to better results than the standard methods for correcting for class underrepresentation.

2. Methods

2.1. Datasets and Tasks

We conduct our study on four different clinical datasets (Table 1). Two of these datasets are publicly available on the UCI Machine Learning Repository: "Pima Indian **Diabetes**" [6], Wisconsin **Breast** Cancer [7]. A third dataset is from the randomized clinical trial NCT00079274 of a **Colon** cancer [8] treatment, obtained via Project Data Sphere [9]. The fourth one is an internal dataset that was collected at Lausanne University Hospital from patients who have undergone a **Surgery** procedure in the Head and Neck Surgery department. Ethics approval for the use of the Surgery dataset has been given by the CER-VD and signed by Pr. Pierre-André Michaud (vote number: 2023-01258). We deliberately choose these clinical datasets to be diverse, and to represent different degrees of class imbalance:

Table 1. Datasets used in this study.

Dataset	*Records*	*Features*	*Target class proportion*	*Prediction Task*
Surgery	130	62	22 %	Post-surgery remission
Colon	2969	23	23%	Post-treatment mortality
Diabetes	768	9	35 %	Diabetes
Breast	569	32	38 %	Breast cancer

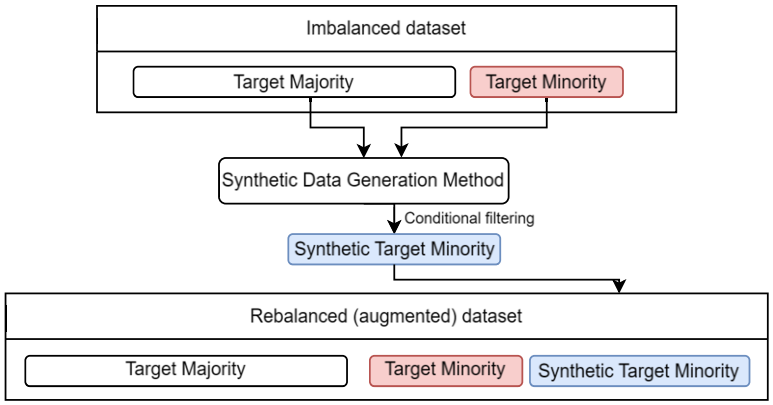

Figure 1. Minority class augmentation using synthetic data generators.

2.2. Correcting for Underrepresentation

To correct for underrepresentation, we use augmentation methods that range from classical method to more complex synthetic data generation models, to augment minority class samples. After that, we train a predictive model on the augmented dataset (see Figure 1). Following this procedure, we assess the performance of two standard machine learning models used in such settings—logistic regression (LR) and random forests (RF)—across the augmentation techniques. We use classical methods as baselines: Random Oversampling (**ROS**) [2] and **SMOTE** [3], and compare them to synthetic data augmentation using Bayesian Networks (**BayNet**) [10], **CTGAN** [11], Normalizing Flows (**NF**) [12] and **TVAE** [11]. To measure the effectiveness of a predictive model, we measure the F1 score for the target—the harmonic mean between precision and recall—a common metric employed in the setting of class imbalance. The metric captures two crucial performance measures in our setting as the healthcare outcome prediction tasks both require the accurate identification of the outcome (precision), and the ability to detect as many true cases as possible (recall). To evaluate the method in an unbiased way, we perform a stratified random split of the available dataset into train and test subsets. We augment and train the model on the train subset and evaluate the F1 score of the predictive model on the test subset. We repeat this evaluation process 50 times to ensure the robustness of the results.

3. Results

We observed no effect on the model performance after data augmentation neither with the standard methods nor the synthetic data augmentation methods on the Breast dataset (one-way dependent t-test p-value > 0.08). We thus focus on the Diabetes, Colon, and Surgery datasets.

In Figure 2, we present two key comparisons to highlight the effects of data augmentation on model performance over these datasets. First, we compare models trained without any augmentation to those utilizing the best-performing augmentation technique. Then, we contrast the best baseline method with the best synthetic augmentation approach. Instances where these differences are statistically significant are marked accordingly.

On the Surgery dataset, we observe evidence of improvement of the model performance over standard training when using Normalizing Flows although there is still no evidence that any of the synthetic data generation methods improves over the baselines. In our analysis of all the datasets, we consistently observe that synthetic data generation methods do not outperform the baseline methods. As a particular case, on the Colon dataset with the LR (Logistic Regression), we see that the baseline method (ROS) attains even higher performance than any of the synthetic methods with $p < 0.0001$, showcasing how synthetic might not necessarily improve model performance better than simple baselines.

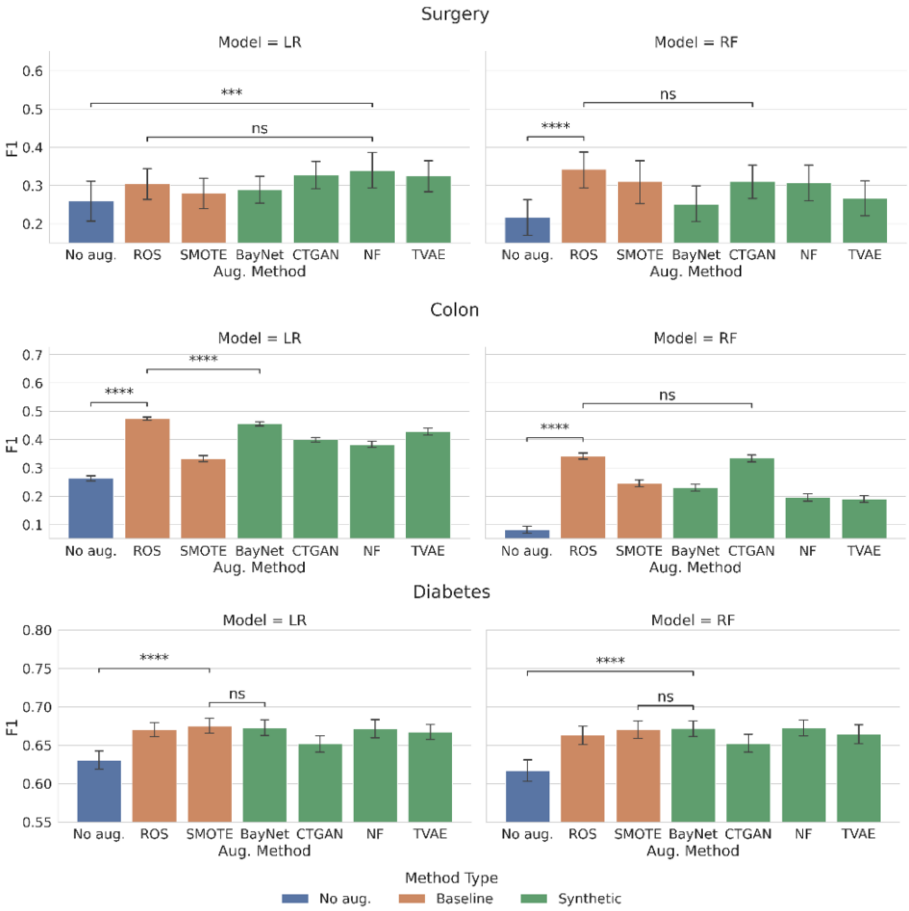

Figure 2. Results on the Surgery, Colon and Diabetes datasets. We indicate the statistical significance of the F1 score differences between selected method pairs for each dataset, determined via paired two-sided t-tests with Benjamini-Hochberg multiplicity adjustment. Specifically, we compare "no augmentation" with the best-performing method among the augmentation techniques, and then additionally compare the best baseline method against the top synthetic augmentation approach. We mark comparisons with ***ns*** when the test p-value > 0.05, with *** when $p < 0.001$, and with *** when $p < 0.0001$.

4. Discussion

The findings of this study shed light on the effectiveness of synthetic data generation techniques to improve predictive modeling performance on small imbalanced clinical datasets. Although some methods demonstrate promising results, e.g., Normalizing Flows improve the performance on the Surgery dataset, none of the differences in performance demonstrated statistically significant improvement compared to the baseline methods over multiple replications on randomly chosen training data subsets.

Our results challenge a common belief regarding the efficacy of synthetic data as a tool for data augmentation. Despite its potential benefits, the results of this study suggest that synthetic data may not always yield significant improvements on realistic, small datasets which are abundant in clinical practice. Even if they do, these complex and

computationally expensive methods might not outperform simple but strong baseline methods such as oversampling of the minority class. By highlighting the limitations of synthetic data and comparing its performance to standard class bias correction methods, we showcase the importance of critically evaluating the applicability of synthetic data in any given research context, and the importance of considering strong baselines.

One of the limitations of our study is the relatively small number of datasets and scenarios considered. Although the study aimed to provide a comprehensive analysis across multiple datasets and models, the scope of our investigation was constrained by the availability of suitable clinical datasets and computational resources. Future research could expand on these findings by incorporating additional datasets and exploring a broader range of experimental scenarios to further validate the generalizability of this paper's conclusions, e.g., with larger datasets.

References

[1] Yeung AWK, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health [Internet]. 2023 Sep 26 [cited 2024 Apr 3];11. Available from: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1196596/full

[2] Idrissi BY, Arjovsky M, Pezeshki M, Lopez-Paz D. Simple data balancing achieves competitive worst-group-accuracy. In: Proceedings of the First Conference on Causal Learning and Reasoning. PMLR; 2022. p. 336–51.

[3] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002 Jun 1;16(1):321–57.

[4] van Breugel B, van der Schaar M. Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data. arXiv; 2023 [cited 2024 Mar 14]. Available from: http://arxiv.org/abs/2304.03722

[5] Jordon J, Szpruch L, Houssiau F, Bottarelli M, Cherubin G, Maple C, et al. Synthetic Data -- what, why and how? [Internet]. arXiv; 2022 [cited 2024 Apr 3]. Available from: http://arxiv.org/abs/2205.03257

[6] Kahn M. Diabetes [Internet]. [cited 2024 Mar 18]. Available from: https://archive.ics.uci.edu/dataset/34

[7] William Wolberg OM. Breast Cancer Wisconsin (Diagnostic) [Internet]. 1993 [cited 2024 Mar 18]. Available from: https://archive.ics.uci.edu/dataset/17

[8] National Cancer Institute (NCI). A Randomized Phase III Trial of Oxaliplatin (OXAL) Plus 5-Fluorouracil (5-FU)/Leucovorin (CF) With or Without Cetuximab (C225) After Curative Resection for Patients With Stage III Colon Cancer. clinicaltrials.gov; 2020 Apr. Report No.: NCT00079274.

[9] Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, et al. The project data sphere initiative: accelerating cancer research by sharing data. The Oncologist. 2015 May;20(5):464-e20.

[10] Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Mach Learn. 1995 Sep 1;20(3):197–243.

[11] Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2019.

[12] Rezende D, Mohamed S. Variational Inference with Normalizing Flows. In: Proceedings of the 32nd International Conference on Machine Learning. PMLR; 2015. p. 1530–8.