Synthetically Rebalancing Healthcare Datasets via Conditional DDPM

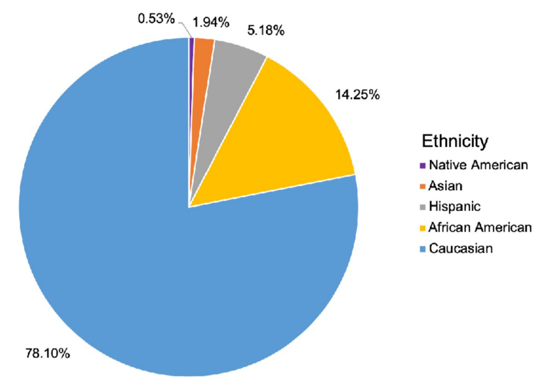
Keira Behal [1], Alina Chen [1], Caleb Fikes[2], Sophia Xiao[1]

[1] Emory University, [2] Rice University

Introduction

Problem

In recent years, machine learning algorithms have become increasingly important in healthcare for tasks like disease prediction, diagnostics, and treatment optimization. However, the training data for these tasks usually contains inherent biases or underrepresentation of certain groups, particularly minority communities.



Consequences

This bias issue is a significant concern in healthcare, where fair and equitable treatment is crucial. This imbalance can lead to inequitable health outcomes, with minority groups receiving less accurate diagnoses or treatment recommendations due to their underrepresentation in the training data.

Evaluating Classifier Fairness

Metric	Formula
Precision	$\frac{T_p}{T_p+F_p}$
Recall (Sensitivity)	$\frac{T_p}{T_p+F_n}$
F1 Score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$
AUROC	$\int_0^1 ROC(x)\,dx$

Methodology

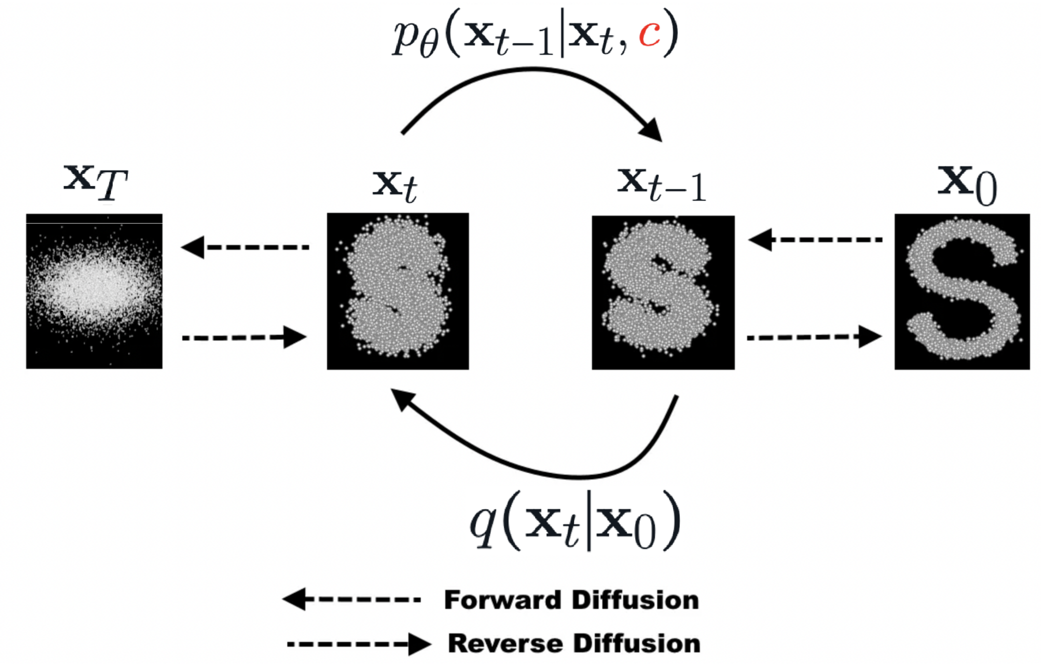
The MCRAGE algorithm, uses synthetic rebalancing to create a training set that equally represents all classes, encouraging demographic and classification parity.

Algorithm 4.1 MCRAGE

Let $s_1, \ldots, s_L$ be protected attributes.
$\bar{s} \leftarrow \Pi_\ell^L s_\ell$ (Cartesian Product)
$s \leftarrow \text{serialize}(\bar{s})$
$K \leftarrow \text{len}(s)$
$\hat{\pi}_k \leftarrow \mathbb{P}(s = k)$ for all $k \in \{1, \ldots, K\}$.
$k^* \leftarrow \arg\max_k \hat{\pi}_k$.
Train CDDPM $p_\theta(x_0|x_T, s = k)$ on original data
for $k \in \{1, \ldots, K\}$ **do**
  $x_k \leftarrow n(\hat{\pi}_{k^*} - \hat{\pi}_k)$ samples from $p_\theta(x_0|x_T, s = k)$.
end for
return $\{x_1, \ldots, x_K\}$

This approach ensures similar distribution of outcomes and fair evaluation of classifier performance for each subgroup, promoting intersectionally equitable performance.

CDDPM



Forward Diffusion Process	$q(\mathbf{x}_t|\mathbf{x}_0) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$ Where $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \Pi_{s=1}^t \alpha_s$
Reverse Denoising Process	$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, c), \beta_t \boldsymbol{I})$ $\mathbf{x}_t|\mathbf{x}_0 = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
Training	$L_{\text{simple}} = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)\|^2$

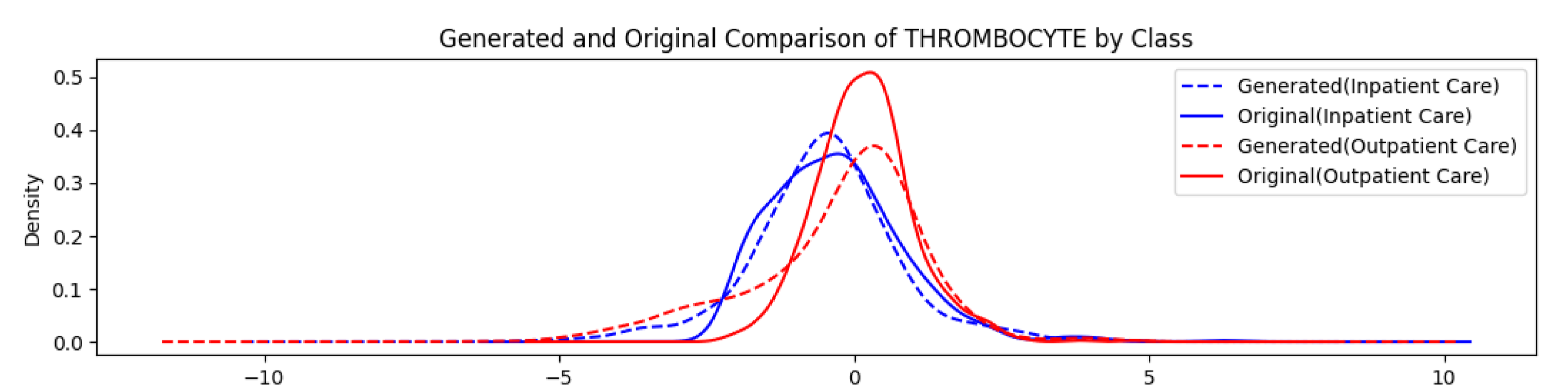
Benefits of CDDPM: Stable training, high-fidelity samples [3], generation of a specific class

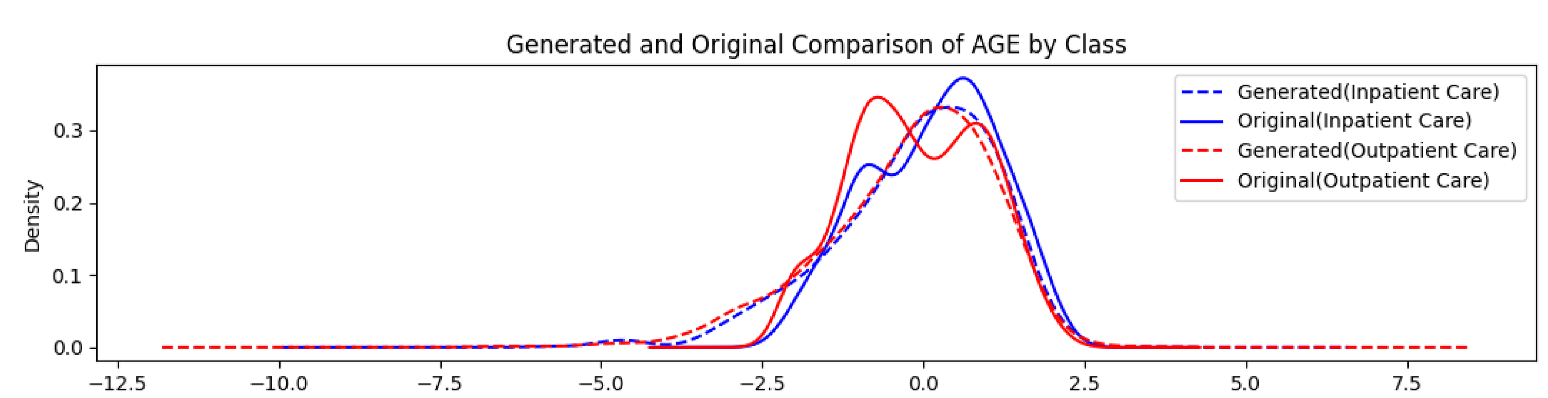
Theory of Convergence

Assume the following assumptions hold:

1. The sampling distribution π is supported on some compact subset $\mathcal{M} \subset \mathbb{R}^d$.
2. The noise schedule $t \mapsto \beta_t$ is continuous.
3. There is a continuous function $s : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ which efficiently approximates $\nabla \log p_t(x_t)$ for all t.

Under these assumptions, one can show that $\mathbf{W}_1(\mathcal{L}(\mathbf{Y}_T), \pi)$ is bounded above in terms of iteration complexity and thus the distribution of samples generated by a DDPM converges to the true sampling distribution where $\mathbf{W}_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma} d(x, y)$ [1].





Results

Dataset

The dataset consists of Electronic Health Records collected from a private hospital in Indonesia, containing 8 different laboratory test results for 3309 patients. [4]

Fairness Evaluation:

	Imbalanced	SMOTE [2]	Augmented	Balanced
Accuracy	60.423	70.039	70.01	70.468
F1 Score	0.00	0.598	0.635	0.599
AUROC	0.5	0.68	0.69	0.68

Synthetic Sample Quality Visualizations:

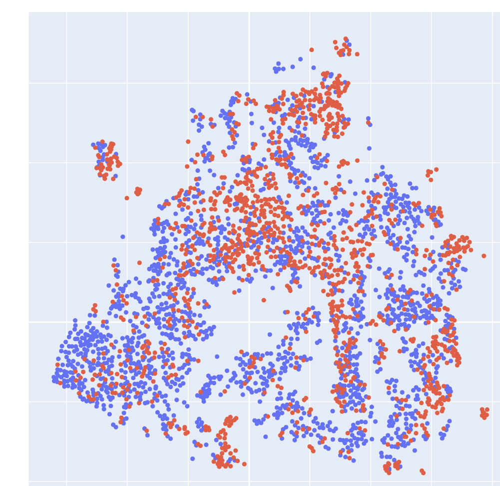


Figure 1: Manifold Projection of Original Dataset

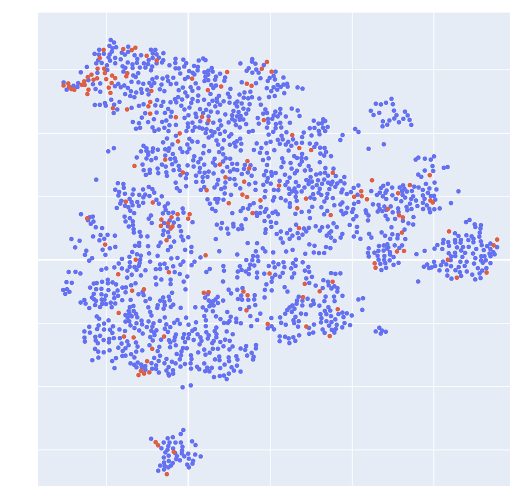


Figure 2: Manifold Projection of Imbalanced Dataset

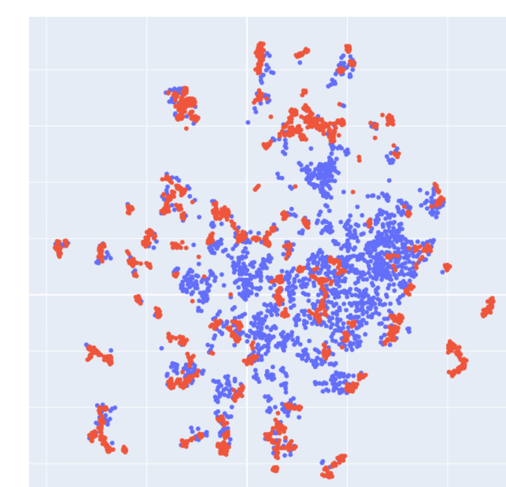


Figure 3: Manifold Projection of SMOTED Dataset

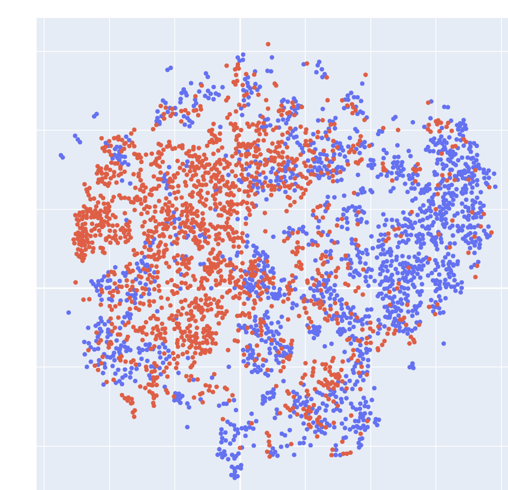


Figure 4: Manifold Projection of Augmented Dataset

Conclusions: In general, classifiers trained on synthetically balanced datasets performed better than those trained on unbalanced datasets. In comparison to SMOTE, MCRAGE produced higher quality samples with better classifier performance.

Acknowledgements

This work is sponsored by the US NSF Research Experience for Undergraduates program. We would like to thank Dr. Yuanzhe Xi for his mentorship during the REU program. We would also like to acknowledge Huan He and Tianshi Xu for their collaboration and insight.

References

[1] V.D. Bortoli. *Convergence of denoising diffusion models under the manifold hypothesis.* 2023.
[2] K. W. Bowyer et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: (2011).
[3] J. Ho et al. *Denoising Diffusion Probabilistic Models.* Neurips, 2020.
[4] https://www.kaggle.com/datasets/manishkc06/patient-treatment-classification.