Mitigating Bias in Medical Data Through a Machine Learning Approach Utilizing Generative Adversarial Networks to Create Diverse Datasets for Underrepresented Groups

Gabriel Xu and Avnith Vijayram

Thomas Jefferson High School for Science and Technology

SCIENCE FAIR RESEARCH REPORT

Abstract

In recent decades, many Machine Learning (ML) models have been utilized to solve problems in medicine, such as analyzing medical images and diagnosing diseases. However, a major obstacle towards building more equitable ML applications is implicit bias against certain groups of people in medical datasets. Especially in the field of medicine, it is essential that ML model training data is both diverse and complete.

We analyzed three public medical datasets with patient information and split the patients in each dataset into several groups based on factors such as ethnicity, gender, and age. Then, we utilized Conditional Tabular Generative Adversarial Networks (CTGANs) to supplement each group with augmented data, providing additional samples for groups with lesser representation. For each group, we trained the same Random Forest Classifier (RFC) model on both the original and augmented data and compared their performances using accuracy and recall metrics.

In datasets with more imbalance between groups, we generally saw a more pronounced improvement in performance (in both accuracy and recall metrics) for the RFC model trained on the augmented data as compared to the original data. Additionally, majority groups generally showed less improvement after augmenting their dataset, and their accuracy and recall did not change at all in several cases.

In our study, we demonstrated how augmenting training data with CTGANs can create a more robust dataset and improve the efficacy and equitability of ML models. Future research can address specific use cases and improve on our results with a more specialized generative model.

Background

The rapid development of Machine Learning (ML) algorithms and models in recent decades has led to an increase in applications of Artificial Intelligence (AI) built to address problems in medicine and healthcare, such as analyzing medical images and diagnosing diseases. However, as ML continues to revolutionize the field of medicine, it is also important to consider the risks associated with it. A major limitation of ML models that is especially crucial in the field of medicine is the possibility that models may reinforce bias present in training data.

Generative Adversarial Networks (GANs) are a recently developed machine learning technology used for generative modeling, which is the process of generating new data that is similar to an original dataset. A GAN consists of two sub-models: a generator that creates new examples to mimic the original data, and

a discriminator, which tries to distinguish between the generated data and real data. These two models are trained together until the generator begins to fool the discriminator more than half of the time, at which point we conclude that the generator has started to produce realistic data.

GAN models were originally developed with image and text generation in mind. However, a major portion of medical data is tabular, consisting of columns of recorded observations for each patient, which makes it difficult to apply standard GANs directly. In 2019, researchers developed Conditional Tabular Generative Adversarial Networks (CTGANs), a technology that specializes in generating tabular data and outperforms conventional GANs and related algorithms at creating synthetic data. Due to its recent conception, CTGANs and their potential applications have not yet been fully investigated.

Purpose

A major obstacle to building more equitable machine learning applications in healthcare is the implicit bias present in medical data. For instance, several ethnic groups have historically been underrepresented in biomedical datasets that should have been representative of the human population. Because ML models learn from large datasets, models trained on biased data often end up reinforcing that bias. In healthcare, these biases may lead to unfair and even dangerous consequences for certain groups of patients, since a model may be more likely to make a misdiagnosis if it was trained on incomplete data.

One way bias manifests in a dataset is when there are significantly more samples for certain groups. For instance, in the publicly available Breast Cancer dataset we used, the White ethnicity group makes up more than 80% of the data. This same pattern repeats in the Patient Survival and Stroke datasets across several factors, including ethnicity, gender, and age.

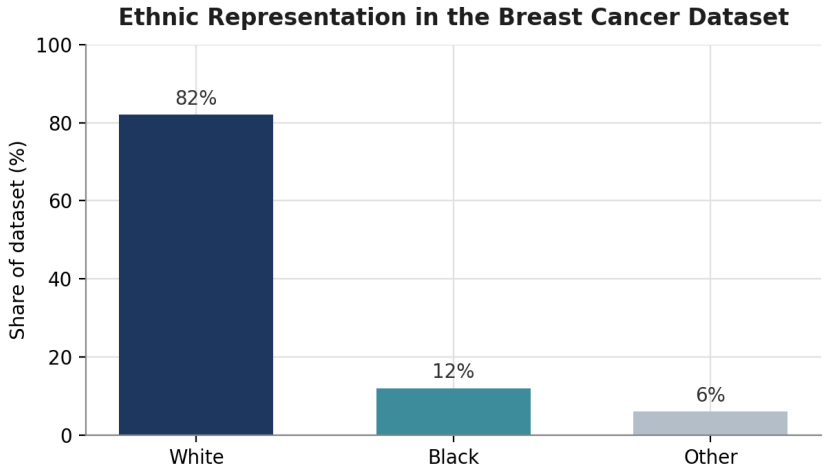

Figure 1. Representative class imbalance in the Breast Cancer dataset, where a single ethnic group accounts for the large majority of samples.

Our main research question is: given a limited dataset where implicit bias arises from a lack of data for certain groups of people, how can we use CTGANs to generate data that is more representative? We aim to design a CTGAN model that can generate reliable and representative samples to make up for the lack

of data on underrepresented groups, and in doing so support the further unbiased advancement of ML applications in medicine.

Procedure

For each of our three datasets, we first preprocessed and cleaned the data, which included removing missing values and encoding categorical variables. We then determined patient groups based on ethnicity, gender, age, or a combination of those factors; one of our groups was the entire dataset. Each group was further separated into training and testing data. A CTGAN model was fitted to the training data and used to create augmented training data, adding new rows with generated values for labels such as age, marital status, T stage, N stage, and tumor size.

We then trained the same Random Forest Classifier (RFC) model on both the original and augmented data and evaluated both on the test data using accuracy and recall metrics. To check whether augmentation led to improvements across groups, we also evaluated each random forest model using test data drawn from other groups.

Figure 2. Overview of the experimental pipeline applied to each of the three datasets.

Datasets

We used three publicly available medical datasets, each exhibiting imbalance across one or more demographic factors:

- **Breast Cancer** — imbalanced primarily by ethnicity (White ethnicity > 80% of samples).
- **Stroke Prediction** — imbalanced across factors including gender and age.
- **Patient Survival** — imbalanced across several demographic factors.

Statistical Analysis (ANOVA)

To analyze our results, we conducted two-sample ANOVA tests with $\alpha = 0.05$. For each dataset, we carried out two ANOVAs: one compared RFC accuracy scores within each patient group after augmentation, and the other compared RFC accuracy across all combinations of training and testing groups. Statistically significant results ($p < 0.05$) are highlighted in green below.

Within Groups

Within Groups	Breast Cancer	Stroke	Survival
Patient Groups	**0.0049**	**0.009394**	**0.002105**
Augmentation	0.3423	**0.01109**	0.208

Across Groups

Across Groups	Breast Cancer	Stroke	Survival
Patient Groups	**4.085e-10**	**7.635e-9**	**3.201e-17**
Augmentation	**0.0139**	0.133	**0.006393**

Cells shaded green denote statistically significant results at $\alpha = 0.05$.

Discussion

In our study, we demonstrated how augmenting training data with CTGANs can create a more robust dataset and improve the efficacy and equitability of ML models. In datasets with more imbalance between groups, we generally saw a more pronounced improvement in performance, in both accuracy and recall, for the RFC model trained on the augmented data as compared to the original data. Additionally, majority groups generally showed less improvement after augmentation, and their accuracy and recall did not change at all in several cases.

Our study was limited mainly by the large amount of data and computational power required for training GANs. Several of our patient groups across multiple datasets did not contain enough samples to train our CTGAN model most effectively. In addition, generative models are known for being very computationally intensive, and the free software we used to build our models was not as efficient as we would have liked.

In conclusion, our research shows promise for specific health-related datasets, but cannot necessarily be generalized to all areas where machine learning models are applied. Future research can address specific use cases and improve on our results with a more specialized generative model.

References

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN. *Advances in Neural Information Processing Systems (NeurIPS) 32*. proceedings.neurips.cc

Breast Cancer dataset (Kaggle). kaggle.com/datasets/reihanenamdari/breast-cancer

Stroke Prediction dataset (Kaggle). kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Patient Survival dataset (Kaggle). kaggle.com/datasets/mitishaagarwal/patient