Mitigating Bias in Medical Datasets: A Comparative Analysis of Generative Adversarial Networks (GANs) Based Data Generation Techniques*

Mohamed Ashik Shahul Hameed[1,*,†], Asifa Mehmood Qureshi[1,*,†] and Abhishek Kaushik[1,*,†]

[1]Regulated Software Research Centre (RSRC), Dundalk Institute of Technology, Dundalk, Ireland

Abstract

The increasing use of Artificial intelligence (AI) in the medical domain has highlighted a critical issue: bias in datasets. Biases in medical datasets can lead to skewed predictions, unfair clinical decisions, incorrect diagnoses and poor generalisation of AI models. Very often, these biases are the consequence of imbalance in the dataset. Generative Adversarial Networks (GANs) have appeared to be a promising solution for solving the data imbalance issue. Synthetic data can help mitigate bias by balancing the dataset for sensitive attributes as well as for class labels. However, the efficiency of different GAN variants in mitigating bias remains unexplored in the medical domain. This paper investigates and compares various GAN variants to identify the most effective approach to producing balanced data. In this study, we evaluated different variants of GAN on three medical datasets with the aim of contributing to the development of more fairer and inclusive AI models in the medical domain. The study shows that the performance of the Machine Learning (ML) model improves when the dataset is balanced using synthetic data samples. Moreover, the MedGAN variant performs better when compared with other variants of GAN.

Keywords

Bias, fairness, medical datasets, GANs, TGAN, CTGAN, MedGAN, MC-MedGAN

1. Introduction

Bias in Artificial Intelligence (AI) models refers to AI systems that produce biased results that reflect and amplify human prejudices within a community, encompassing past and contemporary social injustices [1]. These biases when replicated in medical datasets can have life-threatening consequences due to incorrect diagnosis or treatment recommendations [2, 3]. For example, German researchers built a skin cancer detection system using neural networks in 2016. The system was able to detect 95% of melanoma cases accurately. It was trained on 10,000 skin images and outperformed 58 dermatologists. Later, it was found that the data was highly dominated by white skin images and did not generalise well to a diverse population [4]. These biases can be handled at pre-processing, algorithmic level or in the post-processing stages of an AI model development [5]. Handling bias would help achieve fair models that do not discriminate against different groups and treat them unfairly [6]. Pre-processing techniques involve handling bias at the data level. One of the widely used techniques to mitigate bias is over-sampling. Over-sampling is the generation of synthetic data that mirrors the characteristics of real-world data. It helps to reduce bias by balancing the representation of different demographic groups so that machine learning models produce reasonable outcomes and generalise well over a diverse population [7]. There are several techniques to generate synthetic data to ensure fairness in medical datasets [8]. These techniques include SMOTE [9], FairSMOTE [10], BorderlineSMOTE [11], and Cluster-based over-sampling [12]. Moreover, deep learning is also widely used to generate artificial data because of its high efficiency and accuracy in generating data. The most commonly used algorithm is the Generative Adversarial Network (GANs) that have gained immense popularity in the research

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland

*Corresponding author.

†These authors contributed equally.

D00253306@student.dkit.ie (M. A. S. Hameed); asifa.mehmood@dkit.ie (A. M. Qureshi)

0009-0009-3062-2971 (M. A. S. Hameed); 0009-0002-4312-353X (A. M. Qureshi)

community [13].

GAN is a deep learning model that mainly consists of two neural networks: a Generator used to generate artificial data and a discriminator that tries to distinguish between real and synthetic data to improve quality. These models were first introduced to process only image data, but later different variants of GAN were proposed to process tabular data as well. These variants include Tabular GAN (TGAN) [14], Conditional Tabular GAN (CTGAN) [15], Medical GAN (MEDGAN) [16], Multi-Categorical GAN (MC-MedGAN) [17] and many more.

In this study, we evaluated various GAN variants including GAN, TGAN, CTGAN, MedGAN, and MC-MedGAN to generate synthetic samples to balance different group representations within medical datasets. The newly balanced dataset was fed into different ML models including Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbour (KNN) to draw a comparison. The GAN models are evaluated on three different medical datasets that consist of gender as a sensitive attribute to balance: the Asthma Disease Dataset [18], the Heart Disease Prediction Dataset [19], and the Cancer Prediction dataset [20]. The performance is evaluated using various metrics i.e., accuracy, precision, F1-score, recall, and Area Under Curve (AUC) scores. Fairness is evaluated using Equal Opportunity (EO) [21], Propensity Score (PS) [22], and Statistical Parity (SP) [23].

2. Motivation

In today's world, AI is an integral part of the healthcare system. The AI model must incorporate transparency and accountability. The goal of this research is to reduce bias in medical datasets that contain inherent biases due to unequal representation of different demographic groups. AI models can become unfair and imbalanced, particularly in the healthcare sector, where underrepresented groups may receive scant care. Bias in medical datasets poses a significant challenge to the reliability of predictive models [24]. This could be critical for healthcare systems since an automated model prediction has a direct effect on patients that affects their mental health, and quality of life or may risk the life of an individual [25] as well it also leads to financial loss [26]. Due to an unbiased dataset, certain populations may receive incorrect diagnoses or treatments as a result of unreliable predictions brought on by bias in datasets. Nonetheless, GANs provide a potentially helpful way to generate AI data that can assist in balancing underrepresented groups in health databases. The aim to explore how GAN-based techniques can eliminate bias through data augmentation and enable more reliable and equitable Machine Learning (ML) models motivates this effort [13]. The comparative study's main goal is to identify the optimal variant to lessen bias in medical datasets. We want to improve the quality of treatment by lowering bias and ensuring that AI systems generate reliable, accurate, and equitable forecasts for a range of demographics. Therefore, the motivation of this study is to investigate different variants of GAN including TGAN, CTGAN, MedGAN, and MC-MedGAN for their efficacy in mitigating bias and improving predictive performance on multiple medical datasets. This work will serve as a foundation for further experimentation on data generation via GAN to mitigate biases.

Hypothesis: GAN-based data generation methods can help to reduce biases and ensure fairness in medical datasets.

The formulated research questions to explore the above hypothesis are as follows:

- Does GAN-based synthetic data generation help reduce biases in medical datasets? If yes, which GAN variant performs better among basic GAN, TGAN, CTGAN, MedGAN, and MC-MedGAN?

The rest of the article is structured as follows: Section 3 highlights some of the recent related work. Section 4 explains the methodology in detail. Section 5 explains the results. Section 6 discusses the hypothesis and research questions and Section 7 concludes the discussion with future work.

3. Related work

GANs have gained significant attention in recent years due to their capability of generating high-quality data. Therefore, this section reviews the recent methodologies that leverage GAN models to generate synthetic data. A study [27] presents the potential of GANs in generating synthetic data from observational health data and discusses some of the unique challenges associated with healthcare datasets, such as concerns about class imbalance. Observational Health Data (OHD) is highly valuable for medical research and health informatics. The use of such data is severely limited because of strict regulations. It highlights that GAN-generated synthetic data can help overcome some of the common challenges, such as bias, privacy and class imbalance. The authors argue that GANs are useful in generating healthcare data to combat the scarcity of high-quality medical datasets. Moreover, to address the challenges of drift and class imbalance of gas detection systems, [28] employed CTGAN for data augmentation. The result shows a significant improvement in the classification accuracy of each class for both Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) thus reducing bias toward the majority class. They conclude that CTGAN provides a feasible solution to generate a balanced dataset.

In another study [29] various variants of GAN including CTGAN, TGAN, and Wasserstein GAN (WGAN) are utilised for the anonymisation of real data through data synthesis. These models were compared for precision, recall, and coverage scores to evaluate the generation of realistic tabular data, handling missing and class-imbalanced data, and ensuring privacy. The results show that, although no GAN method performs best in each evaluation metric, CTGAN and TGAN produce better scores in most of the evaluation metrics. Additionally, in [30] a new variant of GAN called Multi-label Time-series GAN (MTGAN) is proposed to generate sequential Electronic Health Record (EHR) data using a gated recurrent unit with a smooth conditional matrix, while the critic evaluates temporal features using Wasserstein distance for improving the quality of synthetic data. The results show that MTGAN generates realistic EHR data effectively and improves accuracy for uncommon diseases.

The above studies show that GANs have the potential to generate high-quality diverse datasets that can be used to handle bias in real-world datasets. Therefore, to analyse the capabilities of different GAN variants, this study aims to conduct multiple experiments and then assess the fairness within the newly generated synthetic medical datasets.

4. Methodology

Figure 1 shows the systematic methodology diagram used to evaluate the different variants of GANs. First, the data is preprocessed and split into standard train and test sets. Then, the data is fed into the GAN variant to generate synthetic data. The newly generated data is augmented with the real data to balance the number of samples for the sensitive attributes and the output label. Afterwards, ML models are trained on the newly generated data to evaluate the overall performance as well as the fairness of the models.

4.1. Preprocessing

The data preprocessing includes one hot encoding to replace categorical variables with numerical numbers. Afterwards, we applied z-score normalisation on each distinct numerical feature because they did not contain extreme outliers [31]. Normalisation helps to specify each variable within a specified range to simplify the model-learning process [32]. Then, the resulting dataset is split into a 70:30 ratio for train and test sets.

4.2. Generate synthetic data

In order to balance the dataset for the sensitive attribute i.e., Gender and class labels. We employed five GAN architectures: basic GAN, TGAN, CTGAN, MedGAN and MC-MedGAN. These variants are

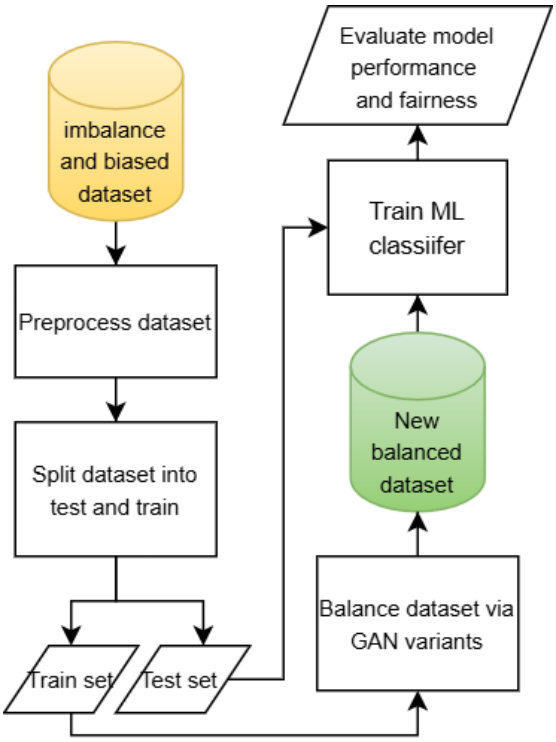


Figure 1: Systematic Methodology Diagram to Evaluate GAN Variants

specifically designed to handle tabular and medical datasets which is the primary focus of our study. GAN is a type of neural network architecture where two networks, a generator, and a discriminator, are trained simultaneously [33, 34, 35]. Tabular GAN is an application-driven variant of the GAN that is designed to generate synthetic tabular data, containing rows and columns like in a spreadsheet or database [33, 14, 36]. The CTGAN is an extension to Tabular GAN that generates synthetic tabular data while taking into consideration the distribution of dependent target variables. This will help associate relations between columns and observe dependence relationships [15]. MedGAN is a specialised version of GAN that generates synthetic data in the medical field, mainly in tabular form containing sensitive information [16]. MC-MedGAN is a variant of MedGAN designed for handling multi-categorical variables, commonly present in medical datasets [17].

4.3. Train ML classifiers

After generating synthetic samples to balance the datasets for sensitive attribute (gender) and class labels, different commonly used ML classifiers including Logistic regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbour (KNN) with default parameters are trained on the newly generated datasets to evaluate the performance of GAN variants.

4.4. Datasets

To evaluate the performance of GAN variants, we used three different medical datasets that contain sensitive attributes. The details of each of these datasets are as follows:

Asthma Disease Dataset: The Asthma Disease Dataset [18] contains a record of 2,392 samples with 28 features. The output label is the diagnosis indicator, which is taken as 0 for the absence and 1 for a positive case. It contains 2,268 samples for class 0 as compared to 124 samples with class label 1. Also, the number of samples for males is 1212 whereas for females the count is 1180.

Heart Disease Prediction Dataset: The Heart Disease Prediction Dataset [19] consists of 13 features and 303 samples. The dataset contains 207 male and 96 female samples.

Cancer Prediction Dataset: The Cancer Prediction Dataset [20] contains 1,500 samples with 8 features. The target variable *'diagnosis'* indicates whether a patient has cancer or not (0 for no cancer and 1 for cancer). The diagnosis distribution shows 943 patients without cancer and 557 with cancer. There are 736 female samples and 764 males in total.

5. Results

The performance is evaluated by training different ML classifiers as mentioned in Section 4. The classifiers are assessed using accuracy, f1-score, precision, recall and AUC. Whereas the fairness of the dataset is evaluated via EO, PS, and SP. EO guarantees that all individuals receive the same treatment and meet the same requirements [21]. PS can be defined as the conditional probability of being exposed to a treatment given the observed covariates [22]. SP is a fairness criterion that requires the probability of a favourable outcome to be the same for each demographic group [23]. Tables 1, Table 2, and Table 3 show each classifier's performance on the original as well as on each generated dataset. It can be seen that MedGAN performs well for the Asthma Disease Dataset and Cancer Prediction Dataset while MC-MedGAN has a better score for the Heart Disease Dataset.

Table 1
Accuracy, F1-score, Precision, Recall and AUC score comparison over the **Asthma Disease Dataset**

Method	Model	Accuracy	F1-score	Precision	Recall	AUC
Original Dataset	LR	0.4989	0.4594	0.4951	0.4285	0.4996
	RF	0.4926	0.5050	0.4901	0.5210	0.4628
	DT	0.4864	0.4029	0.4770	0.3487	0.4559
	KNN	0.4926	0.4840	0.4892	0.4789	0.4903
GAN	LR	0.5381	0.5283	0.5600	0.5000	0.5487
	RF	0.7124	0.6967	0.7667	0.6383	0.7912
	DT	0.6662	0.6862	0.6680	0.7053	0.6648
	KNN	0.5958	0.6128	0.6074	0.6183	0.6047
TGAN	LR	0.5147	0.4892	0.5380	0.4485	0.5254
	RF	0.7267	0.7064	0.7967	0.6345	0.7797
	DT	0.6689	0.6932	0.6666	0.7221	0.6669
	KNN	0.5691	0.5878	0.5827	0.5929	0.5912
CTGAN	LR	0.5103	0.4952	0.5428	0.4553	0.5119
	RF	0.7425	0.7248	0.8309	0.6427	0.7920
	DT	0.6563	0.6952	0.6532	0.7429	0.6509
	KNN	0.5597	0.5720	0.5871	0.5577	0.5786
MedGAN	LR	0.5272	0.5140	0.5472	0.4845	0.5425
	RF	**0.7409**	**0.7233**	**0.8054**	**0.6563**	**0.7883**
	DT	0.6715	0.6980	0.6640	0.7356	0.6694
	KNN	0.5756	0.5898	0.6114	0.5695	0.5858
MC-MedGAN	LR	0.5134	0.5128	0.5167	0.5134	0.5084
	RF	0.6927	0.6916	0.7015	0.6927	0.7635
	DT	0.6226	0.6172	0.6238	0.6226	0.6220
	KNN	0.5711	0.5706	0.5705	0.5711	0.6062

Figure 2, shows the fairness metric performance on the Asthma Disease dataset. The SP, PS, and EO scores improve when the dataset is balanced for class label and gender. MEDGAN has a better performance for all three datasets followed by MC-MedGAN and TGAN. The same performance is observed for the other two datasets. The other graphs are given in Appendix A.

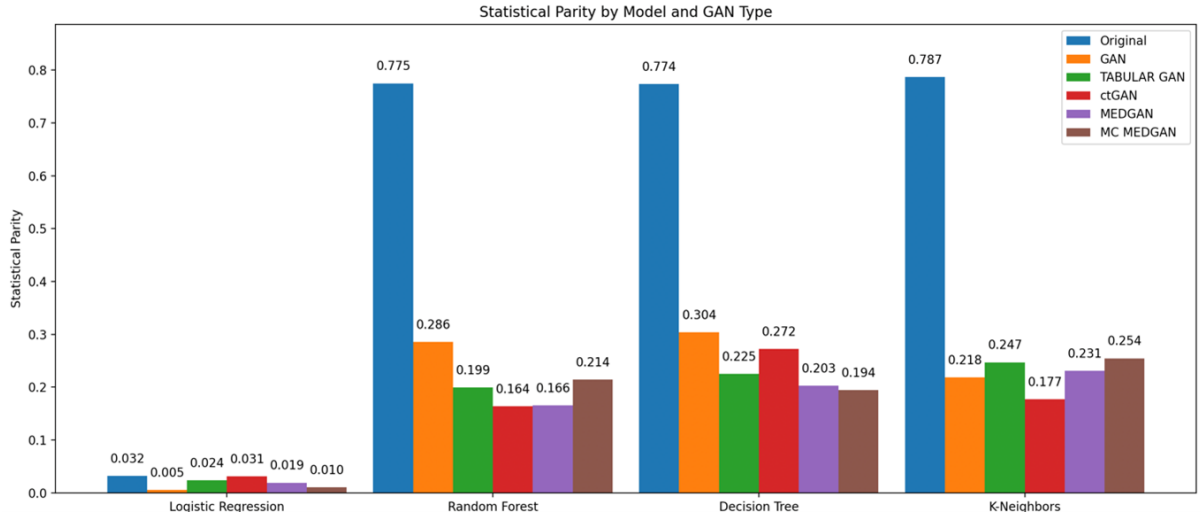
Overall, the results show that balancing the dataset for class labels and sensitive attributes improves the performance as well as the fairness of the model. Among different GAN variants, the MEDGAN produces good results and lower statistical, propensity and equal opportunity scores showing its great capability for reducing bias followed by MC-MedGAN. Moreover, the predictive ability of RF classifiers is better than other classifiers in terms of accuracy, precision, recall, f1-score, and AUC.

Table 2
Accuracy, F1-score, Precision, Recall and AUC score comparison over the **Heart Disease Prediction Dataset**

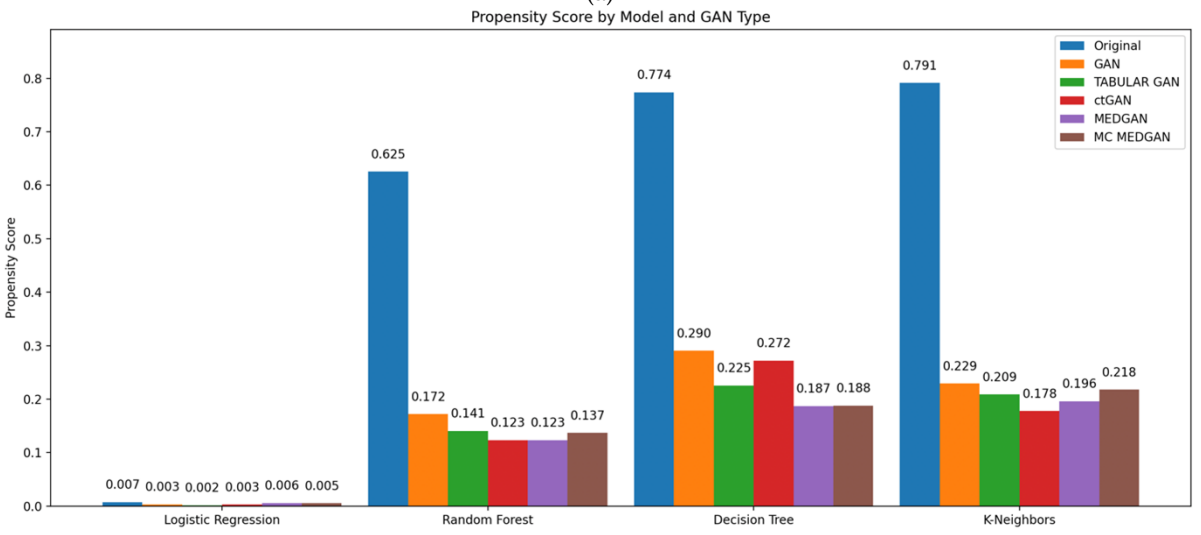
Method	Model	Accuracy	F1-score	Precision	Recall	AUC
Original Dataset	LR	0.7049	0.8125	0.8125	0.8125	0.6522
	RF	0.8032	0.8775	0.8600	0.8958	0.7996
	DT	0.6885	0.7654	0.9393	0.6458	0.8261
	KNN	0.7868	0.8631	0.8723	0.8541	0.7203
GAN	LR	0.7126	0.8166	0.8567	0.8957	0.7039
	RF	0.8915	0.9010	0.9318	0.8723	0.9621
	DT	0.8915	0.8988	0.9523	0.8510	0.8977
	KNN	0.8674	0.8791	0.9090	0.8510	0.9255
TGAN	LR	0.7349	0.8441	0.8205	0.8808	0.7352
	RF	0.8915	0.9010	0.9318	0.8723	0.9621
	DT	**0.9277**	**0.9333**	**0.9767**	**0.8936**	**0.9329**
	KNN	0.8674	0.8791	0.9090	0.8510	0.9137
CTGAN	LR	0.7250	0.8153	0.8500	0.9217	0.6876
	RF	0.9083	0.9197	0.9264	0.9130	0.9766
	DT	0.8250	0.8292	0.9444	0.7391	0.8975
	KNN	0.8750	0.8800	0.9821	0.7971	0.9903
MedGAN	LR	0.7108	0.8891	0.8355	0.8734	0.7340
	RF	0.9036	0.9130	0.9333	0.8936	0.9598
	DT	0.8674	0.8791	0.9090	0.8510	0.9021
	KNN	0.8674	0.8791	0.9090	0.8510	0.9284
MC-MedGAN	LR	0.7746	0.9096	0.9510	0.8382	0.7133
	RF	**0.9277**	**0.9347**	**0.9555**	**0.9148**	**0.9728**
	DT	**0.9277**	**0.9333**	**0.9767**	**0.8936**	**0.9320**
	KNN	0.8433	0.8539	0.9047	0.8085	0.9414

Table 3
Accuracy, F1-score, Precision, Recall and AUC score comparison over the **Cancer Prediction Dataset**

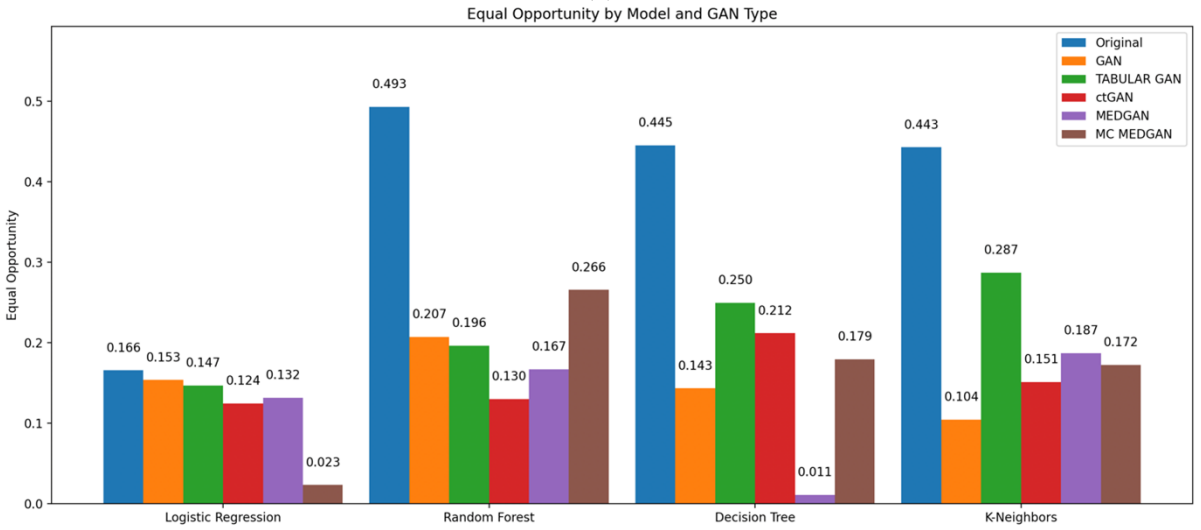
Method	Model	Accuracy	F1-score	Precision	Recall	AUC
Original Dataset	LR	0.6000	0.5714	0.5882	0.5555	0.6846
	RF	0.6133	0.5671	0.6129	0.5277	0.6588
	DT	0.5666	0.5608	0.5460	0.5763	0.6028
	KNN	0.6366	0.5657	0.6635	0.4930	0.6586
GAN	LR	0.6218	0.5776	0.5944	0.6431	0.6898
	RF	0.7527	0.7580	0.7500	0.7661	0.8453
	DT	0.6890	0.7154	0.6656	0.7733	0.6881
	KNN	0.7090	0.7359	0.6798	0.8021	0.8002
TGAN	LR	0.6549	0.5891	0.5920	0.5921	0.6958
	RF	0.7271	0.7478	0.7317	0.7647	0.8331
	DT	0.6849	0.7304	0.6676	0.8062	0.6774
	KNN	0.7161	0.7574	0.6914	0.8373	0.8369
CTGAN	LR	0.6696	0.6458	0.5958	0.5735	0.7680
	RF	0.7287	0.7349	0.7375	0.7323	0.8342
	DT	0.7269	0.7409	0.7224	0.7605	0.7260
	KNN	0.6690	0.7053	0.6498	0.7711	0.7836
MedGAN	**LR**	**0.7745**	**0.6698**	**0.5899**	**0.6563**	**0.7057**
	RF	0.7071	0.7028	0.7230	0.6836	0.7852
	DT	0.6795	0.7119	0.6534	0.7818	0.6782
	KNN	0.7071	0.7371	0.6757	0.8109	0.8123
MC-MedGAN	LR	0.7452	0.5923	0.5977	0.7090	0.7694
	RF	0.7005	0.7011	0.7116	0.6909	0.7782
	DT	0.6635	0.6862	0.6524	0.7236	0.6625
	KNN	0.6543	0.6867	0.6366	0.7454	0.7558



(a)



(b)



(c)

Figure 2: Fairness Assessment for Asthma Disease dataset (a) Statistical Parity (b) Propensity Score, (c) Equal Opportunity

6. Discussion

This section discusses the overall findings of the study in view of the literature review and extensive experimentation conducted to analyse our hypothesis. Based on our research question, the experiments show that classifier performance as well as the fairness metrics score improves when the datasets are balanced for sensitive attributes and class labels. Figure 2 shows the improvement in the fairness scores across each metric when the dataset is balanced via synthetic data generation using GAN variants as compared to the original dataset. Moreover, the analysis of each GAN variant based on performance evaluation using accuracy, precision, F1-score, recall, AUC and fairness metrics via EO, PS, and SP indicates that the MedGAN produces efficient performance followed by MC-MEDGAN across all three datasets. To validate any statistically significant difference between these two methods, we applied a paired t-test on the EO, PS, and SP scores for each of these methods. The p-values for EO, PS, and SP came out to be 0.34, 0.61, and 0.30 respectively. Therefore, we fail to reject our hypothesis and conclude that these two methods are not significantly different. These GAN variants are specifically designed for medical datasets to capture the interdependencies between the different variables to generate synthetic data similar to original data properties [16, 17]. However, further experimentation with other datasets including post-hoc tests will be conducted in future to provide deeper insights into the capability of GAN variants for data generation.

7. Conclusion and future work

In this paper, we tested different types of GANs for their capacity to produce synthetic tabular data to decrease bias in medical datasets. Our key findings are GAN-based models are effective for bias migration and GAN can provide a balanced dataset to produce generalised AI models and provide a solution *AI for all* and *AI for good*. On the other hand, traditional GANs were successful but medical domain-based GANs displayed greater performance in generating high-quality and unbiased data. It drives us to have more specific models in the future. Despite certain advantages of the GAN, we face some obstacles such as evaluation metrics. There is a need to have more standardised and compressive evaluation metrics of this model focused on decreasing bias. The studies in this article suggest that synthetic data can assist in eliminating bias and improve the effectiveness of the classifier. Moreover, MedGAN performs better in terms of SP, PS, and EO. In future, we will extend our work for various variations of GAN focused on refining GAN architecture to adapt the multimodality medical data, bias-sensitive evaluation mechanism and testing the GAN-based techniques in real-world clinical data.

Acknowledgments

This research was managed by the CREATE-DkIT project, supported by the HEA's TU-Rise programme and co-financed by the Government of Ireland and the European Union through the ERDF Southern, Eastern Midland Regional Programme 2021-27 and the Northern Western Regional Programme 2021-27. This research is also partially supported by the Research Ireland under Grant Number 21/FFP-A/9255.

References

[1] N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large language models struggle to learn long-tail knowledge, in: International Conference on Machine Learning, PMLR, 2023, pp. 15696–15707.

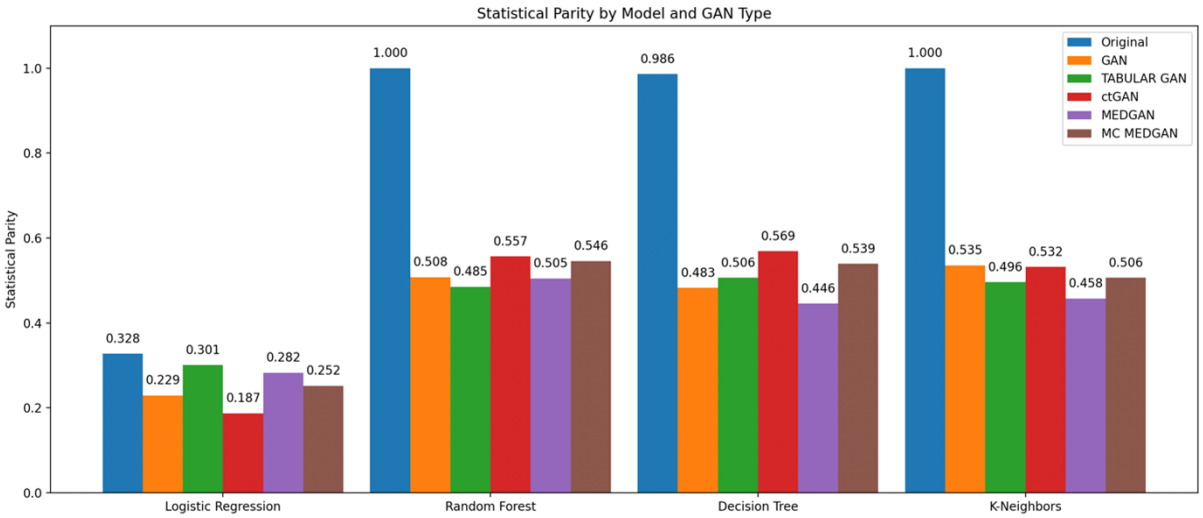
[2] E. Ferrara, Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, Sci 6 (2023) 3.

[3] J. Vaughn, A. Baral, M. Vadari, W. Boag, Dataset bias in diagnostic ai systems: Guidelines for dataset collection and usage, in: Proceedings of the ACM Conference on Health, Inference and Learning, Toronto, ON, Canada, 2020, pp. 2–4.

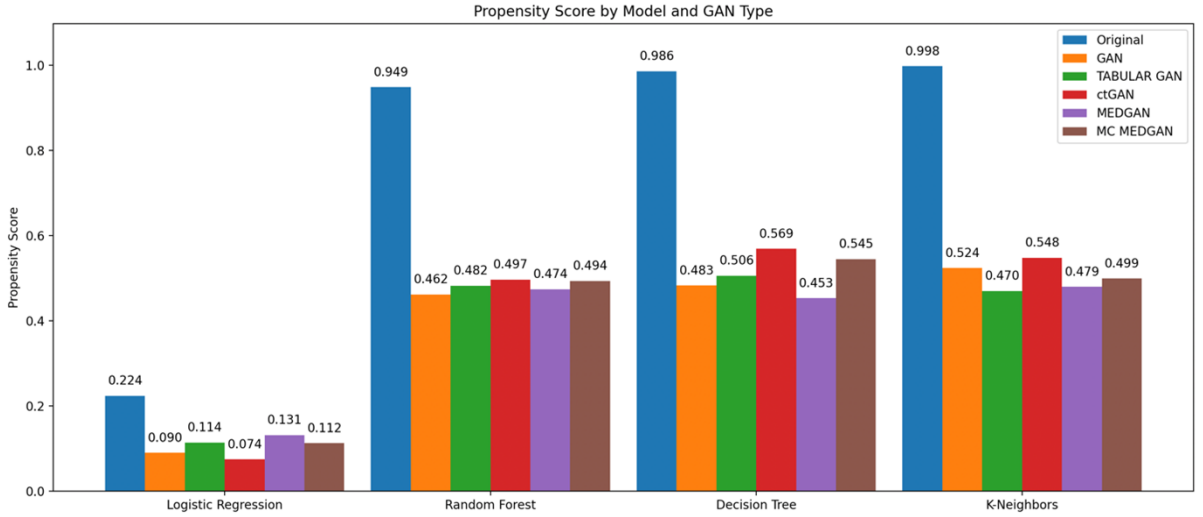
[4] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, et al., Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, Annals of oncology 29 (2018) 1836–1842.
[5] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2021, pp. 1–9.
[6] R. González-Sendino, E. Serrano, J. Bajo, P. Novais, A review of bias and fairness in artificial intelligence (2023).
[7] S. E. Eskimez, D. Dimitriadis, R. Gmyr, K. Kumanati, Gan-based data generation for speech emotion recognition., in: INTERSPEECH, 2020, pp. 3446–3450.
[8] M. A. Shahul Hameed, A. M. Qureshi, A. Kaushik, Bias mitigation via synthetic data generation: A review., Electronics (2079-9292) 13 (2024).
[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
[10] J. Chakraborty, S. Majumder, T. Menzies, Bias in machine learning software: Why? how? what to do?, in: Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering, 2021, pp. 429–440.
[11] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International conference on intelligent computing, Springer, 2005, pp. 878–887.
[12] T. Jo, N. Japkowicz, Class imbalances versus small disjuncts, ACM Sigkdd Explorations Newsletter 6 (2004) 40–49.
[13] A. Figueira, B. Vaz, Survey on synthetic data generation, evaluation methods and gans, Mathematics 10 (2022) 2733.
[14] L. Xu, K. Veeramachaneni, Synthesizing tabular data using generative adversarial networks, arXiv preprint arXiv:1811.11264 (2018).
[15] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, Advances in neural information processing systems 32 (2019).
[16] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, J. Sun, Generating multi-label discrete patient records using generative adversarial networks, in: Machine learning for healthcare conference, PMLR, 2017, pp. 286–305.
[17] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, A. P. Sales, Generation and evaluation of synthetic patient data, BMC medical research methodology 20 (2020) 1–40.
[18] R. E. Kharoua, Asthma disease dataset, 2024. URL: https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset, accessed: 2024-08-23.
[19] K. Ujeniya, Heart disease prediction dataset, 2024. URL: https://www.kaggle.com/datasets/krishujeniya/heart-diseae/data, accessed: 2024-08-23.
[20] R. E. Kharoua, Cancer prediction dataset, 2024. URL: https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset/data, accessed: 2024-08-23.
[21] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).
[22] A. E. Valojerdi, L. Janani, A brief guide to propensity score analysis, Medical journal of the Islamic Republic of Iran 32 (2018) 122.
[23] A. N. Carey, X. Wu, The statistical fairness field guide: perspectives from social and formal sciences, AI and Ethics 3 (2023) 1–23.
[24] J. Gesi, X. Shen, Y. Geng, Q. Chen, I. Ahmed, Leveraging feature bias for scalable misprediction explanation of machine learning models, in: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), IEEE, 2023, pp. 1559–1570.
[25] J. W. Gichoya, K. Thomas, L. A. Celi, N. Safdar, I. Banerjee, J. D. Banja, L. Seyyed-Kalantari, H. Trivedi, S. Purkayastha, Ai pitfalls and what not to do: mitigating bias in ai, The British Journal of Radiology 96 (2023) 20230023.

[26] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin, et al., Revolutionizing healthcare: the role of artificial intelligence in clinical practice, BMC medical education 23 (2023) 689.
[27] J. Georges-Filteau, E. Cirillo, Synthetic observational health data with gans: from slow adoption to a boom in medical research and ultimately digital twins?, arXiv preprint arXiv:2005.13510 (2020).
[28] S. Mahinnezhad, S. Mahinnezhad, K. Kaur, A. Shih, Data augmentation and class imbalance compensation using ctgan to improve gas detection systems, in: 2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), IEEE, 2024, pp. 1–6.
[29] E. S. Shourmasti, Generating Synthetic Health Data Using Machine Learning GAN Methods, Master's thesis, 2022.
[30] C. Lu, C. K. Reddy, P. Wang, D. Nie, Y. Ning, Multi-label clinical time-series generation via conditional gan, IEEE Transactions on Knowledge and Data Engineering (2023).
[31] K. M. Alfadli, A. O. Almagrabi, Feature-limited prediction on the uci heart disease dataset., Computers, Materials & Continua 74 (2023).
[32] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Applied Soft Computing 97 (2020) 105524.
[33] A. Anaissi, Y. Jia, A. Braytee, M. Naji, W. Alyassine, Damage gan: A generative model for imbalanced data, in: Australasian Conference on Data Science and Machine Learning, Springer, 2023, pp. 48–61.
[34] M. Abedi, L. Hempel, S. Sadeghi, T. Kirsten, Gan-based approaches for generating structured data in the medical domain, Applied Sciences 12 (2022) 7075.
[35] Y. Zhang, Z. Wang, Z. Zhang, J. Liu, Y. Feng, L. Wee, A. Dekker, Q. Chen, A. Traverso, Gan-based one dimensional medical data augmentation, Soft Computing 27 (2023) 10481–10491.
[36] B. Wen, Y. Cao, F. Yang, K. Subbalakshmi, R. Chandramouli, Causal-tgan: Modeling tabular data using causally-aware gan, in: ICLR Workshop on Deep Generative Models for Highly Structured Data, 2022.

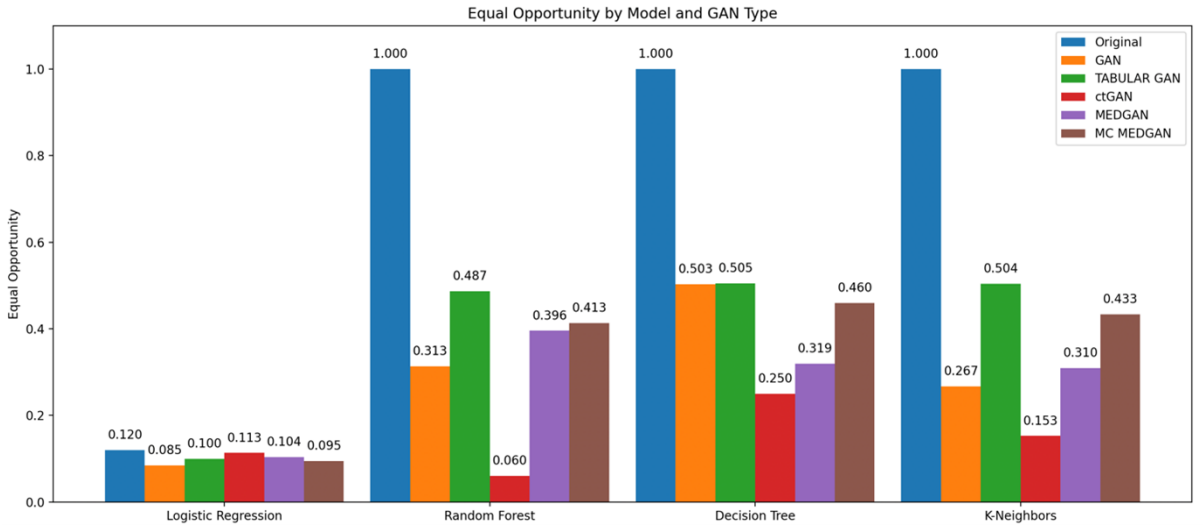
A. Fairness Assessment Graphs



(a)

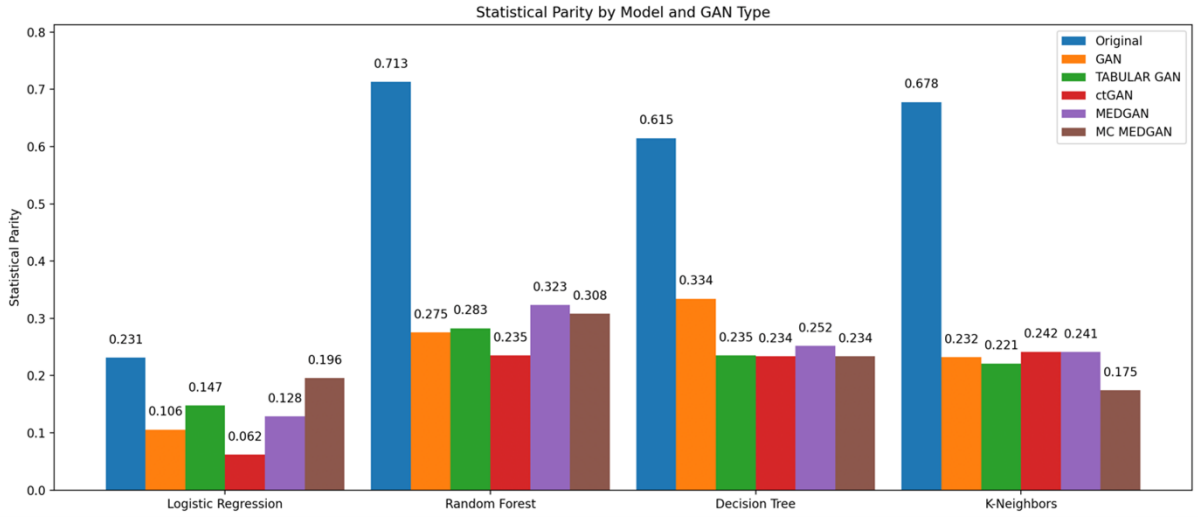


(b)

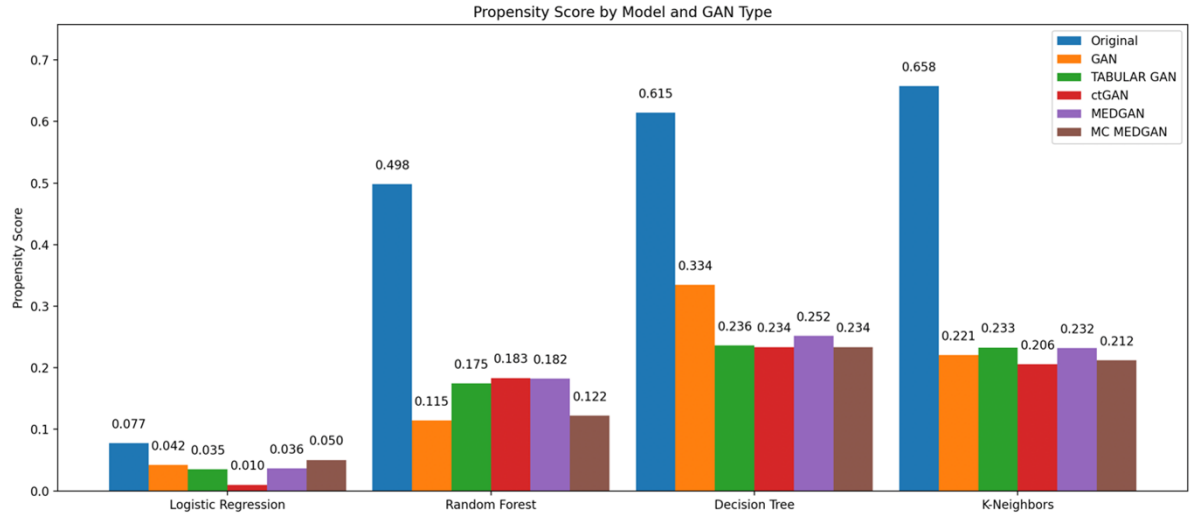


(c)

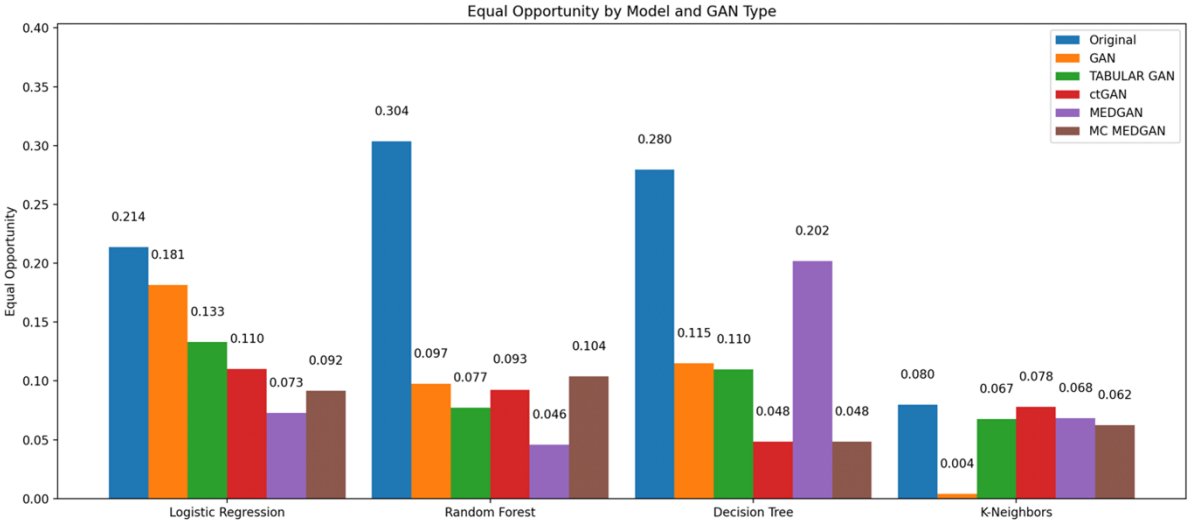
Figure 3: Fairness Assessment for Heart Disease Prediction dataset (a) Statistical Parity (b) Propensity Score, (c) Equal Opportunity



(a)



(b)



(c)

Figure 4: Fairness Assessment for Cancer Prediction dataset (a) Statistical Parity (b) Propensity Score, (c) Equal Opportunity