Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. ICBES 115
DOI: 10.11159/icbes24.115

Reducing Sample Selection Bias in Clinical Data through Generation of Multi-Objective Synthetic Data

Jarren Briscoe, Chance DeSmet, Katherine Wuestney, Assefaw Gebremedhin, Roschelle Fritz, and Diane J. Cook
Washington State University
Pullman, WA USA
{jarren.briscoe, chance.desmet, katherine.wuestney, assefaw.gebremedhin, shelly.fritz, djcook}@wsu.edu

Abstract **-** In the era of data-driven healthcare, identifying, quantifying, and mitigating bias in machine learning is of paramount importance. The impact of fair machine learning is particularly significant when predictions are applied in a clinical setting, where biased predictions can lead to unequal healthcare outcomes. In this paper, we consider the area of biomedical informatics and examine existing bias metrics and introduce a new metric to analyze bias in a smart home dataset. We investigate bias that may occur along sensitive attributes and examine its impact on the machine learning task of activity recognition from the collected data. In a novel approach to bias mitigation, we introduce a multi-objective generative adversarial network that creates synthetic data to mitigate sample bias by enhancing data diversity. We validate these methods using data collected for older adults living in smart homes who are managing multiple chronic health conditions, highlighting the potential of our approach to improve health predictions and outcomes.

Keywords**:** bias metrics, clinical data, generative adversarial network, sample bias, smart homes, synthetic data

1. Introduction

In bioinformatics and biomedicine, the potential of machine learning (ML) to revolutionize healthcare is exciting for many clinicians. Biomedical datasets are increasingly being used to inform clinical decision-making, contributing to the growing field of digital health, which leverages technology to monitor and improve health outcomes [1]–[5]. However, the adoption of these algorithms for critical decision making is still limited. The black-box nature of algorithms frequently necessitates caution for clinicians [6].

Biomedical data often contain inherent biases due to factors such as demographic disparities in data collection, unequal access to healthcare, and historical health disparities. When not properly addressed, these biases can lead to skewed predictions and unequal health outcomes. For instance, a predictive model trained on biased data might disproportionately misclassify certain demographic groups, leading to suboptimal treatment recommendations for those groups. In the worst-case scenario, such biases could exacerbate existing health disparities, undermining the goal of equitable healthcare. Distrust in ML algorithms is heightened by publicized cases where machine learning algorithms yielded prejudicial inferences [7]. Unless ML algorithms are designed to avoid bias along a particular sensitive attribute, they will reflect the prejudices of the data used to train them.

In this paper, we introduce a new bias metric, called *Faceted Disparate Impact* (FDI), that gives a novel quantitative perspective on bias consistent with legal precedent. Using prior metrics and FDI, we analyze bias in a clinical dataset containing smart home data. Finally, we propose a multi-agent generative adversarial network tool, called HydraGAN, to create diverse synthetic data that mitigates bias due to lack of sample diversity. We postulate that the new metric, FDI, is valuable because it considers all outcomes of a classification decision with context. Additionally, our proposed HydraGAN tool, which generates diverse synthetic data, has the potential to enhance the robustness of predictive models in bioinformatics, leading to more accurate and equitable health predictions.

2. Related Work

There have been numerous attempts to reduce machine learning bias. Weighting data points can reduce bias by emphasizing a minority class, while transforming the feature representations may reduce correlation between sensitive attributes and other features [8]. A systematic search of models and hyperparameters can identify the combination yielding

the lowest bias [9]. In the case of digital health, much of the bias is due to a lack of representation for underrepresented groups. Some prior research limit analysis to the data, while others examine how predictions and corresponding actions will affect target groups [10], [11]. Difficulties have been noted in aligning these measures with statistical requirements [12].

In this paper, we propose a multi-agent generative adversarial network (GAN) to generate sample data that improve diversity. Recent synthetic data creation for health applications relies more frequently on GANs [13]. Traditionally, GANs represent two-agent systems. However, the FairGAN architecture [14] balances the generator with two critics, one promoting data realism and the other supporting fairness. Our proposed approach extends these previous works by supporting an arbitrary number of agents, corresponding to a list of optimization criteria for the synthetic data. In this analysis, we harness the power of critics for pointwise realism, distribution realism, and distribution diversity.

3. Clinician-in-the-loop Smart Home Study

We collected continuous ambient sensor data for 22 older adults who are managing two or more chronic health conditions. Because 70% of the world's older adults are managing chronic conditions, the World Health Organization is asking for technology solutions to support these individuals [15]. The goal of this study is to design a clinician-in-the-loop (CIL) smart home that identifies health condition exacerbations using clinician-guided machine learning techniques. Table 1 provides a summary of the study participants.

Table 1: Study participants. F=female, M=male; HS=high school, B=bachelor, M=master, D=doctorate.

Age	89	83	88	75	95	89	81	63	92	79	83	88	90	76	93
Gender	F	M	F	M	F	F	F	F	F	M	F	F	F	F	M
Education	HS	B	HS	D	HS	B	HS	B	HS	D	B	B	B	B	M

We installed a CASAS smart home in a box (SHiB) [16] in the home of each subject for one year. The CASAS SHiB sensors monitor movement, door use, ambient light, and temperature. Nurses met weekly with each subject. Based on these interviews and nurse visual inspection of smart home data, changes in health status related to condition exacerbations were identified. In prior work, we extracted markers from these data that detect flare-ups in symptoms related to conditions such as congestive heart failure, diverticulitis, urinary tract infections, and Parkinson's disease [17].

4. Bias in Activity Recognition

As a first step in analyzing the data, we created a machine learning approach to recognize activities in real time. In this process, a sliding window is moved over data and used as context for the learning algorithm to label the last sensor reading in the window. At least one month of data was manually labelled by research team members (inter-annotator agreement κ=.80). In prior work, we validated that 11 activities (bed-toilet transition, cook, eat, enter home, leave home, hygiene, relax, sleep, wash dishes, work, other) are recognized with accuracy=0.99 [18]. For this work, we categorize the 11 activities into binary classes of active or sedentary activities to train a multi-layer perceptron to predict the patient's activity status. This gives nurses and viewers a more understandable gist of a patient's activities and is easier to tell when something is wrong with a glance. We then use a multi-layer perceptron to predict the patient's activity status. In this paper, we focus our attention on binary activity recognition and analyze bias for this task. This machine learning task is a pivotal component of assessing health state and is central to many other mobile health technologies.

Table 2 summarizes popular bias metrics. As the table shows, many metrics do not reflect all desired properties. To meet this need, we introduce a bias metric called Faceted Disparate Impact (FDI). In our discussion, we distinguish “benefit” from “bias”. While benefit references an advantage that is gained from an action (e.g., a ML prediction), bias reflects the distance between true and expected benefit. FDI is created based on legal precedents established by the US Supreme Court, EU’s Charter of Fundamental Rights, the Canadian Human Rights Act, and the Constitution of India, indicating that a lower admittance rate of a discriminated group warrants further investigation.

Table 2. Popular bias metrics and their properties. Each metric contrasts group *i* with group *j*. A larger score indicates more bias toward group *i*. P=positive, N=negative, n=size of set.

Metric	Formula	Directed	Symmetric	$n > 0$	Resilient	Bounded	Context
Disparate impact [19]	$\frac{\hat{P}_i}{n_i} \div \frac{\hat{P}_j}{n_j}$	✔	×	×	✔	×	×
Predictive parity [20]	$\frac{TP_i}{P_i} - \frac{TP_j}{P_j}$	✔	✔	×	×	✔	×
Treatment equality [21]	$\frac{FP_i}{FN} - \frac{FP_j}{FN_j}$	✔	✔	×	×	×	×
FPR difference (FPRD) [22]	$\frac{FP_i}{N_i} - \frac{FP_j}{N_j}$	✔	✔	×	×	✔	×
TPR difference (TPRD) [22]	$\frac{TP_i}{P_i} - \frac{TP_j}{P_j}$	✔	✔	×	×	✔	×
Equalized odds [23]	$FPRD + TPRD$	✔	✔	×	×	✔	×
Difference in conditional acceptance [22]	$\frac{P_i}{\hat{P}_i} - \frac{P_j}{\hat{P}_j}$	✔	✔	×	×	×	×
Difference in conditional rejection [22]	$\frac{N_i}{\hat{N}_i} - \frac{N_j}{\hat{N}_j}$	✔	✔	×	×	×	×
Difference in positive proportion & labels [22]	$\frac{\hat{P}_i}{n_i} - \frac{\hat{P}_j}{n_j}$	✔	✔	✔	✔	✔	×
Faceted Disparate Impact	$\frac{FP_i - FN_i}{n_i} - \frac{FP_j - FN_j}{n_j}$	✔	✔	✔	✔	✔	✔

The FDI metric is formalized in Eqs. 1-3. Using the legal precedents, we define benefit as a positive (+1) prediction, or desirable outcome, for each individual q. We define b as the overall benefit for any group, estimated as the weighted mean benefit over individuals in the group. Next, we assign weights to the prediction classes, defaulting to a weight of 1 for the positive class and 0 for the negative class. Based on the ground truth values, we compute expected benefit $E[b_q]$ as 1 for points with positive labels and -1 for negative points, resulting in Eq. 1.

$$E[b] = \frac{1}{n}\sum_{\forall b_q} = \frac{1 \cdot P + 0 \cdot N}{n} = \frac{TP + FN}{n} \tag{1}$$

Next, B quantifies the bias for a given group and is calculated as b-$E[b]$. Substituting and simplifying yields:

$$B = \frac{FP - FN}{n} \in [-1,1] \tag{2}$$

To find the bias for group *i* over group *j* (FDI), Eq. 3 calculates the difference of the directional bias for each group.

$$FDI = B_i - B_j = \frac{FP_i - FN_i}{n_i} - \frac{FP_j - FN_j}{n_j \epsilon [-2,2]} \tag{3}$$

We note several benefits of FDI. While some metrics focus solely on benefit or harm (FP or FN), FDI uses all confusion matrix cells. As a result, this metric includes more facets of the comparison between ground truth and predicted values. Additionally, this metric handles class imbalance without impacting calculations, creating metric resiliency. Furthermore, it is directed and symmetric. Importantly for our study, FDI can be adapted for binary classification, multi-class classification,

regression, and ranking. In the class of multi-class classification, the calculation is based on an n-ary confusion matrix, summarized by mean values. For regression, FDI replaces FP and FN with error (e.g., mean absolute error, mean squared error) above and below the desired threshold. When applying to ranking, FDI considers the difference in the ranked positions.

Many clinical data are time series in nature (e.g., EHR entries, lab results, vital signs, or sensor readings). To evaluate bias in such data, we need to extend the bias metrics to apply to time series data. In time series data, we consider each time step as an individual step that repeats benefit or harm, with a corresponding confusion matrix. In this scenario, bias is aggregated over individuals in a group and time steps in the series. For our experiments, we randomly sample 5% of all data to train, leaving 95% to validate and test. This process yields a larger undersampling effect and highlights potential bias.

5. Mitigating Bias with Diverse Synthetic Data

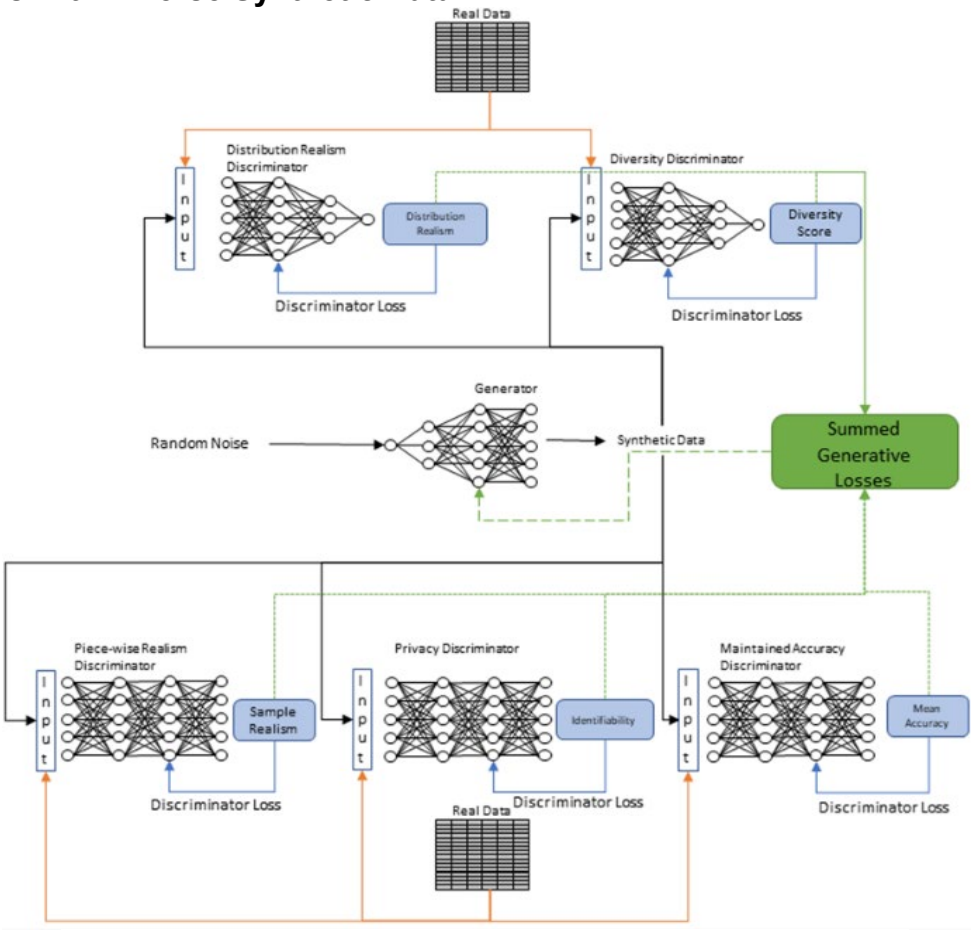

Fig. 1: The HydraGAN multi-agent architecture.

Because researchers recognize the surrogate role offered by synthetic data generators, they create methods to generate increasingly realistic data proxies. We consider the impact of creating realistic, diverse synthetic data for our dataset. What prior approaches lack is the ability to introduce multiple critics, each of which represents a distinct goal of the synthetic data. In some cases, emulating all characteristics of available real data is not the sole, or even desired, outcome. For example, the data may also need to achieve a diversity goal or obfuscate sensitive information. For this, we use HydraGAN [24], a multi-agent generative adversarial network that performs multi-objective synthetic data generation.

We adopt a multi-agent GAN, called HydraGAN, that assigns a discriminator to each data goal (see Fig. 1). Each of the critics separately critique individual or batches of synthetic data points. The generator's loss is the weighted sum of all critic

scores. When the system converges (the weight changes for an epoch are below a threshold value), a Nash equilibrium is formed among the discriminator goals. Here, we focus on three discriminators that minimize loss for traditional pointwise data realism (D_p, Eq. 4), distribution data realism (D_r, minimizing the difference between data distribution characteristics for the real and synthetic data, Eq. 5), and data diversity (minimizing the absolute difference between observed (α) and desired (β) value proportions for a particular feature f, see Eq. 6). Here, x_r and x_g represent batches of real and corresponding synthetic points.

$$\underset{x_r, x_g}{\text{minimize}} \sum_{i \in x_r, x_g} D_\rho(x_{g_i}) + (1 - D_\rho(x_{r_i})) \tag{4}$$

$$\text{minimize}(|\sum_{i \in \alpha_f} |\alpha_{f_i}| \log_2(|\alpha_{f_i}|) - \sum_{i \in \beta_f} |\beta_{f_i}| \log_2(|\beta_{f_i}|)|) \tag{5}$$

(6)

Diversity constraints may be designed to ensure equal representation among alternative groups. As an example, if 90% of a physical data collection represents one value for a sensitive feature (e.g., race) and 10% represents another, the diversity discriminator will move toward a more uniform distribution. Our architecture uses a combination of 1D convolutional layers, learnable positional encoding, and fully connected layers. Our regularization techniques include layer normalizations, instance normalizations, dropout layers, and Gaussian noise. We use the leaky ReLU activation function with a negative slope of 0.2. Each network uses the Adam optimizer with a learning rate of 0.0002, β_1= 0.5, and β_2 = 0.9$. We train with 75 epochs and conduct 100 steps per epoch. Each step contains a mini-batch of 64 time windows, each with a sequence length of 32.

The generator processes the sensitive attribute conditional, and a noise vector and outputs normalized values in time windows. We partition synthetic features to give appropriate output activations. Since our real date-time features are represented with two sine and cosine pairs for the day of year and time of day, the synthetic time features are outputted with a sine activation. Each set of one-hot encoded features (sensor one, sensor two, activity) is then passed through a softmax activation.

6. Experimental Results

We are interested in quantifying the bias contained in our clinical data using traditional metrics and our novel FDI metric. We then analyze bias for the newly generated dataset. While we focus on one dataset, the demographics in our study are like those found in many other clinical studies. A prevalent form of bias in clinical studies is sample selection bias. Many clinical study populations are largely devoid of diversity. As an example, Latinos and Asian Americans are disproportionately underrepresented in clinical trials assessing cognitive decline, comprising only 1%-5% of research participants [25].

We focus on two sensitive attributes: age and gender. In the case of gender, we assess bias for the traditional male and female groups. In the case of age, we assess bias for the older 25% of the sample in comparison with the group containing the younger 75% of the participants. From Table 2, we highlight Disparate Impact (DI), Difference in Conditional Acceptance (DCA), Difference in Proportionate Positives and Labels (DPPI), and Faceted Disparate Impact (FDI). This is a representative set: the remaining metrics yield very similar results to these.

The selected bias metrics focus on a task, in this case activity recognition. To simplify analyses, we aggregate activity categories into two classes: active behavior (bed-toilet transition, cook, eat, enter home, leave home, hygiene, wash dishes) and sedentary behavior (relax, sleep, work, other). For DCA, DPPL, and FDI, a no-bias score is zero. For DI, the no-bias score is one. A closer value to the no-bias score indicates less bias. A value greater than no-bias is bias toward the sedentary categories, while a lower value indicates a bias toward the active categories.

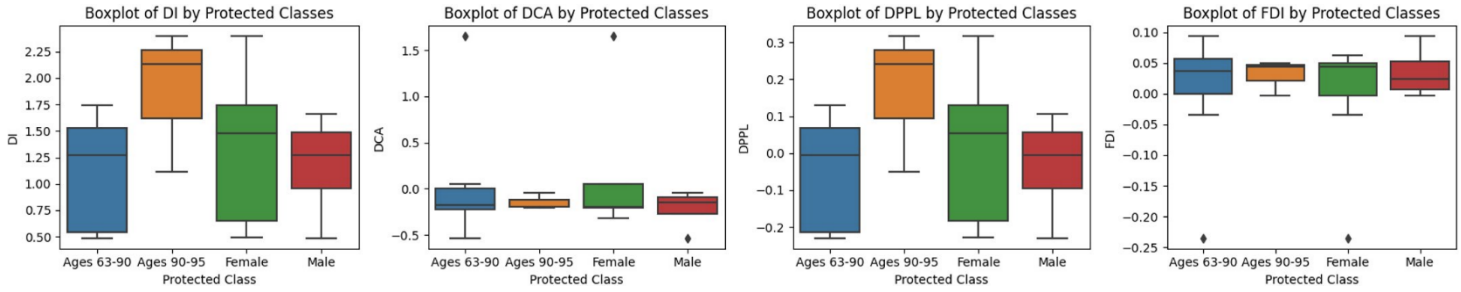

Fig. 2: Boxplots of bias metrics applied to original dataset.

6.1. Original dataset

Fig. 2 depicts boxplots of quantified bias on the original dataset. Since Ages 90-95 receives high scores in DI and DPPL, we see they are more likely to be predicted as sedentary. Referencing FDI, which gives context for correct predictions, we see that these positive predictions are largely correct as the Ages 90-95 FDI score is near zero. This makes sense as our data reflects that these older patients tend to be sedentary more often than the younger class. However, since FDI is positive, we see that there is still a slight bias for predicting the older age group as sedentary more often than they should.

6.2. Expanded dataset

We train the model using synthetic data generated by HydraGAN then quantify the bias by testing on real data. We summarize the bias results in Fig. 3. Here, individual diversity is enforced by querying the generator with each individual's label. This individual diversity also improves the protected class's diversities. In total, 3,072,000 synthetic points are created to be realistic and improve diversity for the underrepresented groups.

We see in Table 3 that the synthetic data mitigates bias overall (the bold cells indicate improvement). For FDI, DPPL, and DCA, a score of 0 indicates no bias. For DI, the ideal score is 1. Over many trials, these scores have a negligible standard deviation.

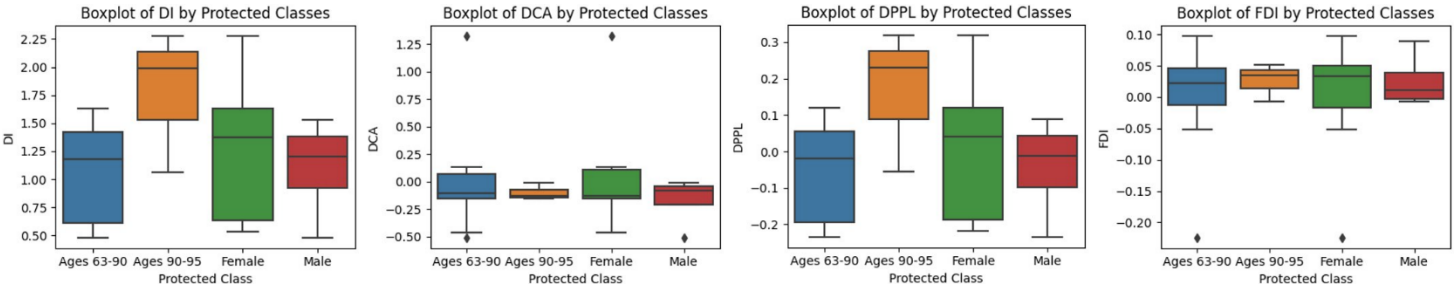

Fig. 3: Boxplots of bias metrics applied to the synthetically-expanded dataset.

Table 3: Comparison of bias results for synthetic and real data for each possible sensitive attribute.

Metric	Data	Ages 63-90	Ages 90-95	Female	Male
FDI	Synthetic	0.002	0.026	-0.001	0.026
	Real	0.004	0.030	-0.001	0.033
	Improves	**0.002**	**0.005**	-0.000	**0.007**
DPPL	Synthetic	-0.056	0.165	0.012	-0.043
	Real	-0.055	0.170	0.012	-0.036
	Improves	-0.002	**0.005**	**0.000**	-0.007
DI	Synthetic	1.060	1.778	1.280	1.103
	Real	1.109	1.891	1.340	1.174
	Improves	**0.049**	**0.113**	**0.060**	**0.071**
DCA	Synthetic	0.003	-0.100	0.046	-0.172
	Real	0.012	-0.151	0.061	-0.219

Table 4: Comparison of synthetic and real data to a uniform distribution using measures of KS (Kolmogorov-Smirnov) statistic, Kullback-Leibler (KL) divergence, and Jensen-Shannon (JS) distance.

	Activities			Individuals		
Metric	**Synthetic**	**Real**	**Improvement**	**Synthetic**	**Real**	**Improvement**
KS statistic	0.863	0.900	**0.027**	0.000	0.733	**0.733**
KL Divergence	1.822	2.302	**0.480**	0.000	0.252	**0.252**
JS Distance	0.652	0.725	**0.073**	0.000	0.250	**0.250**

Table 4 shows that the synthetic data improves diversity. The improvement in uniformity is statistically significant ($p<0.001$) for the Kolmogorov-Smirnov (KS) statistic when comparing distributions for activities. Furthermore, our synthetic data's distribution over individuals reflects distance of zero from a uniform distribution (the desired result) due to the infinitely strong conditional passed to the generator. Reviewing the protected class's bias reductions in Table 3, we furthermore conclude that our synthetic data successfully mitigates bias for age and gender.

7. Discussion and Conclusion

In this paper, we examine biases that may exist in a clinical dataset using smart home sensor data to model activities that are used for health assessment. Metrics of bias are varied yet do not consistently make use of all predicted outcomes. These outcomes lead to advantages for one group over another and so need to be considered in bias analyses. As a result, we not only use traditional metrics, but we also introduce a new metric based on legal precedent, Faceted Disparate Impact. As we show, the FDI metric considers all cells in the confusion matrix and compares predicted labels with ground truth labels, leading to a more comprehensive analysis of bias and fairness.

The experimental results indicate that bias does exist in our data, even for a straightforward task such as activity labelling. Because activity recognition is used as a cornerstone for embedded and mobile technology strategies for health assessment and intervention, even this component necessitates unbiased reasoning and fair treatment of all groups. To potentially mitigate sample bias that results from a lack of diversity in the collected data, we introduce HydraGAN, a multi-agent synthetic data generator. Generating synthetic data with HydraGAN does reduce bias in the data based on multiple metrics.

This is an early analysis of the FDI metric and HydraGAN algorithm to analyze and lessen bias in clinical data. Further validation is needed to assess these contributions on a greater variety of clinical datasets and across additional protected attributes. We also note that HydraGAN can incorporate additional critics that consider metrics such as privacy preservation. Future work will analyze the role these optimization criteria can play in providing more trustworthy machine learning technologies for clinical data assessment and application.

Acknowledgements

This work is supported in part by NINR grant R01NR016732.

References

[1] Q. Li, M. Jiang, and C. Ying, “An assistant decision-making method for rare diseases based on RNNs model,” in

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2022, pp. 2632–2639. doi: 10.1109/BIBM55620.2022.9995519.

[2] X. Guo, Y. Qian, P. Tiwari, Q. Zou, and Y. Ding, "Kernel risk sensitive loss-based echo state networks for predicting therapeutic peptides with sparse learning," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 6–11. doi: 10.1109/BIBM55620.2022.9994902.

[3] Y. Lin, J. Jiang, Z. Ma, D. Chen, Y. Guan, X. Liu, H. You, J. Yang, and X. Cheng, "CGPG-GAN: An acne lesion inpainting model for boosting downstream diagnosis," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 1634–1638. doi: 10.1109/BIBM55620.2022.9995406.

[4] D. Tan, J. Wang, R. Yao, J. Liu, J. Wu, S. Zhu, Y. Yang, S. Chen, and Y. Li, "CCA4CTA: A hybrid attention mechanism based convolutional network for analysing collateral circulation via multi-phase cranial CTA," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 1201–1206. doi: 10.1109/BIBM55620.2022.9995381.

[5] R. Zemouri, N. Zerhouni, and D. Racoceanu, "Deep learning in the biomedical applications: Recent and future status," *Appl. Sci.*, vol. 9, no. 8, 2019, doi: 10.3390/app9081526.

[6] H. Alami, P. Lehoux, Y. Auclair, M. de Guise, M.-P. Gagnon, J. Shaw, D. Roy, R. Fleet, M. Ahmed, and J.-P. Fortin, "Artificial intelligence and health technology assessment: Anticipating a new level of complexity," *J. Med. Internet Res.*, vol. 22, no. 7, p. e17707, 2020.

[7] L. A. Celi, J. Cellini, M.-L. Charpignon, E. C. Dee, and F. Dernoncourt, "Sources of bias in artificial intelligence that perpetuate healthcare disparities - A global review," *PLOS Digit. Heal.*, vol. 1, no. 3, p. e0000022, 2022.

[8] D. Plecko and N. Meinshausen, "Fair data adaptation with quantile preservation," *J. Mach. Learn. Res.*, vol. 21, pp. 1–44, 2020.

[9] A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*, 2019.

[10] T. Speicher, H. Heidari, N. Grgic-Hlaca, and others, "A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 2239–2248.

[11] G. Pleiss, M. Raghavan, F. Wu, and others, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017.

[12] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv Prepr. arXiv1808.00023*, 2018.

[13] K. Baek and H. Shim, "Commonality in natural images rescues GANs: Pretraining GANs with generic and privacy-free synthetic data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7854–7876.

[14] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN: Fairness-aware generative adversarial networks," in *IEEE International Conference on Big Data*, 2018.

[15] R. Yan, X. Liu, J. Dutcher, M. Tumminia, D. Villalba, S. Cohen, D. Creswell, K. Creswell, J. Mankoff, A. Dey, and A. Doryab, "A compuational framework for modeling biobehavioral rhythms from mobile and wearable data streams," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 3, p. 47, 2022.

[16] D. J. Cook, A. Crandall, B. Thomas, and N. Krishnan, "CASAS: A smart home in a box," *IEEE Comput.*, vol. 46, no. 7, pp. 62–69, 2012.

[17] S. Fritz, K. Wuestney, G. Dermody, and D. J. Cook, "Nurse-in-the-loop smart home detection of health events associated with diagnosed chronic conditions: A case-event series," *Int. J. Nurs. Stud. Adv.*, vol. 4, p. 100081, 2022.

[18] S. Aminikhanghahi, T. Wang, and D. J. Cook, "Real-Time change point detection with application to smart home time series data," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 1010–1023, 2019.

[19] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259–268.

[20] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of

the art," *Sociol. Methods Res.*, vol. 50, no. 1, pp. 3–44, 2021.
[21] N. Mehrabi, F. Forstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021, doi: 10.1145/3457607.
[22] S. Das, M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and B. Zafar, "Fairness measures for machine learning in finance," *J. Financ. Data Sci.*, 2021.
[23] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *CoRR*, vol. abs/1610.0, 2016, [Online]. Available: http://arxiv.org/abs/1610.02413
[24] C. DeSmet and D. J. Cook, "HydraGAN: A cooperative agent model for multi-objective data generation," *ACM Trans. Intell. Syst. Technol.*, 2024.
[25] E. Arana-Chicas, F. Cartujano-Barrera, K. Rieth, K. Richter, E. Ellerbeck, L. Cox, K. Graves, F. Diaz, D. Catley, and A. Cupertino, "Effectiveness of recruitment strategies of Latino smokers: Secondary analysis of a mobile health smoking cessation randomized clinical trial," *J. Med. Internet Res.*, vol. 24, no. 6, p. e34863, 2022.