Multi-Objective Synthetic Data to Reduce Sample Selection Bias in Clinical Data

Anonymous

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 propose the use of multi-objective 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.**

Index Terms—**bioinformatics, biomedicine, bias metrics, clinical data, GANs, multi-agent generator adversarial networks, sample bias, smart homes, synthetic data**

I. 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, which forms the basis for many predictive models in healthcare, 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]. The underlying issue in these cases was that the ML algorithm formed inferences based on limited and imbalanced data. Unfairness in machine learning outcomes typically stems from data or algorithm bias. 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 make three main contributions:

1) We introduce a new bias metric, called *Fairness Disparity Index* or FDI, that is consistent with legal precedent and gives a new perspective on bias.
2) Using both existing metrics and FDI, we analyze bias in a clinical dataset containing smart home data linked with activities of daily living and health data.
3) 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.

Detecting, quantifying, and mitigating biases is critical, especially in clinical settings. Because data and ML model biases can result in unforeseen and undesired consequences, we postulate that the new metric, FDI, is valuable because it considers all outcomes of a classification decision with context. With this metric in place, we can further understand and evaluate the impact of algorithmic approaches to reducing sample bias, including the HydraGAN approach of creating realistic, diverse, synthetic data samples.

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. Eliminating health disparities and achieving health equity for all is a key objective of the U.S. government's Healthy People 2030 initiative [8]. Given healthcare AI is applied at a structural level and thus has the potential to do harm at scale, intentional bias mitigation in ML models is a critical step for addressing disparate health outcomes of minoritized individuals and communities. Current disparities in healthcare result in a lack of datasets appropriately representing minority populations thus synthetic datasets are a valuable interim strategy to equalize representation.By developing and validating methods to quantify and mitigate bias in these datasets, we aim to improve the accuracy and fairness of health predictions derived from them.

II. Related Work

There have been numerous attempts to reduce machine learning bias. As with any imbalanced class distribution problem, weighting data points can reduce bias by adjusting the presentation of a minority class. Relabeling points or transforming the feature representations to reduce correlation between sensitive attributes and other features has also been considered [9]. A systematic search of learning algorithms and hyperparameters can identify the combination that yields the lowest measured bias [10]. In the case of digital health models, much of the bias is due to a lack of representation

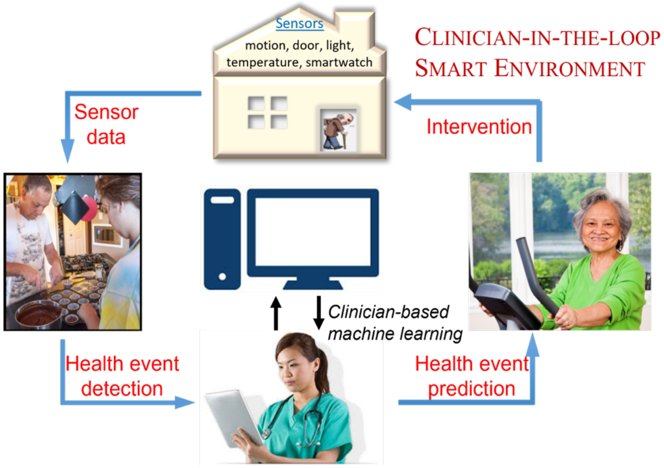

Fig. 1: The framework of the clinician-in-the-loop smart environment.

TABLE I: Study participants.

Age	Gender	Race	Education
89	female	white	high school
83	male	white	bachelor
88	female	white	high school
75	male	white	doctorate
95	female	white	high school
89	female	white	bachelor
81	female	white	high school
63[1]	female	white	bachelor
92	female	white	high school
79	male	white	doctorate
83	female	white	bachelor
88	female	white	bachelor
90	female	white	bachelor
76	female	white	bachelor
93	male	white	master

for one or more underrepresented groups. Related work has also focused on quantifying and describing bias. While some methods analyze the data directly, others look ahead to see how predictions and corresponding actions will affect target groups [11], [12]. Reference [13] highlight the difficulties in aligning many of these measures with statistical requirements.

We introduce a multi-agent generative adversarial network (GAN) to generate needed diverse sample data. Synthetic data generators for clinical applications are numerous and varied. Recently, these rely more frequently on GANs [14]. Traditionally, GANs represent two-agent systems, one agent creating artificial data and the other distinguishing between real and synthetic data. One GAN that also focuses on fair ML is FairGAN [15], which 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.

III. 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 [16]. The goal of this study is to design a clinician-in-the-loop smart home that identifies health condition exacerbations using clinician-guided machine learning techniques (see fig. 1). Table I provides descriptors for the participants analyzed in this paper.

We installed a CASAS (Center of Advanced Studies in Adaptive Systems) "smart home in a box" (SHiB) [17] in the home of each subject for one year. The CASAS SHiB sensors include passive infrared motion detectors to sense motion around the home, magnetic door sensors to detect open/shut status of doors and cabinets, and sensors that report ambient light and temperature levels. We place 2-5 sensors in each area of the home (e.g., bedroom, living room, kitchen, bathroom) using removable adhesive strips. Sensors report changes in state (motion on/off, door open/closed, temperature/light levels) to a computer, which adds timestamps and sensor identifiers, encrypts the readings, and securely transmits the data to an off-site server. The research team includes nurses that meet weekly with each subject. Based on these interviews and nurse visual inspection of smart home data, changes in health status related to condition exacerbations are identified [18].

For a machine learning algorithm to automatically detect condition exacerbations, we first automatically label collected sensor data with the corresponding activities that the resident was performing at that time. This is a critical step because behavior markers are extracted from activity-labeled data that reflect a person's routine activities (e.g., sleep, exercise, work, cook, eat, hygiene, toilet, leave/enter home) and changes in routine. From these markers, we identify clinically-relevant anomalies and identify time periods that correspond to flare-ups in symptoms related to conditions such as congestive heart failure, diverticulitis, urinary tract infections, and Parkinson's disease [19].

IV. Activity Recognition

As a first step in analyzing the data, we created a machine learning approach to recognize activities in real time. While human activity recognition is a popular research topic, this approach is distinct because we recognize activities performed in uncontrolled, unscripted settings, based on continuous data. In this process, a sliding window is moved over data and used as context for the learning algorithm to label the most recent sensor reading in the window. At least one month of data was manually labeled by research team members who assigned an activity category to each sensor reading based on

[1] Estimated from an age range of 60-65.

visual inspection of the data and a house floor plan (inter-annotator agreement $\kappa = 0.80$). To date, 11 activities (bed-toilet transition, cook, eat, enter home, leave home, hygiene, relax, sleep, wash dishes, work, other) are recognized with accuracy=0.99 and macro f1 score=0.99 based on three-fold cross validation [20].

However, 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 use the multi-layer perceptron to predict the patient's activity status.

There are several machine learning models that have been used to detect condition exacerbations and assess health state [19], [21]. For this paper, we focus our attention on activity recognition and will analyze bias in the data and the model for this task. This machine learning task is foundational to the clinician-in-the-loop goal and was a pivotal component of the studies. Furthermore, this task is also central to many other mobile health technologies.

A. The Fairness Disparity Index

As Table II indicates, many metrics do not comprehensively reflect desired properties. To provide a useful analysis of bias in the data, therefore, we introduce the new bias metric called Fairness Disparity Index (FDI). In our discussion, we split the rhetoric into "benefit" and "bias". While benefit references an advantage that is gained from an action (e.g., a ML prediction or strategically sampled data), bias reflects the distance between benefit and expected benefit.

FDI is created based on a modern legal precedent. Specifically, the US Supreme Court, in response to biased treatment of protected groups in work and housing, stated that a lower admittance rate of some discriminated group warrants further investigation even if the discrimination was unintentional [27]. In the context of machine learning, admittance rates reflects a positive class label. We use this precedent to create a metric that specifically computes the benefit difference between groups. FDI is similarly consistent with other precedents established by the Civil Rights Act of 1964, EU's Charter of Fundamental Rights, the Canadian Human Rights Act, and the Constitution of India.

The FDI metric is formalized in (1) to (7). Using the legal precedence, we define benefit as a positive prediction (a desireable outcome) for each individual q in (1).

$$b_q = \begin{cases} 1 & \text{if predicted positive.} \\ -1 & \text{if predicted negative.} \end{cases} \tag{1}$$

Now we define b in (2) as the overall benefit for any group. b is found by the weighted mean of each individual in the group. Next, we assign weights to the prediction classes. We default to giving the positive class a weight of 1 and the negative class a weight of 0. This rule can be updated depending on the use case.

$$b = \frac{1}{n}\sum_{\forall b_q} b_q = \frac{1 \cdot \hat{P} + 0 \cdot \hat{N}}{n} = \frac{TP + FP}{n} \tag{2}$$

Based on the ground truth class values, we compute the expected benefit ($\mathbb{E}[b]$) as:

$$\mathbb{E}[b_q] = \begin{cases} 1 & \text{if labeled positive.} \\ -1 & \text{if labeled negative.} \end{cases} \tag{3}$$

$$\mathbb{E}[b] = \frac{1}{n}\sum_{\forall b_q} \mathbb{E}[b_q] = \frac{1 \cdot P + 0 \cdot N}{n} = \frac{TP + FN}{n} \tag{4}$$

Next, $\mathcal{B}$ quantifies the bias for a given group. This value is calculated as the difference between b and $\mathbb{E}[b]$.

$$\mathcal{B} = b - \mathbb{E}[b] \tag{5}$$

Substituting and simplifying yields:

$$\mathcal{B} = \frac{FP - FN}{n} \in [-1, 1] \tag{6}$$

Now, to find the bias that group i has over group j (FDI), we take the difference of the directional bias that each group holds.

$$FDI = \mathcal{B}_i - \mathcal{B}_j = \frac{FP_i - FN_i}{n_i} - \frac{FP_j - FN_j}{n_j} \in [-2, 2] \tag{7}$$

We note several additional benefits of FDI for our analysis. While some metrics focus solely on benefit or harm (FP or FN), FDI uses all confusion matrix cells (recall that n is the sum of all cells). As a result, this metric includes more context in the quantification, comparison between ground truth and predicted values, and positive and negative classifications. Additionally, this metric handles class imbalance without impacting calculations, creating metric resiliency. Furthermore, it is understandable as it is directed and symmetric.

A property that is important for our ongoing work is that 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. Finally, when applying to ranking, FDI will consider the difference in the expected ranked position. For these reasons, FDI is a particularly well-suited metric for our clinical application.

TABLE II: Popular confusion-matrix-based bias metrics and their properties. Each metric contrasts group i with group j. For all metrics, a greater score indicates more bias toward group i. For the traditionally equality-checking metrics, we instead take the difference to measure bias. Abbreviations: Dir/Direction, Sym/Symbol, Res/Resilience, Bound/Bounded in Range.

Metric	Formula	Directed	Symmetric	$n > 0$	Resilient	Bounded	Context
Disparate Impact [22]	$DI \triangleq \frac{\hat{P}_i}{n_i} \div \frac{\hat{P}_j}{n_j}$	✓	×	×	✓	×	×
Predictive Parity [23]	$PP \triangleq \frac{TP_i}{P_i} - \frac{TP_j}{P_j}$	✓	✓	×	×	✓	×
Treatment Equality [24]	$TE \triangleq \frac{FP_i}{FN_i} - \frac{FP_j}{FN_j}$	✓	✓	×	×	×	×
False-Positive-Rate Difference [25]	$FPRD \triangleq \frac{FP_i}{N_i} - \frac{FP_j}{N_j}$	✓	✓	×	×	✓	×
True-Positive-Rate Difference [25]	$TPRD \triangleq \frac{TP_i}{P_i} - \frac{TP_j}{P_j}$	✓	✓	×	×	✓	×
Equalized Odds [26]	$EO \triangleq FPRD + TPRD$	✓	✓	×	×	✓	×
Difference in Conditional Acceptance [25]	$DCA \triangleq \frac{P_i}{\hat{P}_i} - \frac{P_j}{\hat{P}_j}$	✓	✓	×	×	×	×
Difference in Conditional Rejection [25]	$DCR \triangleq \frac{N_j}{\hat{N}_j} - \frac{N_i}{\hat{N}_i}$	✓	✓	×	×	×	×
Difference in Positive Proportion & Labels [25]	$DPPL \triangleq \frac{\hat{P}_i}{n_i} - \frac{\hat{P}_j}{n_j}$	✓	✓	✓	✓	✓	×
Fairness Disparity Index (see Eq. (7))	$FDI \triangleq \frac{FP_i - FN_i}{n_i} - \frac{FP_j - FN_j}{n_j}$	✓	✓	✓	✓	✓	✓

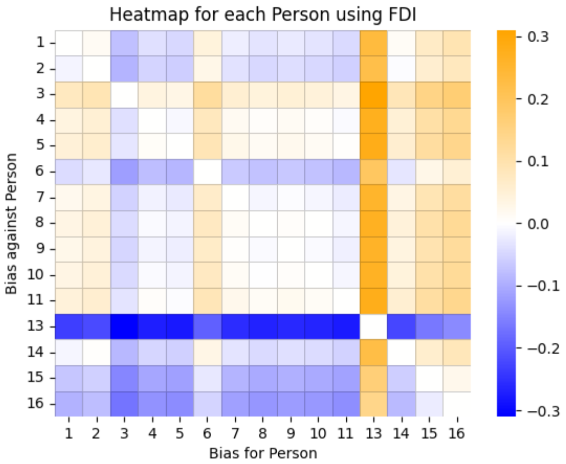

Fig. 2: Heat map illustrating the bias matrix for each person using FDI. Notice that self-bias is zero (the diagonal).

B. Evaluating Bias for Time Series Data

Many clinical data are time series in nature, representing pieces of information collected over time such as EHR entries, lab results, vital signs, or (in our case) sensor readings. To evaluate bias in such data, we need to extend the bias and fairness metrics to apply to time series as well as more traditional i.i.d. data.

We note that group metrics represent an aggregation from individual benefit or bias. In time series data, we consider each time step as an individual step that repeats benefit or harm. A confusion matrix can be created for the corresponding time step. A time series score is computed based on the collection of timestamped confusion matrices. To compute bias for a group, we first create a bias matrix to compare all individuals (see fig. 2). Next, we aggregate the bias for each person using the arithmetic mean (omitting self bias). We then aggregate these by group. Finally, we implement statistical measures over these groupings.

To highlight biases and mitigation, we randomly sample 5% of all data to train, leaving 95% to validate and test. This process yields a larger undersampling effect.

V. Mitigating Bias with Diverse Synthetic Data

Because researchers recognize the surrogate role offered by synthetic data generators, they create methods to generate increasingly realistic data proxies. In this paper, we consider the impact of creating realistic, diverse synthetic data on 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 [28], a multi-agent generative adversarial network that performs multi-objective synthetic data generation.

A. HydraGAN

We adopt a multi-agent GAN, called HydraGAN, that assigns a "head" (critic) to each data goal. 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 critic goals. In other words, HydraGAN seeks to maximize performance for a combination of user-defined goals.

While the original HydraGAN (see fig. 3) uses five heads, we focus on the goals of data realism (considering data distribution) and data diversity. This is because our analysis does not consider privacy as the data is public and we find the accuracy critic to increase bias. Using a piece-wise realism critic with the distribution realism critic for our use case is left for future work.

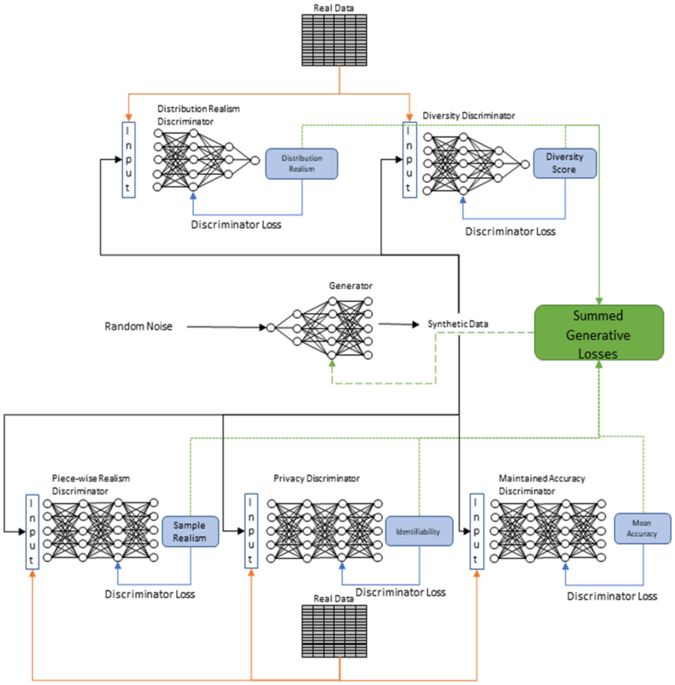

Fig. 3: The HydraGAN multi-agent architecture [28].

TABLE III: Notations used in equations (8) to (12).

Symbol	Definition
C_ρ	Represents the realism critic.
C_d	Represents the diversity critic.
G	Represents the generator.
P_G	Represents the data distribution of the generator.
P_r	Represents the real data distribution.
$P_{\tilde{x}}$	Represents the distribution of randomly sampled points between the real and generated data distributions, also known as the interpolated samples.
W_1	Represents the Wasserstein distance.
GP	Represents the gradient penalty.
λ	Represents the coefficient of the gradient penalty.
t	Represents the conditional, ensuring equal representation for each individual.

Our distribution realism critic C_ρ optimizes the function shown in (10). The diversity critic C_d optimizes the function in (11). Finally, the generator G minimizes the loss defined in (12). We define the notations used in equations (8) thru (12) in Table III.

$$W_1 = \mathbb{E}_{\tilde{x}\sim P_G|t}[C_\rho(\tilde{x},t)] - \mathbb{E}_{x\sim P_r}[C_\rho(x,t)] \tag{8}$$

$$\text{GP} = \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}|t}[(||\nabla_{\hat{x}} C_\rho(\hat{x},t)||_2 - 1)^2] \tag{9}$$

$$L_{C_\rho} = W_1 + \text{GP} \tag{10}$$

As (8) and (10) indicate, the critic attempts to learn the distance between the "real" and "synthetic" data. For the Wasserstein distance, the best possible performance occurs when the two data sets are indistinguishable. We use the gradient penalty (9) to stabilize training [29].

The diversity critic C_d ensures that output from the generator meets externally-imposed constraints on the distribution of a selected feature. Constraints may be designed to ensure equal representation among all the target class values or more greatly emphasize value ranges for a specific feature, providing the ability to achieve the data distribution needed for a given task. 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 critic may be used to achieve a more uniform distribution. C_d attempts to minimize the difference between the synthetic distribution and the desired distribution (typically, a uniform distribution). Thus, C_d approximates the function shown in (11). In this equation, α represents the proportion for each value of feature f in the original dataset and β represents the desired proportion.

In our experiments, we already enforce individual diversity with the conditional t so the activity diversity is promoted using C_d.

$$L_{C_d} = \frac{1}{n}\sum_{f=1}^{n}(\alpha_f - \beta_f)^2 \tag{11}$$

Now using the notations from Table III, and weights w_1 and w_2, we define our generator's loss in (12).

$$L_G = -w_1\mathbb{E}_{\tilde{x}\sim P_G}[C_\rho(\tilde{x},t)] + w_2 L_{C_d}. \tag{12}$$

For our experiments, we set $w_1 = 1$ and $w_2 = 5$.

B. Neural Architecture

Our architecture uses a combination of advanced deep learning techniques, including one-dimensional convolutional layers (1D-CNNs), 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, $\beta_1 = 0.5$, $\beta_2 = 0.9$. 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 takes in 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. Then each set of one-hot encoded features (sensor one, sensor two, activity) is sent through a softmax activation.

The realism critic C_ρ considers real data and the sensitive attribute conditional t. C_ρ then uses an input embedding before passing it on to a learnable positional embedding. A series of 1D-CNNs, fully connected, and regularization layers follow. The output is a single node with no activation function to give the Wasserstein distance (10). The diversity critic takes in the distributions of synthetic activities and the uniform distribution and then returns the diversity score (11).

VI. 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 are focusing on one particular dataset, the demographics in our study are similar to 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 [30]. Lack of adequate data negatively impacts machine learning predictive performance and introduces bias. We aim to help identify this with FDI and help solve it with HydraGAN.

As noted in Table I, all of the study participants are white. Because other groups are not represented here, we focus on two other sensitive attributes for which representation does exist: 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. Rather than evaluate bias for all of the metrics listed in Table II, we select a subset including Disparate Impact (DI), Difference in Conditional Acceptance (DCA), Difference in Proportionate Positives and Labels (DPPl), and Fairness Disparity Index (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 positive value indicates a bias toward the sedentary categories, while a negative value indicates a bias toward the active categories.

While [20] employ random forests for CASAS human activity recognition, our predictions come from a multilayer perceptron (MLP) classifier. We find that the MLP is better able to infer classifications of the real data using the synthetic data. This is likely because the synthetic data's sensors and activities are processed by a softmax activation function while the real data are one-hot encoded. We do not use argmax to one-hot encode the generated data because some information is lost in the process. Empirically, the MLP better transfers knowledge between the synthetic softmax and real one-hot encoded probability density functions.

A. Original Dataset

Fig. 4 depicts boxplots of quantified bias on the original dataset. Since the Ages 90-95 class 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 be.

B. 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. 5. 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 IV that the synthetic data mitigates bias overall (the bold cells indicate improvement). For FDI, DPPL, and DCA, a score of zero indicates no bias. For DI, the ideal score is 1. Over many trials, these scores have a negligible standard deviation. Some improvement results seem off due to rounding.

TABLE IV: Comparing synthetic and real data for each protected class using bias metrics.

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
	Improves	**0.009**	**0.051**	**0.014**	**0.048**

Table V shows that the synthetic data improves individual and activity diversity. With $p < 0.001$, we reject the null hypothesis (that the distributions are equal) for the KS statistic, except for the synthetic individual distribution that we mandated be uniform. Furthermore, our synthetic data's individual distribution receives perfect scores due to the infinitely strong conditional passed to the generator. After reviewing the protected class's bias reductions in Table IV, we conclude that our synthetic data mitigates bias for age and gender.

VII. Discussion and Conclusions

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, Fairness Disparity Index. As we show, the FDI metric considers all cells in the confusion

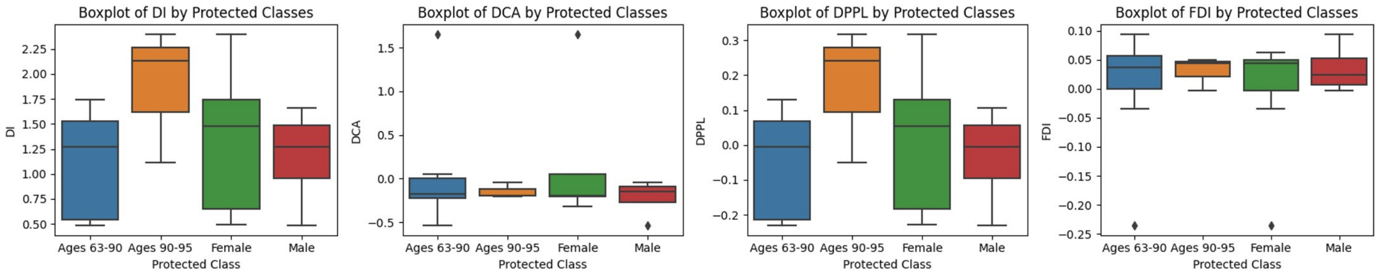

Fig. 4: Boxplots of bias metrics applied to original dataset.

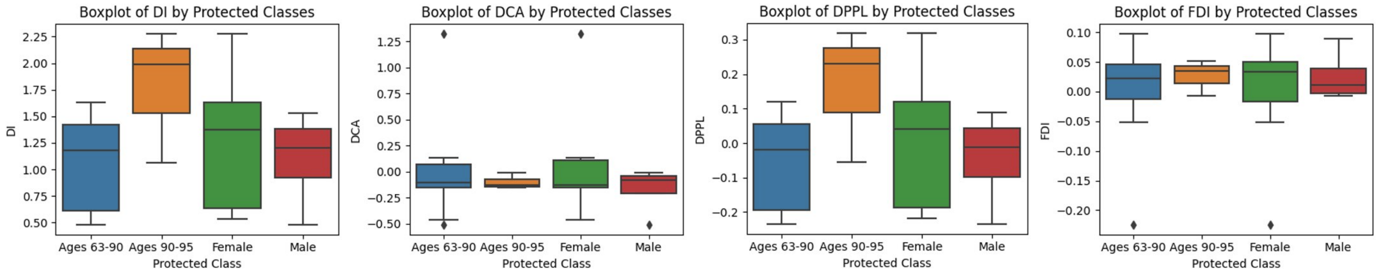

Fig. 5: Boxplots of bias metrics applied to the synthetic dataset.

TABLE V: Comparison of synthetic data and real data to the target uniform distribution.

	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**

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 labeling. 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.

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)*, pp. 2632–2639, 2022.

[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)*, pp. 6–11, 2022.

[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)*, pp. 1634–1638, 2022.

[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)*, pp. 1201–1206, 2022.

[5] R. Zemouri, N. Zerhouni, and D. Racoceanu, "Deep learning in the biomedical applications: Recent and future status," *Applied Sciences*, vol. 9, no. 8, 2019.

[6] H. Alami, P. Lehoux, Y. Auclair, and M. De, "Artificial intelligence and health technology assessment: Anticipating a new level of complexity," *Journal of Medical Internet Research*, vol. 22, pp. 1–22, 2020.

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

[8] Office of Disease Prevention and Health Promotion, "Health Equity in Healthy People 2030." https://health.gov/healthypeople/priority-areas/health-equity-healthy-people-2030.

[9] D. Plecko and N. Meinshausen, "Fair data adaptation with quantile preservation," *Journal of Machine Learning Research*, vol. 21, pp. 1–44, 2020.

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

[11] T. Speicher, H. Heidari, N. Grgic-Hlaca, *et al.*, "A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices," in *ACM SIGKDD International Con-*

ference on Knowledge Discovery and Data Mining, pp. 2239–2248, 2018.

[12] G. Pleiss, M. Raghavan, F. Wu, *et al.*, “On fairness and calibration,” in *Advances in Neural Information Processing Systems*, 2017.

[13] S. Corbett-Davies and S. Goel, “The measure and mismeasure of fairness: A critical review of fair machine learning,” *arXiv:1808.00023v2*, 2018.

[14] 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*, pp. 7854–7876, 2022.

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

[16] Y. Runze, X. Liu, J. Dutcher, M. Tumminia, D. Villalba, S. Cohen, *et al.*, “A compuational framework for modeling biobehavioral rhythms from mobile and wearable data streams,” *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 3, p. 47, 2022.

[17] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, “Casas: A smart home in a box,” *Computer*, vol. 46, no. 7, pp. 62–69, 2012.

[18] A. Ghods, K. Caffrey, B. Lin, K. Fraga, R. Fritz, M. Schmitter-Edgecombe, and D. J. Cook, “Iterative design of visual analytics for a clinician-in-the-loop smart home,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1742–1748, 2019.

[19] 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,” *International Journal of Nursing Studies Advances*, vol. 4, p. 100081, 2022.

[20] S. Aminikhanghahi and D. J. Cook, “Enhancing activity recognition using CPD-based activity segmentation,” *Pervasive and Mobile Computing*, vol. 53, no. 75-89, 2019.

[21] J. Dahmen and D. J. Cook, “Indirectly-supervised anomaly detection of clinically-meaningful health events from smart home data,” *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 2, pp. 1–18, 2021.

[22] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, (New York, NY, USA), p. 259–268, Association for Computing Machinery, 2015.

[23] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, “Fairness in criminal justice risk assessments: The state of the art,” *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[25] S. Das, M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and M. B. Zafar, “Fairness measures for machine learning in finance,” *The Journal of Financial Data Science*, 2021.

[26] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” *CoRR*, vol. abs/1610.02413, 2016.

[27] J. Burger, H. Black, W. O. Douglas, J. M. Harlan II, W. J. Brennan Jr., P. Stewart, B. White, T. Marshall, and H. Blackmun, “Griggs v. duke power co., 401 u.s. 424 (1971),” 1971.

[28] C. DeSmet and D. J. Cook, “HydraGAN: a cooperative agent model for multi-objective data generation,” *ACM Transactions on Intelligent Systems and Technologies*, 2023.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” *Advances in neural information processing systems*, vol. 30, 2017.

[30] E. Arana-Chicas, F. Cartujano-Barrera, K. K. Rieth, K. K. Richter, E. F. Ellerbeck, L. S. Cox, K. D. Graves, F. J. Diaz, D. Catley, and A. P. Cupertino, “Effectiveness of recruitment strategies of Latino smokers: Secondary analysis of a mobile health smoking cessation randomized clinical trial,” *Journal of Medical Internet Research*, vol. 24, no. 6, p. e34863, 2022.