24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems

Adel Abusitta[1] and **Esma Aïmeur**[2] and **Omar Abdel Wahab**[3]

Abstract. In this paper, we propose a new framework for mitigating biases in machine learning systems. The problem of the existing mitigation approaches is that they are model-oriented in the sense that they focus on tuning the training algorithms to produce fair results, while overlooking the fact that the training data can itself be the main reason for biased outcomes. Technically speaking, two essential limitations can be found in such model-based approaches: 1) the mitigation cannot be achieved without degrading the accuracy of the machine learning models, and 2) when the data used for training are largely biased, the training time automatically increases so as to find suitable learning parameters that help produce fair results. To address these shortcomings, we propose in this work a new framework that can largely mitigate the biases and discriminations in machine learning systems while at the same time enhancing the prediction accuracy of these systems. The proposed framework is based on conditional Generative Adversarial Networks (cGANs), which are used to generate new synthetic fair data with selective properties from the original data. We also propose a framework for analyzing data biases, which is important for understanding the amount and type of data that need to be synthetically sampled and labeled for each population group. Experimental results show that the proposed solution can efficiently mitigate different types of biases, while at the same time enhance the prediction accuracy of the underlying machine learning model.

1 INTRODUCTION

The world is facing a historical shift toward adopting Artificial Intelligence (AI) to automate the decision-making process in many sectors, including those of health, transportation and public services. This, however, has led to growing concerns about the bias and discrimination that these systems might produce, which might negatively affect citizens especially those who belong to ethnic and racial minorities. The hazard of bias becomes even more crucial when these systems are applied to critical and sensitive domains such as health care and criminal justice. In fact, biased AI systems are mainly engendered by the data used to feed the training process of the machine learning algorithms [16] [44][52]. Training data can be incomplete, insufficiently diverse, biased, and/or consisting of non-representative samples that are not well (or poorly) defined before use [16], which might lead to biased results and lower accuracy [16] [49]. Obtaining and labeling new data to compensate and overcome these problems is one possible solution to fight against biases. However, it has been shown that such a strategy is largely difficult, costly, privacy-sensitive and dangerous, especially in some critical domains like transportation and health [24] [36].

Many approaches have been recently proposed to fight against bias and discrimination in machine learning systems. The problem of the existing mitigation approaches [40] [42] [35] is that they overlook the fact that the data used to train the machine learning algorithm might be the root cause of unfair results. In particular, these approaches focus on tuning the training algorithms to decrease the chances of producing biased results. Although such a model-based strategy might end up producing fair results, the accuracy of the underlying machine learning algorithm will be largely degraded. In other words, the mitigation will be achieved on the account of the overall prediction accuracy [17] [44] [21]. Besides, when the training data are largely biased, the time needed to complete the training and obtain a fair model will dramatically increase, compared to the case of traditional training algorithms. The reason is that these approaches not only try to minimize the loss function (in order to teach the machine learning model), but also work on minimizing the chances of producing unfair results. Thus, a longer training time is needed to find the suitable parameters for a fair model.

To address the above-mentioned shortcomings, we propose a new framework for mitigating biases in machine learning systems, without degrading their accuracy. The proposed framework is based on conditional Generative Adversarial Networks (cGANs) [43], special versions of the Generative Adversarial Networks (GANs) [25], which have shown unprecedented success in generating high-quality new synthetic data with selective properties. The proposed framework allows the designers of the machine learning systems to estimate the real distribution of the original data pertaining to the targeted population groups (population groups that are victims of biases) through formulating a minimax two-player game [7] [9]. The game is played between two models, which are trained simultaneously, i.e., the Discriminator (*Dis*) and the Generator (*Gen*). *Gen* is trained to capture the data distribution through trying to maximize the probability of *Dis* committing a mistake. On the other hand, *Dis* is trained to maximize the probability that a data sample came from a targeted population group rather than the *Gen*. The training of both *Dis* and *Gen* is repeated over many iterations until a generative model that can generate new synthetic data pertaining to the targeted population groups is obtained. The resulting generative model is then used to synthetically produce new data, which are used to augment the training set so as to compensate and overcome the bias problem. In this way, machine learning algorithms can be trained on these data in order to produce unbiased predictions.

Unlike similar works (e.g., [55]), the proposed model gives the designers of the machine learning systems the flexibility to decide on the amount of data that needs to be synthetically sampled and

[1] McGill University, Canada, email: adel.abusitta@mcgill.ca
[2] University of Montreal, Canada, email: aimeur@iro.umontreal.ca
[3] Université du Québec en Outaouais, Canada, email: omar.abdulwahab@uqo.ca

labeled, taking into account their domain knowledge. The proposed framework is also designed to be integrated into another framework for analyzing and understanding data biases. The objective is to guide the machine learning model designers on the amount and type of data that needs to be synthetically sampled and labeled. This, in turn, minimizes the chances of synthetically generating unnecessary data. Our contributions are summarized as follows. **First**, we propose a new framework for mitigating biases in machine learning systems while at the same time enhancing their overall accuracy. **Second**, we integrate the proposed mitigation framework into an analytical framework for understanding data biases. This allows us to infer the type and amount of data that needs to be synthetically sampled in order to augment the training data. **Finally**, we propose a new framework that gives the designers of the machine learning systems the flexibility to decide on the amount of data that needs to be synthetically sampled and labeled, taking into account their domain knowledge.

2 RELATED WORK

The idea of using adversarial training for mitigating biases in machine learning systems has recently been addressed in several works. For example, Madras et al. [42] propose a "fair" representation of data [39] that can be used by the classifier to generate fair decisions. They employ GANs to ensure that the generated representation of data is fair. Similarly, Louppe et al. [40] propose a new approach called "Pivot-based approach". The framework also uses GANs not to generate new synthetic data but to create a new classifier that guarantees unbiased predictions. The method modifies the GANs design through changing the role of the generator from learning how to generate new synthetic data to a classifier that is used to produce fair results. During the training of GANs, the classifier is optimized and updated based on the prediction losses of the sensitive attributes (Ethnicity, Gender, etc.). The main disadvantage of this approach is that it does not care about the overall accuracy of the classifier during the bias mitigation process. It only cares about reducing the biased results in the classifier. In other words, the mitigation in this approach is achieved on the account of the overall accuracy. In contrast, our framework can reduce biases while at the same time enhancing the overall system's accuracy.

Xu et al. [55] also adopt the GANs with the aim of generating new synthetic fair data, which are then used to train the classifier on how to produce unbiased decisions. For this purpose, another discriminator was used to check if the fairness has been achieved or not. Similar approaches have been proposed in [14], [21] [32] and [38]. These data-driven mitigation approaches suffer from three essential shortcomings. First, they propose to generate new data for each particular population group, thus leading to unnecessary data and unnecessary overhead. Second, these approaches require frequently verifying the machine learning model to check whether the generated data lead to a fair model or not. Third, these approaches are not complemented by any framework for analyzing and understanding data biases. This makes the designers of the machine learning systems unable to efficiently estimate and understand the amount and type of data that need to be synthetically sampled and labeled.

In contrast, our proposed mitigation approach is coupled with a framework for analyzing data biases. This is important to understand the amount of data that needs to be synthetically sampled for each particular population group. Moreover, the proposed framework gives the designers of machine learning systems the flexility to decide on the amount of data that should be synthetically sampled, taking into account both the domain knowledge and prediction accuracy with respect to the original data. As a result, the proposed model enables us to achieve fair machine learning systems while at the same time enhancing the accuracy of the prediction with minimum training overhead.

Celis et al. [17] formulate the adversarial problem as a multi-objective optimization model and try to find the fair model using a gradient descent-ascent algorithm with a modified gradient update step [17]. In fact, their approach is inspired by the work proposed by [57], while adding more robust theoretical foundations. Similarly, Agarwal et al. [11] propose a minimax optimization problem, which is solved using the saddle point methods [37] in order to derive the fair model. Other model-based mitigation approaches also are proposed in [22] [46] [54] [29]. These approaches propose algorithms to find suitable thresholds for trained classifiers so as to ensure equalized and fair odds. In particular, they try to fix the decision boundary in such a way to ensure that the final classifier is fair.

Most of the above-mentioned model-based mitigation approaches do not consider the training data as a potential reason for biased results. Instead, they focus only on modifying the training algorithms to produce fair results. Two main disadvantages can be distinguished in such an approach. First, the mitigation is achieved on the account of the accuracy. Second, the time needed to obtain the fair model is higher than that in traditional training algorithms, especially when the data used for training are largely biased [10] [8]. This is because these models are not only trained to minimize the loss function, but also to minimize the chances of producing unfair results.

3 REAL-WORLD EXAMPLES ON UNFAIRNESS IN MACHINE LEARNING SYSTEMS

Machine learning plays an important role in the economic development of the Information Technology society. This is the case for end users (employees, citizens, clients), service providers (government, corporations of all sectors) and infrastructure providers (Information Technology sector) who all benefit from automated and intelligent systems. Unfortunately, empirical studies reported an inequitable or biased behavior in many recent machine learning-based applications [18]. One should note that bias and discrimination can take on different meanings in different contexts (e.g. politics, psychology, economy). However, in this paper, we define bias as a prejudice in favor or against a person or population group [16].

One concrete example of bias in machine learning comes from Northpointe's tool [2][13], called COMPAS (i.e., Correctional Offender Management Profiling for Alternative Sanctions) [2], which is used to predict whom criminals are most likely to recommit crimes [50]. It has been shown that the COMPAS tool produces biased results. In particular, recent research conducted by ProPublica [3] found that the tool is more likely (i.e., two times more) to incorrectly predict that defendants from persons of color will recommit crimes. Along with the same line, the research also found that the tool is more likely to incorrectly predict that caucasian defendants are less likely to reoffend [3].

Another example is that of the Google Photos application [5], which is used to categorize images through detecting objects contained in them. Recent empirical analysis on this application have shown some racial bias [6]. Also, Amazon's Rekognition [4], which is a cloud-based popular application for facial analysis, was found to produce racial and gender-biased outputs [6].

A final example is that of the Google Word2Vec Model [33], which is used to learn the vector representation of words and is widely used in both research and Nature language processing (NLP) applications

[48]. It has been shown that word2vec word embeddings learnt from huge amounts of text frequently show gender bias. The reason is that the (Euclidean) distance function used to measure the distance between words largely related words like nurse, homemaker with the she pronoun and words like doctor, manager with the he pronoun. As a result, any application built on top of a Word2Vec Model is very likely to be affected by this bias [34].

4 THE PROPOSED FRAMEWORK FOR MITIGATING MACHINE LEARNING BIASES

In this section, we provide the details of our framework proposed for mitigating biases in machine learning systems. We first give some explanations on Generative Adversarial Networks and conditional Generative Adversarial Networks and then present the proposed mitigation model in detail, followed by our framework for analyzing data biases.

4.1 Generative Adversarial Nets and the Conditional Version

Generative adversarial networks (or GANs) is a new generative model that has been proposed by [25]. A generative model can be seen as a way of learning any kind of data distribution using unsupervised learning techniques [12] [30]. Although several generative models have been proposed in the literature such as Deep Belief Network (DBN) [30] and Variational Autoencoder (VAE) [20], GANs have received more attention thanks to their unprecedented ability to generate new synthetic high-quality data compared to the traditional generative models. In fact, GANs consist of two models: a discriminative (*Des*) and a generative (*Gen*) models. *Gen* is trained to capture the data distribution through trying to maximize the probability of *Dis* committing a mistake. On the other hand, *Dis* is trained to maximize the probability that a data sample came from a targeted population group rather than the *Gen*. The training of both the discriminative and generative models is repeated over many iterations until the discriminative model becomes unable to distinguish whether the underlying data is a sample from the data or generated from the generater. This framework is also known as a minimax two-player game [45] [28] [27] and is described formally as follows:

$$min_{Gen} \ \ max_{Dis} V(Gen,Dis) = \\ \mathbf{E}_{x\sim p_{data}(x)} log[Dis(x)] + \mathbf{E}_{z\sim p_z(z)} log[1-Dis(Gen(z))] \quad (1)$$

Formula (1) shows that both the Discriminator and the Generator are trained simultaneously. In particular, the Discriminator is trained to maximize the probability that a data sample came from the training data rather than the Generator. On the other hand, the Generator is trained to capture the data distribution through trying to maximize the probability of the Discriminator committing a mistake [25].

Conditional Generative Adversarial Networks (or cGANs) [43] are a special case of GANs which have shown great success in generating high-quality new synthetic data with selective properties. Although Goodfellow et. al [25] have already indicated in their original work the possibility of training cGANs, their work did not provide theoretical and experimental results to support this claim. cGANs can be achieved through adding a condition c as an input in both *Gen* and *Dis*. The formal description of cGANs is described as follows:

$$min_{Geb} \ \ max_{Dis} V(Gen,Dis) = \\ \mathbf{E}_{x\sim p_{data}(x)} log[Dis(x|c)] + \mathbf{E}_{z\sim p_z(z)} log[1-Dis(Gen(z|c))] \quad (2)$$

Formula (2) looks similar to Formula (1). However, the only difference is that the condition c is considered during the training of both the Generator and the Discriminator. In fact, c could be any type of data or information, for example, class labels or type of data [43].

4.2 The Proposed Model

The proposed mitigation model is based on cGANs. As can be shown in Figure 1, we train *Gen* to synthetically produce new synthetic data based on the Targeted Population Groups (TPG). $TPGs$ represent those population groups against whom the machine learning models produce biased results. The new data generated using the proposed framework are then used to augment the training data (incomplete and biased data). The new data (original data and generated data) will then be used to train the machine learning algorithms.

In the next section, we present a new framework used for analyzing data biases and exploring the $TPGs$. This framework is designed to be integrated into the proposed mitigation approach in order to allow the designers of the machine learning systems to understand the amount and type of data that should be synthetically sampled for each population group. To this end, the objective function of a two-player minimax game is defined as follows:

$$min_{Gen} \ \ max_{Dis} V(Gen,Dis) = \\ \mathbf{E}_{x\sim p_{data}(x)} log[Dis(x|TPG)] + \mathbf{E}_{z\sim p_z(z)} log[1-Dis(Gen(z|TPG))] \quad (3)$$

In Formula (3), both the Discriminator and the Generator are trained simultaneously. Also, the condition TPG is considered during the training.

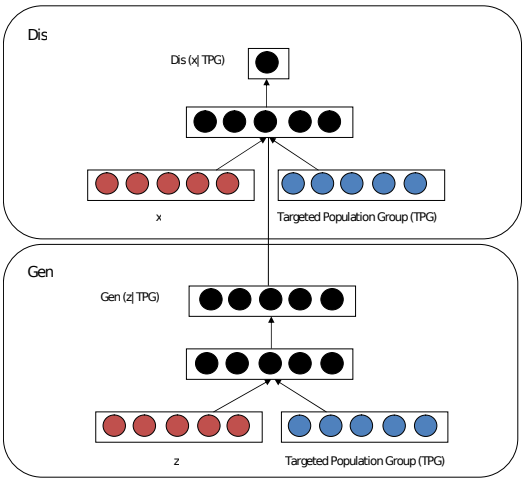


Figure 1: The architecture of our proposed model

Since the standard training of GANs cannot easily converge (i.e., non-convergence problem) [23] and to avoid mode collapse [23], we adopt a Primal-Dual Sub-gradient method to solve this problem. This method is proposed by [19] and can be seen as a Lagrangian perspective of GANs [19]. To this end, we construct a convex optimization problem as follows:

$$maximize \ \sum_{i=1}^{n} p_{data}(x_i|TPG) log(Dis(x_i|TPG)) \\ Subject\ to: \ (1-log(Dis(x_i|TPG)) \geq log(1/2), i=1,...,n \\ \boldsymbol{Dis} \in \mathscr{S}, \quad (4)$$

Algorithm 1: Algorithm for training a generator

Input: Targeted Population Group (TPG)
repeat
  Sample n1 data samples x_i, i= 1, ..., n1 (minibatch sampling)
  Sample n2 noise samples z_i, i= 1, ..., n2 (minibatch sampling)
  for *K steps* **do**
    Update the *Dis* through ascending the stochastic gradient:

$$\nabla_{\theta_{data}}[\frac{1}{n1}\sum_{1}^{n1} log(Dis(x_i|TPG)) + \frac{1}{n2}\sum_{1}^{n2} log(1 - Dis(Gen(z_i|TPG)))] \quad (5)$$

  end
  Update the *Gen* distribution as follows:

$$\tilde{p}_{gen}(x_i|TPG) = p_{gen}(x_i|TPG) - \beta log(2(1 - Dis(x_i|TPG))), i = 1, ...n1 \quad (6)$$

  where β represents some step size and

$$p_{gen}(x_i|TPG) = \frac{1}{n2}\sum_{j=1}^{n2} k_\sigma(Gen(z_j|TPG) - x_i). \quad (7)$$

  Update the Gen through descending the stochastic gradient:

$$\nabla_{\theta_{gen}}[\frac{1}{n2}\frac{1}{n2} log(1 - Dis(Gen(z_j|TPG))) + \frac{1}{n1}\sum_{1}^{n1}(\tilde{p}_{gen}(x_i|TPG) - p_{gen}(x_i|TPG))^2] \quad (8)$$

until ε *elapses*;

where $\mathscr{S}$ is some convex set and the variables are ***Dis***=($Dis(x_1|TPG)$,...$Dis(x_n|TPG)$). Let $p_{gen|TPG}$ = ($p_{gen}(x_1|TPG)$, ... , $p_{gen}(x_n|TPG)$), where $p_{gen}(x_i|TPG)$ is the Lagrangian dual associated with the i-th constraint. Formula (4) shows the Lagrangian version of GANs, which is proposed by [19] to avoid mode collapse problem [23]. The formula adopts a Primal-Dual Sub-gradient method to address this problem.

The proposed training algorithm (Algorithm 1), which is inspired by [19], is based on (4). In Algorithm 1, the targeted population group (TPG) is taken as an input and the goal is to train *Gen* to produce data that cope with the TPG. In the proposed algorithm, the process of updating of *Dis* is similar to the standard *cGAN* training; however, the process of updating *Gen* is different. For the *Gen*, when the data distribution and generated distribution have disjoint supports [26] [19], the *Gen* may not be updated using standard *cGAN* training (7) (8) (9). This is useful to prevent the main source of mode collapse [19]. Note that after a certain fixed period denoted by ε, the whole steps are repeated in order to enable both the *Gen* and *Dis* to learn how to produce new high-quality synthetic data, based on the targeted population group.

4.3 How To Analyze Data Biases?

In the previous section, we proposed a new algorithm (Algorithm 1) for learning how to train the generator on how to create new synthetic data based on a given targeted population group. The algorithm takes as an input a targeted population group in order to learn how to produce new data with respect to that particular group. In this section, we present a new framework that can be used to explore the set of targeted population groups to be used as inputs for Algorithm 1. Note that this framework is inspired by the analysis presented in [51] for detecting biases in machine learning models, while adapting it to our case where we are interested in detecting biases in the data itself rather than in the machine learning model.

The following steps are used for the analysis of data biases. First, select a set of population groups to study if the classifier produces biased results against any of them. Second, train the classifier on the training data. Third, test the classifier by producing results and visualizing the prediction accuracy with respect to each population group. The visualization can be achieved either by showing the probability distribution or by displaying the accuracy obtained for each population group. Finally, analyze these results to see which population group(s) is/are victim(s) of biases.

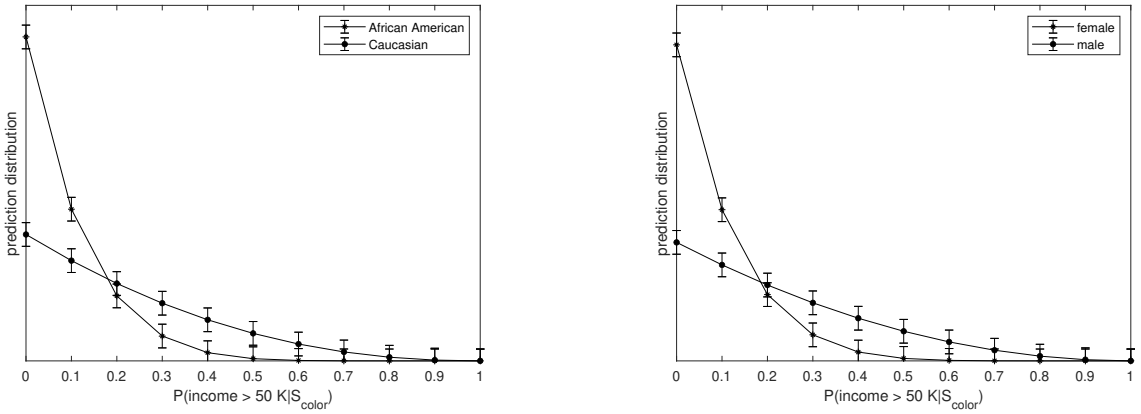
We use the following example to illustrate how does the above-described framework practically work. Consider the adult UCI dataset [56], which is used to predict the salary of a person (below 50K$ or above 50k$). The dataset contains two Sensitive Attributes (*SA*), i.e., Ethnicity and Gender. This leads us to the following four population groups: African American, Caucasian, Female and Male. Although we could have combinations of these population groups (e.g., African American females), we restrict, for the sake of simplicity and without loss of generality, our example to only the above mentioned four population groups.

To determine if the training data are biased or not, we need to test whether a machine learning classifier, that is trained on these data, produces biased results or not. To this end, we trained a neural network classifier on this dataset and analyzed the prediction accuracy, taking into account above mentioned population groups. The results of our testing are given in Figure 2.

Figure 2a shows the distributions of the predicted P(income $>$ 50K$) given the *SA* $S_{Ethnicity}$ = {African American, Caucasian}. The Figure shows that for the ethnicity attribute, the prediction distribution of an "African American" has a large value at the low interval of $[0.1 - 0.2]$ compared to a "Caucasian". These results suggest that when a person is an "African American", the probability that the classifier will predict his/her income below 50K$ is much higher compared to a "Caucasian". Similarly, Figure 2b shows the distributions of the predicted P(income $>$ 50K$) given the *SA* S_{Gender} = {female,male}. The Figure shows that for the gender attribute, the prediction distribution of a "female" has a large value at the low interval of [0.1 -0.2] compared to a "male". These results suggest that when a person is "female", the probability that the classifier will predict her income below 50 K$ is much higher compared to a "male". The results shown in Figure 2 give us a clear indication that the data used for training is incomplete (i.e., the number of Caucasians and males in the dataset is greater than that of African Americans and females). Therefore, we conclude that the targeted population groups that should be used as inputs to Algorithm 1 based on to the above results are: $S_{Ethnicity}$ = {African American} and S_{Gender} = {female}. Simply put, the generator will be trained to generate new African Americans and females.

5 Experimental Evaluation

This section first describes the setup used to evaluate the proposed framework. Then, the performance of the proposed bias mitigation



(a) (b)

Figure 2: (Left) Prediction distribution of the original training data with respect to the ethnicity attribute. (Right) Prediction distribution of the original training data with respect to the gender attribute.

framework is examined. We compared the proposed model with a recent work proposed in [40]. This work is called as a "Pivot-based mitigation approach" and it uses GANs not to generate new synthetic data (like we do) but to create a new classifier that guarantees fairness in predictions. The method makes a modification on the GANs through changing the role of the generator from learning how to generate new synthetic data to a classifier that is used to produce fair results. During the training process of GANs, the classifier is optimized and updated based on the prediction losses of the sensitive attributes (e.g., Ethnicity, Gender, etc.). Although we did not compare our model with [55] and other data-driven mitigation approaches, we mentioned the main shortcomings of these approaches in the related work section.

5.1 Experimental Setup

We implemented the proposed framework using Multilayer Perceptrons (MLPs) with 3 hidden layers. We used the ReLU activation function for both the generators and discriminators. The **two** following datasets were tested: the adult UCI dataset [56] and the Adience dataset [1], which widely used for age and gender prediction. The adult UCI dataset consists of 48842 instances, with 14 attributes [56]. The Adience dataset consists of 26,580 photos distributed in 8 age categories (0 to 2, 4 to 6, 8 to 13, 15 to 20, 25 to 32, 38 to 43, 48 to 53, 60 and above) with the corresponding gender label [1] [47]. Since the adult UCI dataset contains categorial data, we placed in parallel a dense-layer per categorical variable, followed by Gumbel-Softmax activation and a concatenation to get the final output [15] [31] [41]. Prediction performance on the validation dataset is adopted for finding the best hyper-parameter configuration. The results are reported based on a 95% confidence interval. We train a classifier on our dataset using the 10-fold cross-validation model.

5.1.1 Results on the adult UCI dataset

Figure 3 shows the results obtained when applying the proposed framework on the adult UCI dataset. In particular, Figure 3a shows the progress achieved in the prediction distribution compared to Figure 2a. This progress was achieved when we augmented the original data (female) by 85% new data obtained synthetically from the generator. Figures 3b also shows the progress achieved in the prediction distribution, compared to Figure 2b, when we augmented the original data (African American) by 85% new data obtained synthetically from the generator. Note that the proposed framework is flexible in the sense that it enables machine learning designers to control the amount of data (e.g, 85%) that needs to be synthetically added for each population group. This allows the designers to consider the "Domain knowledge" during the data augmentation process.

Table 1 shows a comparison between the proposed approach and a recent work proposed in [40]. Table 1 shows the overall accuracy obtained by the proposed model when training the MLP on the new training data (original data + generated data) with different numbers of Hidden Units (HUs). These results are better than the results obtained using the 'Pivot-based mitigation approach". Our model also yields a better accuracy compared to the baseline. The baseline means that the classifier was trained on the original data without adding new synthetic data. This can be justified by the fact that the data used for training was incomplete (unbalance) and led to biased results, in the sense of having a lower measure of accuracy [16]. The proposed framework overcame this problem through augmenting the training data to mitigate biases and enhancing the prediction accuracy. It is worth mentioned here that the prediction accuracy in Table 1 is obtained with respect to some held-out test dataset (of only real data with no synthetic data).

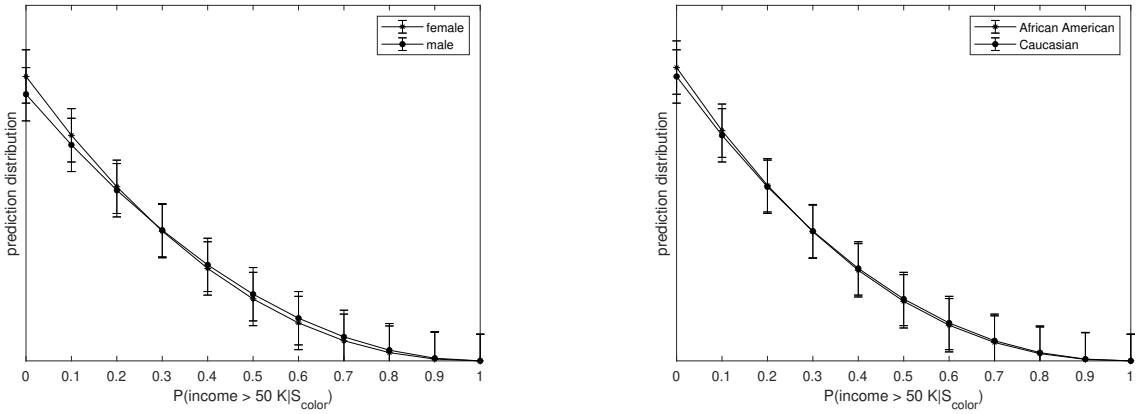
5.1.2 Results on the Adience dataset

Table 2 studies the accuracy of the MLP classifier with respect to a given population group. The results suggest the existence of bias against the women of color. Table 3 shows the progress achieved in the prediction accuracy compared to Table 2 when the training data was augmented with more data on women of color, which were synthetically obtained from the generator (the proposed framework).

Table 4 shows the overall prediction accuracy of the MLP classifier trained on the new training data. These results outperform both the pivot-based classifier and the baseline.

6 LIMITATION

Although the proposed framework has the advantage of mitigating bias in machine learning systems against targeted groups, we cannot claim that our solution fully solves the problem. In fact, bias is a broad and undefined problem, which does not always target members of minority groups (e.g., female). For example, Google conducted a recent study to determine whether the company is underpaying



(a) (b)

Figure 3: (Left) Prediction distribution when 85% of new synthetic data (female) were added to the original dataset. (Right) Prediction distribution when 85% of new synthetic data (African American) were added to the original dataset.

Table 1: Comparison of the prediction accuracy of different approaches (Adult UCI dataset)

	Acc. (300 HUs)	Acc. (500 HUs)	Acc. (700 HUs)	Acc. (900 HUs)
The Proposed Approach	**84.9± 1.14**	**85.1± 1.09**	**85.3± 1.92**	**85.5± 1.15**
Pivot-based Approach	76.1 ± 1.11	76.4 ± 1.84	77.1± 1.23	77.3± 1.78
Baseline	82.0± 1.16	82.3± 1.06	82.6 ± 1.90	82.9 ± 0.88

Table 2: Classification performance with respect to a population group

	Acc. (300 HUs)	Acc. (500 HUs)	Acc. (700 HUs)	Acc. (900 HUs)
Men of color	86.5 ± 0.34	87.68 ± 0.33	87.15 ± 0.22	87.5 ± 0.26
Women of color	**67.8± 0.14**	**67.9± 0.29**	**68.3± 0.36**	**68.4 ± 2.40**
Caucasian men	98.6 ± 0.24	98.7 ± 0.35	98.8 ± 0.19	98.0 ± 0.21
Caucasian women	90.4 ± 0.45	91.3 ± 0.38	91.8 ± 2.02	91.7 ± 0.44

Table 3: Classification performance after the augmentation with the data on women of color (300%)

	Acc. (300 HUs)	Acc. (500 HUs)	Acc. (700 HUs)	Acc. (900 HUs)
Men of color	87.9 ± 0.38	88.1 ± 0.45	88.1 ± 0.84	88.3 ± 0.66
Women of color	**88.1 ± 0.27**	**88.2 ± 0.30**	**88.5 ± 0.34**	**88.6 ± 0.22**
Caucasian men	99.2± 0.31	99.3 ± 0.42	99.5 ± 0.28	99.7 ± 0.37
Caucasian women	91.9 ± 0.66	92.0 ± 0.41	92.3 ± 2.07	92.5 ± 0.23

Table 4: Comparison of overall prediction accuracy (Adience dataset)

	Acc. (300 HUs)	Acc. (500 HUs)	Acc. (700 HUs)	Acc. (900 HUs)
The Proposed Approach	**91.77± 0.29**	**91.9± 0.36**	**92.10± 0.41**	**92.27± 0.25**
Pivot-based Approach	81.71± 0.29	80.43± 0.38	81.01± 0.31	81.37± 0.29
Baseline	85.82± 0.33	86.39± 0.24	86.51± 0.31	86.40± 0.37

women or not. Surprisingly, they found that men were less paid than women even for the same job position [53]. Therefore, we argue that more efforts need to be done to generalize the proposed framework for unpredictable bias cases.

7 CONCLUSION AND FUTURE WORK

This paper presents a new framework for the mitigation of biases in machine learning systems. The proposed framework is based on conditional generative adversarial networks, which allows us to generate new high-quality synthetic data related to the targeted population groups. The proposed framework is integrated into the proposed analytical framework used for understanding of data biases. This allows us to understand the type and amount of data that should be synthetically sampled to augment the training data and overcome the bias problem. The training process then takes place on the new data (original data + generated data). Our model also enables the mitigation to be applied while taking into consideration the knowledge domain. Experimental results show that the proposed framework mitigates the biases against targeted population groups while at the same time enhancing the prediction accuracy of the machine learning classifiers.

As future work, we plan to design an automated mitigation process. In particular, after defining the bias, the system should automatically generate new data and perform unbiased training. The challenge here is to make the system automatically determine the exact amount of data that should be sampled, taking into account the knowledge domain.

ACKNOWLEDGEMENTS

The financial support of the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged. We also would like to acknowledge Dr. Gilles Brassard (University of Montreal), Dr. Kimiz Dalkir (McGill University), Younes Driouiche (Mila), Alexis Tremblay, Amine Belabed and Rim Ben Salem for helpful discussions.

REFERENCES

[1] *The Adience data set*, 2019 (accessed April 2, 2019). https://talhassner.github.io/home/projects/Adience/Adience-data.html#agegender.
[2] *Fairness and bias of the COMPAS algorithm compared to human assessments*, 2019 (accessed November 10, 2019). https://www.hiit.fi/.
[3] *COMPAS Recidivism Risk Score Data and Analysis*, 2019 (accessed November 15, 2019). https://www.propublica.org/.
[4] *Amazon Rekognition*, 2019 (accessed November 17, 2019). https://aws.amazon.com/rekognition/?nc1=h_ls.
[5] *Google Photos*, 2019 (accessed November 17, 2019). https://play.google.com/store/apps.
[6] *Machine learning and bias*, 2019 (accessed November 17, 2019). https://developer.ibm.com/articles/machine-learning-and-bias/.
[7] Adel Abusitta, Martine Bellaiche, and Michel Dagenais, 'On trustworthy federated clouds: A coalitional game approach', *Computer Networks*, **145**, 52–63, (2018).
[8] Adel Abusitta, Martine Bellaiche, and Michel Dagenais, 'An svm-based framework for detecting dos attacks in virtualized clouds under changing environment', *Journal of Cloud Computing*, **7**(1), 9, (2018).
[9] Adel Abusitta, Martine Bellaiche, and Michel Dagenais, 'A trust-based game theoretical model for cooperative intrusion detection in multi-cloud environments', in *2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, pp. 1–8. IEEE, (2018).
[10] Adel Abusitta, Martine Bellaiche, Michel Dagenais, and Talal Halabi, 'A deep learning approach for proactive multi-cloud cooperative intrusion detection system', *Future Generation Computer Systems*, (2019).
[11] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach, 'A reductions approach to fair classification', *arXiv preprint arXiv:1803.02453*, (2018).
[12] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, 'Greedy layer-wise training of deep networks', in *Advances in neural information processing systems*, pp. 153–160, (2007).
[13] Adrienne Brackey, *Analysis of Racial Bias in Northpointe's COMPAS Algorithm*, Ph.D. dissertation, Tulane University School of Science and Engineering, 2019.
[14] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney, 'Optimized pre-processing for discrimination prevention', in *Advances in Neural Information Processing Systems*, pp. 3992–4001, (2017).
[15] Ramiro Camino, Christian Hammerschmidt, and Radu State, 'Generating multi-categorical samples with generative adversarial networks', *arXiv preprint arXiv:1807.01202*, (2018).
[16] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford, 'Ai now 2017 report', *AI Now Institute at New York University*, (2017).
[17] L Elisa Celis and Vijay Keswani, 'Improved adversarial learning for fair classification', *arXiv preprint arXiv:1901.10443*, (2019).
[18] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova, 'Artificial intelligence, bias and clinical safety', *BMJ Qual Saf*, **28**(3), 231–237, (2019).
[19] Xu Chen, Jiang Wang, and Hao Ge, 'Training generative adversarial networks via primal-dual subgradient methods: a lagrangian perspective on gan', *arXiv preprint arXiv:1802.01765*, (2018).
[20] Carl Doersch, 'Tutorial on variational autoencoders', *arXiv preprint arXiv:1606.05908*, (2016).
[21] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, (2015).
[22] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander, 'Satisfying real-world goals with dataset constraints', in *Advances in Neural Information Processing Systems*, pp. 2415–2423, (2016).
[23] Ian Goodfellow, 'Nips 2016 tutorial: Generative adversarial networks', *arXiv preprint arXiv:1701.00160*, (2016).
[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.
[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial nets', in *Advances in neural information processing systems*, pp. 2672–2680, (2014).
[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, 'Improved training of wasserstein gans', in *Advances in Neural Information Processing Systems*, pp. 5767–5777, (2017).
[27] Talal Halabi, Martine Bellaiche, and Adel Abusitta, 'A cooperative game for online cloud federation formation based on security risk assessment', in *2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pp. 83–88. IEEE, (2018).
[28] Talal Halabi, Martine Bellaiche, and Adel Abusitta, 'Toward secure resource allocation in mobile cloud computing: A matching game', in *2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 370–374. IEEE, (2019).
[29] Moritz Hardt, Eric Price, Nati Srebro, et al., 'Equality of opportunity in supervised learning', in *Advances in neural information processing systems*, pp. 3315–3323, (2016).
[30] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, 'A fast learning algorithm for deep belief nets', *Neural computation*, **18**(7), 1527–1554, (2006).
[31] Eric Jang, Shixiang Gu, and Ben Poole, 'Categorical reparameterization with gumbel-softmax', *arXiv preprint arXiv:1611.01144*, (2016).
[32] Faisal Kamiran and Toon Calders, 'Data preprocessing techniques for classification without discrimination', *Knowledge and Information Systems*, **33**(1), 1–33, (2012).
[33] Dhruvil Karani, *Introduction to Word Embedding and Word2Vec*, 2019 (accessed November 17, 2019). https://towardsdatascience.com/.
[34] Matthew Kenney, *Amazon Rekognition*, 2019 (accessed November 17, 2019). http://matthewkenney.site/biases.html.
[35] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba, 'Undoing the damage of dataset bias', in *European Conference on Computer Vision*, pp. 158–171. Springer, (2012).
[36] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, 'Semi-supervised learning with deep generative models', in *Advances in neural information processing systems*, pp. 3581–3589, (2014).
[37] Jyrki Kivinen and Manfred K Warmuth, 'Exponentiated gradient versus gradient descent for linear predictors', *information and computation*, **132**(1), 1–63, (1997).
[38] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris, 'Adaptive sensitive reweighting to mitigate bias in fairness-aware classification', in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 853–862. International World Wide Web Conferences Steering Committee, (2018).
[39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *nature*, **521**(7553), 436, (2015).
[40] Gilles Louppe, Michael Kagan, and Kyle Cranmer, 'Learning to pivot with adversarial networks', in *Advances in neural information processing systems*, pp. 981–990, (2017).
[41] Chris J Maddison, Andriy Mnih, and Yee Whye Teh, 'The concrete distribution: A continuous relaxation of discrete random variables', *arXiv preprint arXiv:1611.00712*, (2016).
[42] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel, 'Learning adversarially fair and transferable representations', *arXiv preprint arXiv:1802.06309*, (2018).
[43] Mehdi Mirza and Simon Osindero, 'Conditional generative adversarial nets', *arXiv preprint arXiv:1411.1784*, (2014).
[44] Tom M Mitchell, *The need for biases in learning generalizations*, Department of Computer Science, Laboratory for Computer Science Research . . . , 1980.
[45] Barry O'Neill, 'Nonmetric test of the minimax theory of two-person zerosum games', *Proceedings of the national academy of sciences*, **84**(7), 2106–2109, (1987).

[46] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger, 'On fairness and calibration', in *Advances in Neural Information Processing Systems*, pp. 5680–5689, (2017).
[47] Pau Rodríguez, Guillem Cucurull, Josep M Gonfaus, F Xavier Roca, and Jordi Gonzalez, 'Age and gender recognition in the wild with deep attention', *Pattern Recognition*, **72**, 563–571, (2017).
[48] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf, 'Transfer learning in natural language processing', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18, (2019).
[49] Surendra K Singhi and Huan Liu, 'Feature subset selection bias for classification learning', in *Proceedings of the 23rd international conference on Machine learning*, pp. 849–856, (2006).
[50] William Dieterich Tim Brennan, *Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)*, 2019 (accessed November 10, 2019). https://www.researchgate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS.
[51] Stijn Tonk, *Towards fairness in ML with adversarial networks*, 2019 (accessed April 2, 2019). https://blog.godatadriven.com/fairness-in-ml.
[52] Antonio Torralba and Alexei A Efros, 'Unbiased look at dataset bias', in *CVPR 2011*, pp. 1521–1528. IEEE, (2011).
[53] Daisuke Wakabayashi, *Google Finds It's Underpaying Many Men as It Addresses Wage Equity*, 2019 (accessed May 2, 2019). https://www.nytimes.com/2019/03/04/technology/google-gender-pay-gap.html.
[54] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro, 'Learning non-discriminatory predictors', *arXiv preprint arXiv:1702.06081*, (2017).
[55] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu, 'Fairgan: Fairness-aware generative adversarial networks', in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, (2018).
[56] Show-Jane Yen and Yue-Shi Lee, 'Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset', in *Intelligent Control and Automation*, 731–740, Springer, (2006).
[57] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, 'Mitigating unwanted biases with adversarial learning', in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. ACM, (2018).