Gradient Reversal Against Discrimination: A Fair Neural Network Learning Approach

Edward Raff
Booz Allen Hamilton
raff_edward@bah.com

Jared Sylvester
Booz Allen Hamilton
sylvester_jared@bah.com

***Abstract*—No methods currently exist for inducing fairness in arbitrary neural network architectures. In this work we introduce GRAD, a new and simplified method for producing fair neural networks that can be used for auto-encoding fair representations or directly with predictive networks. It is easy to implement and add to existing architectures, has only one (insensitive) hyper-parameter, and provides improved individual and group fairness. We use the flexibility of GRAD to demonstrate multi-attribute protection.**

***Index Terms*—fairness, neural networks, ease of use**

I. Introduction

Artificial Neural Network methods are quickly becoming ubiquitous in society, spurred by advances in image, signal, and natural language processing. This pervasiveness leads to a new need for considering the fairness of such networks from many perspectives, including: how they are used, who can access them and their training data, and potential biases in the model itself. There are many reasons for desiring fair classification algorithms. These include legal mandates to be non-discriminative, ensuring a moral or ethical goal, or for use as evidence in legal proceedings [1]. Despite the long-standing need and interest in this problem, there are few methods available today for training fair networks.

When we say that a network is fair, we mean fair with respect to a protected attribute a_p, such as age or gender. Our desire is that a model's predicted label $\hat{y}$ given a feature vector x is invariant to changes in a_p. An initial reaction may be to simply remove a_p from the feature vector x. While intuitive, this does not remove the correlations with a_p that exist in the data, and so the result will still produce a biased model [2].

For this reason we need to devise approaches that explicitly remove the presence of a_p from the model's predictions. We do so in this work by introducing a new method to train fair neural networks. Our approach, termed *Gradient Reversal Against Discrimination* (GRAD), makes use of a network which simultaneously attempts to predict the target class y and protected attribute a_p. The key is that the gradients resulting from predictions of a_p are reversed before being used for weight updates. The result is a network which is capable of learning to predict the target class but effectively inhibited from being able to predict the protected attribute.

GRAD displays competitive accuracy and improved fairness when compared to prior approaches, despite introducing only one new hyper-parameter (which in practice does not need adjustment). Combined with GRAD's flexibility with respect to network architecture, this makes it easier to apply — an important consideration for obtaining practical use [3]. Additionally GRAD is the first, to the authors' knowledge, *neural network*-based approach that can protect multiple attributes simultaneously. Prior works in this space are generally limited to one attribute and require the introduction of multiple hyper-parameters. These parameters must be cross-validated, making the approaches challenging to use. Further, our approach can be used to augment any current model architecture, where others have been limited specifically to auto-encoder architectures.

The rest of our paper is organized as follows. In section II we will discuss the related prior work in building fair neural networks. In section III we will introduce our new approach, which dramatically simplifies the process. Then we will review the evaluation methodology in section IV, followed by our results in section V. We will discuss these results in section VI and then conclude in section VII.

II. Related Work

The first work that explored using neural networks for fairness was the Learning Fair Representations (LFR) approach by Zemel, Wu, Swersky, *et al.* [4]. This seminal approach was based on constructing an auto-encoder combined with a prototype-based projection. Each datum x was mapped to each prototype in a weighted combination, with a constraint that the distribution of points by their protected attribute a_p is equal across the prototypes. Classification was then done by training a normal Logistic Regression model on top of these prototypes. In addition, a loss based on the prediction of the label y from the prototypes was included to encourage task specific performance, making three total loss terms. A hindrance of using the LFR approach in practice is that each of the loss terms introduces its own hyper-parameter, requiring a cumbersome increase to the amount of hyper-optimization that must be done. Zemel, Wu, Swersky, *et al.* also introduced metrics to quantify the fairness of a predictor: one for *group fairness* and one for *individual fairness*. We will use these same measures and discuss them further in section IV, though the other works we evaluate have unfortunately chosen to ignore the individual-fairness metric.

Following the seminal work of Zemel, Wu, Swersky, *et al.*, Louizos, Swersky, Li, *et al.* [5] introduce the Variational Fair Auto-Encoder (VFAE). Their work proposed to treat the

problem of fair prediction as a domain adaptation problem, and that by treating the protected attribute as a new domain one could coax the network to learn a representation invariant to a_p. We will use this same insight in our own design for a fair network.

As the VFAE name alludes to, it extends the Variational Auto-Encoder to the task of fair classification, and again uses Logistic Regression trained on the hidden representation to perform prediction. Fairness is obtained by using the Maximum Mean Discrepancy (MMD) to perform domain adaptation. This introduces two hyper parameters α and β that must be dealt with.

The final pre-existing approach to constructing fair neural networks is Adversarial Learned Fair Representations (ALFR), developed by Edwards and Storkey [6]. Their work combines a Generative Adversarial Network (GAN) [7] with auto-encoding, with supervised prediction of the target attribute y, and with a negative log-loss on the protected attribute. This introduces three new hyper-parameters to balance between the network's auto-encoding, log-loss, and negative log-loss terms, in addition to the GAN-specific hyper-parameter search that must be done to balance the representational power of the generative and adversarial portions of the network. Another issue for practical use is the general convergence challenges that exist with GANs [8].

While LFR and AFRL have an auto-encoding component, they are also tied to predicting a specific attribute. This means they are not as task-flexible as VFAE is (though Zemel, Wu, Swersky, *et al.* did include a brief discussion of results that exclude the y-based terms for task-flexible use). In approaches like VFAE, the construction of a fair representation via just auto-encoding means the hidden representations can be shared and used for multiple predictive tasks, without needing to re-create new representations for each task. Our GRAD approach, with its greater flexibility, can be used in auto-encoding or directly predictive styles. This allows balance between the need for sharing (one encoding for multiple tasks) and better potential accuracy through specificity (one encoding for one task). GRAD is also the only approach that does not mandate an auto-encoding component, which we hypothesize allows for its higher group fairness.

Other non-neural network approaches to fairness have been developed as well. These include attempts to protect attributes by modifying the features [9] and modifying the labels [10] of a dataset. A prevalent strategy is to introduce a new regularization term that penalizes use of the protected attribute [11]–[14], of which our work is a member. In all of these prior works, it is assumed that both the classification task and the attribute to protect are binary, and that only one attribute needs to be protected. The flexibility of GRAD does not constrain it to this type of problem, but we re-use it so that we can compare with prior work.

In subsection V-B we also investigate and discuss the protection of multiple attributes simultaneously. This is a critical feature for practical use of fairness systems. In any situation in which there are sensitive attributes, it is likely that more than one is present. For instance, datasets with demographic information likely contain more than one feature from the eight "protected classes" defined in federal anti-discrimination law in the United States. While other techniques may potentially be adapted to multiple attributes simultaneously the vast majority of prior papers do not discuss how to adjust their techniques to account for this situation, nor do they present the results of doing so. We are aware of only two prior papers which explicitly examine the protection of multiple attributes. Johndrow and Lum [15] mention that their approach is compatible with this goal, but do not discuss it further. Zafar, Valera, Rogriguez, *et al.* [16] showed the results of protecting gender and race concurrently with a modified logistic regression system, but their treatment of the topic was limited to a couple of sentences and a footnote, and presented no analysis or discussion of the results. While this is a valuable contribution, we feel that this is an important practical issue that deserves a more thorough treatment. We are aware of no other work that has tackled this issue, and so we attempt here to add to the discussion of protecting multiple attributes.

III. Gradient Reversal Against Discrimination

We now present our new approach to developing neural networks that are fair with respect to some protected attribute. We call it Gradient Reversal Against Discrimination (GRAD), and it is inspired by recent work in transfer learning. Notably, Ganin, Ustinova, Ajakan, *et al.* [17] introduced the idea of domain adaptation by attempting to jointly predict a target label and a domain label (i.e., which domain did this data instance come from?). By treating the protected attribute as the new domain, we can use this same approach to instead prevent the network from being biased by the protected attribute.

A sketch of the architecture used for GRAD can be seen in Figure 1. After several feature extraction layers the network forks. One branch learns to predict the target y, while the other attempts to predict the protected attribute a_p. We term the portion of the network before the splitting point the "trunk," and those portions after the "target branch" and the "attribute branch." The final loss of the network is sum of the losses of both branches.

$$\ell(y, a_p) = \ell_t(y) + \lambda \cdot \ell_p(a_p) \tag{1}$$

Here, λ determines the relative importance of fairness compared to accuracy. In practice, we find that performance is insensitive to particular choices of λ, and any value of $\lambda \in [50, 2000]$ would perform equivalently. In our experiments we will use $\lambda = 100$ without any kind of hyper-parameter optimization.

The values of both $\ell_t(y)$ and $\ell_p(a_p)$ are calculated and used to determine gradients for weight updates as usual, with one important exception. When the gradients have been back-propagated from the attribute branch they are reversed (i.e., multiplied by -1) before being applied to the trunk. This moves the trunk's parameters *away* from optima in predictions of a_p, crippling the ability to correctly output the protected attribute. Since the target branch also depends on the trunk parameters, it inherits this inability to accurately output the

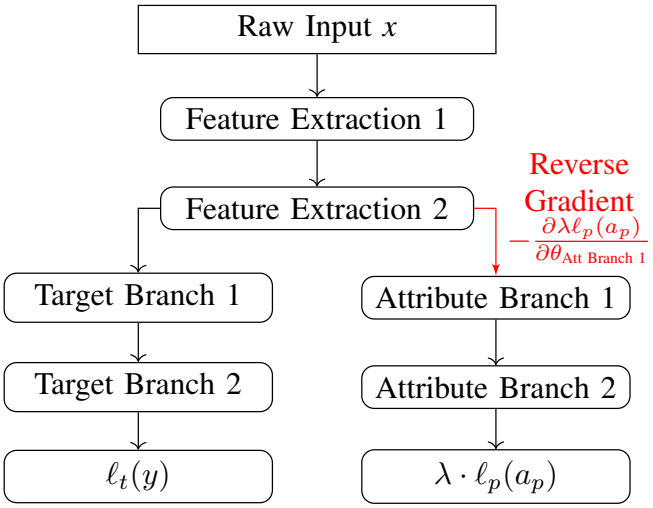

Fig. 1. Diagram of GRAD architecture. Red connection indicates normal forward propagation, but back-propagation will reverse the signs. The value x is the input to the network, and the terminal nodes are the losses that get back-propagated.

value of the protected attribute. No such reversal is applied to the gradients derived from y, so the network's internal state representations are suitable for predicting y but nescient of a_p.

In order to understand why we attempt to have the attribute branch correctly predict a_p and then reverse the resulting gradient, it is instructive to consider why it may be insufficient to set up a loss function which directly punishes the network for correctly predicting a_p. If this were the case, the network could achieve low loss by forming internal representation which are very good at predicting the protected attribute, and then "throw the game" by simply reversing the correct prediction in the penultimate layer. (That is, a potential, reliable strategy to getting the wrong answer is to become very good at getting the right answer, and then lying about what one thinks the answer should be.) If this strategy were adopted then the representations necessary for correctly recovering a_p from x would be available to the target branch when making its prediction of y, which is the situation we aim to prevent.

Architecture Variants

As mentioned above, many of the other neural approaches to fair classification take an auto-encoder or representation learning approach. This approach has its advantages. For instance, it allows the person constructing the fair model to be agnostic about the ultimate task that it will be applied to. Others like ALFR consider a target value directly, and so can not be re-used for other tasks, but may perform better in practice on the specific problem they were constructed for.

Our GRAD approach, thanks to its comparative simplicity, can be used in both formulations. This makes it the only neural network-based approach to fairness that offers both task flexibility and specificity.

GRAD-Auto will designate our GRAD approach when we use an auto-encoder as the target branch's loss. That is to say, if x is the input feature, we will denote $\tilde{x}$ as the feature vector derived from x such that the protected attribute $a_p \notin \tilde{x}$. We then use $\ell_t^{\text{Auto}}(\cdot) = ||h_{\text{target}} - \tilde{x}||_2^2$ as the loss function for the target branch, where h_{target} is the activation vector from the last layer of the target branch. This approach matches the same style as the LFR and VFAE approach, where a hidden representation invariant to a_p is learned, and then Logistic Regression is used on the final activations from the trunk sub-network to perform classification.

GRAD-Pred will designate our task-specific approach, where we use the labels y_i directly. Here we simply use the standard logistic loss $\ell_t^{\text{Pred}}(\cdot) = \log(1+\exp(-y \cdot h_{\text{target}}))$. In this case the target branch of the network will produce a single activation, and the target branch output itself is used as the classifier directly.

Since we are dealing with binary protected attributes, both GRAD-Auto and GRAD-Pred will have the attribute branch of the network use $\ell_p(a_p) = \log(1 + \exp(-a_p \cdot h_{\text{attribute}}))$. We can also protect multiple attributes simultaneously. Considering Z different values to protect, we can use $\ell_p(a_{p_1}, \ldots, a_{pZ}) = \sum_{\forall z \in Z} \log(1 + \exp(-a_{p_z} \cdot h_{\text{attribute}_z}))$.

In the spirit of minimizing the effort needed by the practitioner, we do not perform any hyper-parameter search for the network architecture either. Implemented in Chainer [18], we use two fully-connected layers for every branch of the network (trunk, target & attribute) where all hidden layers have 40 neurons. Each layer will use batch-normalization [19] followed by the the ReLU activation function [20]. Training is done using the Adam optimizer for gradient decent [21]. We train each model for 50 epochs, and use a validation set to select the model from the best epoch. We define best by the model having the lowest Discrimination (see §IV-A) on the validation set, breaking ties by selecting the model with the highest accuracy. When multiple attributes are protected, we use the lowest average Discrimination.

IV. Methodology

There is currently ongoing debate about what it means for a machine learning model to be fair, which metrics should be used, and whether or not they can be completely optimized [22]–[26].

We choose to use the same evaluation procedure laid out by Zemel, Wu, Swersky, *et al.* [4]. This makes our results comparable with a larger body of work, as their approach and metrics have been widely used through the literature (e.g. [12], [14], [27], [28]). We present results for both of the metrics they adopt: Discrimination and Consistency.

A. Metrics

Given a dataset $\{x_1, \ldots, x_n\} \in \mathcal{D}$, we define the ground truth label for the ith datum as y_i and the model's prediction as $\hat{y}_i$. Each are with respect to the binary target label $y \in \{0, 1\}$. While we define both y_i and $\hat{y}_i$, we emphasize that only the predicted label $\hat{y}_i$ is used in the fairness metrics. This is because fairness is not directly related to accuracy by equality of treatment.

Discrimination (also referred to as "demographic parity") is measured by the taking the difference between the average predicted scores for each attribute value, assuming a_p is a binary attribute.

$$\text{Discrimination} = \left| \frac{\sum_{x_i \in \mathcal{D}_{a_p}} \hat{y}_i}{|\mathcal{D}_{a_p}|} - \frac{\sum_{x_i \in \mathcal{D}_{\neg a_p}} \hat{y}_i}{|\mathcal{D}_{\neg a_p}|} \right| \quad (2)$$

Here, $\mathcal{D}_{a_p} \subset \mathcal{D}$ is the subset of data which possesses the sensitive or protected value of a_p, while $\mathcal{D}_{\neg a_p} \subset \mathcal{D}$ does not. Thus $\mathcal{D}_{a_p} \cup \mathcal{D}_{\neg a_p} = \mathcal{D}$ and $\mathcal{D}_{a_p} \cap \mathcal{D}_{\neg a_p} = \emptyset$. Discrimination measures a macro-level quality of fairness, requiring the model to have the same prediction rate for the target value y in each population. As such it is measuring *group*-fairness.

Discrimination is the only fairness metric considered in both the VFAE and ALFR works [5], [6], but Discrimination is not sufficient as a metric to satisfy a desired notion of fairness [24]. There may easily exist sub-populations within a_p and $\neg a_p$ for which the average predictions differ greatly within the sub-populations [15], [29].

To quantify this scenario, the metric of Consistency was introduced as a measure of *individual*-fairness.

$$\text{Consistency} = 1 - \frac{1}{N} \sum_{i=1}^{N} \left| \hat{y}_i - \frac{1}{k} \sum_{j \in k\text{-NN}(x_i)} \hat{y}_j \right| \quad (3)$$

For each datum $x_i \in \mathcal{D}$, we compare its prediction y_i with the average of its k nearest neighbors. Consistency is the average of this score across all points in $\mathcal{D}$.

Because Consistency and Discrimination are independent of the actual accuracy of the method used, we also consider the *Delta* score, where Delta = Accuracy − Discrimination. This gives a combined measure of an algorithm's accuracy that penalizes it for biased predictions.

We use and evaluate Consistency, Discrimination, and Delta in the same manner and on the same datasets as laid out in Zemel, Wu, Swersky, *et al.* [4] so that we can compare our results with prior work. This includes using the same training, validation, and testing splits.

When Protecting Multiple Attributes: In our work, we will also consider the ability to protect multiple attributes simultaneously. When we do so, we will use Discrimination(x) to refer to the Discrimination score with respect to a specific attribute. If we have Z attributes to protect, we use Delta = Accuracy − $Z^{-1} \sum_{z \in Z}$ Discrimination(a_z). That is to say, we extend Delta as the accuracy minus the average Discrimination.

B. Data Sets

To evaluate our work, we will use three classification datasets used by Zemel, Wu, Swersky, *et al.* [4] and one from Edwards and Storkey [6]. The first, second and fourth datasets can be obtained from the UCI repository [30].

- German Credit: This corpus has $n = 1,000$ and 20 features (7 numerical, 13 categorical). The goal is to predict if someone has good or bad credit, and the protected attribute a_p is whether the person is 25 years of age or older.
- Adult Income: Here $n = 45,222$ after removing instances with missing values. There are 14 predictive variables, including the protected attribute of gender (male or female). The goal is predict if an individual's income is $\geq \$50,000$ per year.
- Heritage Health: This is the largest dataset at $n = 147,473$, and was introduced in a Kaggle competition [31]. We use the same 139 features developed by the winning Kaggle team, and the protected attribute is whether the patient is 65 years old or older. The goal is to predict if they will spend one or more days in the hospital this year.
- Diabetes: This dataset has 101,765 rows and attempts to predict if a patient will be re-admitted to the hospital. The protected attribute is race (Caucasian/not Caucasian).

We make a special note that while we compare results on the Diabetes dataset, we are unable to use exactly the same features. We present the results anyway under the belief that imperfect comparison is better than none at all. The original feature set presented by Strack, DeShazo, Gennings, *et al.* [32] has 47 categorical features. However, Edwards and Storkey [6] reported using 235 features without specifying how these features were extracted from the original set. There are three columns in the original Diabetes dataset related to diagnostic hospital codes. If performing a one-hot encoding, the diagnostic feature subset yields 701 unique occurring values, which would already eclipse the number of features reported by Edwards and Storkey. In attempts to be as compatible as possible with Edwards and Storkey, we map each of the values in these three columns to nine high-level groups as outlined in Table 2 of Strack, DeShazo, Gennings, *et al.* [32], and leave all other features as-is. We then perform one-hot encoding to obtain a feature space of 611 features. This is closer to the 235 reported by Edwards and Storkey [6]. As a result, we caution the reader against inferring too much from such comparisons, but present them so as to provide at least some amount of analogy with ALFR.

C. Models Evaluated

As a baseline for comparison against GRAD-Pred and GRAD-Auto, we will consider the same architecture but with the attribute branch removed. This produces a standard neural network, and will be denoted as *NN*. We also consider a standard logistic regression model. These give us some idea what the accuracy, discrimination and consistency would be if no efforts were made toward achieving fair outcomes. For comparison with other fairness-seeking algorithms, we present prior results for fair Logistic Regression (LRF) [11], fair Naive Bayes (NBF) [9], Fair Random Forests (FF) [33], Learning Fair Representations (LFR) [4], Variation Fair Auto-Encoders (VFAE) [5], and Adversarial Learned Fair Representations (ALFR) [6].

For all models on all datasets, we report the metrics as presented in their original publications. The VFAE and ALFR works only show scores in plots [5], [6], so we extract the

TABLE I

FOR EACH DATASET WE SHOW ACCURACY, DELTA, DISCRIMINATION, AND CONSISTENCY FOR OUR NEW METHOD AND PRIOR WORK. NN-* MODELS ARE NEURAL NETWORKS WITH $\lambda = 0$ (I.E. IGNORING FAIRNESS AS A GOAL), WHILE GRAD-* MODELS HAVE $\lambda = 100$. *-AUTO MODELS USE AN AUTOENCODER FORMAT, WHILE *-PRED PERFORM CLASSIFICATION. BEST RESULTS SHOWN IN **BOLD**, SECOND BEST IN *italics*.

	German				Adult				Health			
Algorithm	Acc	Delta	Discr	Cons	Acc	Delta	Discr	Cons	Acc	Delta	Discr	Cons
NN-Auto	*0.7350*	0.5334	0.2016	0.8730	0.7635	0.7191	0.0444	0.9850	**0.8506**	0.7939	0.0567	0.9730
GRAD-Auto	0.6750	0.6296	0.0454	0.8705	0.7554	0.7452	0.0102	*0.9924*	0.8491	**0.8491**	**0.0000**	**1.0000**
NN-Pred	**0.7500**	0.3637	0.3863	0.6945	0.7022	0.6268	0.0754	0.8168	0.8440	0.7511	0.0929	0.9453
GRAD-Pred	0.6750	0.6744	*0.0006*	*0.9705*	0.7543	0.7543	**0.0000**	**1.0000**	*0.8493*	0.8486	*0.0007*	*0.9999*
NBF	0.6888	0.6314	0.0574	0.6868	*0.7847*	*0.7711*	0.0136	0.5634	0.6878	0.5678	0.1200	0.5893
FF	0.7000	**0.7000**	**0.0000**	**1.0000**	0.7530	0.7530	**0.0000**	**1.0000**	0.8474	0.8474	**0.0000**	**1.0000**
LR	0.6790	0.5517	0.1273	0.6950	0.6787	0.4895	0.1892	0.7297	0.7547	0.6482	0.1064	0.7233
LRF	0.5953	0.5842	0.0111	0.8716	0.6758	0.6494	0.0264	0.7766	0.7212	0.7038	0.0174	0.6223
LFR	0.5909	0.5867	0.0042	0.9408	0.7023	0.7018	*0.0006*	0.8108	0.7365	0.7365	**0.0000**	**1.0000**
VFAE	0.7270	*0.6840*	0.0430	—	**0.8129**	0.7421	0.0708	—	0.8490	*0.8490*	**0.0000**	—
ALFR	—	—	—	—	0.8251	**0.8241**	0.0010	—	—	—	—	—

score from the figures.[1] For the Diabetes dataset, we were able to run the Fair Forest algorithm to produce our own scores.

V. RESULTS

Now that we have explained our methodology, we can begin examining the results of our GRAD approach compared to prior works. The majority of methods examined the same datasets used by Zemel, Wu, Swersky, *et al.* [4]. For this reason we group them together, and the results can be seen in Table I. In each column we present Accuracy, Delta (our primary metric), Discrimination, and Consistency (Equation 3). For values unreported in their original work, we show a dash ("—") in the table. Our GRAD approach is shown in the top rows, where "NN" indicates the same network trained with $\lambda = 0$ (i.e., no fairness goal). The bottom seven rows include the other approaches as explained in subsection IV-C.

When we compare the standard neural network (NN) with its GRAD counterpart, we can see that the GRAD approach *always* increases the Delta and Consistency scores, and reduces the Discrimination. This shows its applicability across network types (classifying and auto-encoding). We can even see the GRAD approach improve accuracy on the Adult dataset by 5 percentage points. While we would not expect this behavior (i.e. a negative cost of fairness) in the general case, it is nonetheless interesting and it may indicate that the protected gender attribute of the adult dataset is misleading to the normal network's learning.

Comparing the GRAD algorithms to the other neural networks LFR, VFAE and ALFR, we see that GRAD is usually best or 2nd best in each metric. On both the Adult and Health datasets, it achieves the best Discrimination and Consistency scores compared to any of the algorithms tested. On the German dataset VFAE obtains a higher Delta score by having a high accuracy, though VFAE has 4% discrimination compared to GRAD-Pred's 0.06%. On the Health dataset, GRAD-Auto and GRAD-Pred have near identical results, and differences in the 4th significant figure is the only distinguishing factor. This is overall significantly better than the LFR approach which has an 11 percentage point difference in Accuracy and Delta scores compared to the GRAD approaches. The VFAE algorithm is similarly within a fractional distance, though Consistency is not reported for VFAE.

For the methods where Consistency was reported, we emphasize that GRAD approach reliably produces high Consistency scores regardless of the version, and one of the GRADs always obtained the best Consistency. As discussed in section IV, Consistency is an important metric to quantify that captures information about sub-population level discrimination. It may be possible that ALFR achieves its higher accuracy by being consistent at a macro level while being inconsistent in its predictions at the micro-level. This is not known since it was left unreported, and is the reason we make sure to include Consistency in our results.

A. Diabetes Results

In this section we present the results on the Diabetes dataset. As mentioned in section IV, the results for FLR and ALFR are taken from Edwards and Storkey [6]. For this reason their values have an asterisk (*) to indicate that are not perfectly comparable.

The results are shown in Table II, where we see that LFR and ALFR appear to both have superior accuracy compared to the neural networks presented in this work. ALFR appears to have comparable Discrimination, though GRAD-Auto produces the lowest Discrimination score. The GRAD approach does not look as competitive in this table, but we suspect the performance difference is due to the feature disparity.

We draw this conclusion in part because NN-Pred, which is a normal fully connected network, obtains an accuracy lower than the LFR or ALFR approaches. It would be unusual to expect adding the fairness constraint to any classifier would

[1]This was done using the website https://apps.automeris.io/wpd/. Both papers are on arXiv.org, but the LaTeX source does not have the values in question. Authors were contacted for each paper but did not reply.

significantly *increase* accuracy. If it was the case that the fairness acted as a regularizer, we would expect NN-Pred to have increased accuracy as well, as we observed on the Adult and Health datasets. Altering the network size of NN-Pred did not significantly change these results. Thus we believe the better performance of LFR and ALFR is due to the unspecified feature construction used in Edwards and Storkey [6].

Considering just our GRAD approach, we do see that it continues to successfully reduce Discrimination and increase Consistency in all cases. This makes GRAD successful in its goal of improving the fairness of the naive neural networks.

B. Multiple Protected Attributes

In all but one prior works that we are aware of, it is assumed that there is only *one* attribute that needs to be protected. This is, however, a myopic view of the world. All of the protected attributes that have been tested individually in this work, like age, race and gender, may co-occur and interact with each other in a single corpus. (Why would one need to protect age when predicting credit score, and gender when predicting income, but not vice versa? Yet this is exactly what the standard benchmarks would have one do.) We show this interaction between potentially protected attributes in Table III using the Diabetes dataset, which has both race and gender as features in the corpus. In this case GRAD-Pred and GRAD-Auto are protecting race and gender attributes. GRAD-Pred-R shows the results for protecting only race, and GRAD-Pred-G shows for only protecting gender. GRAD-Auto follows the same convention.

Since Discrimination is computed with respect to specific attributes, in the table we show the discrimination scores with respect to both of the protected attributes. Since we have two protected attributes a_{p_1} and a_{p_2}, we compute Delta = Accuracy $-($ Discrimination(a_{p_1}) + Discrimination(a_{p_2}) $)/2$. In doing so, we can see that when two protected variables are present, the GRAD approach is able to reduce Discrimination and increase Delta for both the auto-encoder and the standard softmax predictive network. GRAD-Pred also continues to increase the Consistency with respect to the naive neural network.

TABLE II
ACCURACY, DELTA, DISCRIMINATION, AND CONSISTENCY FOR GRAD AND PRIOR WORK ON THE DIABETES DATASET. BEST RESULTS IN **BOLD**, SECOND BEST IN *italics*. ASTERISK (*) INDICATES RESULTS USING A DIFFERENT FEATURE SET.

Algorithms	Acc	Delta	Discrm	Cons
NN-Auto	0.5735	0.5323	0.0412	0.6411
GRAD-Auto	0.5851	0.5848	**0.0003**	0.6404
NN-Pred	0.6286	0.5868	0.0418	0.6464
GRAD-Pred	0.5844	0.5824	0.0020	*0.7538*
FF	0.5390	0.5385	*0.0005*	**0.9974**
LFR*	*0.6413*	*0.6271*	0.0142	—
ALFR*	**0.6537**	**0.6524**	0.0013	—

TABLE III
ACCURACY, DELTA, DISCRIMINATION (WITH RESPECT TO RACE AND GENDER), AND CONSISTENCY FOR OUR NEW METHOD ON THE DIABETES DATASET. BEST RESULTS IN **BOLD**, SECOND BEST IN *italics*. LAST FOUR ROWS SHOW GRAD MODELS WHEN ONLY RACE (R) OR GENDER (G) ARE PROTECTED.

			Discrimination		
Algorithms	Acc	Delta	Race	Gender	Cons
NN-Auto	0.5735	0.5392	0.0412	0.0275	0.6411
GRAD-Auto	0.5765	0.5723	*0.0055*	**0.0030**	0.6288
NN-Pred	**0.6286**	*0.5848*	0.0418	0.0458	*0.6464*
GRAD-Pred	*0.5980*	**0.5949**	**0.0028**	*0.0034*	**0.7180**
GRAD-Auto-R	0.5851	0.5749	0.0003	0.0201	0.6404
GRAD-Auto-G	0.5640	0.5143	0.0981	0.0013	0.6093
GRAD-Pred-R	0.5844	0.5478	0.0020	0.0713	0.7538
GRAD-Pred-G	0.5941	0.5526	0.0785	0.0045	0.6849

Comparing GRAD-Pred with GRAD-Pred-R and GRAD-Pred-G is critical to show that protecting both attributes simultaneously provides a significant benefit. On the Diabetes dataset, we see the model increase its discrimination with respect to gender when only race is protected. Similarly, when we protect gender, discrimination with respect to race increases. Explicitly protecting both is the only safe way to reduce discrimination on both.

The model shifting to leverage other protected features is not surprising. In this case, race and gender are not correlated features, so penalizing the use of one does not directly penalize the other. When we penalize a feature which provides information, the model must attempt to recover discriminative information in other (potentially non-linear) forms from the other features. Since the other protected attribute is not correlated, the model takes no penalty when it increases its use of that attribute. Thus the importance and utility of GRAD to protect both simultaneously is established.

We also note that the penalty for protecting multiple attributes is not necessarily higher than protecting a single attribute. For the auto-encoding approach, GRAD-Auto obtains nearly identical accuracy on the Diabetes corpus as just NN-Auto, but with a considerably better Delta score and reduced discrimination. If we needed a general purpose feature representation to use for multiple tasks, the cost in this scenario was minimal.

C. Robustness to λ

We have discused so far that a benefit of the GRAD approach is a simplicity in application due to the having only one hyper-parameter λ. We now show that this value λ is largely robust to the value used. In Figure 2 we plot the Accuracy, Discrimination, and Consistency as a function of λ for values in the range $[1, 2000]$.

The largest variation comes from the German dataset, though Discrimination and Accuracy are have less variation for $\lambda \geq 50$. The Consistency score instead has some variability. This variation is not entirely unexpected given the small size of the

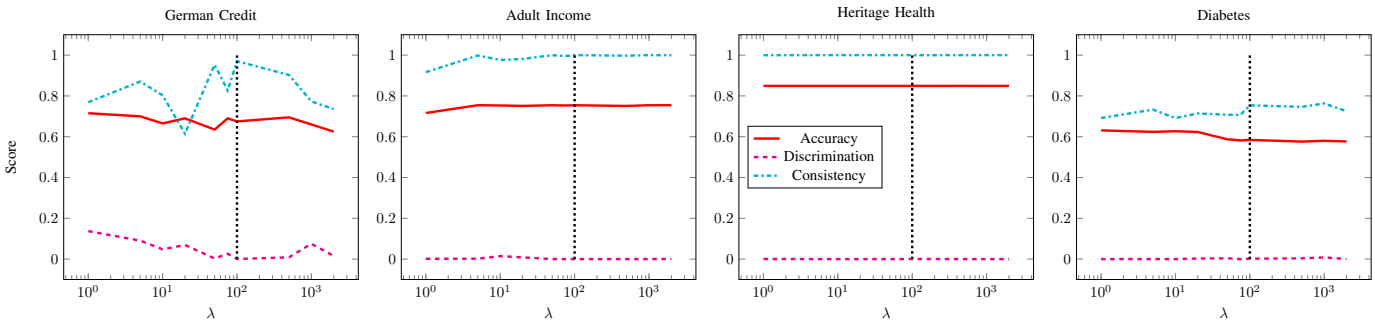

Fig. 2. Plots show the performance of GRAD-Pred as a function of λ on the x-axis (log scale). The y-axis shows Accuracy, Consistency (higher is better) and Discrimination (lower is better). A dashed vertical black line shows the value $\lambda = 100$ used in all experiments. Above each plot is the dataset used. All plots share the same legend.

German dataset, which contains only $n = 1000$ samples total. This is an un-ideal use case for neural networks in general which have historically performed best when an abundance of data is available.

The Adult and Health datasets are more representative of the GRAD approach's normal behavior. On the Adult dataset, we see results stabilize after $\lambda \geq 10$. The Health dataset looks flat through the entire plot. This is not in fact the case, but the variation is on the order of 10^{-3}, making it visually indiscernible.

GRAD's performance on the Diabetes dataset is also consistent, though has a slight change as λ increases. For one large range $\lambda \in [1, 20]$, the Accuracy of GRAD-Pred is slightly higher at $\approx 62\%$, but Consistency somewhat lower at ≈ 0.70. The model then becomes progressively fairer and stabilizing at $\lambda \in [50, 2000]$ (which contains our default value of $\lambda = 100$) with an increased Consistency of ≈ 0.75, but slightly decreased accuracy of $\approx 58\%$. This is all despite negligible impacts to the Discrimination metric.

We believe this example highlights the importance of using both Discrimination *and* Consistency in evaluating model fairness. The GRAD approach penalizes any ability to predict the protected attribute a_p. In this case there was still some sub-population discrimination with respect to a_p when $\lambda \leq 20$. If one looked only at Discrimination, one may erroneously conclude that a smaller value of λ was better due to comparable Discrimination but improved Accuracy.

VI. Discussion

We believe we have shown GRAD is competitive with prior methods in all metrics of interest: Accuracy, Delta, Discrimination, and Consistency. It is not uniformly superior to current algorithms for building fair classifiers, however comparison has been hindered due to incomplete reporting.

The GRAD approach does have the novel benefit that it can be applied to *any* currently used neural network, and requires no additional hyper-parameters in practice. Approaches like LFR, VFAE, and ALFR all constrain the user to a particular network style and type, which may not be appropriate for any particular problem. In contrast, GRAD is completely agnostic to network architecture and could be immediately used with CNNs and RNNs as well. As neural networks are applied to larger and more diverse problems, we believe GRAD will be faster to apply (since it does not require significant new hyper-parameters) and easier to apply (since it does not force the user into a particular architecture type).

A. Task Flexibility and Specificity

Another item of importance, as mentioned in section II, is the choice between an approach's task specificity and flexibility. GRAD-Auto allows us to satisfy flexibility, and GRAD-Pred specificity. This allows us to make a trade-off that others can not perform. We can see the value in this from a quantitative perspective in Table I, where GRAD-Pred has improved Delta, Discrimination, and Consistency scores compared to GRAD-Auto on the German and Adult datasets.

LFR and VFAE are task flexible: they learn a single representation that can be shared and potentially used for multiple predictive tasks. Our results clearly show that despite lesser performance, GRAD-Auto is still competitive with LFR and VFAE, though there is no definitive "best" approach by these metrics. However, if greater accuracy is needed — using LFR or VFAE leaves no options but to re-optimize the hidden representation and bias its hyper parameters toward a task-specific goal. In actuality, this is how their training was done [4], where GRAD-Auto had no hyper parameters tuned toward the Delta, Discrimination, or Consistency metrics. In this regard GRAD-Auto is better attuned to this flexibility scenario than these prior approaches.

In regards to the auto-encoding approach, we draw the reader's attention to the fact that GRAD-Auto's discrimination is usually reduced dramatically by switching to GRAD-Pred, yet is in-line with VFA and LFR's Discrimination scores as well. We hypothesize that this may be an intrinsic weakness of the auto-encoding approach, as the auto-encoder must learn to re-produce multiple features, any number of which may be correlated with a_p. This increases the network's incentive to retain the feature as its value grows with the number of correlated variables.

B. Protecting Multiple Attributes

Protecting multiple attributes simultaneously is an important problem for future consideration, simply because there are multiple attributes that are common place and must be protected for ethical or legal reasons (such as race, gender, age, religious identity, etc.) [1]. A naive approach to embed multiple attribute protection within algorithms that are designed to protect only a single attribute is to create a new dummy attribute that represents every possible combination of the protected attributes' values. However this creates a combinatorial explosion in the state space, and does not allow for protecting attributes that might be continuous in nature (such as age). As such we feel it is critical that we discuss, as a community, how we will begin to evaluate the effectiveness of our methods for protecting multiple attributes simultaneously. While one could argue that many prior works could be "easily extended" to handle this case, it is not a forgone conclusion that they will work well. More importantly, we need some agreed upon method of determining effectiveness at multi-attribute protection.

We have evaluated GRAD using multiple protected attributes on the Diabetes dataset, which has two features that one typically desires to protect: race and gender. This shows GRAD's capability and the ease with which it can be applied, but does not sufficiently cover the space of possible scenarios. There could be many more attributes to protect, of varying combinations of discrete and continuous natures. One option is to simply collect more datasets and define all attributes that are worthy of protection. At first glance this appears to immediately solve the dilemma by evaluating with respect to a (hopefully) representative set of datasets and attributes in need of protection.

However, as we consider future algorithms for the multiple-attribute protection task, it may be that different algorithms are impacted by differing factors in non-trivial ways that require further contemplation. We argue that two in particular are of immediate consequence to future work and impact each other:

- There may be a large number Z of attributes to protect. Different methods may do best for small-to-large values of Z, and so it is likely that evaluating over a range of $2 \leq Z' \leq Z$ values is necessary.
- The correlations between two different protected attributes a_{p_i} and a_{p_j} may impact performance. This could over-regularize the model through additive effects, or cause other non-correlated attributes to be ignored.

The need to test many sub-sets of protected features to evaluate a range $Z' \in [2, Z]$ creates a combinatorial explosion of possible sub-sets of attributes to look at. Which sub-pairs of attributes to look at may have dramatically different results depending on the aforementioned correlation issue. For example, if a_{p_i} and a_{p_j} were highly correlated, we might see that protecting one implicitly protects the other, resulting in no net-benefit from protecting both simultaneously. This is the behavior we would expect from GRAD.

When explicitly protecting multiple highly correlated attributes, we may see the cumulative impact result in an over-regularization of the problem. We will use GRAD's behavior as an example to illustrate the possible effects. Consider the hypothetical situation where $a_{p_1}, a_{p_2}, \cdots, a_{p_K}$ which are all highly correlated with each other. The contributions of each of the K protected attributes would have similar values for the gradient, meaning the final gradient would have its magnitude increased by a factor of K. This could result in destabilizing the training of the network, as the contribution of the Attribute branch to the Feature Extraction branch would overshadow that of the Target branch.

This correlation problem makes the sub-set selection and evaluation more challenging than a standard combinatorial problem. Evaluation would ideally take into account the correlation of attributes to different degrees (uncorrelated, mild, and strongly correlated) to elucidate any potential weakness in each approach. Simultaneously, combinations of varying degrees of correlation are equally valid and likely to occur in practice. This does not begin to consider potentially non-linear interactions between a larger set of protected attributes.

These are all issues we have not yet seen discussed with regards to the application of fairness imbued algorithms in practice. As such we believe determining *how* to evaluate multi-attribute protection is an important research item on its own for future work.

VII. Conclusions

We have introduced GRAD, an approach for building fair neural networks that can be used to augment any network architecture. GRAD does not mandate the auto-encoding approach of prior work or additional, cumbersome hyper-parameters. GRAD is competitive with prior work, and often delivers superior fairness through low discrimination.

The GRAD approach of appending an extra output branch with reversed gradients can be used to augment any arbitrary neural network. This includes both auto-encoder networks and classification/regression networks, which allows GRAD to be used when either flexibility (in the former case) or specificity (in the latter) is desired. This simplicity of implementation — in terms of both architecture and paucity of hyper-parameters — lowers the cost to the practitioner. As such we expect it will be useful for increasing the quantity of fairness-seeking solutions in the world.

Finally, we believe that protecting multiple features concurrently has been insufficiently addressed in the prior literature despite the prevalence of this requirement in the real world. We make a contribution towards redressing this imbalance by expressly demonstrating the protection of multiple attributes simultaneously and including significant discussion on the issue.

References

[1] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 05, pp. 582–638, Nov. 2014, ISSN: 0269-8889. DOI: 10.1017/S0269888913000039. [Online].

Available: http://www.journals.cambridge.org/abstract_S0269888913000039.

[2] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware Data Mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08, New York, NY, USA: ACM, 2008, pp. 560–568, ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401959. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401959.

[3] J. Sylvester and E. Raff, "What About Applied Fairness?" In *Machine Learning: The Debates (ML-D) organized as part of the Federated AI Meeting (FAIM 2018)*, 2018. [Online]. Available: http://arxiv.org/abs/1806.05250.

[4] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., ser. Proceedings of Machine Learning Research, vol. 28, Atlanta, Georgia, USA: PMLR, 2013, pp. 325–333. [Online]. Available: http://proceedings.mlr.press/v28/zemel13.html.

[5] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The Variational Fair Autoencoder," in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: http://arxiv.org/abs/1511.00830.

[6] H. Edwards and A. Storkey, "Censoring Representations with an Adversary," in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: http://arxiv.org/abs/1511.05897.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[8] L. Mescheder, S. Nowozin, and A. Geiger, "The Numerics of GANs," in *Advances In Neural Information Processing Systems 30*, 2017. [Online]. Available: http://arxiv.org/abs/1705.10461.

[9] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*, IEEE, Feb. 2009, pp. 1–6, ISBN: 978-1-4244-3313-1. DOI: 10.1109/IC4.2009.4909197. [Online]. Available: http://ieeexplore.ieee.org/document/4909197/.

[10] B. T. Luong, S. Ruggieri, and F. Turini, "k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11, New York, NY, USA: ACM, 2011, pp. 502–510, ISBN: 978-1-4503-0813-7. DOI: 10.1145/2020408.2020488. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020488.

[11] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware Learning Through Regularization Approach," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ser. ICDMW '11, Washington, DC, USA: IEEE Computer Society, 2011, pp. 643–650, ISBN: 978-0-7695-4409-0. DOI: 10.1109/ICDMW.2011.83. [Online]. Available: http://dx.doi.org/10.1109/ICDMW.2011.83.

[12] Y. Bechavod and K. Ligett, "Learning Fair Classifiers: A Regularization-Inspired Approach," in *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1707.00044.

[13] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A Convex Framework for Fair Regression," in *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1706.02409.

[14] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling Attribute Effect in Linear Regression," in *2013 IEEE 13th International Conference on Data Mining*, IEEE, Dec. 2013, pp. 71–80, ISBN: 978-0-7695-5108-1. DOI: 10.1109/ICDM.2013.114. [Online]. Available: http://ieeexplore.ieee.org/document/6729491/.

[15] J. E. Johndrow and K. Lum, "An algorithm for removing sensitive information: Application to race-independent recidivism prediction," *ArXiv preprint arXiv:1703.04957*, 2017.

[16] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, Fort Lauderdale, FL, USA: PMLR, 2017, pp. 962–970. [Online]. Available: http://proceedings.mlr.press/v54/zafar17a.html.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial Training of Neural Networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Jan. 2016, ISSN: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=2946645.2946704.

[18] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a Next-Generation Open Source Framework for Deep Learning," in *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015. [Online]. Available: http://learningsys.org/papers/LearningSys_2015_paper_33.pdf.

[19] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of The 32nd International Conference on Machine Learning*, vol. 37, 2015, pp. 448–456.

[20] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

[21] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference On Learning Representations*, 2015.

[22] M. Skirpan and M. Gorelick, “The Authority of "Fair" in Machine Learning,” in *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1706.09976.

[23] E. García-Martín and N. Lavesson, “Is it ethical to avoid error analysis?” In *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1706.10237.

[24] M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

[25] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[26] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *ArXiv preprint arXiv:1609.05807*, 2016.

[27] V. Landeiro and A. Culotta, “Robust Text Classification in the Presence of Confounding Bias,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI’16, AAAI Press, 2016, pp. 186–193. [Online]. Available: http://dl.acm.org/citation.cfm?id=3015812.3015840.

[28] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, “Decoupled classifiers for fair and efficient machine learning,” in *FAT ML Workshop*, 2017. DOI: 1707.06613. [Online]. Available: https://arxiv.org/pdf/1707.06613.pdf.

[29] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness Through Awareness,” in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS ’12, New York, NY, USA: ACM, 2012, pp. 214–226, ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255. [Online]. Available: http://doi.acm.org/10.1145/2090236.2090255.

[30] M. Lichman, *UCI Machine Learning Repository*, 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[31] *Heritage Health Prize*, 2012. [Online]. Available: https://www.kaggle.com/c/hhp.

[32] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” *BioMed Research International*, vol. 2014, pp. 1–11, 2014, ISSN: 2314-6133. DOI: 10.1155/2014/781670. [Online]. Available: http://www.hindawi.com/journals/bmri/2014/781670/.

[33] E. Raff, J. Sylvester, and S. Mills, “Fair Forests: Regularized Tree Induction to Minimize Model Bias,” in *AAAI / ACM conference on Artifical intelligence, Ethics, and Society*, 2018. [Online]. Available: http://arxiv.org/abs/1712.08197.