ABCinML: Anticipatory Bias Correction in Machine Learning Applications

Abdulaziz A. Almuzaini
Department of Computer Science, Rutgers University
New Brunswick, NJ, USA
abdulaziz.almuzaini@rutgers.edu

Chidansh A. Bhatt
Thomas J. Watson Research Center, AI & Hybrid Cloud, IBM
Yorktown Heights, NY, USA
chidansh.amitkumar.bhatt@ibm.com

David M. Pennock
Department of Computer Science, Rutgers University
New Brunswick, NJ, USA
dpennock@dimacs.rutgers.edu

Vivek K. Singh
School of Communication and Information, Rutgers University
New Brunswick, NJ, USA
v.singh@rutgers.edu

ABSTRACT

The idealization of a static machine-learned model, trained once and deployed forever, is not practical. As input distributions change over time, the model will not only lose accuracy, any constraints to reduce bias against a protected class may fail to work as intended. Thus, researchers have begun to explore ways to maintain algorithmic fairness over time. One line of work focuses on *dynamic learning*: retraining after each batch, and the other on *robust learning* which tries to make algorithms robust against all possible future changes. Dynamic learning seeks to reduce biases soon *after* they have occurred and robust learning often yields (overly) conservative models. We propose an anticipatory *dynamic learning* approach for correcting the algorithm to mitigate bias *before* it occurs. Specifically, we make use of anticipations regarding the relative distributions of population subgroups (e.g., relative ratios of male and female applicants) in the next cycle to identify the right parameters for an importance weighing fairness approach. Results from experiments over multiple real-world datasets suggest that this approach has promise for anticipatory bias correction.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; **Batch learning**.

KEYWORDS

classification, fairness, algorithmic bias

ACM Reference Format:
Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock, and Vivek K. Singh. 2022. ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3531146.3533211

1 INTRODUCTION

Machine learning (ML) algorithms have been widely used in numerous aspects of our daily lives ranging from simpler tasks such as digit recognition to more complicated and sensitive tasks such as risk assessments [6], job hiring [38], loan lending [17], sentiment analysis [5, 29], facial analysis [4, 10], college admission [43, 50] and health care [3, 40]. Past work has often focused on optimizing the performance (i.e., accuracy) by devising various methods and frameworks, while implicitly ignoring the ethical aspects of the ML models. Recently, the artificial intelligence (AI) research community has started to examine the fairness aspects in the trained state-of-the-art ML applications and significant bias has been detected in various domains. For instance, risk assessment tools were found to discriminate against black people [6], facial recognition algorithms had higher false positive rate for minority groups including black and hispanic groups [10], and natural language processing (NLP) models, such as hate speech detection and sentiment analysis models were found to be implicitly biased against different groups of the public [5, 24]. Several in-depth surveys cover these and related findings [8, 11, 16, 23, 35, 36, 39, 48].

Although various fairness approaches have been proposed recently in the literature to tackle the issue of bias, the majority of them deal with the conventional static ML training process. Unfortunately, this approach might not generalize well in the future since the world is *non-stationary*, and that could lead to model failures. For instance, a language model that is trained on past data might struggle to generalize well for future data that they have not seen before (e.g., "COVID-19" and "universal lockdown") [34]. Also, in scenarios where models learn from future users' inputs (i.e., users behaviors are fed back to a model) without any checks or balances, this could cause the model to learn spurious correlations or even mirror racist or toxic language over time.

To address the aforementioned problems in both the accuracy and fairness aspects, researchers have proposed various solutions to help ML models generalize well in the future. Currently, there are two main approaches that deal with bias dynamics and early prevention methods: (1) bias detection in a dynamic learning paradigm

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533211

followed by a correction [26, 34], and (2) early prevention using generalization methods such as *robustness* or *domain adaptation* to prepare a model for any issues or shifts that might occur in the future (e.g., ensuring "worst-case optimal" results) [2, 31, 42, 44]. Although these methods are reasonable and effective, they suffer from some limitations. In dynamic learning, bias might be corrected *after* a user gets affected which might be harmful for that user and her group. In the case of early prevention via robustness, the model will not have access to future data or retraining, hence these approaches are likely to be either ineffective or overly conservative.

In scenarios where models interact with the public (e.g., dynamic modeling), it is better to detect bias early and then apply the mitigation methods before it exacerbates. Such techniques (like almost all bias reduction approaches) often do not eliminate bias completely but can significantly reduce its scale, which can be useful in practical settings. In an idealized case, having access to future data will allow the algorithm to tune parameters to obtain desired accuracy and fairness. In practical scenarios, a perfect estimation of future data is impossible. However, there are multiple domains where future data is not completely independent of past data and certain macro-properties of the data follow predictable patterns (e.g., monotonic increase in female representation in college applicants).[1]

Here we assume that we do not have access to specific instances of future data, but certain macro-properties of the data, for example the relative distributions of population sub-groups (e.g., male, female, wealthy, poor) can be estimated with reasonable accuracy for the next time instance. For instance, while estimating the exact feature description of every new college application for the next year might be very difficult, estimating the relative percentage of applicants from the unprivileged group for the next cycle might be possible. Identifying the relative distributions for the privileged (and unprivileged) groups opens the door for a number of data preprocessing techniques, wherein relative weights or normalizations are undertaken based on the group representations.

Lastly, while there have been multiple metrics for quantifying bias proposed in the past literature [36], most of them have implicitly assumed a static world model. Especially, in ML-fairness literature most of the commonly used metrics (e.g., accuracy and error rate parity) focus on scores derived from the confusion matrix where results across time are collapsed into a single representation. Hence, with the growing interest in temporal aspects of bias, we posit that there is a need to *reify* time in fairness metrics, and newer metrics like temporal stability are critical for evaluating ML models.[2]

Our main contributions in this paper are to:

- add empirical evidence to the literature demonstrating that: (a) bias fluctuates frequently and is rarely stable, and (b) static learning is more likely to not generalize well for both accuracy and fairness,
- propose a framework for dynamic learning along with the anticipation component that can help mitigate bias *before* it happens, and
- propose newer metrics for quantifying bias in temporally evolving settings.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the related work. Then, Section 3 describes the proposed method along with the limitations of the existing baselines. The experimental setup and results are provided in Sections 4 and 5. Finally, in Section 6, a summary and future directions are shared.

2 RELATED WORK

In the last few years, significant research effort has been devoted to fairness in machine learning and as a result, several definitions, metrics and methods have been proposed to address both bias detection and mitigation aspects [11, 22, 35, 36]. These measurements might be domain specific and unfortunately have limitations. Several authors provide mathematical proofs of the impossibility of simultaneously satisfying different proposed metrics [8, 12, 15, 30].

Following bias measurements, researchers proposed various mitigation methods which include: *pre-processing* in which the dataset has to be corrected (i.e., modified to support fairness) prior to the modeling [28], *in-processing* where a model itself is being corrected by using constrained optimization or penalty methods [7, 51], and *post-processing* where the model's prediction distributions are modified in order for the model to be fair [21].

Addressing the limitations of static learning, recently Lazaridou et al. [34] examined the performance of a language model performance when a model is trained on past data and used to generalize to future data. The model performs worse on data that are far away from the training periods which leads to a failure in the temporal generalization aspect. Proposed solutions to mitigate this kind of issue are by retraining or adapting the model repeatedly.

Applying generalization methods by training a model that is robust to various distribution shifts, researchers examine the effects of robustness on fairness. Iosifidis et al. [25, 26] apply pre-processing and distribution shifts methods in streaming classification framework in which the dataset or the model has to be corrected at each time step so the model stays stable. Singh et al. [44] utilize causal learning to build a model that is insensitive to distribution shifts that might occur in the features (i.e., covariate shift). Rezaei et al. [42] try to mitigate bias under covariate shift as well by using pre-processing and in-processing methods simultaneously to build a robust and a fair model. Lastly, applying invariant risk minimization (IRM) instead of the conventional empirical risk minimization (ERM) method forces the model to learn features that are invariant to any distribution shifts (i.e., forcing the model to not rely on spurious relationships) [2].

3 PROPOSED METHOD

Our approach includes an anticipation component to better address fairness in dynamic learning. In dynamic learning, unlike static learning, the model keeps updating its knowledge as new data arrives. We assume there are two disjoint batches that have data from different sets such as: the current and the future. A model learns during the current period and once the future period becomes available, a model uses it to measure its performance. If the performance shows some sort of generalization issue, the model

[1] https://www.theatlantic.com/ideas/archive/2021/09/young-men-college-decline-gender-gap-higher-education/620066/
[2] Code: https://github.com/Behavioral-Informatics-Lab/ABCinML

needs to be updated. This kind of issue is referred to as distribution shift [34, 47].

Distribution shifts can be divided into multiple classes, namely concept, features or temporal shift [47]. In our formulation, we assume a temporal shift exists in the data and this type of shift is what causing the model to be unfair. We examine a specific example of temporal shift which is *selection bias* where different batches of the data have different ratios of both the class labels and the sensitive attributes. Thus a model trained with such data might learn well with the majority groups and not with the minority groups. This imbalance is likely going to lead to the issue of bias. Specifically, we assume (and empirically validate) that we are able to estimate the relative distributions of the class labels and the sensitive attributes at an accuracy level that is good enough to support bias correction.

To mitigate this issue of *selection bias*, we want to encourage the model to learn from both groups equally regardless of the data imbalance. To do so, we adapt a *pre-processing* method developed by [28] which is a *reweighing* method that gives weights to different groups based on their representations in the dataset with the class labels in order to force the model to learn fairly. Different from [28], we not only correct the dataset based on the current representations but also apply a correction based on the relative distributions that we expect in the future.

3.1 Preliminaries

We define a random variable V, representing the input variables (i.e., covariates). We can divide V into (A, X), where A is a binary random variable representing the *sensitive* attribute and X represents the *non-sensitive* attributes. Also, we represent the ground truth by a binary random variable Y. Each instance $v \in V$ has a label $y \in Y$ and a sensitive attribute $a \in A$ such that: $y = \{y^+, y^-\}$, where *+, (-)* represents *positive class (negative class, respectively)*, and $a = \{a^+, a^-\}$ where *+, (-)* represents *privileged (unprivileged, respectively)*. We utilize a function $f\colon V \longrightarrow \{y^+, y^-\}$, representing a binary classifier.

We assume that the data arrive sequentially in batches $\{B_1, B_2, ..., \}$ in which each batch B_t has a collection of j instances drawn from V such as $B_t = (v_1^t, v_2^t, ..., v_j^t)$ and t represents the time dimension. Following the dynamic learning settings, we have disjoint sets representing the (B_t) and (B_{t+1}) in which we refer to as *current* and *future* batches, respectively. Specifically, B_t represents the current data we train f on, (i.e., $f_t(B_t)$), whereas B_{t+1} is the future data that we use to evaluate f_t, (i.e., $f_t(B_{t+1})$). Since B_t and B_{t+1} are disjoint, we assume the distribution is *non-stationary*, i.e., $P_{B_t}(V, Y) \neq P_{B_{t+1}}(V, Y)$.

3.2 Pre-Processing Method

To mitigate the *discrimination* in the dataset, we adapt a popular pre-processing method to reduce the discrimination in the dataset before applying the learning model [28]. The *Reweighing* method assigns different weights w for each sub-populations with regards to their representations in the dataset. Hence, positive outcome instances for the unprivileged group should be valorized, while negative outcome instances for the unprivileged group can be given lower weights. Specifically, $P(A = a^-, Y = y^+)$ will have higher *weights* compared to $P(A = a^+, Y = y^+)$, whereas $P(A = a^-, Y = y^-)$ will have lower *weight* compared to $P(A = a^+, Y = y^-)$. Therefore, for each B_t, the weights assigned as follows:

$$W_{B_t}(A, Y) = \frac{P_{expected}(A, Y)}{P_{observed}(A, Y)} \tag{1}$$

where $P_{expected}(A, Y)$ can be estimated from the dataset as the following:

$$\begin{aligned} P_{expected}(A, Y) &= P(A) \times P(Y) \\ &= \frac{|\{A = a\}|}{|B_t|} \times \frac{|\{Y = y\}|}{|B_t|} \end{aligned}$$

and $P_{observed}(A, Y)$ would be:

$$P_{observed}(A, Y) = \frac{|\{A = a, Y = y\}|}{|B_t|}$$

Therefore, we could use any learning models which permits applying these weights in their frameworks.

3.3 Models

In this section, we provide details of the modeling techniques used in the experiment:

0) Vanilla setting: Train once for accuracy, Don't mitigate bias, Test sequentially. In this baseline, we don't address fairness at all and we want to examine the behavior of bias through time (i.e., whether it is stable or fluctuating). While simple, this is the most common setting used in current machine learning implementations.

1) Static setting: Train once for accuracy and bias mitigation, Test sequentially. We simulate the typical static learning where a model is only trained and corrected once on a static dataset then deployed. Thus, a model might be initially *discrimination-free* but fails in the future due to changes in the underlying distribution.

At the beginning of the training, we assume B_t arrives with the corresponding features and labels (i.e., V and Y, respectively) and to *mitigate* bias before training, we apply the *reweighing* method on this batch to get $W_{B_t}(A, Y)$. Using B_t and its corresponding weight W_{B_t}, we train a classifier $f_t(B_t, W_{B_t})$. Then, we use f_t to evaluate incoming batches $\{t + 1, t + 2, ..., n\}$.

2) Dynamic setting: Train for accuracy and bias mitigation sequentially, Test sequentially. Addressing the limitations of baselines (0) and (1), we overcome this issue of training once by re-training the model *continuously* every time a batch arrives. By doing so, we keep the model up-to-date and as a result, we reduce the effects of the temporal and the distribution shift.

3) Anticipatory Bias Correction (ABC). We propose an anticipatory model that utilizes future estimates of the upcoming batches. In a variety of applications, especially when a dataset doesn't follow the i.i.d assumptions, the underlying distribution might show some behaviors that might be forecasted. In this work, we utilize a basic yet effective forecasting model to anticipate some *macro-properties* about the future batches. Specifically, we use a *Moving Average* model to *anticipate* the incoming batch's relative distributions in $\tilde{B}_{t+1}$, i.e., $P_{\tilde{B}_{t+1}}(A = a, Y = y)$ as following:

$$P_{\tilde{B}_{t+1}}(A, Y) = \frac{P_{B_t}(A, Y) + P_{B_{t-1}}(A, Y) + ... + P_{B_{t-S}}(A, Y)}{S} \tag{2}$$

Algorithm 1 Anticipatory Bias Correction

```
procedure ABC( $B_{1:t}(A,Y),\ S,\ \alpha$ )    ▷ data until time $t$, S: window length, $\alpha$: smoothing factor
    $P_{\tilde{B}_{t+1}}(A,Y) \leftarrow \frac{P_{B_t}(A,Y) + P_{B_{t-1}}(A,Y) + \ldots + P_{B_{t-S}}(A,Y)}{S}$    ▷ forecast the relative distributions for batch $\tilde{B}_{t+1}$
    $W_{B_t}(A,Y) \leftarrow Reweighing$    ▷ use Eq. (1) to get weights for batch $B_t$
    $W_{\tilde{B}_{t+1}}(A,Y) \leftarrow Reweighing$    ▷ use Eq. (1) to get weights for batch $\tilde{B}_{t+1}$
    $W_{New} \leftarrow \alpha \times W_{B_t}(A,Y) + (1-\alpha) \times W_{\tilde{B}_{t+1}}(A,Y)$    ▷ acquire the new weight
    return $f_t(B_t, W_{New})$    ▷ learn a new classifier with the new weight
end procedure
```

where S represents the window's length and $\tilde{B}_{t+1}(A,Y)$ is the estimated distribution for batch $t+1$. Note that we do not assume that the actual data points from B_{t+1} are available or estimated. Rather, we hypothesize that the data points from B_t, when weighed according to the anticipated ratios for B_{t+1} would already be useful in mitigating the bias levels in B_{t+1}. Using Eq.(1) and Eq.(2), we will be able not to only mitigate bias for the current time step but also for the future as well. Doing so, will help us *prevent* bias before it shows up in the output of the algorithm. (See Algorithm 1).

To apply the reweighing method, at time t: we will have $W_{B_t}(A,Y)$ and $W_{\tilde{B}_{t+1}}(A,Y)$, weight estimates for the current and the future data, respectively. We combine these *weights* to have a new weight W_{New} as the following:

$$W_{New} = \alpha \times W_{B_t}(A,Y) + (1-\alpha) \times W_{\tilde{B}_{t+1}}(A,Y) \tag{3}$$

where $\alpha \in [0,1]$. We apply this weighted approach to balance the weights between current data and the future data. Lower α will put much emphasis on the future data, whereas higher α focuses on the current data. Lastly, we build a classifier using the the new weight, i.e., $f_t(B_t, W_{New})$.

3.4 Models Assessments

To assess the model discrimination, we utilize three different bias metrics that are commonly used in the fairness literature [35]. However, as these metrics build upon confusion matrices that collapse variations over time into a single representation, we complement these traditional snapshot metrics with some newer temporal metrics.

3.4.1 Snapshot metrics.

Statistical Parity Difference (Δ S.P) measures the disparity of being assigned to a positive class for individuals from different groups. In other words, a fair model requires the predictions to be statistically independent from the sensitive attributes:

$$\Delta\ S.P = |P(\hat{Y}=1|A=0) - P(\hat{Y}=1|A=1)| \tag{4}$$

Equal Opportunity (Δ TPR) measures the disparity of the True Positive Rate (TPR) for individuals from different groups. In other words, a fair model requires equal TPR for individuals from different groups:

$$\Delta\ TPR = |P(\hat{Y}=1|Y=1,A=0) - P(\hat{Y}=1|Y=1,A=1)| \tag{5}$$

Predictive Equality (Δ FPR) measures the disparity of the False Positive Rate (FPR) for individuals from different groups. In other words, a fair model requires equal FPR for individuals from different groups:

$$\Delta\ FPR = |P(\hat{Y}=1|Y=0,A=0) - P(\hat{Y}=1|Y=0,A=1)| \tag{6}$$

All the above equations measure the absolute difference where a lower value indicates a fair model and a larger value indicates a biased model.

For the model accuracy, we use the **Area Under the ROC Curve (AUC)** since it is robust to class imbalance. Often, it is impossible to achieve a low value for all the fairness measures at the same time since different metrics are domain-specific and have potentially contradictory assumptions [15].

3.4.2 Temporal Metrics.

To allow for a richer understanding of bias in temporally evolving settings, we propose some temporal metrics to examine bias. Following related literature in fairness in ML, time series analysis, and capturing the performance of dynamical systems [18, 20, 49], we posit that the metrics should be able to capture the following aspects:

- the worst case performance of the system
- the fluctuations in the performance of the system

These metrics can be helpful in capturing the underlying behaviors in which bias rate can change, and thus can be used for evaluation and monitoring purposes. Just like there exist multiple metrics to measure aspects related to accuracy (e.g., AUC, true positive rate, precision, recall, f-measure), we expect a subset of these temporal bias metrics to be used in a given scenario based on the application priorities. Following substantive existing literature in the space, we use Δ (i.e., a disparity treatment between two sensitive groups) as the starting point to quantify bias. This Δ can be operationalized over any of the metrics that capture performance (e.g., accuracy, TPR, statistical parity etc.) as appropriate in the given context.

Worst case performance.

The bias level of the system may fluctuate over time, hence in many scenarios it is important to understand the worst case bias level.

Maximum Bias (MB) is defined as the maximum bias observed across all time steps N:

$$MB = \max_{1 \leq i \leq N} \Delta_i \tag{7}$$

MB quantifies the most biased performance that can be expected from the system if we cannot control the time at which the user will interact with the system. A stable and a fair model is expected to have a low MB.

Fluctuation in performance.

An ideal ML system will have low bias and that level of bias will not change dramatically over time. Dramatic changes over time will yield very different performances to different users who may happen to use the system on neighboring time points.

Temporal Stability (TS) is defined as the average absolute deviation of consecutive bias windows:

$$TS = \frac{1}{N} \sum_{i=2}^{N} |\Delta_i - \Delta_{i-1}| \tag{8}$$

Since we are using a dynamic learning approach, consecutive time steps are expected to have yield similar performance. Therefore, we propose a metric the examine the *adjacent/local fluctuation* with respect to the previous time steps. If the model were to be stable and fair, we expect a lower estimate.

Maximum Bias Difference (MBD) is defined as the maximum bias difference between consecutive bias windows:

$$MBD = \max_{2 \leq i \leq N} |\Delta_i - \Delta_{i-1}| \tag{9}$$

This metric is examining the *sudden change* that might happen in the bias rate during modeling. A larger sudden change may suggest a phase transition or mishap that needs to be looked into. A *stable* fair model will persistently generate a lower change over consecutive time windows and hence will have a lower MBD.

Table 1: Datasets overview

Dataset	Period	No. Samples	Sensitive	Target
Funding	2005-2013	612,262	Poverty-Level	Funded?
Toxicity	2015-2017	60,287	Gender	Toxic?
Adult	2014-2018	636,625	Race	>=50K?

4 EXPERIMENTS

Our approach is applicable to datasets that have a temporal dimension in order to test the temporal shift and generalization problems. Unfortunately, most of the current fairness benchmarking datasets do not have the temporal aspect, i.e., the data does not come with explicit timestamp. However, there are three temporal datasets that have been used recently for fairness applications and we focus on these for this work.

4.1 Datasets

We validate our approach with the following datasets that have a temporal dimension.

Funding is a dataset provided by DonorsChoose[3], an organization that facilitates educational projects funding posted by teachers in the United States and encourages local community members to support teachers' projects by donation. The 2014 Data Mining and Knowledge Discovery competition (KDD) publicly released the dataset to encourage ML practitioners to build a predictive model to find which projects are more likely to be funded in the future based on information/attributes related to the projects, schools, teachers and funding [1, 33]. To address the fairness aspect, we utilize the "poverty-level" feature as a sensitive attribute in order to see if there is any discrimination with respect to the wealth level of the school district. We use data records from 2005-2013 and employ monthly temporal modeling. We utilize a window of 3 for forecasting (e.g., using the first three months to forecast some knowledge about the 4^{th} month).

Civil Comments Toxicity[4] is a corpus of comments collected for 2 years (2015-2017) in order to address bias in the toxicity classification applications. The dataset has been used to address the spurious correlations between the sensitive attribute and the probability of toxicity. The prediction task is to predict if a comment is toxic or not. To measure the fairness aspect, the task is to measure bias in comments in which an identity or an ethnicity has been mentioned. In this task, we focus on the gender bias [2, 9]. We experiment with a monthly temporal modeling in which a window of 5 has been used.

Adult is an extension of the popular 1994 adult dataset that have been widely used in the ML-fairness literature [32]. We used a newly released version in which the dataset is collected between 2014-2018 and span across US states [14]. The prediction model is to classify whether a person's income will be more than 50K based on the demographic attributes [41]. We only utilize a subset of the dataset from the state of California and we focus on the racial bias (i.e., "white" v.s. "black"). For this dataset, we apply a yearly temporal modeling since the monthly measurement is not available. Because of the small time-range, we use a window of 2 for forecasting.

The datasets' information is summarized in Table 1.

5 RESULTS AND DISCUSSION

Since, in this experiment we are utilizing dynamic learning as apposed to the traditional static learning, we use a re-training approach. Having large data size might allow the model to learn features adequately by being exposed to a variety of data points. Therefore, we train on the current batch and we test on the next. We employ a growing window approach for the training dataset to maximize the learning opportunities for the model.

To evaluate the model, we restrict the learning hypothesis domain to be a simple classifier, namely "Logistic Regression". For each dataset, we use feature representations suitable for each applications. Specifically, for the Funding dataset, we preprocess the data and apply feature engineering methodology to keep only useful features as suggested by [46] yielding 113 features. In the case of the Toxicity dataset, we use a pre-trained word embedding (i.e., Word2vec-100d) to represent linguistic features [37]. Lastly, we use all the 12 features provided by the authors for the Adult dataset [14].

5.1 Temporal Variation

To examine the temporal dynamics for both the accuracy and bias, we plot the performance of the abovementioned approaches over time for all datasets. Figure 1 shows the baselines' results along

[3] https://www.donorschoose.org/

[4] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

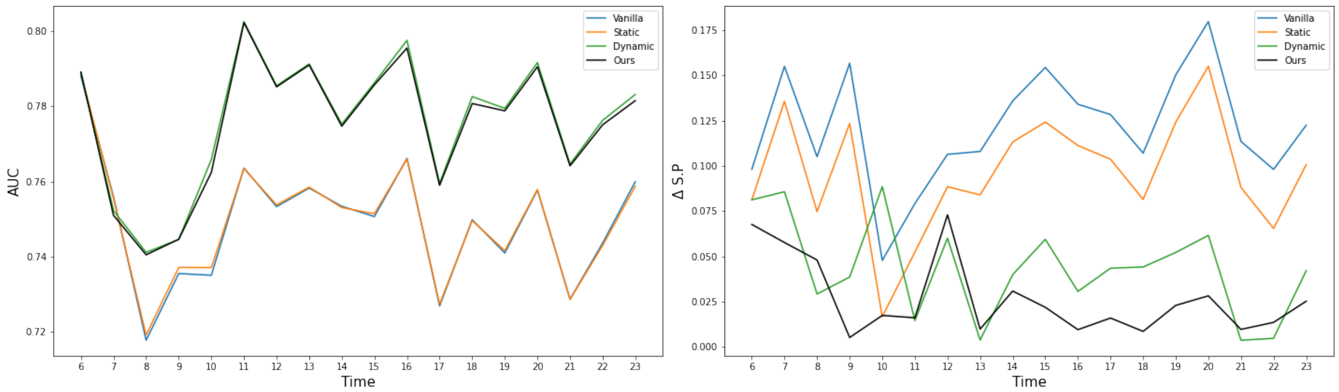


Figure 1: Changes in accuracy (AUC) and fairness (Δ S.P) over time for the toxicity dataset

with our proposed method for accuracy (AUC) and the primary fairness metric considered (Δ S.P) in the Toxicity dataset. (Results for other datasets and using other fairness metrics are presented in the Supplementary Material). As can be seen, the accuracy of the model (AUC) changed noticeably over time. For baselines 0 (vanilla) and 1 (static), which are the most common approaches for ML implementations, the AUC varied between 0.79 and 0.72. A growing window and the dynamic learning approach (baseline 2) yielded a higher accuracy and that performance was virtually matched by the proposed approach. There is also noticeable fluctuation in bias levels for baseline 0 (range from 0.175 to 0.050). In effect, the results add empirical evidence to the literature demonstrating that: (a) bias is not a static entity and fluctuates frequently, and (b) vanilla and static learning (baselines 0,1) are not likely to generalize well for both accuracy and fairness [13, 14, 26, 27].

Further, we notice that baseline 2 performs better (i.e., provides lower bias levels and higher accuracy levels) than baselines 0 and 1. The proposed approach yields lower bias level compared to baseline 2 while maintaining accuracy at levels comparable to baseline 2.

5.2 Impact of Future Estimation

A key question in this work is to study the impact of future estimation on the accuracy and fairness levels of the algorithms. For ease of interpretation and comparison with existing work, we first quantify fairness based on popular (snapshot) metrics and then discuss the impact on proposed temporal metrics in the next subsection.

Table 2 and Figure 2 show the results of the three baselines and the proposed approach. (Note that the α for the proposed approach is chosen for its best performance on the Δ S.P). As can be seen, the baselines 0 and 1 struggle to mitigate bias *on average*. Baseline 2 is helpful in decreasing the bias discrimination across all the considered metrics with a noticeable reduction. The proposed approach yields the lowest value of bias with regard to Δ S.P across all datasets (Fig. 2 - left plots). Conceptually, this can be interpreted as importance weighing approach trying to mimic the sampling procedures by having equal representations of different groups with regard to the class labels. Additionally, the proposed approach has successfully reduced the bias across all measurement in the Funding dataset (Fig. 2 (a)) and for two out of the three metrics in the other datasets. As suggested by prior literature, different bias metrics may not always be reduced in the same settings.

We report the impact of α (i.e., relative importance given to current data or the future estimation (Eq. 3)) in Table 3. The value of α ranged from 0 to 1 in which lower value means that the model is more focused on mitigating the bias using the future estimates, whereas a higher value is the opposite. Traditional systems engineering approaches such as Kalman Filtering suggest that different applications and contexts would require different level of importance to be given to the learned parameters from past or current data and the estimates of future data [45]. Here we also found different applications have different behaviors. In case of the Funding and Adult datasets, the models perform better when focusing more heavily on the current data with slight reliance on the future estimates (α=0.9). In contrast, in applications that are more prone to distribution shift, such as Toxicity Classification, the model performs better when giving more importance to future estimates (α=0.0) [2].

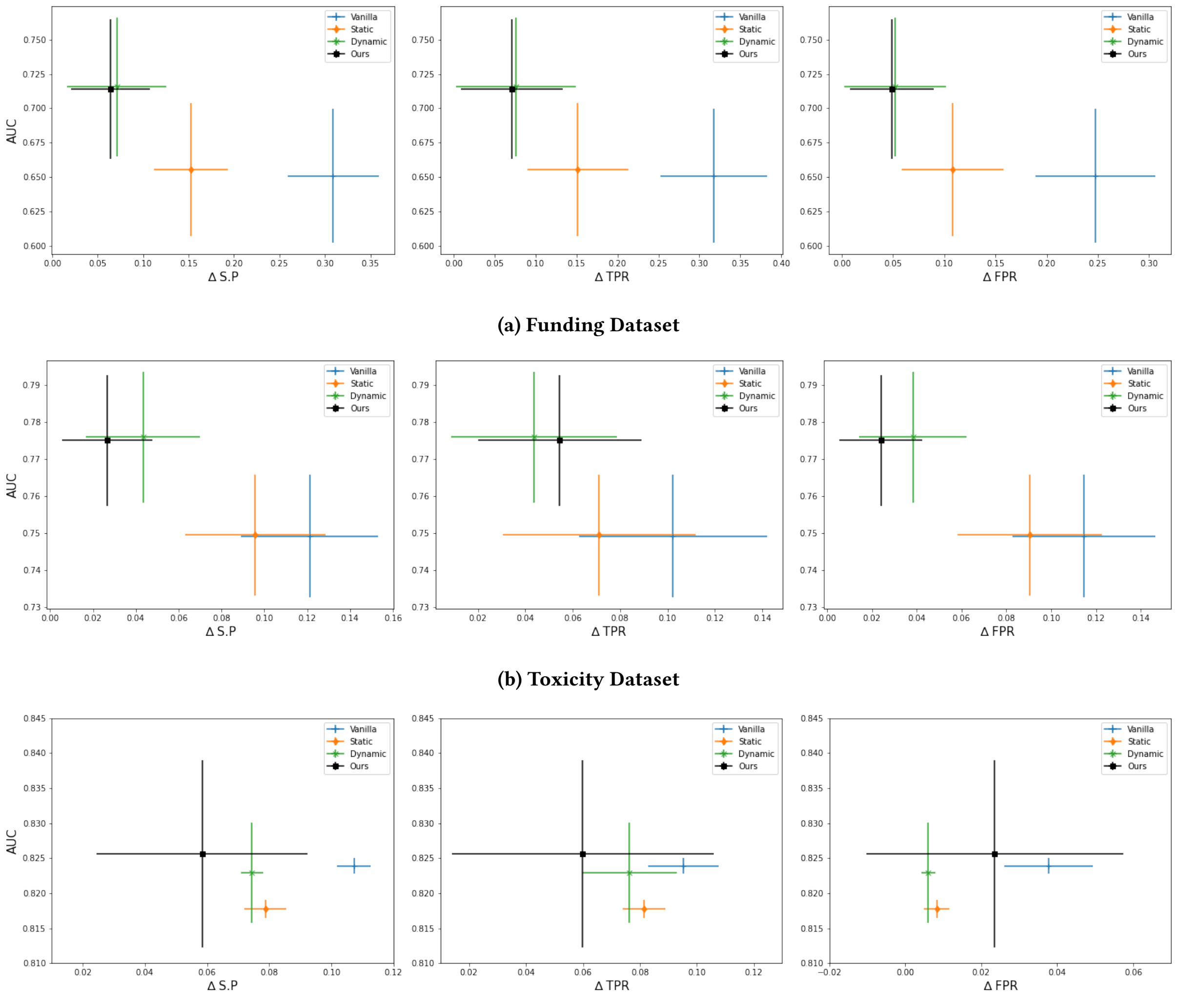
5.3 Temporal Metrics Evaluation

To evaluate our approach with respect to the proposed temporal metrics, we used only two datasets (Funding and Civil Comments Toxicity) since they have a longer temporal window. (The Adult dataset has only 5 time windows at yearly resolution). Results are provided in Table 4 (lower scores are better for each metric; they indicate more fairness and/or more stability). The proposed approach yields a better worst case (lower MB) performance compared to the other baselines in both datasets. Additionally, the proposed approach has lower fluctuation measures with the Toxicity dataset (i.e., lower TS and MBD) but not with the Funding dataset in which the first baseline is performing slightly better (i.e., lower MBD). In all, the proposed approach yields the best performance in 5 of the 6 scenarios (dataset + metric) considered.

As previously discussed in Section 5.2, in this work the parameters were chosen to reduce Δ S.P, which provides a reasonable

Table 2: Results for performance of different baselines and the proposed approach in different applications. Results reported for accuracy (AUC) and popular 'snapshot' fairness metrics.

	Funding				Toxicity				Adult			
	AUC ↑	Δ S.P ↓	Δ TPR ↓	Δ FPR ↓	AUC ↑	Δ S.P ↓	Δ TPR ↓	Δ FPR ↓	AUC ↑	Δ S.P ↓	Δ TPR ↓	Δ FPR ↓
Vanilla	0.651	0.309	0.318	0.248	0.749	0.121	0.102	0.115	0.824	0.107	0.095	0.038
Static	0.655	0.153	0.152	0.108	0.749	0.096	0.071	0.090	0.818	0.079	0.082	0.008
Dynamic	**0.716**	0.071	0.076	0.052	**0.776**	0.043	**0.044**	0.038	0.822	0.074	0.076	**0.006**
Ours	0.714	**0.064**	**0.071**	**0.049**	0.775	**0.027**	0.054	**0.024**	**0.826**	**0.058**	**0.060**	0.024



(a) Funding Dataset

(b) Toxicity Dataset

(c) Adult Dataset

Figure 2: Results of experimental evaluation for different datasets. Y-axis shows accuracy (AUC) and the X-axis shows a bias metric. Average scores for different approaches are shown as points with with standard deviation shown as a bar. Best models are those that lie in the top left portion of the figure.

Table 3: Results for different values of α and its effects on our approach. Lower value of α focuses on future, whereas higher value focuses on the current.

	Funding				**Toxicity**				**Adult**			
	AUC ↑	Δ S.P ↓	Δ TPR ↓	Δ FPR ↓	AUC ↑	Δ S.P ↓	Δ TPR ↓	Δ FPR ↓	AUC ↑	Δ S.P ↓	Δ TPR ↓	Δ FPR ↓
0.0	0.713	0.107	0.115	0.084	*0.775*	***0.027***	*0.054*	***0.024***	0.820	0.088	0.088	0.012
0.1	0.714	0.101	0.108	0.077	0.775	0.028	0.051	0.025	0.823	0.094	0.095	0.018
0.2	0.713	0.095	0.104	0.074	0.775	0.030	0.048	0.027	0.820	0.087	0.083	0.013
0.3	0.714	0.088	0.097	0.069	0.775	0.031	0.045	0.028	0.817	0.078	0.080	0.008
0.4	0.714	0.083	0.092	0.063	0.775	0.032	0.044	0.030	0.822	0.089	0.090	0.013
0.5	0.715	0.079	0.088	0.063	**0.776**	0.034	0.041	0.032	0.821	0.093	0.0901	0.023
0.6	0.714	0.076	0.084	0.059	**0.776**	0.036	0.042	0.032	**0.825**	0.101	0.098	0.026
0.7	0.714	0.072	0.081	0.056	**0.776**	0.037	0.042	0.034	0.822	0.073	0.071	0.008
0.8	0.714	0.067	0.074	0.054	**0.776**	0.039	**0.040**	0.035	0.822	0.092	0.092	0.015
0.9	*0.714*	***0.064***	***0.071***	***0.049***	**0.776**	0.042	0.042	0.037	***0.825***	***0.058***	***0.059***	*0.023*
1.0	**0.716**	0.071	0.076	0.052	**0.776**	0.043	0.044	0.038	0.823	0.074	0.076	**0.006**

performance in terms of multiple traditional fairness metrics as well as the temporal fairness metrics.

Table 4: Results of proposed temporal fairness metrics experimented with two datasets.

	Funding			**Toxicity**		
	MB ↓	TS ↓	MBD ↓	MB ↓	TS ↓	MBD ↓
Vanilla	0.430	0.036	**0.148**	0.179	0.035	0.109
Static	0.261	0.034	0.158	0.155	0.037	0.107
Dynamic	0.349	0.039	0.218	0.088	0.030	0.074
Ours	**0.180**	**0.029**	0.164	**0.072**	**0.018**	**0.063**

6 CONCLUSION AND FUTURE WORK

In this paper, we examine the applicability of using an early prevention approach to mitigate bias in advance by experimenting with three different real-world ML applications. We compare our approach to the traditional models that have been widely used in the fairness literature and evaluate the advantages of the anticipatory correction approach. Additionally, we also propose newer fairness metrics that would be suitable when dealing with temporally evolving settings.

Although we have a used a simple model for future estimation, we are able to see the effect of this approach on bias reduction. The proposed approach yielded best results in terms of most (though not all) metrics across different real world datasets. This trend was consistent across traditional as well as proposed temporal fairness metrics. Some degree of variation in results is consistent with past research suggesting the difficulty in reducing different bias metrics simultaneously. A possible approach suggested in the literature is to identify a primary metric for making prioritization depending on the context [33, 35].

The work described has some limitations. It focuses on a single pre-processing based bias reduction approach and works with a single machine learning approach over three datasets. Yet, by utilizing dynamic learning there are multiple sources of bias that could be involved in the ML pipelines such as uncontrolled data points quality in each time step as well as the model itself. Additionally, since we are learning in a sequential fashion, the distribution of the protected and unprotected group can switch (i.e., what was considered a majority in the past batches might become a minority in the future batch [19]) but we mitigate this issue by applying a re-training model.

Our future work will focus on investigating the nuances that could lead to this variation in model training by understanding the dynamics for such applications. Besides that, we will utilize a more robust anticipation model and a range of the fairness mitigation methods to understand their applicability in anticipatory bias correction. Our approach is one of the earliest attempts in *anticipatory bias prevention* and we hope that it will encourage the research community to undertake more sophisticated efforts in this direction.

ACKNOWLEDGMENTS

This material is in part based upon work supported by the National Science Foundation under Grant SES-1915790.

REFERENCES

[1] [n. d.]. KDD Cup 2014 - Predicting Excitement at DonorsChoose.org. https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose. Accessed: 2021-07-1.

[2] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485* (2020).

[3] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. 2020. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3529–3530.

[4] Jamal Alasadi, Ahmed Al Hilli, and Vivek K Singh. 2019. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*. 19–25.

[5] Abdulaziz A Almuzaini and Vivek K Singh. 2020. Balancing Fairness and Accuracy in Sentiment Detection using Multiple Black Box Models. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*. 13–19.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159.

[7] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological*

Methods & Research 50, 1 (2021), 3–44.

[9] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*. 491–500.

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[11] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[12] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[13] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[14] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *arXiv preprint arXiv:2108.04884* (2021).

[15] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[16] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[17] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2020. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (October 1, 2020)* (2020).

[18] Christopher G. Harris. 2020. Methods to Evaluate Temporal Cognitive Biases in Machine Learning Prediction Models. In *Companion Proceedings of the Web Conference 2020*. 572–575.

[19] Heitor Murilo Gomes, Jesse Read, Albert Bifet, Jean Paul Barddal, and João Gama. 2019. Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 6–22.

[20] Swati Gupta, Akhil Jalan, Gireeja Ranade, Helen Yang, and Simon Zhuang. 2020. Too Many Fairness Metrics: Is There a Solution? *Available at SSRN 3554829* (2020).

[21] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[22] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass* 15, 8 (2021), e12432.

[23] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[24] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[25] Vasileios Iosifidis and Eirini Ntoutsi. 2020. FABBOO-Online Fairness-Aware Learning Under Class Imbalance. In *International Conference on Discovery Science*. Springer, 159–174.

[26] Vasileios Iosifidis, Thi Ngoc Han Tran, and Eirini Ntoutsi. 2019. Fairness-enhancing interventions in stream classification. In *International Conference on Database and Expert Systems Applications*. Springer, 261–276.

[27] Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2021. Online Fairness-Aware Learning with Imbalanced Data Streams. *arXiv preprint arXiv:2108.06231* (2021).

[28] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[29] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[30] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[31] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.

[32] Ron Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. 202–207.

[33] Hemank Lamba, Kit T Rodolfa, and Rayid Ghani. 2021. An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings. *ACM SIGKDD Explorations Newsletter* 23, 1 (2021), 69–85.

[34] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, et al. 2021. Pitfalls of Static Language Modelling. *arXiv preprint arXiv:2102.01951* (2021).

[35] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the applicability of ML fairness notions. *arXiv preprint arXiv:2006.16745* (2020).

[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[37] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

[38] Claire Cain Miller. 2015. Can an algorithm hire better than a human. *The New York Times* 25 (2015).

[39] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).

[40] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[41] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, and Eirini Ntoutsi. 2021. A survey on datasets for fairness-aware machine learning. *arXiv preprint arXiv:2110.00530* (2021).

[42] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian Ziebart. 2020. Robust Fairness under Covariate Shift. *arXiv preprint arXiv:2010.05166* (2020).

[43] Maria Veronica Santelices and Mark Wilson. 2010. Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review* 80, 1 (2010), 106–134.

[44] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.

[45] Vivek K Singh, Pradeep K Atrey, and Mohan S Kankanhalli. 2008. Coopetitive multi-camera surveillance using model predictive control. *Machine Vision and applications* 19, 5 (2008), 375–393.

[46] Aparicio Sofia. 2020. Predicting Excitement at DonorsChoose.org. https://github.com/SofiaAparicio/KaggleComp_DonorsChoose.

[47] Amos Storkey. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning* 30 (2009), 3–28.

[48] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[49] Irene Teinemaa, Marlon Dumas, Anna Leontjeva, and Fabrizio Maria Maggi. 2018. Temporal stability in predictive process monitoring. *Data Mining and Knowledge Discovery* 32, 5 (2018), 1306–1338.

[50] Austin Waters and Risto Miikkulainen. 2014. Grade: Machine learning support for graduate admissions. *Ai Magazine* 35, 1 (2014), 64–64.

[51] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.