Fair Active Learning

Hadis Anahideh [1] **Abolfazl Asudeh** [2]

Abstract

Bias in training data and proxy attributes are probably the main reasons for bias in machine learning. ML models are trained on historical data that are biased due to the inherent societal bias. This causes unfairness in model outcomes. On the other hand, collecting labeled data in societal applications is challenging and costly. Hence, other attributes are usually used as proxies for labels that are already biased and result in model unfairness.

In this paper, we introduce fair active learning (FAL) for mitigating machine bias. Considering a limited labeling budget, FAL carefully selects data points to be labeled in order to balance between the model performance and its fairness quality. Our comprehensive experiments, comparing traditional active learning with FAL on real datasets, confirm a significant improvement of fairness for models trained using FAL, while maintaining the model performance.

1. Introduction

Data-driven decision making plays a significant role in modern societies. Data science and advanced computational methods have enabled it to make wise decisions and to make societies more just, prosperous, inclusive, and safe. With this unique opportunity, however, comes to a lot of responsibilities as improper development of data science technologies can not only fail to but also make matters worse. For example, judges in the US use background information of the individuals for setting bails or to sentence criminals. This is useful as it can lead to safer societies, but at the same time has the potential to significantly impact the lives of individuals if not developed properly. For instance, the recidivism scores provided for the judges are highly criticized to be racist, since it turns out they assign higher risks to African American individuals (Angwin et al., 2016).

[1]University of Illinois at Chicago, Illinois, US
[2]University of Illinois at Chicago, Illinois, US
. Correspondence to: Hadis Anahideh <hadis@uic.edu>.

Machine learning (ML) is in the center of data-driven decision making as it provides insightful unseen information about phenomena based on available observations. For example, ML has enabled "precisely" classifying emails to spam or regular email, based on the text and other features an email contains. Image classification, traffic prediction, personal assistive tools, and product recommendations are a tiny fraction of other examples ML has made significant advancements.

Machine learning has also been used for evaluating individuals and making decisions about human beings and society. In this context, similar to other applications, ML models try to learn the system (individuals and society) based on some observations (historical data). Blindly applying machine learning without paying attention to societal impacts, however, can lead to significant issues such as biased decision making, resulting in racism or sexism. The following are two major reasons for this to happen:

- *Bias in training data:* ML models use the background information of individuals and information that are biased due to the historical discrimination against different groups in society. For example, redlining is a historical systematic denial of services against (mainly) specific racial communities, leaving its footprint up to the day and the existing data records (Jan, 2018). Gender bias in Data (Perez, 2019), including health care (Pley & Keeling, Sep. 2019) is yet another example of bias in training data.

- *Proxy attributes:* labeled data is the corner-stone of supervised learning. Yet, there often exists no (or limited) labeled data in societal applications. For example, the aforementioned recidivism scores (used by judges) are supposed to show how likely an individual is to commit a crime in the future. Another example can be highlighted in the context of college admission where the goal is to admit students that are likely to be most successful in the future. Training data with such information is either not available or there exists a very limited number of labeled data. As a result, other available attributes are usually used as proxies for these attributes. For instance, "getting arrested by cops" may be considered as a proxy for committing a crime, GPA for future success in college.

Machine learning models rely heavily on the data and the data can be highly biased. The decisions made from the output of a model based on a biased labeled dataset can result in discrimination and unfairness. A job platform can rank less qualified male candidates higher than more qualified female candidates, (Lahoti et al., 2019). As a result, a new paradigm of fair machine learning has emerged. Fairness has a different definition is measured in various ways. To further clarify these points, in the following we provide two running example applications that we shall use during the paper for clarification. We also use these examples for setting up experiments on real-world data as proof of concept and to validate our proposal.

Example 1. *To help judges make a wise decision when setting bails, a company is interested to create a model to predict recidivism; that is, how likely an individual is to commit a crime in the future. Suppose the company has access to the background information of some criminal defendants* [1]*. The collected data, besides being limited in the count, is not labeled. That is because the information about whether or not an individual will commit a crime in the future is not available at the time of trial. Considering a time window, labeling an individual in the dataset is possible by checking the background of the individual within the time window after being released. This, however, is not easy, may be associated with a cost, and may need expert efforts for data integration and entity resolution.*

Example 2. *A loan consulting company is about to create models that help financial agencies identifying "valuable customers" that will pay their loans off on time. The company has collected a dataset of customers who have received a loan in the past few years. In addition to the demographic information, the dataset includes background information such as education level and income of individuals. Unfortunately, at the time of approving loans, it is not known if the customer will pay his/her debt on time, and hence, the data is not labeled. Still, the company has hired experts that given the information of an individual who has received a loan in the past, can check his/her background and identify if the payments have been made on time. Of course, considering the costs associated with a background check, it is not viable to freely label the customer information.*

Both of the above examples use historical data for building their models that, as we shall investigate in our experiments, are biased. For instance, the `income` in Example 2 is known to include gender bias (Jones, 1983) or `prior count` in Example 1 is racially biased (Angwin et al., 2016). Also, both datasets are unlabeled. A data scientist may decide to use the (problematic) proxy attributes for training the model.

[1] Such information is provided by the counties' sheriff offices in the US. For instance, ProPublica used information obtained from the Broward County Sheriff's Office for the COMPAS dataset. `https://bit.ly/36CTc2F`

Another alternative is (depending on the available budget) to randomly label a subset of data and use it for training. A more responsible solution, however, is applying *active learning* and using expert oracles for gradually labeling the data. In these cases, allowing the model to choose from its unlabeled training set will minimize the labeling cost. Active learning (Settles, 2009) sequentially chooses the unlabeled instances to be labeled by an oracle, if there is a value-added by that instance to the approximated model. The model trained through an active learning framework is performing well with less and selective training data. There are different sampling strategies, which select instances from the pool of unlabeled data points according to their informativeness. The most commonly used measure of informativeness is the classification uncertainty, where the classifier is most uncertain about the label of an instance. Active learning has been extensively studied in machine learning which we shall further elaborate in § 2.2.

Unfortunately, as the best of our knowledge, none of the existing work in active learning consider the societal impacts of these models such as *fairness* and solely optimize for maximizing accuracy. This is our objective in this paper. We aim to design an active learning technique to generate fair outcomes. As we shall further elaborate in § 2.3, we define the notion of fairness with respect to sensitive attributes such as `race` and `gender`. Focusing on the group fairness (Dwork et al., 2012; Li & Cropanzano, 2009), we consider a model fair if its outcome does not depend on the sensitive attributes. That is, we adopt demographic parity, one of the popular fairness measures (Kusner et al., 2017; Dwork et al., 2012). Although we consider model independence as our measure of fairness, in § 5, we shall show how to extend our results for other notions of fairness based on separation and sufficiency (Barocas et al., 2019).

In this paper, we *introduce fairness in active learning* for constructing a fair learning model with limited labeled data. We carefully provide theoretical background and propose the fair active learning (FAL) to balance model performance and fairness. We conduct comprehensive experiments on real datasets to show that performing active learning while considering the fairness constraint can significantly improve the fairness of a classifier while not significantly impacting its performance. As we shall present in § 4, our experiment results across different fairness metrics confirm improvements in fairness by around 50% without major reductions in model performance. In summary, our contributions are as following:

Summary of contributions.

- Carefully formalizing terms and background, we present different fairness measures based on model independence.

- We introduce fairness in active learning, an iterative approach that incorporates the fairness measure in its sample selection unit and constructs a fair predictive model as a result.
- We propose the expected fairness measure for unlabeled sample points based on the best-known estimate of the function.
- We conduct comprehensive experiments on real-world data, considering different fairness metrics based on model independence. Our results show an improvement of around 50% in fairness measures while not significantly impacting the model performance.
- We discuss how to extend our framework for different fairness measures.

In the remainder of this paper, we start with a background on active learning and our fairness model in § 2. In § 3, we will present the Fair Active Learning (FAL) framework, which enables practitioners to build a fair ML model with limited labeled training data. The proposed framework is flexible to incorporate different measures for model performance and fairness. We will provide our comprehensive experiment results in§ 4. In § 5, we will discuss how to extend our proposal for different fairness measures and conclude the paper in § 6.

2. Background

2.1. Data Model

We assume the existence of a (training) dataset $\mathcal{D}$ with n instances, each consisting of d features $\{x_1, x_2, \cdots, x_d\}$. For a data point $P_i \in \mathcal{D}$, we use the notation $X_j^{(i)}$ to refer to the value of $X^{(i)}$ on feature x_j. We also assume each data point is associated with at least one sensitive attribute S. As we shall further explain in § 2.3, sensitive attributes such as `gender` and `race` are non-ordinal categorical attributes used in the fairness model. We use the notation $S^{(i)}$ to refer to the sensitive attributes of $P_i \in \mathcal{D}$. Without loss of generality and to simplify the explanations, unless explicitly stated, we assume S is a single sensitive attribute. Still, we would like to emphasize that our techniques are not limited to the number of sensitive attributes. Each data point P_i is also associated with a label attribute $y^{(i)}$ with K possible values $\{y_1, \cdots, y_K\}$. We assume the labels of the data points are initially unknown. In § 2.2, we will explain how to obtain the label of a data point P_i. At any moment during the training process, the subset $\mathcal{L} \subseteq \mathcal{D}$ of data points for which labels are known is referred to as *labeled pool* and the rest of them $\mathcal{U} = \mathcal{D} - \mathcal{L}$ are called as *unlabeled pool*. We note that every entry in $\mathcal{U}$ is identified by the pair $\langle X^{(i)}, S^{(i)} \rangle$, while every entry in $\mathcal{L}$ is the triple $\langle X^{(i)}, S^{(i)}, y^{(i)} \rangle$.

2.2. Learning Model

Pool-based active learning for classification was introduced in (Lewis & Gale, 1994). The objective is to learn a classifier function $C : \mathbb{R}^d \rightarrow \{0, \cdots, k\}$ that maps the feature space to the labels. We use $\hat{y} = C(X)$ to refer to the predicted label for X. Recall that data points are initially unlabeled. Active learning assumes the existence of an *expert oracle* that given a data point P provides its labels. Labeling, however, is costly and there usually is a *limited labeling budget* B. Using the sampling budget, one can randomly label B data points and use them for building a classifier. The challenge, however, is to wisely exhaust the budget to build the most accurate model.

Different sampling strategies have been proposed in the context of active learning. Uncertainty sampling (Lewis & Gale, 1994) is probably the most common strategy in active learning for classification. It selects data points for labeling such that the model variance maximally is minimized. It chooses the point $P \in \mathcal{U}$ that the current model is least certain about its label. At every iteration of the process, let the classifier $C(.)$ be the current model, based on the set of labeled data $\mathcal{L}$. For every data point $P_i \in \mathcal{U}$ with features $X^{(i)}$, let $\mathbb{P}(y = k|X^{(i)})$ be the posterior probability that its label $y^{(i)}$ be k based on $C(.)$. By maximizing the uncertainty, active learning select the points that are close to the decision boundary, where we are least certain about the class label. Uncertainty can be defined in different ways (Settles, 2009). In general, uncertainty sampling refers to maximum entropy (Shannon, 1948). Equation 1 denotes the entropy formulation for classifying $X \in U$ with $y \in \{0, 1\}$ based on the probability obtained from the classifier $C(.)$.

$$X^* = \underset{\forall X \in \mathcal{U}}{\text{argmax}} \; -\sum_{k=1}^{K} \mathbb{P}(y = k|X) \log \mathbb{P}(y = k|X) \quad (1)$$

Algorithm 1 presents the active learning algorithm, using Equation 1. Iteratively, the algorithm selects a point from $\mathcal{U}$ to be labeled next. It uses the classifier trained in the previous step to obtain class probabilities $\mathbb{P}(.)$ (initially, all the probabilities are equal), and to calculate the entropies. The algorithm passes the selected point to the labeling oracle, acquires its label, and adds the point to the set $\mathcal{L}$: the set of labeled instances. It uses $\mathcal{L}$ to train the classifier C_t, where t shows the current iteration. This process continues until the labeling budget gets exhausted.

2.3. (Un)Fairness Model

Following many of the existing work, while founding our model on societal norms of fairness (Barocas et al., 2017),

[2]Alternatively, one could use correlation: $|corr(\hat{y}, S)|$.

	$I(\hat{y};S)$	$\lvert cov(\hat{y},S)\rvert$ [2]
difference	$\lvert\mathbb{P}(\hat{y}=1\mid S=0)-\mathbb{P}(\hat{y}=1\mid S=1)\rvert$	$\lvert\mathbb{P}(S=1\mid\hat{y}=1)-\mathbb{P}(S=1)\rvert$
ratio	$1-\min\left(\frac{\mathbb{P}(\hat{y}=1\mid S=0)}{\mathbb{P}(\hat{y}=1\mid S=1)},\frac{\mathbb{P}(\hat{y}=1\mid S=1)}{\mathbb{P}(\hat{y}=1\mid S=0)}\right)$	$1-\min\left(\frac{\mathbb{P}(S=1\mid\hat{y}=1)}{\mathbb{P}(S=1)},\frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1\mid\hat{y}=1)}\right)$

Table 1. Difference metrics for measuring Demographic disparity.

Algorithm 1 Active Learning

for $t=1$ to B **do**
  $X^* = \underset{\forall X\in\mathcal{U}}{\text{argmax}}\ -\sum_{k=1}^{K}\mathbb{P}(y=k|X)\log\mathbb{P}(y=k|X)$
  $y=$ label X^* using the labeling oracle
  add $\langle X^*,y\rangle$ to $\mathcal{L}$
  train the classifier $C_t(.)$ using $\mathcal{L}$
end for
return $C_t(.)$

we develop our fairness model on the notion of *model independence* or *demographic disparities* (Barocas et al., 2019; Žliobaitė, 2017; Narayanan, 2018; Zafar et al., 2017; Asudeh et al., 2019), also referred by terms such as group fairness (Dwork et al., 2012), statistical parity (Dwork et al., 2012; Simoiu et al., 2017), and disparate impact (Barocas & Selbst, 2016; Feldman et al., 2015; Ayres, 2005). Although our main focus in this paper is on fairness based on model independence, in § 5 we shall show how to extend our framework for other measures based on separation ($\hat{y}\perp S\mid y$) and sufficiency ($y\perp S\mid\hat{y}$) (Barocas et al., 2019).

Of course, disparities in the model does not necessarily imply that the designers intentionally want them to arise. The problem occurs since these models fully rely on (biased) historical data for learning a system; hence, the historical disparities in the data causes the (unintentional) bias in the model. We believe that machine learning practitioners are responsible to intervene in the modeling process (in different learning stages) and mitigate the disparities of the model.

Given a classifier $C(.)$ and a random point $\langle X,S\rangle$, labeled as $\hat{y}=C(X)$, demographic parity holds iff $\hat{y}\perp S$ (Barocas et al., 2017; 2019). Consider a binary classifier and think of $\hat{y}=1$ as "acceptance" – in Example 2, the group that receive a loan. Demographic parity is the condition that requires the acceptance rate to be the same in all groups of S (Barocas et al., 2019). Under demographic parity, for a binary classifier and a binary sensitive attribute, all of the followings hold:

1. $\mathbb{P}(\hat{y}=1|S=0)=\mathbb{P}(\hat{y}=1|S=1)$: The probability of acceptance is equal for members of different demographic groups. For instance, in Example 1 members of different demographic groups have an equal chance for being classified as low risk.

2. $\mathbb{P}(S=1|\hat{y}=1)=\mathbb{P}(S=1)$: If the population ratio of a specific group is ρ, the ratio of this group in the accepted class is also ρ. For instance, in Example 2, let ρ be the ratio of females in the applicants' pool. Under demographic parity, females will have the same ratio of ρ in the set of admitted for the loan.

3. $I(\hat{y};S)=0$: Mutual information is the measure of mutual dependence between two variables. Under demographic parity $\hat{y}$ and S are independent, hence their mutual information is zero. That is, the conditional entropy $H(S|\hat{y})$ is equal to $H(S)$.

4. $cov(\hat{y},S)=0$: Under demographic parity $\hat{y}$ and S are independent. As a result, the covariance $cov(\hat{y},S)$ is equal to zero.

These, however, hold when demographic parity holds. A disparity (or unfairness) *measure* can be defined using any of the above quantities. Table 1 summarizes some of the ways the disparity can be measured. As stated in the first row of the table, mutual information and covariance (or correlation) provide two natural measures. Another way of quantifying the disparity is by subtracting the probabilities (Row 2 of Table 1) or the ratios between probabilities (Row 3 of Table 1). Consistent with Row 2, we defined the ratio-based measures such that zero is the maximum fairness and the measure is in the range [0,1].

Recall that we consider the societal norms, which is not always quite aligned with statistical measures. Due to the societal discrimination against some minority groups, social data is usually "biased" (Olteanu et al., 2019). Hence, actions known as reverse discrimination (such as affirmative action) are taken to increase the presence of underrepresented groups in the outcomes. Such fairness guidance is usually provided by law. That can be viewed as a higher acceptance probability for certain protected groups of sensitive attributes. In this paper, we do not limit ourselves to any of the (demographic disparity) fairness measures and give the user the freedom to provide a customized measure. That is, *we are agnostic to the choice of fairness measure*. Also, although we used a single binary sensitive attribute for the explanations and measures, we would like to mention that a measure can be defined over multiple non-binary sensitive attributes with overlapping protected groups. Again, we allow the data scientist to provide such customized measures and our model is agnostic to the choice of fairness measure based on demographic parity. Generalizing our notion of fairness to any measure based on demographic disparity, in the rest of the paper, we use the notation $\mathcal{F}(S, C(.))$ to refer to the (user-provided) fairness measure. Also, when S is clear by context, we simplify the notation to $\mathcal{F}(C(.))$.

Finally, algorithmic fairness can be achieved by intervention at pre-processing, in-processing, or post-processing strategies (Friedler et al., 2019). This paper provides an *in-process* strategy for fairness.

3. Fair Active Learning

Smartly selecting samples to label, active learning has the potential to mitigate algorithmic bias. We note that biased training data cause ML models to be unfair. Hence, mitigating bias in the input data can resolve the fairness issues in the model. For instance, consider a linear regression model in form of $\hat{y} = \theta^\top X$. Also, let the covariance between the model outcome and a sensitive attribute S be the measure of fairness. Theorem 1 shows that $cov(S, \hat{y})$ only depends on $cov(S, X)$ and θ.

Assumption 1. There is a linear relationship between X and y. $\hat{y} = \theta^\top X$.

Theorem 1. *Under the Assumption 1.* $cov(S, \hat{y}) = \theta^\top cov(S, X)$.

Proof.

$$\begin{aligned}
cov(S,\hat{y}) &= E[S\hat{y}] - E[S]E[\hat{y}] \\
E[S]E[\hat{y}] &= \mu_S E\big[\sum \theta_i x_i\big] = \mu_S \sum \theta_i \mu_{x_i} \\
&= \theta_1 \mu_S \mu_{x_1} + \theta_2 \mu_S \mu_{x_2} + \cdots + \theta_d \mu_S \mu_{x_d} \\
E[S\hat{y}] &= E\big[S \sum \theta_i x_i\big] = E\big[\sum S\,\theta_i x_i\big] \\
&= E\big[S\,\theta_1 x_1 + S\,\theta_2 x_2 + \cdots + S\,\theta_d x_d\big] \\
&= E\big[S\,\theta_1 x_1\big] + E\big[S\,\theta_2 x_2\big] + \cdots + E\big[S\,\theta_d x_d\big] \\
&= \theta_1 E\big[S\,x_1\big] + \theta_2 E\big[S\,x_2\big] + \cdots + \theta_d E\big[S\,x_d\big]
\end{aligned}$$

$$\begin{aligned}
\Rightarrow & cov(S,\hat{y}) = \\
& \theta_1 E\big[S\,x_1\big] + \theta_2 E\big[S\,x_2\big] + \cdots + \theta_d E\big[S\,x_d\big] \\
& - (\theta_1 \mu_S \mu_{x_1} + \theta_2 \mu_S \mu_{x_2} + \cdots + \theta_d \mu_S \mu_{x_d}) \\
=& \theta_1 (E\big[S\,x_1\big] - \mu_S \mu_{x_1}) + \cdots + \theta_d (E\big[S\,x_d\big] - \mu_S \mu_{x_d}) \\
=& \sum_{i=1}^{d} \theta_i cov(S, x_i) = \theta^\top cov(S, X)
\end{aligned} \tag{2}$$

□

Following the above discussion, we note that the samples selected for labeling in an active learning framework can significantly impact the demographic parity of the trained model. Of course, achieving high model performance is the major objective in any machine learning setting, including active learning. However, blindly optimizing for model performance without considering fairness is problematic. On the other hand, optimizing for fairness, the model may lose its purpose for accuracy. For instance, in Example 2, consider a model that randomly (with equal probabilities) approves customers for a loan without even considering their qualifications. It is easy to see that this model satisfies demographic parity since the outcome probability is (random hence) independent from all features including S. However, such a model loses its purpose since its outcome does not provide any information about how qualified a customer is.

To balance the fairness and accuracy trade-off, in this paper, we introduce *fairness in active learning* to develop fair classifiers for applications with limited labeled data, similar to Example 1 and Example 2. The goal is to minimize the misclassification error as well as demographic disparity (c.f. § 2.3). Similar to regular active learning algorithms, Fair Active Learning (FAL) is an iterative process that every time selects a sample from the unlabeled pool $\mathcal{U}$ to be added to the labeled pool $\mathcal{L}$. Different from other active learning models, however, FAL considers both fairness and misclassification error as the optimization objective for the sampling step. That is, to choose the next sample to be labeled, it chooses a one that contributes to both model accuracy and fairness at the same time. Specifically, following the literature, for a sample point $\langle X^{(i)}, S^{(i)}\rangle \in \mathcal{U}$, we choose Entropy $\mathcal{H}_{t-1}(y^{(i)})$ for misclassification error, while considering demographic disparity $\mathcal{F}(C_t^i(.))$ for fairness — $C_t^i(.)$ is the classifier trained on $\mathcal{L}$, after labeling the point $\langle X^{(i)}, S^{(i)}\rangle$ and $\mathcal{H}_{t-1}(y^{(i)})$ is the entropy of the $y^{(i)}$ based on the current model $C_{t-1}(.)$. Formally, the multi-objective optimization between misclassification error and fairness is as in Equation 3 shows:

$$\max_{\langle X^{(i)}, S^{(i)}\rangle \in \mathcal{U}} \quad \alpha\, \mathcal{H}_{t-1}(y^{(i)}) + (1-\alpha)\big(1 - \mathcal{F}(C_t^i(.))\big) \tag{3}$$

The coefficient $\alpha \in [0, 1]$ is the user-provided parameter that allows her to tune the trade-off between fairness and model performance. Values closer to 1 put more emphasize on model performance, while smaller values of α put more importance on fairness (We experimentally observe in § 4 that setting $\alpha = 0.6$ resulted in models with accuracy almost the same as regular active learning while it dropped the demographic disparity by 50%). Also, the values of entropy and fairness are standardized to the same scale before getting combined in Equation 3.

We note that sometimes fairness is in the form of a hard

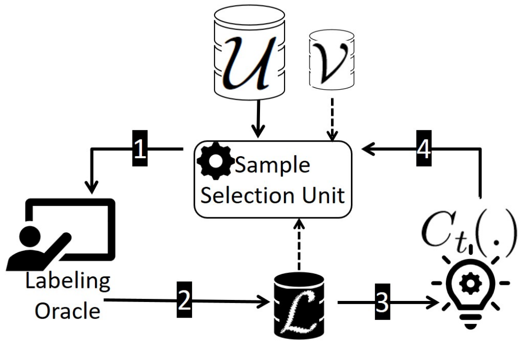

Figure 1. FAL framework

constraint, usually, requested by law. For instance, in Example 2, the requirement can be that at least 20% of candidates approved for a loan should be female.

The optimization formulation in such settings can be formulated as follows:

$$\begin{aligned} \max \quad & \mathcal{H}_{t-1}(y^{(i)}) \\ s.t. \quad & \\ & \mathcal{F}(C_t^i(.)) \le \epsilon \\ & \langle X^{(i)}, S^{(i)} \rangle \in \mathcal{U} \end{aligned} \tag{4}$$

Alternatively, one would consider misclassification error as hard constraint while optimizing for fairness. Both these models can be reformulated as unconstrained optimizations (as in Equation 3) using Lagrange multipliers (Rockafellar, 1993).

Having discussed the optimization function, next we will provide the details of our FAL framework and its components.

3.1. Framework

At a high level, FAL is an iterative approach similar to regular active learning approaches. As shown in Figure 1, the central component of FAL is the sample selection unit (SSU) that picks an unlabeled point $\langle X^{(i)}, S^{(i)} \rangle$ from $\mathcal{U}$ and asks the labeling oracle to provide its label. The labeled point $\langle X^{(i)}, S^{(i)}, y^{(i)} \rangle$ then gets added to $\mathcal{L}$, the set of labeled points. The set of labeled points are then used to train $C_t(.)$: the model at iteration t. Then at iteration $t+1$, SSU uses $C_t(.)$ and picks the next point to be labeled. This process continues until the labeling budget gets exhausted.

The label selection unit is in charge of selecting the next point to be labeled. To do so, it uses Equation 3 to balance the trade-off between fairness and misclassification error. A problem, however, is that at the time of evaluating a point, we still do not know its label as it belongs to $\mathcal{U}$. On the other hand, to calculate the fairness, we need to know what the model parameters would be after adding the current point which requires the label of the point. In other words, in order to pick the next point to be labeled according to Equation 3, we need *all* data points in $\mathcal{U}$ to be already labeled! This has contradiction with the fact that the points in $\mathcal{U}$ are unlabeled.

To resolve this issue, we replace the fairness of the model after adding a point, to the "*expected fairness*" after adding $\langle X^{(i)}, S^{(i)} \rangle$ to $\mathcal{L}$. That is, we use Equation 5 for sample selection.

$$\max_{\langle X^{(i)}, S^{(i)} \rangle \in \mathcal{U}} \quad \alpha\, \mathcal{H}_{t-1}(y^{(i)}) + (1-\alpha)\big(1 - E\big[\mathcal{F}(C_t^i(.))\big]\big) \tag{5}$$

Figure 2 shows how SSU computes the expected fairness after labeling a point $\langle X^{(i)}, S^{(i)} \rangle$. It uses Equation 6 for computing the expected fairness:

$$E\big[\mathcal{F}(C_t^i(.))\big] = \sum_{j=1}^{K} \mathcal{F}(\overline{C_j}(.))\mathbb{P}(y=j|X^{(i)}) \tag{6}$$

where $\overline{C_j}(.)$ is the model trained using $\mathcal{L} \cup \{\langle X^{(i)}, S^{(i)}, y_j \rangle\}$.

Following Figure 2, for every point $P^{(i)} = \langle X^{(i)}, S^{(i)} \rangle$ in the unlabeled pool, the SSU considers different value of $y_0 \cdots y_K$ as possible labels for $P^{(i)}$. For every possible label y_j, it generates the intermediate model $\overline{C_j}(.)$ trained using $\mathcal{L} \cup \{\langle X^{(i)}, S^{(i)}, y_j \rangle\}$.

Next, SSU needs to evaluate the fairness of $\overline{C_j}(.)$. We note that the set of labeled points $\mathcal{L}$ has a different distribution from the underlying data distribution. That is because the labeled points are carefully selected from $\mathcal{U}$ and hence are not unbiased samples from the underlying data distribution. As a result, even though the model is trained using these data points, $\mathcal{L}$ cannot be used for evaluating the fairness of the model. On the other hand, the points in the unlabeled pool $\mathcal{U}$ are expected to be representative of the environment, i.e., the underlying data distribution. Therefore, to create a dataset for evaluating the fairness of a model, we draw random uniform samples from $\mathcal{U}$ and put them in a *verification set* $\mathcal{V}$. The verification set is created once and used in different iterations FAL.

For every possible outcome y_j for the point $\langle X^{(i)}, S^{(i)} \rangle$, SSU requires to compute its probability. At every iteration t, the model $C_{t-1}(.)$ created in the previous step shows the "best known" estimates for the outcome probabilities, according to the labeled data points. Hence, SSU uses $C_{t-1}(.)$ for computing $\mathbb{P}(y=j|X^{(i)})$.

Having the fairness measures and the probabilities for each of possible outcomes, the expected fairness is computed by

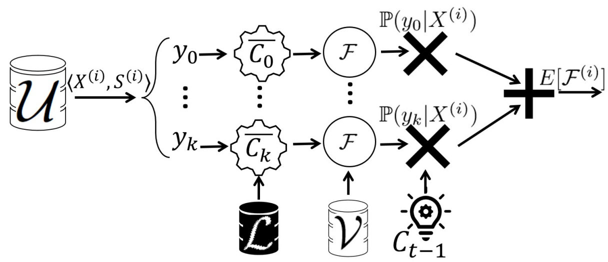

Figure 2. Sample Selection Unit: Evaluation of a point $\langle X^{(i)}, S^{(i)}\rangle \in \mathcal{U}$ for fairness

aggregating $\mathcal{F}(\overline{C_j}(.))\mathbb{P}(y=j|X^{(i)})$ for different values of j (Equation 6).

Finally, having computed the expected fairness for each of the data points in $\mathcal{U}$, SSU finds the point that optimizes Equation 5, and passes it to the labeling oracle as the next point to be labeled. Algorithm 2 shows the pseudo-code of FAL. It uses the function of Algorithm 3 for computing the expected fairness.

Algorithm 2 Fair Active Learning

for $t = 1$ to B **do**
  $max = 0$
  for $i = 1$ to $|\mathcal{U}|$ **do**
    $H = -\sum_{k=1}^{K} \mathbb{P}(y=k|X)\log\mathbb{P}(y=k|X)$
    $F = \textbf{ExpF}(\langle X^{(i)}, S^{(i)}\rangle, \mathcal{V}, \mathcal{L}, C_{t-1}(.))$
    $obj = \alpha H + (1-\alpha)F$
    if $obj > max$ **then**
      $max = obj$
      $\langle X^*, S^*\rangle = \langle X^{(i)}, S^{(i)}\rangle$
    end if
  end for
  $y =$ label X^* using the labeling oracle
  add $\langle X^*, S^*, y\rangle$ to $\mathcal{L}$
  train the classifier $C_t(.)$ using $\mathcal{L}$
end for
return $C_t(.)$

Algorithm 3 ExpF
input: $\langle X^{(i)}, S^{(i)}\rangle, \mathcal{V}, \mathcal{L}, C_{t-1}(.)$

$sum = 0$
for $j = 1$ to K **do**
  train $\overline{C_j}(.)$ using $\mathcal{L} \cup \{\langle X^{(i)}, S^{(i)}, y_j\rangle\}$
  compute $\mathcal{F}(\overline{C_j}(.))$ using $\mathcal{V}$
  $sum = sum + \mathcal{F}(\overline{C_j}(.))\mathbb{P}(y=j|X^{(i)})$
end for
return sum

4. Experiments

4.1. Experiments Setup

The experiments were performed on a Linux machine with a Core I9 CPU and 128GB memory. The algorithms were implemented using Python 3.7.

Datasets.

COMPAS[3]: published by Probuplica (Angwin et al., 2016), this dataset contains information of juvenile felonies such as marriage status, race, age, different types of prior convictions, and the charge degree of the current arrest. We normalized data for the inputs and outputs to be mean zero and unit variance. We consider `sex` and `race` as sensitive attributes. Defining the default fairness measures on `race`, we filtered dataset to black and white defendants. The dataset contains 5875 defendants, after filtering. Following the standard practice (Corbett-Davies et al., 2017; Mehrabi et al., 2019; Dressel & Farid, 2018; Flores et al., 2016), we use two-year violent recidivism record as the true label of recidivism, $y^{(i)}$, for each individual. We set $y^{(i)} = 1$ if the two year recidivism is greater than zero and $y^{(i)} = 0$ otherwise.

4.2. Algorithms Evaluated

We evaluate the performance of three approaches and evaluate the performance of each through experiments. We use the standard logistic regression as the classifier in all cases.

Fair Active Learning (FAL). FAL, Algorithm 2, as described in § 3, has an optimization core, which maximizes the linear combination of entropy and expected fairness measures for sample selection. For each unlabeled instance in $\mathcal{U}$, the expected fairness ExpF is calculated as in Algorithm 3. Briefly speaking, we add each unlabeled instance to the labeled pool with a predicted label each time and calculate the expected fairness. The two components of the optimization objective function, entropy, and expected fairness are on different scales. Consequently, assigning α value directly on these measures does not have an equal impact due to the magnitude of the measures. Let $H^{\min}$, $H^{\max}$, $F^{\min}$, and $F^{\max}$ be the minimum entropy for the points in $\mathcal{U}$, maximum entropy, minimum expected fairness, and maximum expected fairness, respectively. To ensure that the measures have the same format and the trade-off between two components is consistent, we standardize entropy and one minus expected fairness as in Equations 7 and 8, correspondingly.

$$H_i = \frac{\mathcal{H}_{t-1}(y^{(i)}) - H^{\min}}{H^{\max} - H^{\min}}, \forall \langle X^{(i)}, S^{(i)}\rangle \in \mathcal{U} \quad (7)$$

[3] https://bit.ly/35pzGFj

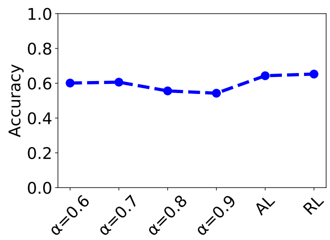

Figure 3. Accuracy score of FAL-$Fairness_{11}$ with different α values vs. AL and RL

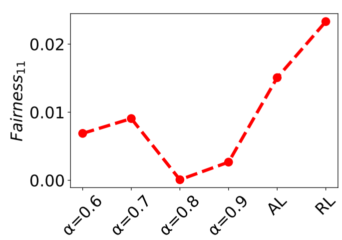

Figure 4. $Fairness_{11}$ of FAL-$Fairness_{11}$ with different α values vs. AL and RL

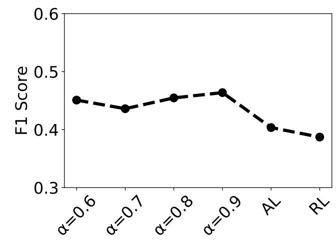

Figure 5. F1-score of FAL-$Fairness_{11}$ with different α values vs. AL and RL

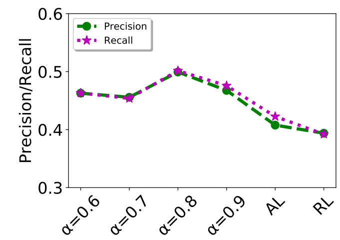

Figure 6. Precision/Recall of FAL-$Fairness_{11}$ with different α values vs. AL and RL

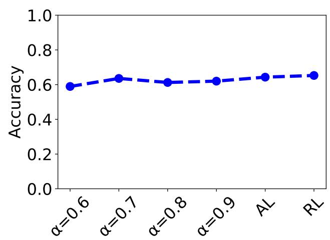

Figure 7. Accuracy score of FAL-$Fairness_{12}$ with different α values vs. AL and RL

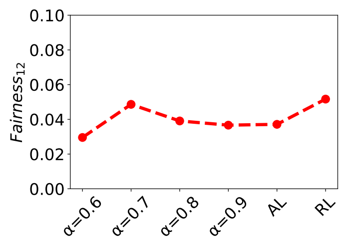

Figure 8. $Fairness_{12}$ of FAL-$Fairness_{12}$ with different α values vs. AL and RL

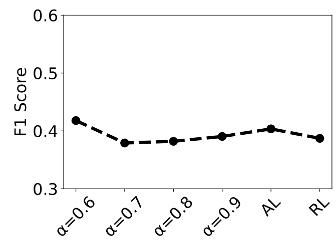

Figure 9. F1-score of FAL-$Fairness_{12}$ with different α values vs. AL and RL

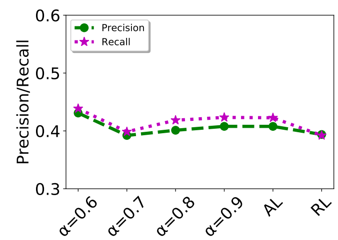

Figure 10. Precision/Recall of FAL-$Fairness_{12}$ with different α values vs. AL and RL

$$\tilde{F}_i = \frac{F^{\max} - E\big[\mathcal{F}(C_t^i(.))\big]}{F^{\max} - F^{\min}}, \forall \langle X^{(i)}, S^{(i)} \rangle \in \mathcal{U} \quad (8)$$

Consequently, Equation 5 can be rewritten as:

$$\max \quad \alpha\, H_i + (1-\alpha)\tilde{F}_i, \forall \langle X^{(i)}, S^{(i)} \rangle \in \mathcal{U} \quad (9)$$

Active Learning (AL). In AL, Algorithm 1, the sample selection is merely based on minimizing the miss-classification error. To do so, we calculate the entropy of each unlabeled instances in $\mathcal{U}$ and choose the point with maximum entropy to label next. In AL, the focus of the sampling is on the prediction accuracy of the classifier.

Random Labeling (RL). As briefly mentioned in § 1, the baseline approach for the limited labeling context is to randomly label a subset of points in $\mathcal{U}$. In RL, we ask the labeling oracle to provide the label of B samples and use them to train the classifier.

4.3. Performance Evaluation

We evaluate the performance of FAL, AL, and RL, with different demographic disparity functions defined in Table 1. We labeled the Fairness measures as $Fairness_{ij}$, which is consistent with the corresponding cell that each measure is located in Table 1. For instance, $Fairness_{11}$ refers to Mutual Information located in cell [1,1] (first row and first column) of Table 1. We study the trade-off between the accuracy and demographic disparity, by changing the coefficient α in Equation 5. We issued 10 different random splits for each scenario and considered the accuracy, F1-score, Precision/Recall, and the corresponding demographic function used in the optimization step, as evaluation measures.

In addition to Active Learning, Algorithm 3.1, we compare our results against Random Labeling (random labeling baseline). In the plots, we label the standard active learning algorithm as AL and the Random Labeling baseline is labeled as RL. We run the experiments on the different random split of datasets while varying the coefficient α for different Fairness metrics in Table 1. We study the values of $\{0.6, 0.7, 0.8, 0.9\}$ for α, plus AL and RL.

For COMPAS dataset, we split the dataset into training $\mathcal{U}$ 60%, verification $\mathcal{V}$ 20%, and testing 20%. We specify the maximum labeling budget to 400 based on the preliminary results, where the performance leveled off. In each FAL and AL scenario, we start with five labeled points and sequentially select points to label, until the 400 budget is exhausted.

Summary of results: The experiment results are provided in Figures 3–38. We label each FAL scenario as FAL-$Fairness_{ij}$. In summary, the experiment results confirm the effectiveness of our proposal, since FAL could improve fairness significantly by reducing the disparities by around 50% in different scenarios (Fig-

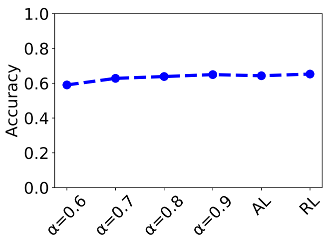

Figure 11. Accuracy score of FAL-$Fairness_{21}$ with different α values vs. AL and RL

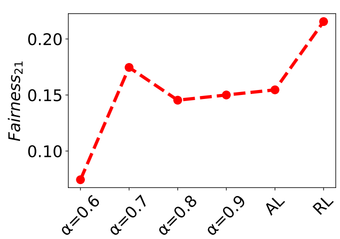

Figure 12. $Fairness_{21}$ of FAL-$Fairness_{21}$ with different α values vs. AL and RL

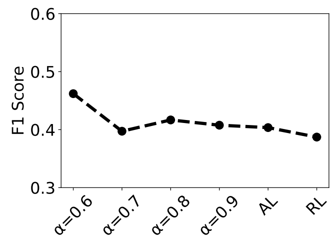

Figure 13. F1-score of FAL-$Fairness_{21}$ with different α values vs. AL and RL

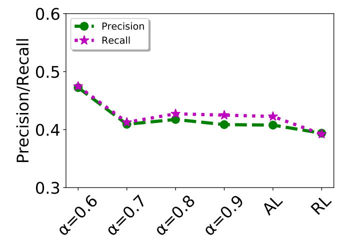

Figure 14. Precision/Recall of FAL-$Fairness_{21}$ with different α values vs. AL and RL

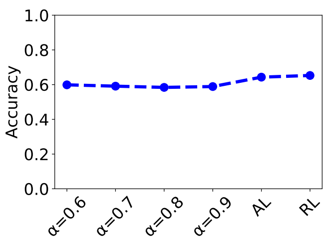

Figure 15. Accuracy score of FAL-$Fairness_{22}$ with different α values vs. AL and RL

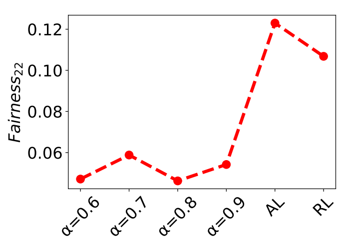

Figure 16. $Fairness_{22}$ of FAL-$Fairness_{22}$ with different α values vs. AL and RL

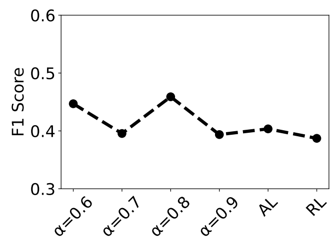

Figure 17. F1-score of FAL-$Fairness_{22}$ with different α values vs. AL and RL

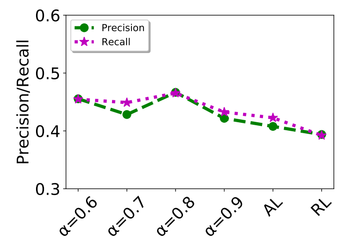

Figure 18. Precision/Recall of FAL-$Fairness_{22}$ with different α values vs. AL and RL

ures 4,8,12,16,20, and24) while model performance and accuracy (Figures 3, 7, 11, 15, 19, and 23) remained almost unchanged. Below, we explain different cases in detail.

Figures 3–6 corresponds to employing Mutual Information, $Fairness_{11}$, in optimization step. We note that adding the fairness measure to the optimization for candidate selection does not significantly impact the overall accuracy of the model across different α values. Compared to AL also, the accuracy does not drop dramatically. As we observe, the demographic disparity of the model though, dropped by almost 50% compared to AL and RL, especially when the weight of the fairness in higher, $\alpha = 0.6$. F1-score and precision/recall have an overall decreasing trend reaching toward AL and RL, with a peak at $\alpha = 0.9$ that can be explained with the randomness effect.

Moving to second fairness measure, Figures 7–10 shows the results using $Fairness_{1,2}$, the covariance metric as fairness metric in optimization. We note that the accuracy of the model across different scenarios has negligible differences as well as the covariance measure. F1-score and precision/recall are quite stable across different scenarios.

Figures 11–14 provide the results where we use demographic error measure $Fairness_{21}$ for the optimization fairness component. Here, there is an insignificant change in the accuracy score as the α value changes. We can observe the same overall increasing pattern for the fairness metric with the best fairness value occurring at $\alpha = 0.6$. It is worth mentioning that, the F1-score and precision/recall have their best value compared to other scenarios with different fairness metrics.

Employing $Fairness_{22}$ in the optimization process leads to the results presented in Figures 15–18. A very small change can be seen in the accuracy score across different α values and the same increasing fairness pattern with the best at $\alpha = 0.6$ is observable. F1-score and precision/recall have a less stable pattern in this scenario. Results provided in Figures 19–22 correspond to the $Fairness_{31}$ metric applied in optimization process. Although the fairness measure has an increasing trend across different α values the accuracy does not change, remarkably. F1-score and precision/recall are fairly stable across α changes.

Figures 23–26 presents the results for the case $Fairness32$ is applied as the fairness metric in optimization. Reasonably stable accuracy with an increasing fairness trend can be noticed. There is an F1-score drop at $\alpha = 0.7$, which is resulted from the precision drop under that scenario.

Next, we study the trend of accuracy and fairness as the labeled instances increases under different fairness scenario. We provide the results for a single run at $\alpha = 0.6$. Each pair in figures 27–38 corresponds to the accuracy and fairness measure, correspondingly. Overall, as the number of sample points increase the accuracy of the model increases. Also, the fairness component tries to maintain the model fairness over time which is observed in our results. Initially, the

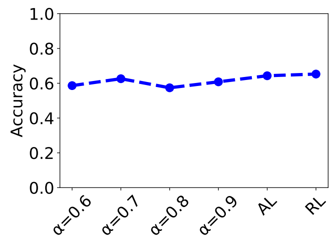

Figure 19. Accuracy score of FAL-$Fairness_{31}$ with different α values vs. AL and RL

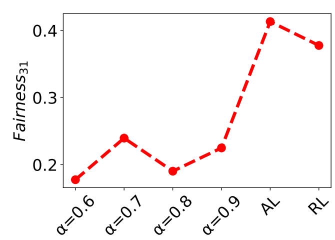

Figure 20. $Fairness_{31}$ of FAL-$Fairness_{31}$ with different α values vs. AL and RL

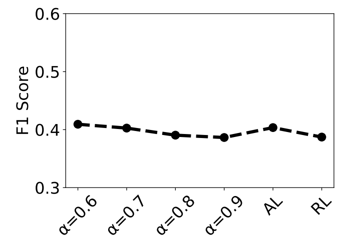

Figure 21. F1-score of FAL-$Fairness_{31}$ with different α values vs. AL and RL

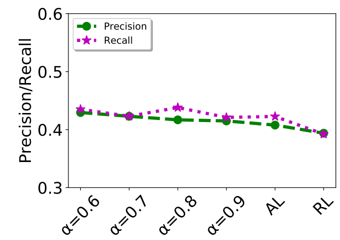

Figure 22. Precision/Recall of FAL-$Fairness_{31}$ with different α values vs. AL and RL

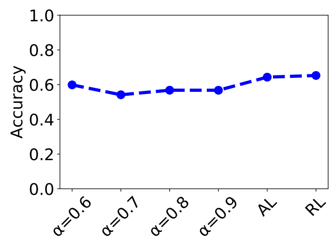

Figure 23. Accuracy score of FAL-$Fairness_{32}$ with different α values vs. AL and RL

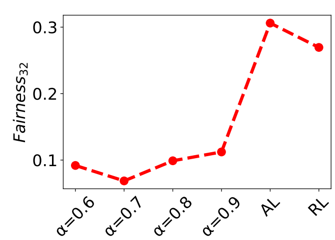

Figure 24. $Fairness_{32}$ of FAL-$Fairness_{32}$ with different α values vs. AL and RL

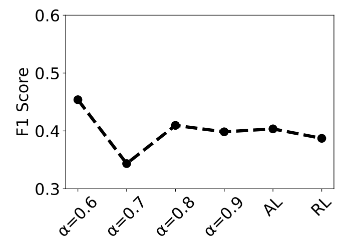

Figure 25. F1-score of FAL-$Fairness_{32}$ with different α values vs. AL and RL

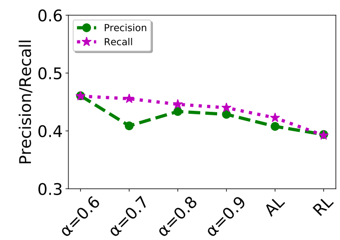

Figure 26. Precision/Recall of FAL-$Fairness_{32}$ with different α values vs. AL and RL

models are not reliable as the number of training samples were small. This resulted in an unstable behaviour of the models both for accuracy and fairness in initial steps. The trends become more stable after a few iterations that models become more reliable.

5. Discussions

So far in this paper, we considered the fairness measures defined based on the notion of independence ($\hat{y} \perp S$). Here we discuss how our results can be extended for other measures based on separation ($\hat{y} \perp S \mid y$) and sufficiency ($y \perp S \mid \hat{y}$) (Barocas et al., 2019), such as predictive parity, error rate balance, and accuracy equity (Narayanan, 2018).

FAL follows balancing fairness and misclassification error, using Equation 5. Obviously the entropy term $\mathcal{H}_{t-1}(y^{(i)})$ does not depend on the choice of fairness measure. Also, the abstract fairness term $E\big[\mathcal{F}(C_t^i(.))\big]$ is not limited to a specific definition. However, despite being abstract, computing the expected fairness (based on separation or sufficiency) is challenging.

Looking at Figure 2, recall that we use the verification set $\mathcal{V}$ for estimating the fairness of a model. As an unbiased sample set from $\mathcal{U}$, $\mathcal{V}$ follows the underlying data distribution and, hence, can be used for demographic disparity. However, this set cannot be used for estimating fairness according to separation or sufficiency since its instances are not labeled. On the other hand, the pool of labeled data is not representative of the underlying data distribution.

Fortunately, in order to extend our framework for other fairness measures based on the true label, it is enough to label $\mathcal{V}$. Once the points in $\mathcal{V}$ are labeled, it is easy to see that our framework works as-is for any such fairness measure. We understand that since the labeling budget is limited, labeling $\mathcal{V}$ may reduce the number of instances we can use for training the model. One resolution is to limit $\mathcal{V}$ to a small set and accept the potential error in estimating the probabilities. How small $\mathcal{V}$ can be to still provide accurate-enough estimations, besides other resolutions, are interesting questions that we will consider for future work.

6. Conclusion

In this paper, we developed a fair active learning framework. Our framework computes the expected fairness for each unlabeled sample point and selects the one that maximizes a weighted linear combination of accuracy score and the expected fairness measure. We carried out experiments with different fairness measures across different weights. The results confirmed that our proposed FAL frameworks work properly in terms of building a fair model without majorly affecting its performance. The comparison with the standard active learning shows an improvement in the fairness quality of the constructed classifier by about 50%.

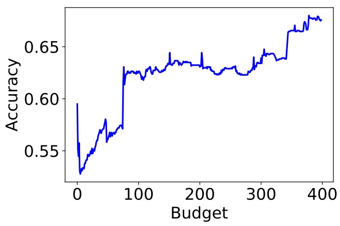

Figure 27. Accuracy score of FAL-$Fairness_{11}$ vs. labeling budget

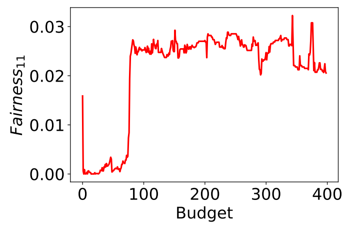

Figure 28. $Fairness_{11}$ of FAL-$Fairness_{11}$ vs. labeling budget

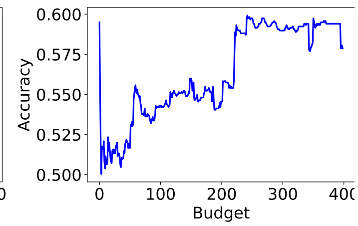

Figure 29. Accuracy score of FAL-$Fairness_{12}$ vs. labeling budget

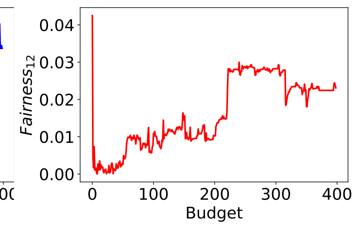

Figure 30. $Fairness_{12}$ of FAL-$Fairness_{12}$ vs. labeling budget

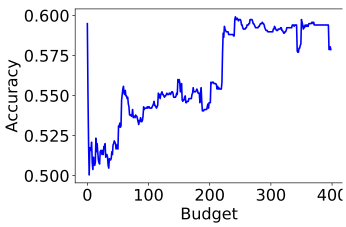

Figure 31. Accuracy score of FAL-$Fairness_{21}$ vs. labeling budget

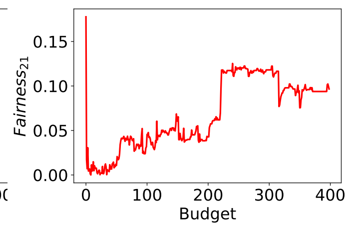

Figure 32. $Fairness_{21}$ of FAL-$Fairness_{21}$ vs. labeling budget

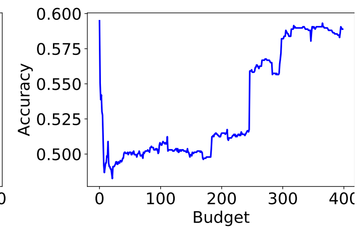

Figure 33. Accuracy score of FAL-$Fairness_{22}$ vs. labeling budget

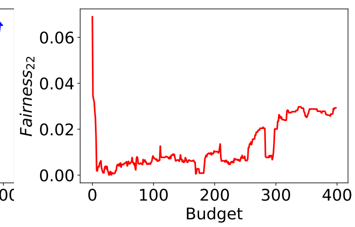

Figure 34. $Fairness_{22}$ of FAL-$Fairness_{22}$ vs. labeling budget

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, 2016. URL https://bit.ly/2s0UMfA.

Asudeh, A., Jagadish, H., Stoyanovich, J., and Das, G. Designing fair ranking schemes. *SIGMOD*, 2019.

Ayres, I. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of" included variable" bias. *Perspectives in biology and medicine*, 48 (1):68–S87, 2005.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *NIPS Tutorial*, 2017.

Barocas, S., Hardt, M., and Narayanan, A. Fairness and machine learning: Limitations and opportunities. fairmlbook.org, 2019.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338. ACM, 2019.

Jan, T. Redlining was banned 50 years ago. it's still hurting minorities today. Washington Post, 2018.

Jones, F. L. Sources of gender inequality in income: what the australian census says. *Social Forces*, 62(1):134–152, 1983.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

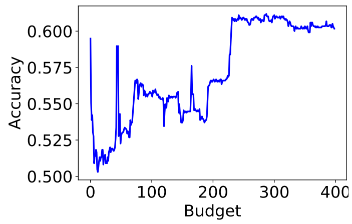

Figure 35. Accuracy score of FAL-$Fairness_{31}$ vs. labeling budget

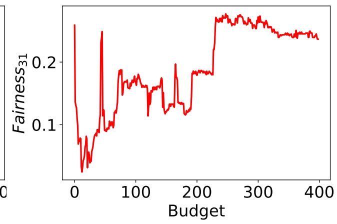

Figure 36. $Fairness_{31}$ of FAL-$Fairness_{31}$ vs. labeling budget

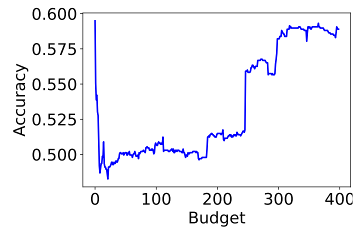

Figure 37. Accuracy score of FAL-$Fairness_{32}$ vs. labeling budget

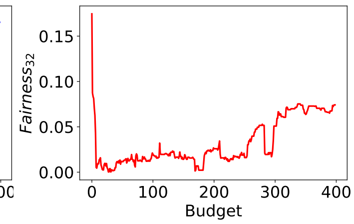

Figure 38. $Fairness_{32}$ of FAL-$Fairness_{32}$ vs. labeling budget

Lahoti, P., Gummadi, K. P., and Weikum, G. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1334–1345. IEEE, 2019.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *SIGIR'94*, pp. 3–12. Springer, 1994.

Li, A. and Cropanzano, R. Fairness at the group level: Justice climate and intraunit justice climate. *Journal of management*, 35(3):564–599, 2009.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

Olteanu, A., Castillo, C., Diaz, F., and Kiciman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

Perez, C. C. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Random House, 2019.

Pley, C. and Keeling, A. Gender bias in health AI - prejudicing health outcomes (or getting it right!). *Women in Global Health*, Sep. 2019.

Rockafellar, R. T. Lagrange multipliers and optimality. *SIAM review*, 35(2):183–238, 1993.

Settles, B. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

Simoiu, C., Corbett-Davies, S., Goel, S., et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31 (4):1060–1089, 2017.