ReSGrAL: Fairness-Sensitive Active Learning

Nina van Liebergen[1,2,*], Marianne Schaaphok[1] and Giovanni Sileno[2]

[1] *Netherlands Organisation for Applied Scientific Research (TNO) - Data Science, The Hague, The Netherlands*
[2] *University of Amsterdam, Amsterdam, The Netherlands*

Abstract

The use of machine learning models for decision support in public organizations is generally constrained by limited labeled data and the high cost of labeling. Additionally, models used by the public sector have been shown to express various biases (e.g., towards gender or ethnicity), highlighting the urgency to address fairness concerns. Although active learning has proven to be useful in efficiently selecting instances for labeling (and thus reducing the impact of the first issue), its impact on fairness is still unclear. The present work has a two-fold objective. First, it aims to experimentally study the relationship between active learning and fairness. Second, it explores fairness-sensitive methods for active learning, proposing two novel variations, Representative SubGroup Active Learning (ReSGrAL) and Fair ReSGrAL. Our experiments show that, in general, active learning can increase model unfairness beyond the dataset bias, and thus caution is needed when using active learning in sensitive contexts. Fortunately, we also show that techniques like ReSGrAL can mitigate unfairness without sacrificing accuracy.

Keywords

Active Learning, Fairness, Machine Learning, Bias Mitigation

1. Introduction

Although data-driven decision support is promoted for its potential to enhance efficiency, the implementation of machine learning models in public organizations often encounters challenges related to the availability of labeled data and the risk of producing biased, discriminatory, unfair outcomes [1, 2]. Understanding how these two issues intersect and addressing them effectively is therefore of foremost importance.

Looking at the first issue only, obtaining high-quality labeled data is known to be generally costly and time-consuming, leading to datasets that are insufficient or only partially labeled [1]. Fortunately, active learning has emerged as a promising method to address the scarcity of labeled data by selectively focusing the labeling effort on the most informative instances in an unlabeled pool [3]. Orthogonally to this dimension, current decision support systems in the public sector have been found to exhibit various forms of biases and unfairness[1], especially towards sensitive groups. In the Ethics Guidelines for trustworthy AI (2019) of the High-level Expert Group on AI of the European Commission[2], it is explicitly stated that *unfair bias must be avoided”*; governments are obligated to respect all applicable laws and values when developing and adopting AI systems.

To make the matter further complex, traditional methods for evaluating fairness (generally targeting group-level fairness) may not adequately detect biases within subgroups [4]. Unexpectedly, the intersection of active learning and the evaluation of bias and fairness is relatively understudied. However, there are two studies that apply active learning for training a fair model, Fair active learning (FAL) [5] and PANDA [6], although they consider two different use scenarios (FAL aims to train a model from scratch, PANDA starts from an existing, possibly biased, labeled subset).

AEIB 2024: Workshop on Implementing AI Ethics through a Behavioural Lens | co-located with ECAI 2024, Santiago de Compostela, Spain - Workshop AIEB

[*] Corresponding author.
nina.vanliebergen@tno.nl (N. v. Liebergen)

[1] See e.g. WIRED, Inside the Suspicion Machine”, https://www.wired.com/story/welfare-state-algorithms/ (2023, March), NOS, *DUO mag algoritme niet gebruiken totdat meer bekend is over mogelijke discriminatie*”, https://nos.nl/op3/artikel/2480024 and *Belastingdienst gebruikte omstreden software om fraude op te sporen vaker*” https://nos.nl/artikel/2426009

[2] Ethics Guidelines for Trustworthy AI (8 April 2019). Shaping Europe’s digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

To contribute to reducing this research gap, we aim to investigate further the relationship between active learning and fairness. We examine state-of-the-art active learning strategies (uncertainty sampling and representative sampling), acknowledge the different types of unfairness in machine learning models and assess the influence of active learning, focusing on its performance of subgroups. Additionally, the research explores existing adaptations of active learning strategies and proposes two new variations (simple, yet effective in our experimental settings), ReSGrAL and Fair ReSGrAL, to mitigate unfairness while using active learning.

The paper proceeds as follows. Section 2 summarizes the theoretical background on fairness, bias, and active learning, identifying the research gap we address. Section 3 presents the methodology we used to approach our two research questions: (1) the evaluation of active learning in a fairness-sensitive context, (2) ReSGrAL, the novel method we propose. Section 4 presents details on our experiments (datasets, evaluation metrics, sampling strategies, and technical implementation). Section 5 reports the results. Section 6 offers a higher-level view on the research findings, their limitations, and implications.

2. Background and Related Works

Unacceptable Bias Any machine learning model is subject to some form of informational bias, since the model is meant to capture patterns observable within a given dataset. For instance, when a classifier is trained, it tries to learn how to decide which data instances belong to which class. Looking at the wider pipeline related to machine-learning models, different types of biases are identified in the literature [7, 2]. They can occur in the input data (e.g. *historical bias, sampling bias, measurement bias and evaluation bias*), in the algorithm itself, or be dependent on the application of the algorithm.

Unfortunately, some of the patterns captured in the machine learning may be unjustified, since they are based on and/or realize (socially, legally) *unacceptable* discrimination [8]. In general, an unacceptable bias in a model can arise from direct discrimination or from indirect discrimination. Direct discrimination occurs when protected attributes directly lead to unfavorable outcomes, such as denying a mortgage based on sex. In contrast, indirect discrimination is subtler, as outcomes are not directly tied to protected attributes but correlate with other non-protected ones (eg. ZIP codes).

Algorithmic Definitions of Fairness When a bias is detected or suspected, the model is possibly unfair. The number of existing interpretations of fairness, within philosophy and computer science, shows that a single, concrete definition is missing, even impossible [2]. This makes it hard to test "unfairness" in an absolute sense. However, different definitions can be tested to have a more informed decision. Commonly used fairness definitions are *statistical/demographic parity*, *equal opportunity* and *equalized odds* [7, 2]. These fairness metrics are generally applied to groups with shared (protected) attributes. However, fairness in machine learning can be conceptualized at multiple levels: group fairness, individual fairness and subgroup fairness. Group fairness refers to the equitable treatment of groups, ensuring that decisions do not disproportionately disadvantage any one group. Individual fairness, on the other hand, seeks to provide consistent predictions for individuals who are similar. This aims to ensure that like cases are treated alike, preserving fairness at the most granular level of analysis. However, achieving both group and individual fairness can be challenging as efforts to optimize for one can inadvertently compromise the other. This conflict is exemplified in scenarios where adjustments to enhance group fairness may lead to outcomes where similar individuals are treated differently, violating individual fairness. This phenomenon is known as *fairness gerrymandering* [4].

Furthermore, a model may appear fair when assessing broad categories like gender or race separately but still exhibit biases against intersections of these categories. Intersectional perspectives highlight the necessity of considering multiple identity factors simultaneously to truly assess and address fairness [**?**], but even more, it shows the importance of understanding the sources of oppression in the dataset to identify the origins of discrimination [9]. Subgroup fairness is centered on subgroups typically defined by combinations of attributes [2]. Identifying which combinations are relevant can be methodologically demanding. The solutions observed in the literature range from relying on experts specifying the sensitive profiles, or automating the process based techniques such as clustering or relying on clustering. Independently of how subgroups are selected, a sub-group centered analysis help in revealing biases that might not be detected when analyzing fairness only at the general group level. Decision consistency at sub-group level also favours individual-based fairness.

Sampling In domains concerning the general public, access to the data of the total population is generally impossible (nor desirable). There are different methods known when only limited labeling of data is possible. For training a model, one could use a sample, a subset of the total population, selected in order to be representative of the larger population [10], as well to be useful to train a robust algorithm. *Random Sampling* treats individuals equally, reducing selection bias, though it may lack representativeness (minority groups have fewer chances to be extracted) and may lead to inaccurate decisions towards unrepresented groups. In contrast, *Stratified Sampling* divides the population into subgroups based on common characteristics, ensuring diversity but potentially reducing overall accuracy.

Active Learning Another sampling strategy is *active learning*, which selects data so that the model learns the most while using the least labels as possible, thus minimizing the cost for labeling data [11]. Typically, active learning is used to train a model when only a small amount of data is labelled and the rest of the (relatively large) dataset is unlabelled. By applying this method, samples of an unlabelled dataset are adaptively selected for labeling based on an *acquisition function* which ranks them in order of *importance* to train the model. Instead of batch/offline learning, where the model is trained on the data at once, active learning has a sequential experimental design. There are three main sampling strategies known: informativeness-based (or uncertainty based), representativeness-based and a combination of the two.

Informativeness-based strategies focus on selecting instances that provide most information with respect to the knowledge currently reified by the model. When training a binary classifier, one selects the instances for which the probabilities are close to 0.5. This is also called uncertainty sampling based on confidence. From now one, we refer to this sampling strategy as uncertainty sampling.

Following a representative-based active learning approach, one includes the data distribution in the selection process. One approach of representative sampling is to select the instances that are the closest to other data instances. The representativeness of an unlabeled data instances is calculated in the following way:

$$I(x) = \frac{1}{|X_u|} \sum_{x' \in X} \text{sim}(x, x') \tag{1}$$

where $\text{sim}(x, x')$ is a similarity function. There are multiple formulas for calculating the similarity (or the contrary of the distance) between points. We selected an Euclidean distance as it is the most commonly used. We refer to this sampling strategy as representative sampling.

Some studies combine the informativeness and representativeness of data instances to form an acquisition function. For instance, QUIRE (2014) [12] implements the combination as solving a min-max optimization problem. The selection is based on maximizing the informativeness of the selected sample by training a model on the labelled instances and selecting the most uncertain instances (min). The measure of representativeness is based on the prediction accuracy based on the unlabelled instances (max).

$$\text{QUIRE}(x) = \alpha \cdot \text{Uncertainty}(x) + (1 - \alpha) \cdot \text{Diversity}(x)$$

Where α is a hyperparameter controlling the balance between uncertainty and diversity (with $0 \leq \alpha \leq 1$), $\text{Uncertainty}(x)$ is the uncertainty score of instance x, $\text{Diversity}(x)$ is the diversity score of instance x. The higher the Diversity, the lower the representativeness of the instance.

Fairness-sensitive Active Learning By tinkering with the acquisition function, active learning can in principle be adapted to become a tool to increase/decrease any metrics, including the (un)fairness of a model. Indeed, in the literature two active-learning methods sensitive to fairness are known: Fair active learning (FAL) [5] and PANDA [6]. These two proposals consider however two different scenarios. FAL aims to train a fair model from scratch: it iteratively selects the most optimal (regarding accuracy and fairness) instance for labeling starting from zero instances. FAL proposes several strategies to balance accuracy and fairness, including an optimization to manage the computational complexity involved in fairness calculations. The fairness metrics used, such as equalized odds or demographic parity, are adjustable according to needs. The results show that FAL (specifically FAL-α) gains a higher fairness rate (here: demographic parity calculated with mutual information) than other active and passive learning algorithms while handing in some accuracy.

Where the authors of FAL try to find the most optimal datapoint for labeling using the expected unfairness reduction from the unlabelled datapoints, the authors of PANDA [6] focus on information learned by the already labelled datapoints. The setting is therefore different than FAL since in this setting a batch of instances is already labelled. The algorithm PANDA (standing for "Parity-constrAiNeD metA active learning") uses a meta-learning approach: it learns the most optimal query strategy for selecting new labels subject to a parity constraint. Their results show that PANDA performs the best of all the active learning methods (including FAL) in terms of fairness/accuracy trade-off in this setting.

For the rest of our study, we focus on FAL since this approach fits the best in our setting (having almost no labelled data available). Besides, FAL is the most accessible and the most cited fairness improving active learning method.

3. Methodology

We aim to investigate the effectiveness of various active learning methods through three different setups. Figure 3.1 presents the three data processing pipelines for our three experiments. The source code and datasets used for our experiments are added as attachment.

3.1. Evaluation of Active Learning

The first pipeline serves to assess various state-of-the-art active learning techniques to understand their impact on model accuracy and fairness. Abiding by the standard pipeline, the process includes the following steps: (i) import the unlabeled dataset; (ii) perform active learning to rank the unlabeled data instances; (iii) obtain the label for the selected (highest-ranked) instance, add the labeled instance to the labeled dataset and retrain the model on the new labeled dataset; (iv) at every iteration, evaluate the performance (accuracy and unfairness) on the total dataset and on the performance per different subgroup.

3.2. Representative Subgroup Active Learning

Traditional active learning overlooks subgroup disparities. We therefore extended the traditional pipeline for *Representative SubGroup Active Learning* (ReSGrAL). This method integrates the principles of stratified sampling within an active learning framework to enforce equitable data representation across predefined subgroups. We hypothesized that this approach may reduce the speed at which accuracy and fairness metrics evolve, potentially offering a more balanced trade-off between these two objectives. ReSGrAL is defined by the following steps: (i) import the unlabeled dataset; (ii) identify subgroups (via domain expertise or statistical methods); (iii) perform active learning independently to each subgroup; (iv) label the most informative samples from each subgroup and train a model based on the labelled instances; (v) reassess the model's performance after incorporating new data, focusing on both overall and subgroup-specific metrics. When considering N subgroups, ReSGrAL selects N datapoints at once. The amount of iterations will thus be N times smaller.

Selection of Subgroups Critical to the success of ReSGrAL is the accurate identification of subgroups. This involves understanding which segments of data might provide information possibly contributing

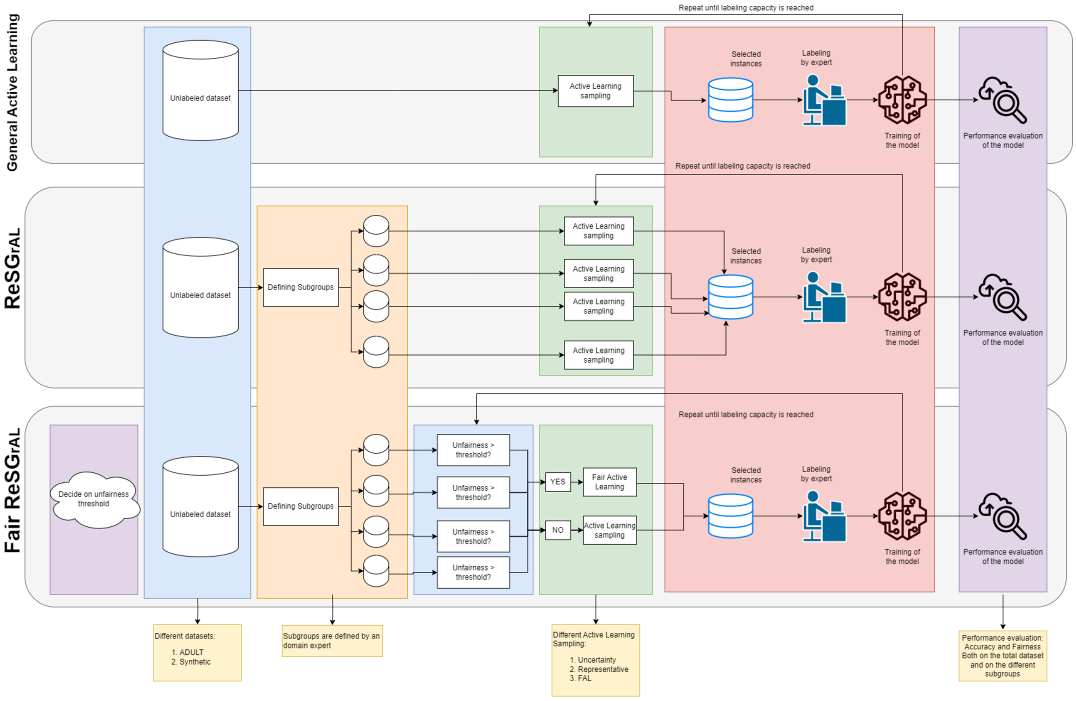

Figure 1: Design of the three methods for experiments: above the general active learning pipeline, in the middle ReSGrAL and below Fair ReSGrAL.

to unfair bias. Subgroups can be selected based on insights from domain experts, analysis of proxy attributes or through clustering algorithms designed to detect nuanced patterns of bias within the data.

3.3. Fair ReSGrAL

Our third pipeline, named Fair ReSGrAL, considers a method which combines the insights we gained from the previous experiments for the training of an accurate and fair model on subgroups. It is a modified version of ReSGrAL that uses a threshold to determine whether the sampling selection for a given group will be based on accuracy or on fairness. The method takes the following steps: First, a threshold on unfairness is selected by the domain expert (for example, demographic disparity for gender is not acceptable when higher then 0.4). Secondly, active learning based on accuracy is performed on every subgroup, measuring unfairness. When a subgroup reaches the given threshold of unfairness, instead of selecting samples by active learning based on accuracy, samples are selected that decrease the unfairness (active learning based on fairness).

We hypothesize that Fair ReSGrAL will not only enhance overall accuracy but also reduce unfairness more effectively compared to methods that do not adapt based on subgroup-specific performance metrics. This adaptive approach is designed to offer a practical solution to the challenge of maintaining model fairness without excessive compromise on decision-making accuracy.

4. Experiments

In this section, we detail the experimental setup used to evaluate the relationship between active learning and fairness, as well as to assess the efficacy of the proposed methods. We utilize two distinct datasets for our evaluations: an artificial dataset designed specifically for controlled testing conditions and the Adult Income dataset[3], a widely recognized in the fairness literature.

[3] UCI repository, https://bit.ly/2GTWz9Z

4.1. Datasets

Synthetic Dataset We first applied our method on a synthetic dataset in order to have a more robust evaluation in a controlled experimental setting. The dataset represents an employee database with the attributes of the dataset consisting of *Province, Income, Occupation, Time Employed, Fraud* and *Gender*. In one part of the dataset we implement a historical bias: in group 3, women are deemed more fraudulent than men. We implement the bias as indirect discrimination: we make attribute *Occupation* a proxy for the sensitive attribute (*Gender*). For instance, in a certain society, there may be relatively more female teachers, while most of the doctors may be male; this is an example of how *Occupation* can be correlated with the sensitive attribute. When *Occupation* (and thus indirect *Gender*) provides information of the true value *Fraud*, the dataset contains a form of indirect discrimination.

Adult Income Dataset This real dataset includes diverse attributes such as age, education and occupation. We select the proxy attribute *occupation* for the sensitive attribute *gender*, since this has the highest relevant correlation value. Since we only want to focus on one sensitive attribute, we remove the attribute *race*. Besides, we remove all the data instances that have at least an unknown value. After the data preprocessing, the dataset consists of 45222 data instances. Occupation has 14 different values. Since the occupation *Armed-forces* consist of only 14 data instances, we added this group to the *other-services* group. Therefore, we remain with 13 different groups.

Simulating an Unlabeled Dataset For both datasets, we simulate an unlabeled dataset by hiding the true value of *income* (a binary value, high or low). When we *sample* and *label* the data instances based on our different acquisition functions, we retrieve the specific value from the database and add this to the labelled dataset where the model will be trained on.

4.2. Evaluation Metrics

To produce a quantitative assessment, we use two primary metrics.

Accuracy This common metric evaluates the percentage of correctly predicted instances out of the total dataset. It helps gauge the learning efficiency and effectiveness of the active learning algorithms.

Unfairness We measured unfairness by demographic parity difference (DPD). Demographic parity requires that the probability of a person in a sensitive class being classified positively needs to be equal to the probability of the total population being classified positively. The protected and the unprotected groups should therefore have the same positive rates. The demographic parity difference is measured as follows:

$$\mathrm{DPD} = |P(Y = 1|S = 1) - P(Y = 1|S = 0)| \tag{2}$$

The terms $P(Y = 1|S = 0)$ and $P(Y = 1|S = 1)$ denote the probabilities of a positive outcome for non-sensitive and sensitive groups, respectively. Ideal demographic parity occurs when these probabilities are nearly equal, indicating fairness.

4.3. Sampling Strategies

Our experiments incorporate three active learning strategies besides random sampling. Therefore, the sampling strategies that we use are:

Random Sampling This baseline method involves selecting unlabeled data instances randomly for labeling. It serves as a control/baseline to evaluate the effectiveness of more strategic sampling methods.

Active Learning Based on Uncertainty Sampling Focused on informativeness, this strategy selects instances for which the model's prediction is most uncertain. Typically, these are instances near the decision boundary (i.e., with predicted probabilities around 0.5), assumed to provide most information if labeled.

Active Learning Based on Density Sampling Under this representative-based approach, samples are chosen based on their proximity to other data points in the feature space, calculated using Euclidean distance. This method aims to select instances that are central within clusters of the dataset, hypothesizing that such samples are more representative of the dataset's overall structure.

Fair Active Learning (FAL) - Estimation of FAL Prioritizing fairness, FAL selects instances that are likely to reduce unfairness in the model's predictions. Due to its intensive computational requirements[4], full-scale implementation was not feasible; instead, a reduced evaluation was conducted on a 20% subsample of the dataset. The experiment was conducted 30 times on 30 distinct subsamples of the dataset, and comparisons were also made using these specific subsets. We henceforth denote this method as the *Estimation of FAL*. In contrast, for the synthetic dataset, the execution of FAL was feasible.

4.4. Technical Implementation

Experiments were implemented using Python with logistic regression models for the synthetic dataset and random forest classifiers[5] for the Adult Income dataset. Active learning scenarios were deployed in a stream-based format, where initial models were trained with a small set of randomly selected labeled points. Iterative updates to the model were made as new labels were acquired based on the active learning strategy in use. Performance was evaluated against accuracy and unfairness, with results averaged over 30 randomized trials to ensure robustness. This experimental design allowed us to dissect the impacts of each active learning strategy on both the fairness and accuracy of predictive models.

5. Results

This section presents the results of the experiments conducted on the Adult Income dataset. For reasons of space, we present only the results for the Adult Income (from here on named *adult*, to simplify) dataset, as more representative of real-world scenarios, and we only show the most relevant figures for answering our research questions.

5.1. Learning Curves and Sampling Efficiency

Our analysis starts with an evaluation of the learning curves associated with random, uncertainty, and representative sampling strategies illustrated in Figure 2. These curves show how each strategy progresses in terms of accuracy and fairness as more data points are sampled. Generally, all strategies eventually converge to the baseline performance, which is expected in a typical active learning setup. However, significant differences in the speed and pattern of convergence offer insights into the efficiency and effectiveness of each strategy. From Figure 2, it is evident that representative sampling does not significantly outperform random sampling in terms of accelerating learning, particularly in accuracy. This suggests that the distribution of the Adult dataset might not have well-defined high-density regions, which can limit the effectiveness of representative sampling strategies. Conversely, uncertainty sampling shows a more pronounced impact on learning rates, particularly for unfairness metrics, indicating its potential in environments where quick learning of complex patterns is crucial.

[4] for every possible unlabeled data point to label, the unfairness reduction needs to be calculated. For only 16 samples from 20% of the adult dataset, this takes already 16 seconds

[5] We use the models from the python library Scikit-learn, https://scikit-learn.org/stable/

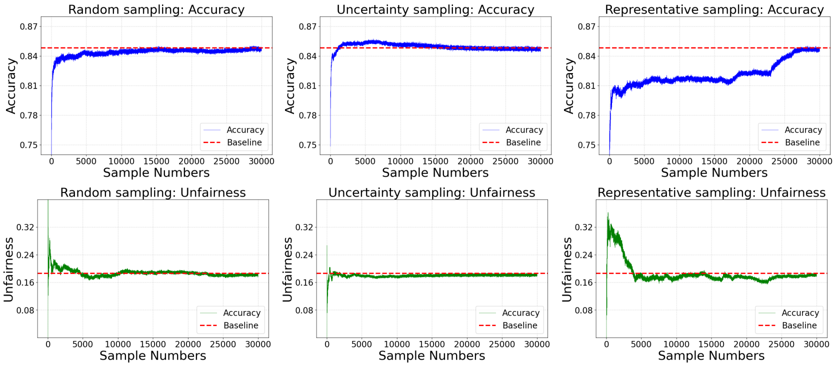

Figure 2: Learning curve accuracy (above) and fairness (below) for every General Active Learning Sampling Strategy on all the datapoints of the adult dataset (approx. 30k instances). Learning curves obtained from a single train/test split.

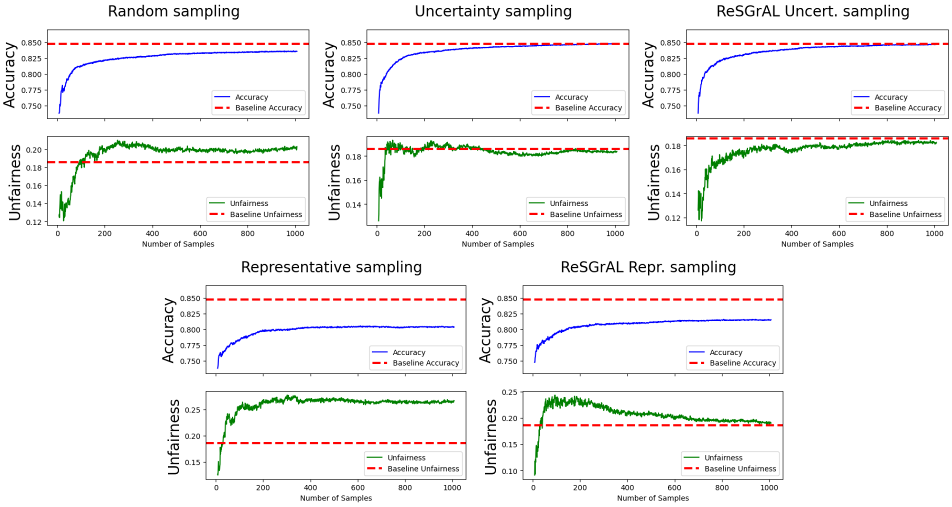

Figure 3: Mean learning curves obtained from 30 train/test splits: accuracy vs unfairness for every sampling strategy.

5.2. Interplay Between Accuracy and Unfairness

As detailed in Figure 3, there is not a consistent correlation between accuracy improvements and reductions in unfairness across sampling strategies. In fact, some strategies, like uncertainty sampling, can increase model unfairness faster than they improve accuracy. This highlights the challenge in balancing these two aspects of model performance and suggests there may be value to introduce strategies that explicitly account for fairness during the sampling process.

5.3. Subgroup Performance Variability

Diving deeper into the subgroup analyses (Figure 4), we find that unfairness does not increase uniformly across all subgroups. This variability stresses the importance of considering subgroup-specific dynamics when implementing active learning strategies, particularly in heterogeneous populations. Our results indicate that while some subgroups may benefit from reduced unfairness, others might experience an increase, especially under sampling strategies not specifically designed to mitigate unfairness.

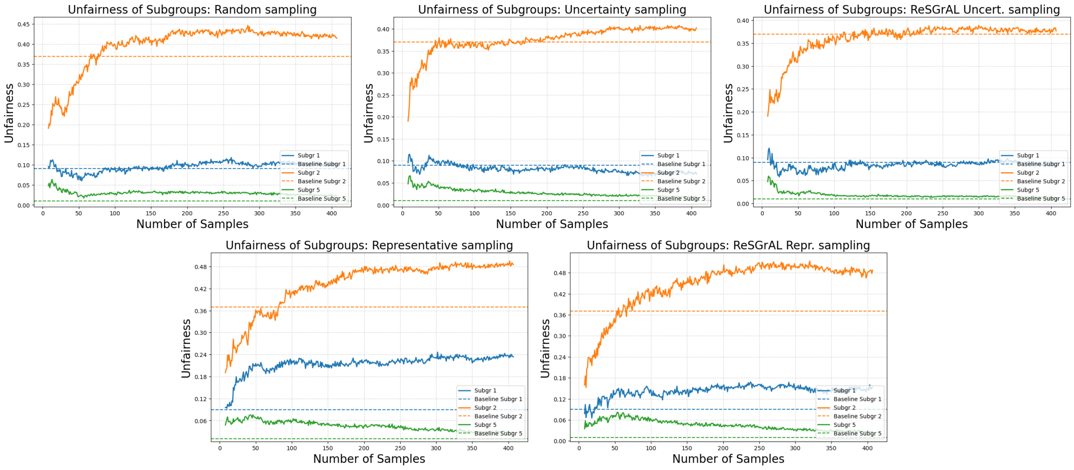

Figure 4: Mean unfairness learning curves obtained from 30 train/test splits on 3 sub-groups (best, median, lowest following unfairness baselines), with baselines for every sampling strategy.

5.4. Results of (Fair) ReSGrAL

ReSGrAL combines stratified sampling with active learning. We compare its representative and uncertainty sampling variants to their general counterparts, observing that ReSGrAL maintains accuracy while reducing unfairness more slowly (Figure 3). Similar outcomes are shown in Figure 5, where we only show uncertainty sampling since this sampling technique outperforms representative sampling on the adult dataset as seen before.

Fair ReSGrAL modifies ReSGrAL to prioritize unfairness reduction once a threshold is exceeded. It performs comparably to ReSGrAL in accuracy and unfairness (Figure 5), and both models demonstrate lower unfairness but similar accuracy when compared to uncertainty sampling. Additionally, in our experiments Fair ReSGrAL vastly outperforms FAL in computational speed, requiring only 20 seconds compared to FAL’s 2000 second.

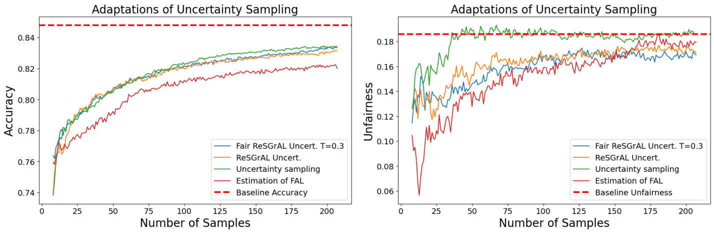

Figure 5: Mean learning curves obtained from 30 train/test splits: Accuracy (left) and unfairness (right) for uncertainty sampling, ReSGrAL uncertainty sampling, estimation of FAL and Fair ReSGrAL (threshold is set to 0.3).

Subgroup analyses show Fair ReSGrAL improves unfairness in groups above the threshold, though some variance exists across subgroups (Figure 6). This figure shows 5 of the 13 subgroups that have the highest unfairness baselines, thus are the most interesting to see regarding their learning curve of unfairness.

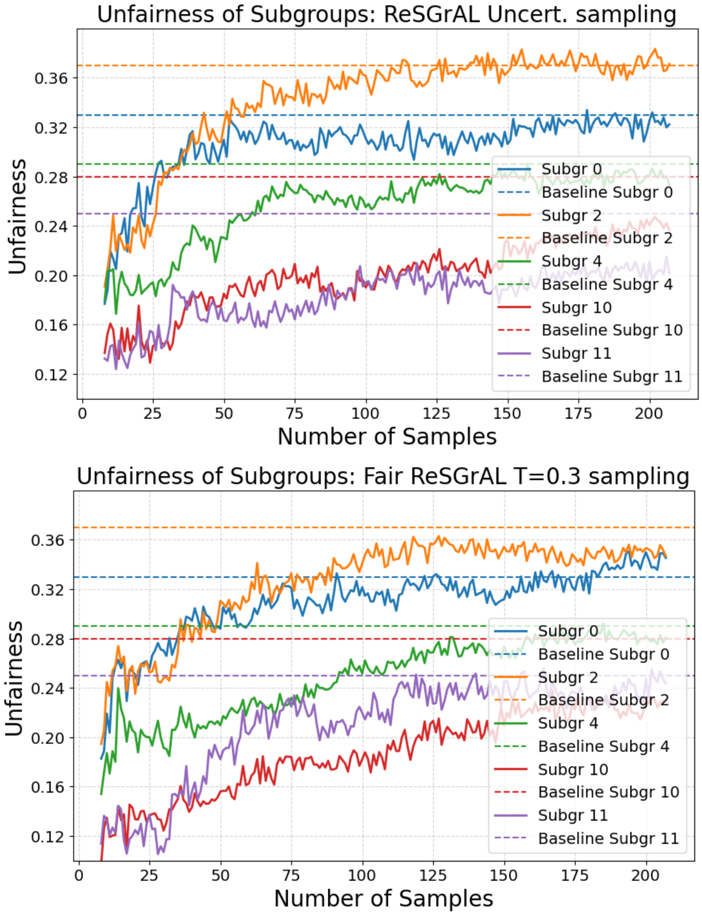

Figure 6: Mean unfairness learning curves for 30 train/test splits: 5/13 Subgroups with the highest unfairness baselines. Top: ReSGrAL, Down: Fair ReSGrAL (threshold = 0.3)

6. Perspectives

In the context of developing models for decision support in the public domain, recent literature shows that active learning proves to be a valuable solution for partially labelled datasets [11]. However, this study emphasizes the potential significant impact of active learning on the unfairness of model outcomes. It may be possible that a model's learning curve for unfairness may increase faster than its accuracy during training, and in some cases, that the model's unfairness may even surpass the inherent unfairness of the entire dataset. Public organisations need to be aware of the consequences of sampling.

One notable aspect of our study is the relatively modest scale of impact of the sampling strategies in terms of reducing unfairness (here: decreasing the demographic parity difference) and enhancing accuracy. For example, when analysing Figure 3, the differences of unfairness between sampling strategies are in the range of 10%[6]. Nonetheless, the findings demonstrate a consistent and significant trend. To fully assess the practical implications, it is essential to delve deeper into the real-world consequences of these improvements. Future research should investigate whether these effects are substantial enough to lead to meaningful differences in decision-making processes.

Fair ReSGrAL's results, on the other hand, emphasize that choosing specific instances to improve fairness in a particular subgroup of a dataset can lead to a reduction in unfairness within that subgroup, without causing any adverse impact on accuracy on the total dataset. Furthermore, it acts as an active

[6] In Figure 3, representative sampling gains an unfairness of 0.275 after 200 samples, while, after the same amount of sampling, the unfairness of ReSGrAL uncertainty is 0.175.

learning approach focused on fairness while requiring significantly less computational time than state-of-the-art-proposals as FAL. Rather than creating a new model for each conceivable label across all unlabeled data in every subgroup, it primarily requires significant computation time for the unlabeled data within groups that exhibit unfairness exceeding a predefined threshold.

However, a notable limitation of the comparison between Fair ReSGrAL and FAL [5] has indeed been caused by computational time. Because of that, we were unable to fully apply FAL on the adult dataset, which comprises over 30,000 instances. Additionally, extensive testing across all train/test splits would be essential for a thorough evaluation of FAL's performance. Yet, our results are clear in terms of computational efficiency, which also is a crucial dimension for practical applications.

Differences were noted between synthetic and adult datasets. The synthetic dataset supported our hypothesis, showing aligned accuracy and unfairness curves under active learning. However, the adult dataset, a real-world data set with complex interrelated attributes, displayed a different pattern: active learning exacerbated its inherent bias. Unlike the synthetic data, which had minimal noise and few correlations except with the sensitive attribute, the adult dataset's complexity necessitates examining specific characteristics and their impact on active learning effectiveness.

Experimentally, our work is based on simulating unlabeled datasets, by hiding the decision labels and unveiling it only mimicking the presence of an expert only once a data point is selected. In real deployment, this may cause problems, as experts may not accept to evaluate cases which are selected in an opaque fashion (to them), or with a heuristic which they deem not adequate. Additionally, not having a baseline reference, it is not evident to assess whether the convergence is going how expected. The method should be expanded in practice to introduce further guarantees.

Lastly, the context of representation bias [2] in initial datasets offers a fertile ground for future research. For example, biases in municipal algorithms where men are randomly inspected and women only on indication of fraud [7], show how non-representative sampling skews perceived behavior patterns. This real-world relevance highlights the need to study active learning's effects in similar scenarios to understand and mitigate biases from the onset.

References

[1] M. Favier, T. Calders, S. Pinxteren, J. Meyer, How to be fair? a study of label and selection bias, Machine Learning (2023) 1–24.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[3] P. Kumar, A. Gupta, Active learning query strategies for classification, regression, and clustering: a survey, Journal of Computer Science and Technology 35 (2020) 913–945.

[4] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: International conference on machine learning, PMLR, 2018, pp. 2564–2572.

[5] H. Anahideh, A. Asudeh, S. Thirumuruganathan, Fair active learning, Expert Systems with Applications 199 (2022) 116981.

[6] A. Sharaf, H. Daume III, R. Ni, Promoting fairness in learned models by learning to active learn under parity constraints, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2149–2156.

[7] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in: Equity and access in algorithms, mechanisms, and optimization, 2021, pp. 1–9.

[8] F. Kamiran, I. Žliobaitė, Explainable and Non-explainable Discrimination in Classification, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 155–170. URL: https://doi.org/10.1007/978-3-642-30487-3_8. doi:10.1007/978-3-642-30487-3_8.

[7] As occurred with an algorithm used by the municipality of Rotterdam. (See e.g. WIRED, "Inside the Suspicion Machine", https://www.wired.com/story/welfare-state-algorithms/)

[9] Y. Kong, Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 485–494.

[10] A. S. Acharya, A. Prakash, P. Saxena, A. Nigam, Sampling: Why and how of it, Indian Journal of Medical Specialties 4 (2013) 330–333.

[11] H. Hino, Active learning: Problem settings and recent developments, arXiv preprint arXiv:2012.04225 (2020).

[12] S.-J. Huang, R. Jin, Z.-H. Zhou, Active learning by querying informative and representative examples, Advances in neural information processing systems 23 (2010).