Look and You Will Find It: Fairness-Aware Data Collection through Active Learning

Hilde Weerts[1,†], Renée Theunissen[1,†] and Martijn C. Willemsen[1]

[1]*Eindhoven University of Technology, The Netherlands*

Abstract

Machine learning models are often trained on data sets subject to selection bias. In particular, selection bias can be hard to avoid in scenarios where the proportion of positives is low and labeling is expensive, such as fraud detection. However, when selection bias is related to sensitive characteristics such as gender and race, it can result in an unequal distribution of burdens across sensitive groups, where marginalized groups are misrepresented and disproportionately scrutinized. Moreover, when the predictions of existing systems affect the selection of new labels, a feedback loop can occur in which selection bias is amplified over time. In this work, we explore the effectiveness of active learning approaches to mitigate fairness-related harm caused by selection bias. Active learning approaches aim to select the most informative instances from unlabeled data. We hypothesize that this characteristic steers data collection towards underexplored areas of the feature space and away from overexplored areas – including areas affected by selection bias. Our preliminary simulation results confirm the intuition that active learning can mitigate the negative consequences of selection bias, compared to both the baseline scenario and random sampling.

Keywords

selection bias, algorithmic fairness, active learning, machine learning

1. Introduction

Machine learning models are often trained on data sets subject to *selection bias*: non-random selection of instances from the population. When selection bias is related to sensitive characteristics such as gender and race, it can result in fairness-related harm. For example, in banking, fraud detection models are typically trained on labeled transaction data, which can be affected by social biases against sensitive groups. Transactions of customers belonging to those groups may be flagged more often as suspicious, creating the appearance of a relatively high number of fraudulent transactions compared to other groups - even when the true fraud rates are similar. As aptly put by a quote popularly attributed to Sophocles: *look and you will find it - what is unsought will go undetected.* Unaddressed, the consequences can be severe: fairness-related harm linked to selection bias has been observed in various cases such as fraud detection in welfare benefit applications [1], predictive policing [2], and medical diagnosis [3].

CEUR Workshop Proceedings (CEUR-WS.org)

Given these implications, it seems imperative to understand and mitigate the harmful effects of selection bias in machine learning. In this paper, we present the results of a preliminary simulation study which explores the effectiveness of *active learning* approaches in mitigating fairness-related harm caused by selection bias. Active learning approaches aim to select the most informative instances from unlabeled data. We hypothesize that this characteristic steers data collection towards underexplored areas of the feature space and away from overexplored areas – including areas affected by selection bias. Our simulation results confirm the intuition that active learning can mitigate the negative consequences of selection bias, compared to both the baseline scenario and random sampling.

The remainder of this work is structured as follows. In Section 2, we further detail the problem of selection bias and the moral argument that motivates our work. In Section 3, we motivate our hypothesis that active learning can be helpful to mitigate harmful effects of selection bias. In Section 4, we provide a brief overview of related work. Section 5 details our experiment setup and the results are presented in Section 6. Section 7 concludes the paper.

2. The Unfairness of Selection Bias

A major assumption in machine learning is that the training data is representative of the population from which it is drawn. In practice, training data is often not sampled at random, skewing the distribution of the training data set [4]. In particular, when positives are rare, (manual) labeling of instances is expensive, and resources are limited, the selection of instances to be labeled is typically informed by domain expertise or the output of existing statistical models. In other words, non-random selection is leveraged to achieve reasonable predictive performance with a small number of labeled instances. However, selection bias in historical decision-making policies can be affected by social biases and structural injustice, resulting in an unequal distribution of labeled data across sensitive groups [e.g., 1, 3].

Both over- and underrepresentation can result in fairness-related harm. In the context of clinical prediction models, the underrepresentation of demographic groups in clinical data sets can lead to underdiagnosis. For example, it is well-known that women have been historically underrepresented in clinical trials. As a result, there exists limited knowledge regarding adverse effects, benefits, and risks of treatments for women [3], which negatively affects healthcare outcomes. In the context of fraud detection, selection bias can result in an unequal distribution of burdens, where marginalized groups are overrepresented and disproportionately scrutinized. For example, in 2021, it was brought to light that the Dutch Tax and Customs Administration had unlawfully processed Dutch citizenship in fraud investigations of childcare benefit applications and even explicitly included Dutch citizenship as a risk factor in the risk assessment model [1, 5]. As a result, applicants who had a nationality other than Dutch were overly scrutinized through manual processing of the application and many of them were wrongfully accused of fraud.

The adverse consequences of selection bias are further amplified when the predictions of existing systems affect the selection of new labels, resulting in a feedback loop. For example, when a fraud detection model is trained on a data set in which a sensitive group is overrepresented, the model is more likely to generate alerts of potentially fraudulent activities for members of

this group, resulting in higher testing and a relatively larger number of positives. When the model is retrained on the newly available data, selection bias is reinforced. This type of feedback loop has been previously characterized in different domains [e.g., 6], most prominently in the context of criminal justice [2, 7, 8, 9, 10].

If the goal is to achieve reasonable predictive performance, some form of selection bias seems unavoidable under limited resources.[1] However, from a moral perspective, a disproportionate distribution of burdens or benefits caused by the misrepresentation of sensitive groups in the training data violates basic principles of equality [11]. Moreover, as shown by the case of the Dutch childcare benefit scandal, a failure to mitigate this form of selection bias may not survive legal scrutiny under EU law.[2]

In conclusion, we argue that practitioners have a responsibility to mitigate the harmful effects of selection bias related to sensitive group membership.

3. Fairness-Aware Data Collection through Active Learning

Active learning is a form of semi-supervised machine learning, where the objective is to achieve high predictive performance with a small number of labeled samples [13]. To this end, active learning methods iteratively query an oracle (e.g., a human domain expert) to label specific data points that are expected to improve the predictive performance of the machine learning model. Once the labels are obtained, the model is retrained on a data set including the newly labeled data. This process is repeated several times. Active learning is typically used when it is too costly or time-consuming to label all data instances manually.

Active learning approaches can generally be divided into two main categories. *Pool-based* approaches select one instance (*screening-based*) or multiple instances (*batch-mode*) from a pool of unlabeled data instances. In streaming settings, *streaming-based* approaches determine whether a new instance is labeled or discarded on a one-by-one basis as new data arrives. In this setting, unlabeled instances are not revisited. In this work, we focus primarily on pool-based approaches, as this setting best matches the typical scenarios where selection bias can occur, such as fraud detection.

An important component of pool-based active learning approaches is the sampling approach that determines which sample of data instances will be selected for labeling by the oracle. One of the most common sampling approaches is *uncertainty sampling*, in which data instances are ranked based on how uncertain the algorithm is regarding the true label of the data instance, typically informed by the confidence score of the machine learning model [14].

Returning to the problem of selection bias, we hypothesize that active learning could potentially mitigate fairness-related harm caused by selection bias related to sensitive group membership. In particular, uncertainty sampling is expected to steer data collection towards underexplored areas of the feature space and away from overexplored areas – including areas

---

[1]Of course, the mere notion of limited resources and the perceived value of predictive performance already embed important value judgments. In particular, a straightforward alternative mitigation strategy would be to increase the amount of available resources such that random selection is feasible.

[2]We would like to emphasize that EU law is highly contextual and judicial decisions related to the Dutch childcare benefits scandal cannot be readily applied to all cases of unfairness caused by selection bias. We refer to Weerts et al. [12] for a more elaborate overview of EU non-discrimination law in the context of algorithmic fairness.

affected by selection bias. For example, consider a scenario where some group *A* is historically overrepresented in a data set compared to group *B*. A model trained on this data set is likely to produce more uncertain confidence scores for instances in group *B* compared to instances that belong to group *A*. As a result, the active learning algorithm is more likely to query instances in group *B*, counteracting selection bias.

Previous research points towards the potential effectiveness of active learning in the context of unfairness caused by selection bias. While Branchaud-Charron et al. [15] do not explicitly study the problems of selection bias and feedback loops, the authors do find that active learning approaches can effectively improve fairness measured by several fairness metrics compared to random sampling. Additionally, Richards et al. [16] do not explicitly consider fairness, but do show that active learning is more effective at mitigating selection bias unrelated to sensitive group membership than other techniques such as importance weighting. While neither of these works fit our setting exactly, the results are promising.

Active learning could have several advantages compared to existing fairness-aware machine learning (fair-ml) approaches. First of all, selection bias is addressed directly at its source: during data collection. As such, we expect the approach to be more effective compared to technical interventions that address harmful consequences through fairness constraints during training. Second, if active learning is shown to consistently and effectively mitigate selection bias, this implies that unfairness can be mitigated *without having access to sensitive features* – an important concern in many efforts towards fair outcomes. However, we would like to emphasize that even if active learning is effective, it will not be a panacea. Selection bias is primarily a concern in scenarios where social bias and structural injustice are prevalent. In such contexts, selection bias is unlikely to be the *only* source of downstream unfairness and additional interventions are necessary to ensure equitable outcomes.

4. Related Work

Researchers have developed a plethora of fair-ml approaches aimed at mitigating unfairness, ranging from pre-processing the data that obscure undesirable associations [e.g., 17], enforcing fairness constraints during model training [e.g., 18], and post-processing (predictions of) existing models [e.g., 19]. A common denominator of these approaches is that fairness is formulated as an optimization task, where the objective is to achieve high predictive performance under some quantitative fairness constraint, such as a maximum difference in error rates across sensitive groups.

Our work is related to sampling-based pre-processing techniques [e.g., 20], which use specific sampling schemes to satisfy particular fairness constraints. However, different from our work, these approaches typically assume a fully labeled training data set. More closely related to our work are fair-ml approaches that adapt or complement existing active learning approaches. For example, the Fairness-aware Active Learning (FAL) framework [21] aims to create a balanced data set by selecting new data instances based on total accuracy and improvement of demographic parity and equalized odds. Similarly, Parity-Constrained Meta Active Learning (PANDA) [22] uses meta-learning to learn a selection policy that optimizes accuracy and fairness constraints, outperforming random sampling, uncertainty sampling, and FAL in terms of

accuracy and two commonly used fairness metrics.

All of the above-mentioned fair-ml approaches are proposed as generic tools to mitigate unfairness via quantitative fairness constraints and most of them do not make explicit assumptions about the nature of the bias that lies at the root of unfairness. We argue that formulating fairness as a black-box optimization task has several limitations. While fairness metrics can be useful indicators of potential fairness-related harm, they often fail to capture more nuanced notions of equality, rendering them poor optimization constraints [11]. Moreover, fair-ml approaches attempt to address fairness primarily during the modeling stage of the development process, which often fails to meaningfully address the biases and design choices at the root of fairness-related harm [23]. In contrast to traditional fair-ml approaches, our work thus falls in a line of recent work [e.g., 8, 9, 24] that shifts focus towards understanding and mitigating specific types of biases directly.

5. Simulation Setup

We use a simulation study to explore the potential effectiveness of active learning for mitigating the harmful effects of selection bias related to sensitive group membership. The source code of the simulation is available on Github.[3]

Data Set and Pre-processing The data set that is used for this research is a simulated fraud data set created using Sparkov Data Generation [25] and was published under the CC0 1.0 public domain license [26]. The simulated data set consists of 1.8 million transactions of which roughly 9000 are fraudulent, resulting in a fraud rate of 0.005. The data set contains several features, including demographics (e.g., gender, date of birth) and characteristics of the transaction (e.g., transaction number, amount, time, category).

To make it easier to visualize and observe the harmful effects of selection bias, non-fraudulent transactions are under-sampled such that the data set has a ground-truth fraud rate of 0.1 for both males and females. Further pre-processing consists of dropping several features (e.g., identifiers such as `trans_num` and the client's `first` and `last` name), merging similar categories (e.g., `grocery_net` and `grocery_pos` into one category `grocery`), one-hot-encoding of categorical features (e.g., the transaction `category`), and transforming `dob` to a new feature `age`.

Simulation Design We simulate a scenario in which a machine learning model is trained on a data set affected by selection bias. The selection bias in the initial training data is simulated by sampling different levels of observed fraud of two sensitive groups, *males* and *females*.

Subsequently, the model is retrained daily, based on all the labeled data that has been collected up until that point in time. Newly labeled data originates from two sources: (1) transactions that are labeled *organically* through alerts of the fraud detection model (i.e., these are the transactions for which the model outputs the highest confidence scores), (2) transactions labeled through *explorative* sampling (e.g., via active learning).

Note that this setup implies several important assumptions. First of all, it is assumed that labels are not noisy: fraud analysts are always able to accurately determine the ground-truth

---

[3]https://github.com/reneetheunissen/fraud_detection

labels. In practice, this assumption may not hold, especially when human annotators use different levels of scrutiny for different types of alerts. Furthermore, it is assumed that each label has equal annotation costs, resulting in a fixed amount of alerts that can be labeled each day. Additionally, we assume no external sources of labels, such as notifications from customers. That is, we assume that all observed fraudulent transactions are discovered through either fraud detection system alerts or explorative sampling. Finally, we assume that the data set is not subject to concept drift.

Simulation Parameters   The simulation has four main parameters that are varied across simulation scenarios: the *observed fraud rates* for males and females, the alert rate, the exploration rate, and the machine learning model class.

- The **observed fraud rate** indicates the proportion of all transactions that are labeled as fraudulent for a particular subgroup. The initial *observed fraud rates* for males and females differs across simulation scenarios.
- The **alert rate** is defined as the proportion of incoming daily transactions that will be labeled. In other words, this parameter determines how many transactions are added to the training data set each day.
- The **exploratory rate** is the proportion of alerts that are labeled through exploratory sampling. All other alerts are labeled organically through alerts of the fraud detection model. The *exploratory rate* represents the balance between exploration of unlabeled data and identifying fraudulent transactions.
- We train two types of **machine learning models** of different levels of complexity: logistic regression models and random forest classifiers. We leverage the implementations in scikit-learn [27]. All models are trained using the default parameters in scikit-learn 1.2.

Explorative Sampling Approaches   We evaluate the following three explorative sampling approaches.

- **Biased data.** *The model is trained based solely on instances labeled organically through alerts.* This approach serves as the most extreme baseline where no exploration occurs at all. This scenario is unlikely to occur in practice, as fraudulent transactions are typically also discovered via external data sources such as notifications from customers or (manual) investigations by fraud analysts.
- **Random sampling.** *The model is trained based on both alerts and explorative sampling via random sampling.* This approach serves as a baseline where some exploration is done, but not via active learning. We hypothesize that while exploration will be able to mitigate harmful effects of selection bias compared to the *biased data* approach, the approach may not identify many fraudulent transactions due to the highly imbalanced nature of the data set.
- **Uncertainty sampling.** *The model is trained based on both alerts and explorative sampling via uncertainty sampling.* We hypothesize that active learning via uncertainty sampling performs best, as it steers exploration towards underexplored areas of the feature space and away from overexplored areas – including areas affected by selection bias.

Evaluation Metrics   Selection bias can be viewed as a form of *measurement bias* in the target variable: *observed* base rates are an imperfect proxy of *true* base rate [28]. Consequently, metrics computed over the observed labels can be a poor measure of the true characteristics of the machine learning model. In practice, we only have access to observed rates. In our simulation, however, we can compute metrics over the ground-truth target variable, allowing us to investigate the true effects of selection bias.

In particular, we present the following statistics. The *true fraud rate* (TFR) indicates the ground-truth proportion of positives. The *observed fraud rate* (OFR) indicates the proportion of predicted positives out of all instances until that point in time. The *rate of predicted positives* (RPP) shows the proportion of predicted positives. Additionally, we compute the following predictive performance metrics: *false positive rate* (FPR), *false negative rate* (FNR), *false discovery rate* (FDR), *false omission rate* (FOR), classification *accuracy* (ACC).

Experiments   The initial training data set contains 6500 instances in all scenarios, which is created through random sampling from fraudulent and non-fraudulent transactions for males and females according to the desired *observed fraud rate*. Each day, 3340 new transactions arrive. In each scenario, the simulation runs for 25 days. We perform two experiments.
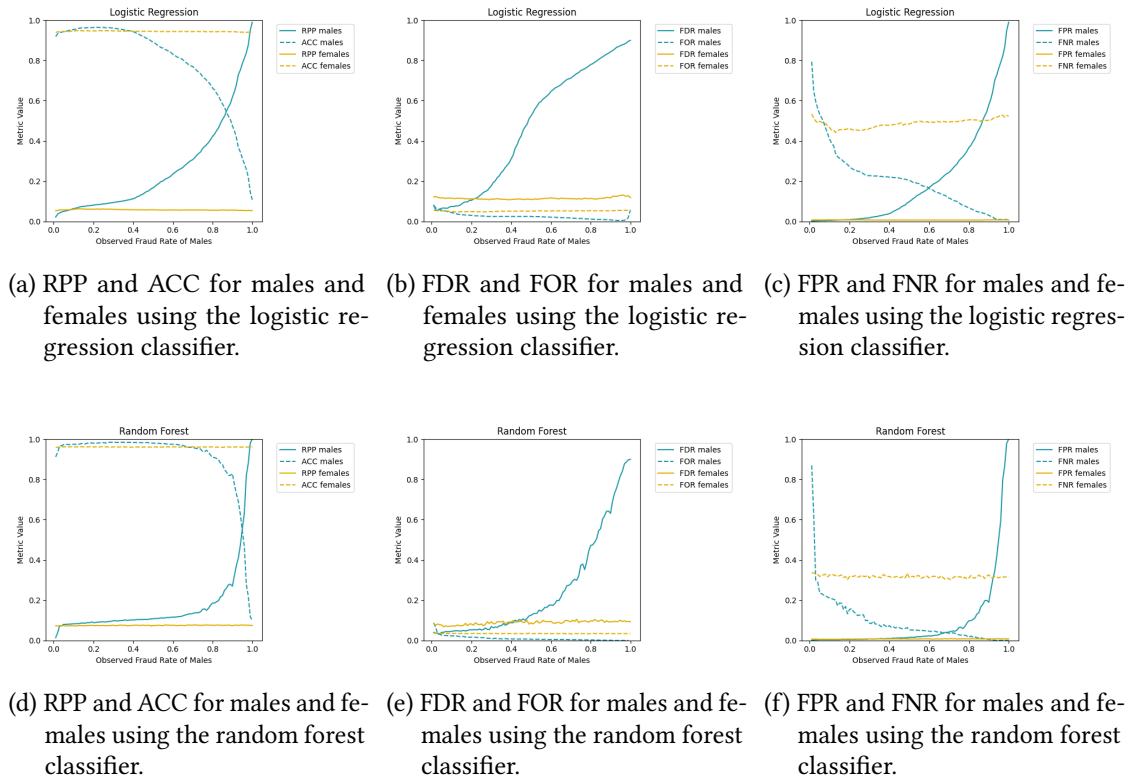
1. **The Effect of Selection Bias.** In this set of simulations, we explore the effects of selection bias in the *biased data* baseline scenario.

   - *Observed fraud rates.* To study the effects of the magnitude of the initial selection bias, we perform a set of simulations where the initial *observed fraud rate* for females is constant (0.05) and the *observed fraud rate* of males in the initial training data is varied between the range of 0 and 1.
   - *Alert rates.* We vary the *alert rate* to investigate the effect of the magnitude of selection bias over time. Across simulations, the alert rate is varied between the values 0.01, 0.05, and 0.10.

2. **The Effect of Exploration.** In this experiment, we study to what extent exploration through *random sampling* and *uncertainty sampling* can mitigate harmful effects of selection bias. Across simulations, the *exploratory rate* is varied between the values 0.10, 0.25, and 0.50. In this experiment, the *alert rate* is set to 0.05. *observed fraud rate* is set to 0.3 for the overrepresented group and 0.05 for the underrepresented group.

6. Results

6.1. Experiment 1: The Effect of Selection Bias

High Observed Fraud Rates Lead to False Alerts and Overscrutinization of the Overrepresented Group   Our first simulation shows the effects of selection bias in the initial training data (Figure 1). We observe that a high *observed fraud rate* leads to false alerts and overscrutinization of the overrepresented group, while a low *observed fraud rates* leads to undetected fraud. As expected, a high *observed fraud rate* results in an increase of the FPR, FDR, and RPP, which leads to more transactions incorrectly being predicted as fraudulent, while a
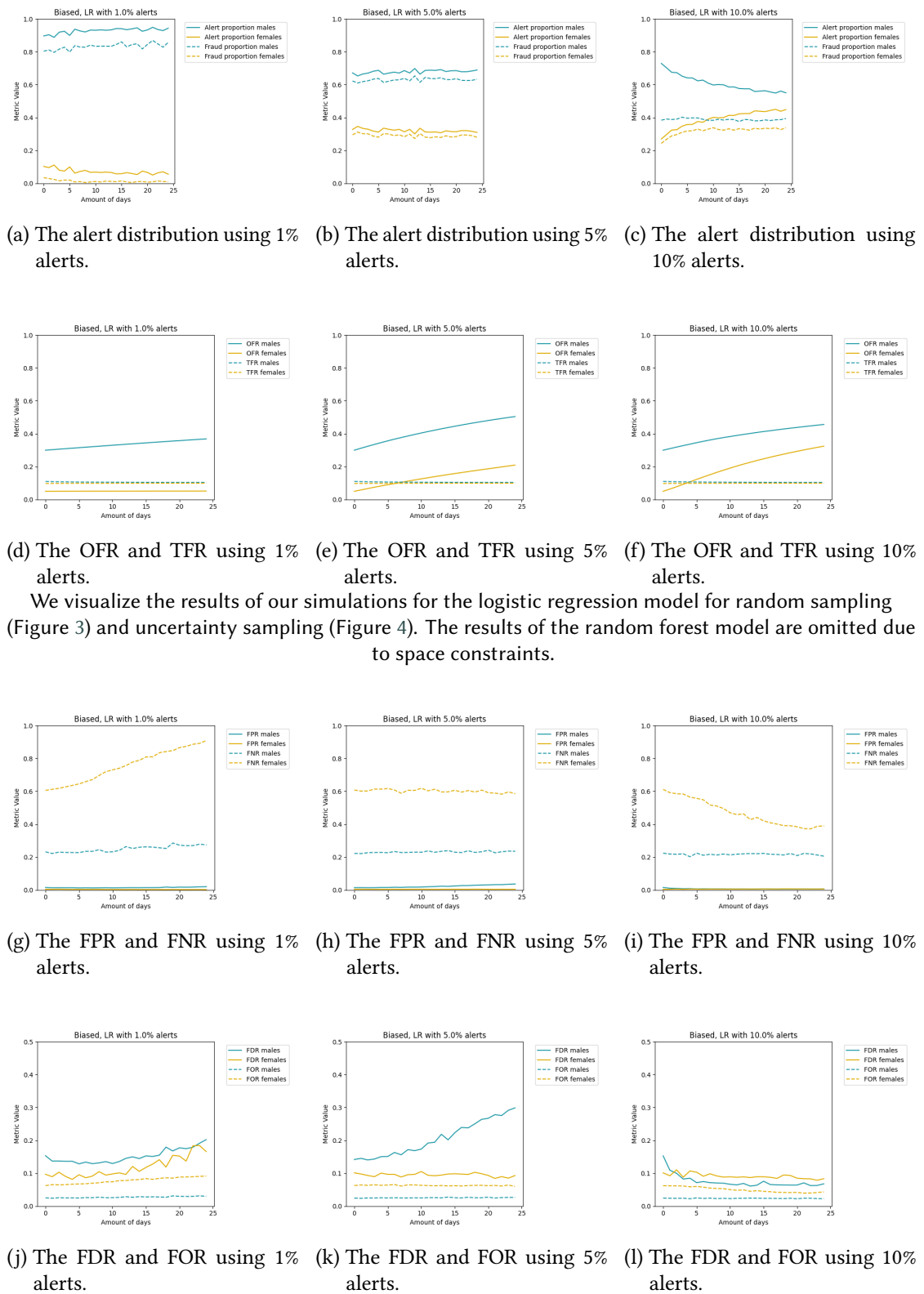
low *observed fraud rate* results an RPP that is lower than the true proportion of positives (0.10), which leads to many undetected fraudulent cases.

(a) RPP and ACC for males and females using the logistic regression classifier.

(b) FDR and FOR for males and females using the logistic regression classifier.

(c) FPR and FNR for males and females using the logistic regression classifier.

(d) RPP and ACC for males and females using the random forest classifier.

(e) FDR and FOR for males and females using the random forest classifier.

(f) FPR and FNR for males and females using the random forest classifier.

Figure 1: The impact of the *observed fraud rate* of males in the initial training data on different metrics as measured after one prediction.

Increasing the Number of Labeled Instances Reduces Inequalities Figure 2 shows the effect of different *alert rates* over time. Low *alert rates* increases the inequality between the over and underrepresented groups. Moreover, we observe that these effects are amplified over time. In particular, low *alert rates* do not offer a great variety of alerts, resulting in further overscrutinization of the overrepresented group. The divergence of the alert distribution is the driving factor behind the development of other metrics (OFR, TFR, FPR, FNR, FDR, and FOR). When the alert rate is high, harmful effects of selection bias are dampened. The wider variety of alerts allows for organic correction of the high initial *observed fraud rates* as well as other metrics.

(a) The alert distribution using 1% alerts.

(b) The alert distribution using 5% alerts.

(c) The alert distribution using 10% alerts.

(d) The OFR and TFR using 1% alerts.

(e) The OFR and TFR using 5% alerts.

(f) The OFR and TFR using 10% alerts.

We visualize the results of our simulations for the logistic regression model for random sampling (Figure 3) and uncertainty sampling (Figure 4). The results of the random forest model are omitted due to space constraints.

(g) The FPR and FNR using 1% alerts.

(h) The FPR and FNR using 5% alerts.

(i) The FPR and FNR using 10% alerts.

(j) The FDR and FOR using 1% alerts.

(k) The FDR and FOR using 5% alerts.

(l) The FDR and FOR using 10% alerts.

Figure 2: The evolution of the metrics in the *biased data* scenario over time, as *alert rates* are varied between 0.01, 0.05, and 0.10.

6.2. Experiment 2: The Effect of Exploration

Uncertainty Sampling Outperforms Random Sampling Figures 3 and Figure 4 show the results of our simulations for random sampling and uncertainty sampling, respectively. The results of the random forest model are omitted due to space constraints. While a low *exploratory rate* of random sampling improves the distribution of alerts between the over and underrepresented group, the improvement is unable to improve predictive performance metrics. Indeed, without improving the *observed fraud rate*, selection bias persists and the RPP will continue to decrease while the FNR will continue to increase.

Uncertainty sampling, on the other hand, is able to counter negative effects of selection bias and decrease disparities in predictive performance between groups. Uncertainty sampling leads to a better alert distribution with balanced alerts (both fraudulent and non-fraudulent) of both the over- and underrepresented group. In this way, the *observed fraud rate* is corrected over time through the additional exploratory data.

As can be expected, higher exploratory rates (up to 0.50) result in higher levels of mitigation of harmful effects: a better alert distribution, decreased disparities in the FPR, FNR, FDR, and FOR, and increase in accuracy.
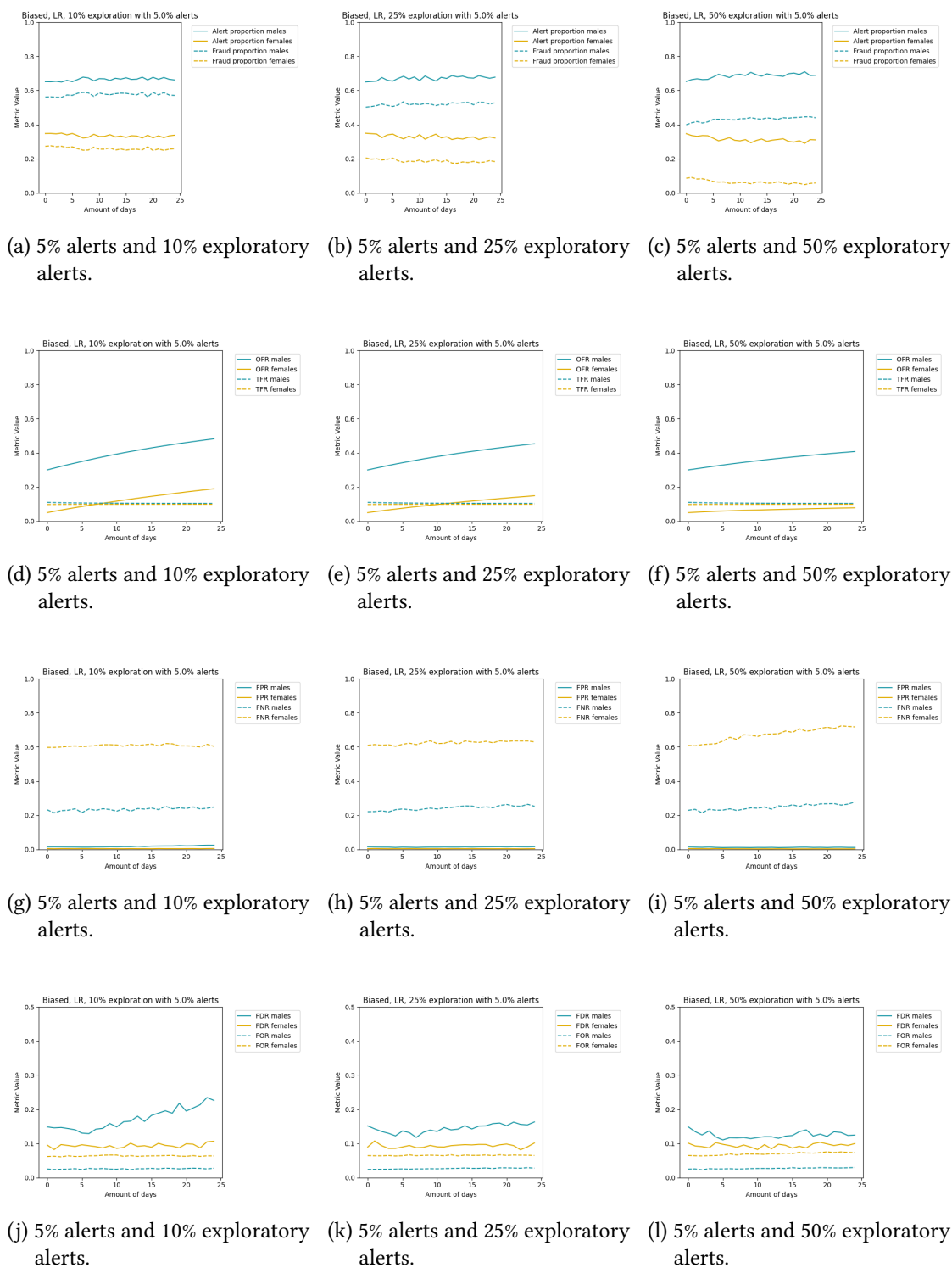
7. Conclusion

In this work, we have studied the problem of selection bias related to sensitive features and analyzed the effectiveness of active learning to mitigate the ensuing harmful effects.
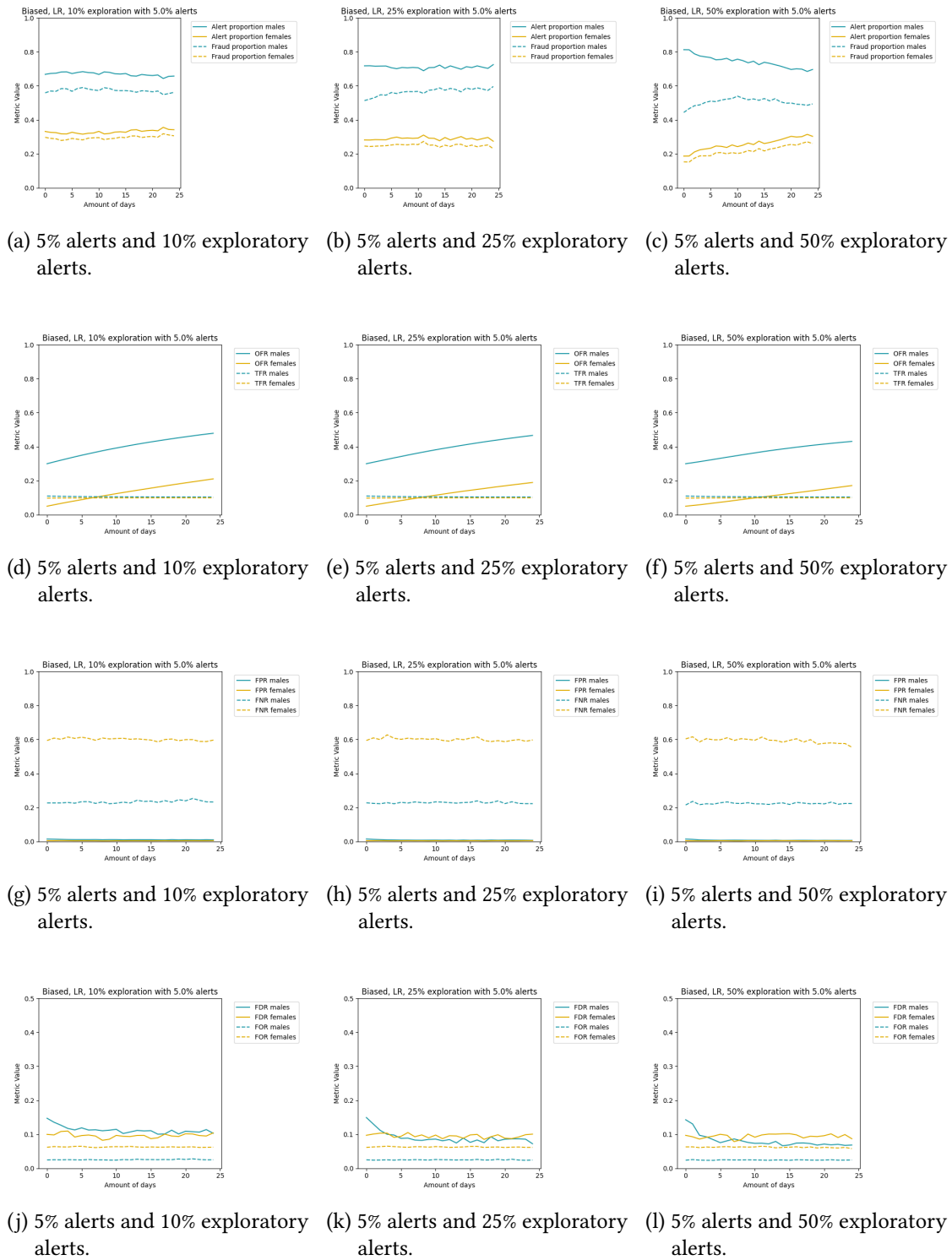
The results of our simulations confirm that in the absence of interventions, selection bias is reinforced over time, resulting in an increase in scrutinization as well as an increase in the number of false positives for overrepresented groups. Moreover, our experiments show preliminary evidence that uncertainty sampling can mitigate disparities between groups caused by selection bias and, in the studied setting, outperforms exploration via random sampling.

Our results also imply that limited resources are an important bottleneck in the organic correction of disparities in *observed fraud rates*, as increasing the number of labeled instances reduces inequalities between sensitive groups even in absence of interventions. These results suggest that the problem of mitigating selection bias can be seen as analogous to the well-known exploration-exploitation trade-off. That is, under limited resources, decision-makers are required to balance the mitigation of selection bias through exploration and the exploitation of existing models for identifying positives in production.

The preliminary results presented in this paper open up many directions for future work. First of all, we see several ways in which our simulation study can be extended, including the replication of our experiments on more (real-world) data sets, uncertainty quantification via repeated experiments, and relaxation of the simplifying assumptions in our simulation, such as the lack of noisy labels, equal annotation cost, absence of external sources of labels, and a lack of concept drift. In particular, the effects of using online or adaptive learners could be further explored. Additionally, future work could focus on evaluating alternative active learning sampling techniques. In particular, an important limitation of uncertainty sampling is that it relies solely on the model's confidence score as a proxy for uncertainty. This could result in repeated selection of instance types that are inherently difficult to predict based on the available

(a) 5% alerts and 10% exploratory alerts.

(b) 5% alerts and 25% exploratory alerts.

(c) 5% alerts and 50% exploratory alerts.

(d) 5% alerts and 10% exploratory alerts.

(e) 5% alerts and 25% exploratory alerts.

(f) 5% alerts and 50% exploratory alerts.

(g) 5% alerts and 10% exploratory alerts.

(h) 5% alerts and 25% exploratory alerts.

(i) 5% alerts and 50% exploratory alerts.

(j) 5% alerts and 10% exploratory alerts.

(k) 5% alerts and 25% exploratory alerts.

(l) 5% alerts and 50% exploratory alerts.

Figure 3: Plots showing the evolution of the alert distribution for the biased data using logistic regression and random sampling as baseline mitigation technique.

(a) 5% alerts and 10% exploratory alerts.

(b) 5% alerts and 25% exploratory alerts.

(c) 5% alerts and 50% exploratory alerts.

(d) 5% alerts and 10% exploratory alerts.

(e) 5% alerts and 25% exploratory alerts.

(f) 5% alerts and 50% exploratory alerts.

(g) 5% alerts and 10% exploratory alerts.

(h) 5% alerts and 25% exploratory alerts.

(i) 5% alerts and 50% exploratory alerts.

(j) 5% alerts and 10% exploratory alerts.

(k) 5% alerts and 25% exploratory alerts.

(l) 5% alerts and 50% exploratory alerts.

Figure 4: Plots showing the evolution of the metrics for the biased data using logistic regression and uncertainty sampling as active learning technique.

features, even when an additional instance is unlikely to improve the predictive performance of the model. Finally, we envision future work that tackles the development of active learning techniques that are specifically designed to tackle selection bias related to sensitive group membership. In particular, more research is needed to identify a suitable trade-off between exploration through active learning and exploitation of organically generated alerts across scenarios.

References

[1] Nederlandse Autoriteit Persoonsgegevens, Onderzoeksrapport Belastingdienst/Toeslagen - De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag, 2021. URL: https://www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf.

[2] K. Lum, W. Isaac, To predict and serve?, Significance 13 (2016) 14–19. doi:https://doi.org/10.1111/j.1740-9713.2016.00960.x.

[3] V. Daitch, A. Turjeman, I. Poran, N. Tau, I. Ayalon-Dangur, J. Nashashibi, D. Yahav, M. Paul, L. Leibovici, Underrepresentation of women in randomized controlled trials: a systematic review and meta-analysis, Trials 23 (2022). doi:10.1186/s13063-022-07004-2.

[4] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, G. Bontempi, Credit card fraud detection: A realistic modeling and a novel learning strategy, IEEE Transactions on Neural Networks and Learning Systems 29 (2018) 3784–3797. doi:10.1109/TNNLS.2017.2736643.

[5] Nederlandse Autoriteit Persoonsgegevens, Besluit tot boeteoplegging, 2021. URL: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/boetebesluit_belastingdienst.pdf.

[6] A. Sinha, D. F. Gleich, K. Ramani, Deconvolving feedback loops in recommender systems, Advances in neural information processing systems 29 (2016).

[7] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, S. Venkatasubramanian, Runaway feedback loops in predictive policing, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 160–171.

[8] N. Kallus, A. Zhou, Residual unfairness in fair machine learning from prejudiced data, in: International Conference on Machine Learning, PMLR, 2018, pp. 2439–2448.

[9] R. Fogliato, A. Chouldechova, M. G'Sell, Fairness evaluation in presence of biased noisy labels, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2325–2336.

[10] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, D. Tzovaras, Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2302–2314. doi:10.1145/3531146.3534644.

[11] H. Weerts, L. Royakkers, M. Pechenizkiy, Does the end justify the means? On the moral justification of fairness-aware machine learning, arXiv preprint arXiv:2202.08536 (2022).

[12] H. Weerts, R. Xenidis, F. Tarissan, H. P. Olsen, M. Pechenizkiy, Algorithmic unfairness

through the lens of EU non-discrimination law, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2023. doi:10.1145/3593013.3594044.

[13] B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[14] D. D. Lewis, J. Catlett, Heterogeneous uncertainty sampling for supervised learning, in: Machine learning proceedings 1994, Elsevier, 1994, pp. 148–156.

[15] F. Branchaud-Charron, P. Atighehchian, P. Rodríguez, G. Abuhamad, A. Lacoste, Can active learning preemptively mitigate fairness issues?, arXiv preprint arXiv:2104.06879 (2021).

[16] J. Richards, D. Starr, H. Brink, A. Miller, J. Bloom, N. Butler, J. James, J. Long, a. Rice, Active learning to overcome sample selection bias: Application to photometric variable star classification, The Astrophysical Journal 744 (2011) 192. doi:10.1088/0004-637X/744/2/192.

[17] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2011) 1–33. doi:10.1007/s10115-011-0463-8.

[18] M. B. Zafar, I. Valera, M. G. Rogriguez, K. P. Gummadi, Fairness Constraints: Mechanisms for Fair Classification, in: A. Singh, J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 962–970.

[19] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016) 3315–3323.

[20] F. Kamiran, T. Calders, Classification with no discrimination by preferential sampling, in: Proc. 19th Machine Learning Conf. Belgium and The Netherlands, volume 1, Citeseer, 2010.

[21] H. Anahideh, A. Asudeh, S. Thirumuruganathan, Fair active learning, Expert Systems with Applications 199 (2022) 116981. doi:https://doi.org/10.1016/j.eswa.2022.116981.

[22] 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, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2149–2156. URL: https://doi.org/10.1145/3531146.3534632. doi:10.1145/3531146.3534632.

[23] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 59–68. doi:10.1145/3287560.3287598.

[24] L. Guerdan, A. Coston, K. Holstein, Z. S. Wu, Counterfactual prediction under outcome measurement error, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1584–1598.

[25] H. Brandon, Sparkov data generation, 2022. URL: https://github.com/namebrandon/Sparkov_Data_Generation.

[26] K. Shenoy, Credit card transactions fraud detection dataset, 2020. URL: https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine

Learning Research 12 (2011) 2825–2830.

[28] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 375–385. doi:10.1145/3442188.3445901.