Université de Montréal

Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics

par

Charan Reddy

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Discipline

July 04, 2022

© Charan Reddy, Année de la thèse

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics

présenté par

Charan Reddy

a été évalué par un jury composé des personnes suivantes :

Irina Rish
(président-rapporteur)

Sarath Chandar Anbil Parthipan
(directeur de recherche)

Golnoosh Farnadi
(membre du jury)

Résumé

Le succès des modèles d'apprentissage en profondeur et leur adoption rapide dans de nombreux domaines d'application ont soulevé d'importantes questions sur l'équité de ces modèles lorsqu'ils sont déployés dans le monde réel. Des études récentes ont mis en évidence les biais encodés par les algorithmes d'apprentissage des représentations et ont remis en cause la fiabilité de telles approches pour prendre des décisions. En conséquence, il existe un intérêt croissant pour la compréhension des sources de biais dans l'apprentissage des algorithmes et le développement de stratégies d'atténuation des biais. L'objectif des algorithmes d'atténuation des biais est d'atténuer l'influence des caractéristiques des données sensibles sur les décisions d'éligibilité prises. Les caractéristiques sensibles sont des caractéristiques privées et protégées d'un ensemble de données telles que le sexe ou la race, qui ne devraient pas affecter les décisions de sortie d'éligibilité, c'est-à-dire les critères qui rendent un individu qualifié ou non qualifié pour une tâche donnée, comme l'octroi de prêts ou l'embauche. Les modèles d'atténuation des biais visent à prendre des décisions d'éligibilité sur des échantillons d'ensembles de données sans biais envers les attributs sensibles des données d'entrée. La difficulté des tâches d'atténuation des biais est souvent déterminée par la distribution de l'ensemble de données, qui à son tour est fonction du déséquilibre potentiel de l'étiquette et des caractéristiques, de la corrélation des caractéristiques potentiellement sensibles avec d'autres caractéristiques des données, du décalage de la distribution de l'apprentissage vers le phase de développement, etc. Sans l'évaluation des modèles d'atténuation des biais dans diverses configurations difficiles, leurs mérites restent incertains. Par conséquent, une analyse systématique qui comparerait différentes approches d'atténuation des biais sous la perspective de différentes mesures d'équité pour assurer la réplication des résultats conclus est nécessaire. À cette fin, nous proposons un cadre unifié pour comparer les approches d'atténuation des biais. Nous évaluons différentes méthodes d'équité formées avec des réseaux de neurones profonds sur un ensemble de données synthétiques commun et un ensemble de données du monde réel pour obtenir de meilleures informations sur le fonctionnement de ces méthodes. En particulier, nous formons environ 3000 modèles différents dans diverses configurations, y compris des configurations de données déséquilibrées et corrélées, pour vérifier les limites des modèles actuels et mieux comprendre dans quelles configurations ils sont sujets à des défaillances. Nos résultats montrent que le biais des modèles augmente à mesure que les ensembles de données deviennent plus déséquilibrés

ou que les attributs des ensembles de données deviennent plus corrélés, le niveau de dominance des caractéristiques des ensembles de données sensibles corrélées a un impact sur le biais, et les informations sensibles restent dans la représentation latente même lorsque des algorithmes d'atténuation des biais sont appliqués. Résumant nos contributions - nous présentons un ensemble de données, proposons diverses configurations d'évaluation difficiles et évaluons rigoureusement les récents algorithmes prometteurs d'atténuation des biais dans un cadre commun et publions publiquement cette référence, en espérant que la communauté des chercheurs le considérerait comme un point d'entrée commun pour un apprentissage en profondeur équitable.

Mots clés: Équité dans l'apprentissage automatique, Atténuation des biais, Apprentissage des représentations, Évaluation du modèle d'équité, Apprentissage en profondeur équitable, Équité contradictoire

Abstract

The rapid use and success of deep learning models in various application domains have raised significant challenges about the fairness of these models when used in the real world. Recent research has shown the biases incorporated within representation learning algorithms, raising doubts about the dependability of such decision-making systems. As a result, there is a growing interest in identifying the sources of bias in learning algorithms and developing bias-mitigation techniques. The bias-mitigation algorithms aim to reduce the impact of sensitive data aspects on eligibility choices. Sensitive features are private and protected features of a dataset, such as gender of the person or race, that should not influence output eligibility decisions, i.e., the criteria that determine whether or not an individual is qualified for a particular activity, such as lending or hiring. Bias mitigation models are designed to make eligibility choices on dataset samples without bias toward sensitive input data properties. The dataset distribution, which is a function of the potential label and feature imbalance, the correlation of potentially sensitive features with other features in the data, the distribution shift from training to the development phase, and other factors, determines the difficulty of bias-mitigation tasks. Without evaluating bias-mitigation models in various challenging setups, the merits of deep learning approaches to these tasks remain unclear. As a result, a systematic analysis is required to compare different bias-mitigation procedures using various fairness criteria to ensure that the final results are replicated. In order to do so, this thesis offers a single paradigm for comparing bias-mitigation methods. To better understand how these methods work, we compare alternative fairness algorithms trained with deep neural networks on a common synthetic dataset and a real-world dataset. We train around 3000 distinct models in various setups, including imbalanced and correlated data configurations, to validate the present models' limits and better understand which setups are prone to failure. Our findings show that as datasets become more imbalanced or dataset attributes become more correlated, model bias increases, the dominance of correlated sensitive dataset features influence bias, and sensitive data remains in the latent representation even after bias-mitigation algorithms are applied. In summary, we present a dataset, propose multiple challenging assessment scenarios, rigorously analyse recent promising bias-mitigation techniques in a common framework, and openly disclose this benchmark as an entry point for fair deep learning.

Keywords: Fairness in machine learning, Bias mitigation, Representation learning, Fairness model evaluation, Fair deep learning, Adversarial fairness

Contents

List of tables

List of figures

Liste des sigles et des abréviations

ML	Machine Learning
DL	Deep Learning
MLP	Multi-Layer Perceptron
NN	Neural Network
CNN	Convolutional Neural Network
FC	Fully Connnected Network
LRELU	Leaky Rectified Linear Unit
MNIST	Modified National Institute of Standards and Technology
CI-MNIST	Correlated and Imbalanced MNIST
VAE	Variational Auto-Encoder
GAN	Generative Adversarial Networks

LAFTR	Learning Adversarially Fair and Transferable Representations
CFAIR	Conditional Learning of Fair Representations
FFVAE	Flexibly Fair VAE
BER	Balanced Error Rate
DP	Demographic Parity
EqOdd	Equality of Odds
EqOpp	Equality of Opportunity

Remerciements

This thesis would not have been possible without many people. First and foremost, I am incredibly grateful to Prof. Sarath Chandar; he has been supportive throughout my master's. He lets us freely explore different areas and provides his valuable guidance at every stage of this journey. His emphasis on perfection dramatically motivates us to push ourselves beyond our limits. I will forever be thankful to Sarath for supporting me throughout my master's, starting from my supervisor change to exploring various projects at Mila and letting me be part of the unique Chandar Research Lab (CRL) community. I was fortunate enough to be a part of great talent from CRL, who supported me throughout my time. They gave me valuable input many times and steered projects in the right direction when I lost track.

I was fortunate to collaborate with talented researchers Samira Shabanian, Soroush Mehri from Microsoft Research, Sina Honari from EPFL, Adriana Romero-Soriano from Facebook Research, and Deepak Sharma from Mcgill. I thank Benjamin Fish for reading the thesis multiple times and providing insightful input. I also thank Elliot Creager for providing valuable information on modeling fairness algorithms. I am thankful to Abdelrahman, Fernando Diaz and Philip Bachman for their feedback during this project.

I want to express profound gratitude to my talented labmates at Chandar Research Lab, including Janarthanan, Gabriele, Louis, Simon, Abdelrahman, Mojtaba, Ali, Doriane, Andreas, Mohamed, Akilesh, and Hadi. During many group meeting presentations, they provided me with valuable comments to enhance the quality of my work. I would especially like to thank Abdelrahman for reviewing my thesis, which helped shape the thesis it is now. I was lucky to be around a talented peer group at Mila, including Soumye, Makesh, Dhaivat, Nikolaus, Mostafa, Rey, Soroosh with whom I have had numerous interactions shaped my views about academia and research. I want to thank my co-authors for this work, Samira, Sina, Adriana, Soroush, Deepak, and my supervisor Sarath without whom this thesis would not have materialized. I am deeply grateful to Deepak for supporting me on this project, from designing experiments to proofreading the paper and reviewing codes; he was by my side throughout this project.

I am deeply thankful to Sarath Chandar and Alain Tapp for supporting me with funding throughout my master's. I am grateful to the University of Montreal for the Bourse-C scholarship, which exempted me from paying international tuition at the university.

Montreal wouldn't have been fun without my roommates Soumye Singhal, Makesh Narsimhan, and Siddhartha Saxena, who had been very supportive throughout my masters, and we went through hard COVID times together. I consider myself fortunate to have found such incredible and super intelligent friends.

Most importantly, I am grateful to my co-founder Ali Mrani Alaoui; I started my startup during my master's, startup work was so hectic, and I had to manage writing my thesis while working 15 * 7 hours on the startup. My co-founder gave me full support to pursue both of my works.

I am foremost grateful to my family for supporting me throughout my master's. Here I take an opportunity to extend my gratitude to my parents and my sister for their unconditional love and moral support. Their enormous faith in my decisions and continuous encouragement kept me going during many moments of crisis. At last, I would like to thank all my school and college friends, who kept checking on me throughout the Covid-19 pandemic, I can not list all the names here, but you are always on my mind.

Chapter 1

Introduction

Deep Learning (DL) (Goodfellow et al., 2016) has been immensely successful in achieving groundbreaking performance across a wide range of Machine Learning (ML) (Mitchell et al., 1997) tasks including image classification (Simonyan and Zisserman, 2015, He et al., 2015), language modelling (Mikolov et al., 2013, Vaswani et al., 2017, Devlin et al., 2019), speech processing (Hinton et al., 2012), and object recognition (Krizhevsky et al., 2012). The success of DL models and their quick adoption in many application domains, particularly those involving decision-making using the predictions of these models, has raised questions about the fairness of these models when deployed in the real world. Recent studies (Agrawal et al., 2018, Kay et al., 2015, Lu et al., 2020, Buolamwini and Gebru, 2018, Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018) have highlighted the biases encoded by representation learning algorithms and have questioned the reliability of such approaches to make decisions.

Therefore, this thesis aims to provide a unified framework to benchmark bias-mitigation approaches and perform an in-depth analysis of existing methods leveraging the proposed framework. Using this framework, DL approaches can be rigorously tested for biases before being deployed in various target domains.

1.1. Machine Learning

In the words of Tom Mitchell, Professor at CMU (Mitchell et al., 1997), ML is the study of computer algorithms that improve automatically through experience.

The learning step of ML usually proceeds through one of three broad types: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. The critical difference between the approaches is that supervised learning uses labeled data to help predict outcomes while unsupervised and reinforcement learning does not. However, there are nuances between the three approaches and critical areas in which one outperforms the other.

1.1.1. Supervised Learning

Supervised learning (Mitchell et al., 1997) is a machine learning method that uses dataset labels. These datasets are used to train or "supervise" algorithms so that they can accurately identify data or predict outcomes. Labels refer to the value being predicted by the models. The model tests its accuracy and learns over time using inputs and outputs. The domain of supervised learning is further divided into two subcategories based on the nature of the tasks:

(1) Classification problems, as the name suggests, deal with identifying given data to one of the predetermined dataset labels. They employ an algorithm to assign test data into certain categories accurately. Such supervised learning algorithms, for example, can be employed in the real world to classify whether a patient has a disease or not. Classification methods include linear classifiers, support vector machines, decision trees, and random forests.

(2) Regression is another form of supervised learning method which employs an algorithm to deduce the relationship between dependent and independent variables. They can be thought of as function approximators. Regression models help forecast numerical values based on various data sources, such as sales revenue estimates for a particular company. Linear regression, logistic regression, and polynomial regression are some popular regression algorithms.

1.1.2. Unsupervised Learning

Unsupervised learning (Mitchell et al., 1997) analyses and learns patterns from unlabeled data sets using machine learning methods. The lack of human supervision for data tagging makes them unsupervised. These strategies aim to drive the machine to develop a compact internal picture of its surroundings through mimicry, a basic form of human learning and then generate imaginative content from it. Unsupervised learning is mainly used for clustering, association, and dimensionality reduction tasks.

(1) Clustering is a data mining technique that groups unlabeled data into groups based on similarities and differences. K-means clustering algorithms, for example, divide related data points into groups, with the K value indicating the size and granularity of the grouping. This method is useful for market segmentation, image compression, and other applications.

(2) Association is another form of unsupervised learning method, which employs several rules to discover relationships between variables in a dataset. These techniques are commonly employed in market basket analysis and recommendation engines, such as *Customers who bought this item also bought* suggestions.

(3) Dimensionality Reduction is utilised when the number of features (or dimensions) in a dataset is too large. It keeps the data integrity while reducing the amount of data inputs to a tolerable size. This technique is frequently employed in the data preprocessing stage. One example would be autoencoders eliminating noise from visual data to improve picture quality.

1.1.3. Reinforcement Learning

Reinforcement learning (Sutton and Barto, 2005) is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. In reinforcement learning, artificial intelligence faces a game-like situation. The computer employs trial and error to devise a solution to the problem. To get the machine to do what the programmer wants, the artificial intelligence gets either rewards or penalties for the actions it performs. Its goal is to maximize the total reward.

1.1.4. Comparing Machine Learning Algorithms

In Supervised Learning, the algorithm *learns* from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models are more accurate than unsupervised and reinforcement learning models, they necessitate human interaction to label all the data correctly. On the other hand, unsupervised learning models uncover unlabeled data's structure. Reinforcement learning models function independently to discover a similar structure but optimize for a goal by experimenting with multiple paths. It's worth noting, however, that even for unsupervised learning and reinforcement learning, validating output variables and rules of optimization still necessitates human intervention.

The use of labelled data has generally proved beneficial for model accuracy. But acquisition, cleaning, and labeling of datasets might not be practical and even feasible in some cases. This is where semi-supervised methods come in handy. These methods utilize a small-sized labeled dataset and a significantly larger unlabelled dataset.

In this thesis, we focus on the supervised learning problem in a classification setup of predicting the output decision eligibility, which is a binary 0 or 1 label in our setup. The eligibility is the criteria that makes an individual qualified or unqualified for a given task, such as providing loans or hiring.

1.2. Deep Learning

Deep Learning (DL) (Goodfellow et al., 2016) is a relatively recent paradigm within ML. It enables data representation and hierarchical learning through some sequential layers of abstraction. DL models learn to extract relationships from raw data using Neural Networks (McCulloch and Pitts, 1943). Neural Networks can learn these relationships from the expected outputs and correct their predictions based on the backward propagation algorithm (Kelley, 1960). This has paved the way for DL to be used to develop a broader set of techniques for representation learning, i.e., automatically extracting features and variable inter-relationships from raw data.

1.3. Representation Learning

Representation learning is about learning representations of the data that make it easier to extract useful information when building classifiers or other predictors (Bengio et al., 2013). In the case of probabilistic models, a good representation often captures the posterior distribution of the underlying explanatory factors for the observed input. A good representation is also useful as input to a supervised predictor. These representations have reduced the need for manual feature engineering, a necessary component in traditional ML systems. As a result, DL models have been able to outperform classical ML models in many complex tasks, particularly in the computer vision domain.

1.4. Bias in Representation Learning

Recent studies (Agrawal et al., 2018, Kay et al., 2015, Lu et al., 2020, Buolamwini and Gebru, 2018, Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018) have highlighted the biases encoded by representation learning algorithms and have questioned the reliability of such approaches in making decisions. For example, (Cohen et al., 2018) outline the biases exhibited by learning algorithms whose goal is to match the data distribution in an adversarial setting, by demonstrating how biases in the input data distributions can lead to domain translation models such as CycleGANs (Zhu et al., 2017), pix2pix (Isola et al., 2016) *hallucinating* image features that never existed in the original image. This problem of bias extends beyond the field of image translation. Similar findings have been revealed in the context of visual question answering (Agrawal et al., 2018), image search tasks (Kay et al., 2015), language models (Lu et al., 2020) and gender classification (Buolamwini and Gebru, 2018). As a result, there is increasing interest in understanding the sources of bias in learning algorithms and developing bias-mitigation strategies.

There is increasing literature studying how biased datasets can bias learning algorithms to discriminate (Cohen et al., 2018, Agrawal et al., 2018, Madras et al., 2018, Zhao et al., 2020,

Bolukbasi et al., 2016, Mehrabi et al., 2019b, Hooker, 2021). While some approaches focus on developing an objective for fair prediction (Madras et al., 2018), some concentrate on developing restrictions on the architecture to prevent the model's reliance on priors (Agrawal et al., 2018). Some approaches tweaked the algorithms for learning compact representations of datasets that allowed the prediction to be flexibly fair (Creager et al., 2019).

These recent approaches to bias mitigation (Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018) focus on designing models that better fulfill specific fairness criteria while maintaining the model's performance.

These studies have led researchers to propose new evaluation tools and datasets (Buolamwini and Gebru, 2018, Hardt et al., 2016a, Jones et al., 2020, Kusner et al., 2017, Dwork et al., 2012, Bellamy et al., 2018a) to identify potential error rate gaps among different sub-groups in data.

1.5. Bias-Mitigation Algorithms

The goal of bias-mitigation algorithms is to mitigate the influence of sensitive data features on the made eligibility decisions. Sensitive features are private and protected features of a dataset, such as gender or race, which should not affect output decisions of eligibility. Eligibility is the criteria that make an individual qualified or unqualified for a given task, such as providing loans or hiring. Bias mitigation models aim at making eligibility decisions on dataset samples without bias towards the input data's sensitive attributes.

The difficulty of bias-mitigation tasks is often determined by the dataset distribution, which in turn, is a function of the potential label and feature imbalance, the correlation of potentially sensitive features with other features in the data, and the perhaps inevitable distribution shift from training to the development phase, to name a few. We argue that their merits remain unclear without evaluating bias-mitigation models in various challenging setups.

In addition to the challenges associated with each dataset and task, the current state of the bias-mitigation literature hinders method comparisons due to inconsistencies in the experimentation and dataset setups.

Given the importance of proposed algorithms and possibly tangible adverse effects when in production, we advocate for a rigorous and unified evaluation protocol to assess their capabilities. We argue the need for a systematic analysis that would benchmark different bias-mitigation approaches under the perspective of varying fairness metrics to ensure replication of concluded results. This would help elucidate the most promising research contributions and possible future avenues to explore.

1.6. Issues with Bias-Mitigation Algorithms

To evaluate bias-mitigation models, some contributions have emerged to probe ML systems at different levels and reduce discrimination from the perspective of modeling. (Friedler et al., 2019) assess the fairness of pre-processing, in-processing, and post-processing ML approaches. (Verma and Rubin, 2018a) evaluate how fair an off-the-shelf logistic regression model is, given a set of fairness metric definitions. Contrary to these works, we consider *DL-based bias-mitigation models*. In particular, due to the popularity and further usage of in-processing adversarial methods, we take some promising approaches from this category and assess their merits in challenging setups. (Du et al., 2020) provide a review of sources of bias in DL models and different approaches applied to them; however, they do not evaluate and contrast models experimentally. To the best of our knowledge, we provide the first benchmark to systematically compare DL-based bias-mitigation approaches in varied and increasingly challenging scenarios, particularly in imbalanced and correlated dataset setups, to bring further insights into the working of bias-mitigation methods and propose new frameworks for evaluating them.

Given the importance of bias-mitigation approaches and the severe implications of their misuse, we intentionally try to push these models to their breaking point by creating various *challenging datasets*. We make the following observations through extensive experiments:

(1) When the correlation between the eligibility and the sensitive attribute in the training data increases, models tend to exploit it and become more biased in their predictions. These biases are further accentuated as the sensitive features become more predominant.

(2) When a group is under-represented, bias can arise due to imbalance or scarceness of the data, both of which affect existing models in different ways. Moreover, as the under-represented group becomes proportionally more imbalanced, the models act more biased.

(3) Bias-mitigation models do not *completely* remove sensitive information from their latent representations. Instead, they successfully reduce the bias in their results by designing loss functions that balance different subgroups.

(4) The robustness to random seeds is model dependent, with some models exhibiting high variance in their results, making the choice of random seed a source of potential bias.

1.7. Contributions

We start with introducing a synthetic dataset that facilitates creating challenging scenarios by controlling data imbalance or correlation among eligibility and sensitive or non-sensitive attributes. Contrary to real datasets, where the different components of data generation cannot be controlled independently, this synthetic dataset enables the soft modification of dataset characteristics, by

changing one component and keeping all others unchanged, which in turn allows the study of different sources and levels of bias in the data. We also consider a real and commonly used dataset, the *Adult dataset* (Dua and Graff, 2017), to investigate the impact of our settings on real data. The *Adult dataset*, also known as the *Census Income dataset*, aims to predict whether an individual's income exceeds \$50k/year based on census data, such as age, sex, work, race, etc. In this case, We adapt the dataset characteristics by altering the binarization process of its sensitive attributes, obtaining variations of the data reminiscent of those explored in the synthetic dataset we introduced.

We then provide an in-depth analysis of baselines and recent bias-mitigation models, leveraging both datasets mentioned above. In particular, we evaluate three promising bias-mitigation models – and seven variants – together with two baseline models. The analysis is performed by considering a unified set of fairness metrics and reporting results by carrying extensive hyper-parameter search in all cases, ensuring that conclusions can be attributed to modeling or loss choices. In doing so, we train about 3000 models in increasingly difficult scenarios. We transition from balanced dataset setups towards challenging imbalanced and correlated setups, where the eligibility criterion is correlated with sensitive or non-sensitive attributes.

To summarize, We make the following contributions:

- We provide a dataset with controllable sets of features and correlation levels to facilitate research in bias mitigation in a wide range of scenarios.
- We propose challenging test setups to evaluate bias-mitigation models by considering increasingly imbalanced and correlated scenarios and performing a rigorous analysis of existing methods, showing there remains much room for improvement.
- We release a benchmarking codebase composed of seven state-of-the-art models and two baselines to the community for reproducible evaluation of bias-mitigation algorithms.

1.8. Publication

Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero, Samira Shabanian, and Sina Honari. In Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks, Volume 1, 2021.

Contributions

- I came up with the idea when I started collaborating with Microsoft Research. Prof. Sarath Chandar provided significant support, gave me full freedom to pursue the idea, and encouraged me during the entire process.

- Experiment setups and analysis were designed by me and with valuable feedback from Samira Shabanian and Sina Honari.
- I wrote all the code for algorithms, performed experiments, and reported results in the published paper and the thesis.
- Deepak Sharma and Soroush Mehri helped me in running some experiments.
- I also wrote the entire paper with valuable help from Adriana Romero-Soriano, Samira Shabanian, and Sina Honari.
- In the entirety, I have contributed nearly all of the work from the start of the project to the publication and presented the work at NeurIPS.

Affliations of the authors

- Charan Reddy: Mila - Quebec AI Institute, Université de Montréal.
- Deepak Sharma: Mila - Quebec AI Institute, McGill University.
- Soroush Mehri: Microsoft Research Montreal.
- Adriana Romero-Soriano: Facebook AI Research Montreal, McGill University.
- Samira Shabanian: Microsoft Research Montreal.
- Sina Honari: Ecole Polytechnique Fédérale de Lausanne (EPFL).

1.9. Thesis Layout

The rest of the thesis is divided into four different chapters. In Chapter 2, we introduce some of the relevant background concepts and related works. In Chapter 3, we explain the methodology of our proposed benchmark. In Chapter 4, we explain in great detail the experiments we perform. Finally, in Chapter 5, we list certain limitations of the proposed method and conclude the thesis.

It is noteworthy that our analysis is not to undermine the effectiveness of any bias-mitigation method but instead to set the expectations and boundaries for different use cases and encourage the community to investigate models under more extensive scenarios.

Chapter 2

Background

In this chapter, we review several key concepts required to understand the key contributions of this work. We first explore the necessity and development of fairness metrics [2.2] in machine learning and deep learning. Next, we define metrics that we use for our benchmark. We then explore the literature of ML and DL models used for bias-mitigation [2.5]. Finally, we describe the working of the models used in our evaluations in detail.

2.1. Introduction to Fairness

Recent studies, such as (Agrawal et al., 2018, Kay et al., 2015, Lu et al., 2020, Buolamwini and Gebru, 2018, Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018), have exposed the biases incorporated by representation learning algorithms and called into doubt the validity of such decision-making systems. Bias in data can show up in several forms. The most common ones are:

(1) *Historical bias* is a type of prejudice that already exists in the world and has infiltrated our data. Even in ideal sampling contexts and feature selection, this bias can exist, and it is more common in populations that have been historically disadvantaged or excluded.

(2) *Representation bias* occurs due to how we define and sample a population to produce a dataset. For example, datasets acquired using smartphone apps are a type of representation bias, as they may underrepresent lower-income or older age demographics.

(3) *Measurement bias* occurs when choosing or gathering features or labels to employ in prediction models. Easy-to-find data is frequently a noisy proxy for the underlying traits or labels of interest. Furthermore, measuring methods and data quality varies widely amongst groups. As Artificial Intelligence (AI) is utilised for additional applications, such as predictive policing, this bias has the potential to have a significant negative influence on people's lives.

For example, (Cohen et al., 2018) show how biases in the input data distributions can lead to domain translation models like CycleGANs (Zhu et al., 2017), pix2pix (Isola et al., 2016)

hallucinating image features that never existed in the original image. Bias is a problem that isn't limited to image translation, in the context of visual question answering (Agrawal et al., 2018), image search tasks (Kay et al., 2015), language models (Lu et al., 2020) and gender classification (Buolamwini and Gebru, 2018), similar findings have been discovered.

Even if we have perfect data, our modeling methods can introduce bias (Suresh and Guttag, 2021). The most common ways for those biases are:

(1) *Evaluation Bias* develops during the iteration and evaluation of a model. A model's quality is generally tested against specific standards, but it is optimised using training data. Biases can be seen when these benchmarks don't represent the broader population or aren't appropriate for how the model will be used.

(2) *Aggregation bias* occurs when different populations are incorrectly mixed during model creation. Many AI applications have a heterogeneous population of interest, and a single model is unlikely to satisfy all groups.

Even if our model makes accurate predictions, a human reviewer's biases can be introduced while deciding whether to accept or dismiss a model's prediction. A human reviewer, for example, may override a correct model prediction due to their own systemic bias.

As a result, there is increasing interest in understanding the sources of bias in learning algorithms and developing bias-mitigation strategies (Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018).

There is a growing body of knowledge about how biased datasets might cause learning systems to discriminate incorrectly (Cohen et al., 2018, Agrawal et al., 2018, Madras et al., 2018, Zhao et al., 2020, Bolukbasi et al., 2016, Mehrabi et al., 2019b, Hooker, 2021). These recent studies have led researchers to propose new evaluation tools and datasets (Buolamwini and Gebru, 2018, Hardt et al., 2016a, Jones et al., 2020, Kusner et al., 2017, Dwork et al., 2012, Bellamy et al., 2018a), to identify potential error rate gaps among different groups in ML. For example, when CycleGANs were trained to transform MRI images from Flair to T1 types, models had a bias to remove tumors because the target distribution did not have any tumor examples, so the transformation was forced to remove tumors in order to match the target distribution. Using these translated images for medical diagnostic purposes can naturally lead to misdiagnosis of medical conditions.

2.2. Fairness Metrics

We must first quantify fairness to measure and improve the fairness of ML or DL models. Fairness metrics are used to measure and assess the bias tendencies in models, for both comparisons across models and for steadily improving the performance of models concerning fairness.

In recent literature (Gajane, 2017, Verma and Rubin, 2018b, Barocas et al., 2019), a large number of metrics have been proposed, and this is not a failing of the literature but is more a reflection of the multi-faceted nature of the concept of fairness itself. The differences stem from different intuitions among researchers as to the notion of "unfair decisions." While definitions differ on several points, they largely agree on describing bias or unfair decision-making as a tendency to prejudge candidates based on sensitive attributes like age, gender, race, etc. This section provides a brief overview of the potential reasons for bias in ML and DL systems and some of the most common fairness metrics (Barocas et al., 2019) used to benchmark both ML and DL models.

Unawareness Unawareness is a relatively straightforward metric that does not include the sensitive attribute as a feature in the training data. This metric is based on the idea of addressing *disparate treatment*. A decision-making process is said to suffer from *disparate treatment* if it's decisions are even partly based on the values of sensitive attributes. The fundamental limitation of the metric is that in practice, there are usually additional features in the dataset that are not explicitly identified as sensitive outputs. Yet, they serve as proxies of the sensitive attributes. Hence this measure is typically not sufficient to address bias.

Individual fairness Individual fairness is a relatively different notion; unlike group-based fairness, it is individual-based. It was first proposed in Fairness Through Awareness (Dwork et al., 2012), highlighting that individuals should be treated similarly.

Denote O to be a measurable space and $\Delta(O)$ to be the space of the distribution over O. Denote $M : X \to \Delta(O)$ to be a map of each individual to distribution of outcomes. The formulation is then: $D(M(X), M(X')) \leq d(X, X')$, where $X, X' \in R^d$ are two input feature vectors, and D and d are two metric functions on the input space and the output space respectively. The formulation assures that if two vectors X, X' are closer in the input space, then their corresponding outputs $M(X), M(X')$ are closer in the output space.

Predictive Rate Parity (PRP) Predictive Rate Parity, also called Sufficiency (Zafar et al., 2017b) is defined as Y is independent of S, conditional on $\tilde{Y}$ (where Y is the eligibility criteria, $\tilde{Y}$ is the model prediction and S is the sensitive attribute). This is equivalent to satisfying both Positive Predictive Parity (PPP) and Negative Predictive Parity (NPP).

$$\Delta_{PPP} = 1 - P[Y = 1|\tilde{Y} = 1, S = Unprotected] - P[Y = 1|\tilde{Y} = 1, S = Protected], \quad (2.2.1)$$

$$\Delta_{NPP} = 1 - P[Y = 0|\tilde{Y} = 0, S = Unprotected] - P[Y = 0|\tilde{Y} = 0, S = Protected]. \quad (2.2.2)$$

Demographic Parity (DP) Demographic Parity requires that each segment of a sensitive attribute (e.g. gender) should be assigned a positive eligibility outcome at an equal rate as any other segment within the class. Formally, we define Demographic Parity as follows (where $\tilde{Y}$ is the model prediction and *S* is the sensitive attribute) -

$$\Delta_{DP} = 1 - |p(\hat{Y} = 1|S = \text{Protected}) - p(\hat{Y} = 1|S = \text{Unprotected})|. \tag{2.2.3}$$

Higher Δ_{DP} is better, i.e., fairer. When $\Delta_{DP} = 1$, then we say that fairness has been achieved with respect to the sensitive attribute *S*. However, this metric does not take into account the predictive performance of the model on the protected class, i.e., a model can achieve a high Δ_{DP} even by predicting $\hat{Y} = 1$ for random instances of the protected and unprotected classes, not just the correct instances.

Equality of Opportunity (EqOpp) (Hardt et al., 2016b) Given $\tilde{Y}$ is the model prediction, *S* is the sensitive attribute, and *Y* is the target variable, a binary predictor $\tilde{Y}$ is said to satisfy *Equal Opportunity* with respect to sensitive attribute *S* if -

$$EqOpp1 = 1 - P[\tilde{Y} = 1|Y = 1, S = Unprotected] - P[\tilde{Y} = 1|Y = 1, S = Protected], \tag{2.2.4}$$

$$EqOpp0 = 1 - P[\tilde{Y} = 1|Y = 0, S = Unprotected] - P[\tilde{Y} = 1|Y = 0, S = Protected]. \tag{2.2.5}$$

Equality of Odds (EqOdd) (Hardt et al., 2016b) Equality of Odds is defined as the summation of $EqOpp1$ which is the Equality of Opportunity w.r.t $Y = 1$ and $EqOpp0$ which is the Equality of Opportunity w.r.t $Y = 0$ -

$$EqOdd = 0.5 * (EqOpp0 + EqOpp1). \tag{2.2.6}$$

Table 2.1. X, Y, S denote the input, label, and the sensitive attribute. $\hat{Y}$ and p are the model's prediction and the output probability of the model. For all metrics, 1 indicates the perfect and 0 the lowest value.

Fairness Criteria	Formulation	Short form
Demographic Parity	$1 - \|p(\hat{Y} = 1\|S = \text{Protected}) - p(\hat{Y} = 1\|S = \text{Unprotected})\|$	DP
Equality of Opportunity (w.r.t $y = 1$)	$1 - \|p(\hat{Y} = 1\|Y = 1, S = \text{Unprotected}) - p(\hat{Y} = 1\|Y = 1, S = \text{Protected})\|$	EqOpp1
Equality of Opportunity (w.r.t $y = 0$)	$1 - \|p(\hat{Y} = 1\|Y = 0, S = \text{Unprotected}) - p(\hat{Y} = 1\|Y = 0, S = \text{Protected})\|$	EqOpp0
Equality of Odds	$0.5 \times [\text{EqOpp0} + \text{EqOpp1}]$	EqOdd
unprotected-accuracy	$p(\hat{Y} = y\|Y = y, S = \text{Unprotected})$	up-acc
protected-accuracy	$p(\hat{Y} = y\|Y = y, S = \text{Protected})$	p-acc
accuracy	$0.5 \times [\text{up-acc} + \text{p-acc}]$	acc

Table 2.2. Fairness metrics. X, Y, S denote respectively the input sample, the ground truth label, and the sensitive attribute. p is the output probability of the model and $\hat{Y}$ is the model's prediction. For the metrics presented in this table, the sensitive attribute S takes binary values in $\{0, 1\}$.

Fairness Criteria	**Definition**	**Abbreviation**
Group conditioned s-accuracy	$p(\hat{Y} = y \mid Y = y, S = s)$	s-accuracy
s-True positive ((Verma and Rubin, 2018a))	$\lvert\{x \mid \hat{y} = 1 \text{ for } (x, y = 1, S = s) \in X\}\rvert$ where $\lvert \cdot \rvert$ refers to cardinality of a set	s-TP
s-False positive ((Verma and Rubin, 2018a))	$\lvert\{x \mid \hat{y} = 1 \text{ for } (x, y = 0, S = s) \in X\}\rvert$ where $\lvert \cdot \rvert$ refers to cardinality of a set	s-FP
s-False negative ((Verma and Rubin, 2018a))	$\lvert\{x \mid \hat{y} = 0 \text{ for } (x, y = 1, S = s) \in X\}\rvert$ where $\lvert \cdot \rvert$ refers to cardinality of a set	s-FN
s-True negative ((Verma and Rubin, 2018a))	$\lvert\{x \mid \hat{y} = 0 \text{ for } (x, y = 0, S = s) \in X\}\rvert$ where $\lvert \cdot \rvert$ refers to cardinality of a set	s-TN
s-True positive rate ((Friedler et al., 2019)) = s-positive predictive value (s-PPV)	$p(\hat{Y} = 1 \mid Y = 1, S = s)$	s-TPR
s-True negative rate	$p(\hat{Y} = 0 \mid Y = 0, S = s)$	s-TNR
s-False positive rate	$p(\hat{Y} = 1 \mid Y = 0, S = s)$ equivalent to $1 - s$-TNR	s-FPR
s-False negative rate	$p(\hat{Y} = 0 \mid Y = 1, S = s)$ equivalent to $1 - s$-TPR	s-FNR
s-Balanced classification rate	$0.5 \times [p(\hat{Y} = 1 \mid Y = 1, S = s) + p(\hat{Y} = 0 \mid Y = 0, S = s)]$ equivalent to $0.5\times$(s-TPR + s-TNR)	s-BCR
Equality of odds ((Hardt et al., 2016c) & (Beutel et al., 2017)) = Equalized odds ((Hardt et al., 2016c))) = conditional procedure accuracy equality ((Berk et al., 2018)) = disparate mistreatment ((Zafar et al., 2017b))	$p(\hat{Y} = \hat{y} \mid Y = y) = p(\hat{Y} = \hat{y} \mid Y = y, S = s)$ equivalent to $\Big[p(\hat{Y} = 1 \mid Y = 1, S = 1) = p(\hat{Y} = 1 \mid Y = 1, S = 0)$ and $p(\hat{Y} = 1 \mid Y = 0, S = 1) = p(\hat{Y} = 1 \mid Y = 0, S = 0)\Big]$ equivalent to [1-TPR = 0-TPR and 0-TNR = 1-TNR]	-
s-calibration+ ((Friedler et al., 2019))	$p(Y = 1 \mid \hat{Y} = 1, S = s)$	-
s-calibration− ((Friedler et al., 2019))	$p(Y = 1 \mid \hat{Y} = 0, S = s)$	-
Conditional use accuracy equality ((Berk et al., 2018))	$\Big[p(Y = 1 \mid \hat{Y} = 1, S = 1) = p(Y = 1 \mid \hat{Y} = 1, S = 0)$ and $p(Y = 0 \mid \hat{Y} = 0, S = 1) = p(Y = 0 \mid \hat{Y} = 0, S = 0)\Big]$ equivalent to [0-calibration+ = 1-calibration+ and 0-calibration− = 1-calibration−]	-
Calders and Verwer ((Calders and Verwer, 2010a))	$1 - [p(\hat{Y} = 1 \mid S = 1) - p(\hat{Y} = 1 \mid S \neq 1)]$	CV
Demographic parity ((Hardt et al., 2016c) & (Beutel et al., 2017)) = Group fairness ((Dwork et al., 2012)) = statistical parity ((Dwork et al., 2012)) = equal acceptance rate ((Zliobaite, 2015))	$p(\hat{Y}) = p(\hat{Y} \mid S)$ equivalent to 1−CV equivalent to $p(\hat{Y} = 1 \mid S = 1) = p(\hat{Y} = 1 \mid S \neq 1)$	DP
Disparate Impact ((Feldman et al., 2015) & (Zafar et al., 2017a))	$\frac{p(\hat{Y}=1 \mid S\neq 1)}{p(\hat{Y}=1 \mid S=1)}$	DI
Equality of opportunity with respect to y ((Hardt et al., 2016c))	$p(\hat{Y} = \hat{y} \mid Y = y) = p(\hat{Y} = \hat{y} \mid Y = y, S = s)$ Equality of odds is stronger than equality of opportunity	-
False positive error rate balance ((Chouldechova, 2017)) = predictive equality((Corbett-Davies et al., 2017))	$p(\hat{Y} = 1 \mid Y = 0, S = 1) = p(\hat{Y} = 1 \mid Y = 0, S = 0)$ equivalent to $p(\hat{Y} = 0 \mid Y = 0, S = 1) = p(\hat{Y} = 0 \mid Y = 0, S = 0)$ equivalent to 1-TNR = 0-TNR equivalent to [Equality of opportunity with respect to $y = 0$]	-
False negative error rate balance ((Chouldechova, 2017)) = equal opportunity ((Kusner et al., 2017) & (Hardt et al., 2016c))	$p(\hat{Y} = 0 \mid Y = 1, S = 1) = p(\hat{Y} = 0 \mid Y = 1, S = 0)$ equivalent to $p(\hat{Y} = 1 \mid Y = 1, S = 1) = p(\hat{Y} = 1 \mid Y = 1, S = 0)$ equivalent to 1-TPR = 0-TPR equivalent to [Equality of opportunity with respect to $y = 1$]	-
Matthews correlation coefficient	$\frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$	MCC

In this thesis, we use Demographic Parity (DP), Equality of Opportunity for label classes y equal to 0 and 1 (EqOpp0 and EqOpp1), Equality of Odds (EqOdd), and label accuracy (acc) in our experiments to evaluate the following bias-mitigation strategies: Learning adversarially fair

and transferable representations (Laftr) (Madras et al., 2018), Conditional learning of fair representations (Cfair) (Zhao et al., 2020), and Flexibly fair representation learning by disentanglement (Ffvae) (Creager et al., 2019). Table 2.1 presents all fairness metrics and their mathematical formulations used in our evaluations. In Table 2.2, we offer the comprehensive list of fairness metrics taken from the literature, along with their mathematical definitions and abbreviations. In the code-base we released, all these metrics are provided and can be used for evaluation. We use the tool proposed in (Friedler et al., 2019) to compute all of these metrics.

2.3. Bias-Mitigation Methods in Machine Learning

In recent years, there has been a great use of Machine Learning (ML) for a wide range of applications which routinely deal with sensitive attributes such as financial lending (Byanjankar et al., 2015, Malekipirbazari and Aksakalli, 2015), hiring (Hoffman et al., 2018, Bogen and Rieke, 2018), health care (Kourou et al., 2015), education (Oneto et al., 2017, Papamitsiou and Economides, 2014), etc. With this, proportionally, there has been an increase in bias based on sensitive attributes such as race, gender, disabilities, sexual orientation, etc. (Agrawal et al., 2018, Kay et al., 2015, Lu et al., 2020, Buolamwini and Gebru, 2018, Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018). As a result, several methods have been developed to identify, quantify and minimize this unfairness/bias (Oneto and Chiappa, 2020, Mehrabi et al., 2019c, Awasthi et al., 2021, Caton and Haas, 2020, Yu et al., 2021). In this section we elaborate on some of the bias-mitigation models and approaches applied in the context of ML tasks.

Bias-mitigation algorithms can be sub-divided into the following three categories as seen below. Figure 2.1 dicusses the three categories in a standard ML pipeline,

- Pre-processing techniques, that address biases in the data to reduce innate unfairness in the data itself (Bellamy et al., 2018b, d' Alessandro et al., 2017, Brunet et al., 2018, Kamiran and Calders, 2011, Calmon et al., 2017). Pre-processing techniques include reweighting of training samples (Kamiran and Calders, 2011), editing features and labels (Calmon et al., 2017), resampling datasets (Chawla et al., 2002). For example, pre-processing techniques are used in (Brunet et al., 2018) to identify and reduce bias in word embeddings trained on large unsupervised corpora.
- In-processing techniques remove sensitive information from the learned representation space (Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018, Jiang et al., 2019, Zemel et al., 2013, Louizos et al., 2016, Träuble et al., 2020, Beutel et al., 2017, Marx et al., 2019, Kim and Mnih, 2018, Chen et al., 2018, Locatello et al., 2019, Louppe

et al., 2017, Xie et al., 2017). Once an effective learning algorithm has been developed, it is modified to eliminate biases during the model training process.

- Post-processing techniques calibrate predictions given sensitive attributes at inference time (Hardt et al., 2016a, Zhao et al., 2017, Jiang et al., 2019). These methods are particularly advantageous when dealing with large pre-trained models that cannot be re-trained to eliminate biases.

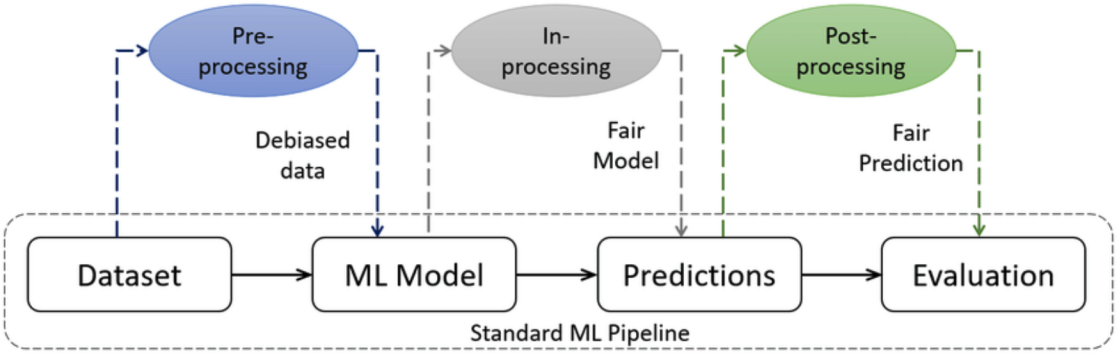


Fig. 2.1. Three categories of bias-mitigation algorithms in a standard ML pipeline

To address the biases, approaches have been suggested to modify algorithms for classification (Huang and Vishnoi, 2020, Goel et al., 2018, Calders and Verwer, 2010b), regression (Berk et al., 2017), clustering (Backurs et al., 2019), community detection (Mehrabi et al., 2019a), dimensionality reduction (Samadi et al., 2018) etc. Due to the emergence of more in-processing models in the deep learning community, we focus on this group of models, as bias-mitigation algorithms are mainly applied directly to the learning models.

2.4. Bias-Mitigation Methods in Deep Learning

Deep learning approaches have become more popular in the fairness community. For example, (Louizos et al., 2016) employ a deep variational autoencoder (Kingma and Welling, 2014) to learn fair latent representations. At the same time, adversarial learning, introduced initially within the framework of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), has also been widely leveraged in the bias-mitigation literature (Madras et al., 2018, Zhao et al., 2020, Creager et al., 2019, Zhang et al., 2018, Träuble et al., 2020, Beutel et al., 2017, Marx et al., 2019, Kim and Mnih, 2018, Chen et al., 2018, Locatello et al., 2019, Louppe et al., 2017, Xie et al., 2017) to make different groups indistinguishable from one another with respect to a sensitive attribute. Due

to more widespread usage of adversarial approaches, we analyse this group of models. Adversarial bias-mitigation techniques can be divided into approaches which: (i) seek to mitigate bias through adversarial training directly applied on the class or target labels (Zhang et al., 2018), where the class label is an indicator of eligibility; (ii) focus on mitigating bias through enforcing group fairness on the learned latent space where the latent is directly used for classification of eligibility (Madras et al., 2018, Zhao et al., 2020, Beutel et al., 2017, Louppe et al., 2017, Xie et al., 2017); and (iii) discard sensitive features for downstream tasks after disentangling the learned latent space into sensitive and non-sensitive features (Creager et al., 2019, Träuble et al., 2020, Marx et al., 2019). This line of research is motivated by the recent development in disentangled representation learning (Louizos et al., 2016, Kim and Mnih, 2018, Chen et al., 2018, Locatello et al., 2019).

Multiple deep learning debiasing algorithms rely on adversarial techniques to mitigate bias with respect to fairness metrics of interests (Madras et al., 2018, Beutel et al., 2017, Zhang et al., 2018, Zhao et al., 2020, Creager et al., 2019).

2.5. Benchmarked biased-mitigation methods

We empirically analyze the performance of multi-layer perceptrons (MLPs) and Convolutional Neural Networks (CNNs) as our baselines with, 1. Laftr model (Madras et al., 2018) which explore adversarial learning to debias representations, 2. Cfair (Zhao et al., 2020) which performs conditional alignment of representations to achieve a better accuracy-fairness trade-off, and, 3. Ffvae (Creager et al., 2019) which disentangles sensitive attributes in the representations. In this section, we give a brief overview of the aforementioned models.

Baseline Model. We use the multi-layered perceptrons (MLPs) and convolutional neural networks (CNNs) as our baseline models, which given an input image x, predict the probability p of the eligibility criteria. This probability is then transformed into a classification prediction $\hat{y}$. These models do not leverage any bias mitigation method and are trained using a standard cross-entropy loss. It is meant to show how fair a baseline deep learning model would perform under different fairness criteria.

Learning Adversarially Fair and Transferable Representations (Laftr). Laftr (Madras et al., 2018) is an adversarial based bias mitigation algorithm within the scope of representation learning. Generally, adversarial models have been leveraged to debias deep learning models. In (Zhang et al.,

2018) discriminator is trained by minimizing the following objective function:

$$\min_{D} \mathbb{E}_{x,y,s\in\mathcal{D}} \mathcal{L}_{\mathcal{S}}(D(C(x),y),s), \tag{2.5.1}$$

where $C : X \to Y$ is a classifier, D is a discriminator, and $\mathcal{L}_{\mathcal{S}}$ indicates the sensitive attribute's loss function, which compares the discriminator's output $D(C(x), y)$ with the sensitive attribute s as seen in Figure. 2.2. In the supervised version of Laftr, given an input x, it first learns a latent encoded representation z that is passed to the discriminator to be debiased. The learned representation is then passed to a classifier to predict the task of interest y. The discriminator is trained by minimizing

$$\mathcal{L}_{fair}^{Laftr} = \mathbb{E}_{x,y,s\in\mathcal{D}} \mathcal{L}_{\mathcal{S}}(D(E(x),y),s), \tag{2.5.2}$$

where $\mathcal{L}_{\mathcal{S}}$ is the adversarial loss, and y is only passed in debiasing models aimed for *equality of odds* and *equality of opportunity* fairness metrics. The encoder and classifier are trained jointly by minimizing

$$\mathcal{L}_{Laftr} = \mathbb{E}_{x,y,s\in\mathcal{D}} \mathcal{L}_{\mathcal{Y}}(C(z),y) - \gamma \mathcal{L}_{fair}^{Laftr}, \tag{2.5.3}$$

where $z = E(x)$ is the encoded feature, passed to both the classifier C and the discriminator D. The first term on the right side of the equation measures the classification loss (denoted as $\mathcal{L}_{cl}^{Laftr}$), and the second term (or the fairness objective) gets the adversarial gradients from the discriminator regarding the sensitive attribute s. Following the original paper, four variants of Laftr model are considered that represent the desired fairness criteria via $\mathcal{L}_{fair}^{Laftr}$. This includes:

(1) Laftr-DP in which the fairness objective is defined as

$$\mathcal{L}_{DP}^{Laftr} = 1 - \sum_{s\in\{0,1\}} \mathbb{E}_{x,s\in\mathcal{D}_s} |D(z) - s|. \tag{2.5.4}$$

(2) Laftr-EqOpp0 in which the fairness objective is considered as

$$\mathcal{L}_{EqOpp0}^{Laftr} = 1 - \sum_{s\in\{0,1\},y=0} \mathbb{E}_{x,s\in\mathcal{D}_s^y} |D(z) - s|. \tag{2.5.5}$$

(3) Laftr-EqOpp1 whose fairness objective $\mathcal{L}_{EqOpp1}^{Laftr}$ is obtained by replacing $y = 1$ in Eq. (2.5.5).

(4) Laftr-EqOdd with the equality of odds fairness objective denoted as $\mathcal{L}_{EqOdd}^{Laftr}$ which is the sum of $\mathcal{L}_{EqOpp0}^{Laftr}$ and $\mathcal{L}_{EqOpp1}^{Laftr}$.

Conditional Learning of Fair Representations (Cfair). Proposed by (Zhao et al., 2020), this model as seen in Figure.2.3 leverages two adversarial networks h_0 and h_1, predicting sensitive attribute s respectively for class labels $Y = 0$ and $Y = 1$. Cfair depends on an objective function

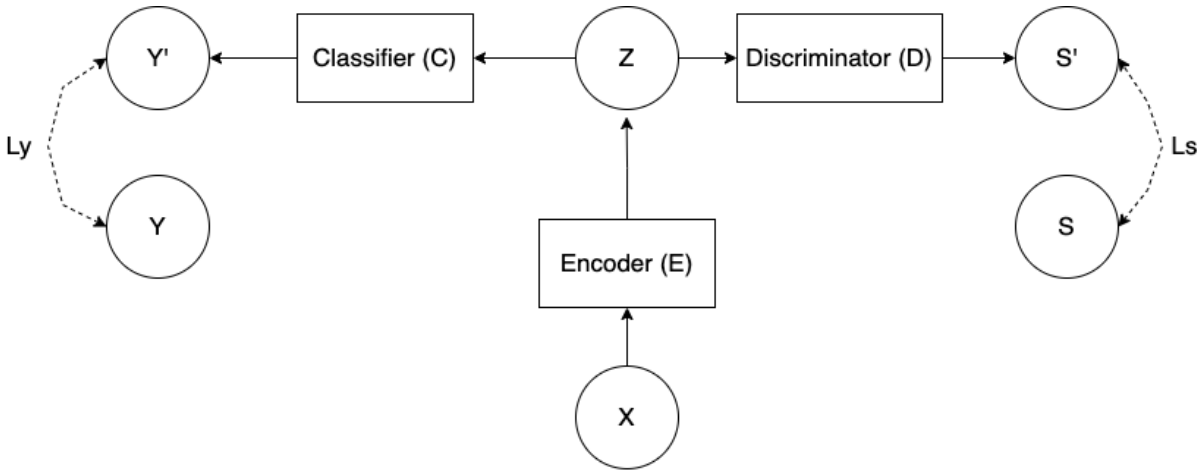


Fig. 2.2. An illustration of learning adversarially fair representations (Laftr) model adapted from (Madras et al., 2018). The variables are data X, latent representations Z, sensitive attributes S, and labels Y. The Encoder E maps X to Z, the Classifier C predicts Y' from Z, and the Discriminator D predicts S' from Z. And Loss functions $\mathcal{L}_{\mathcal{Y}}$ calculates classification and $\mathcal{L}_{\mathcal{S}}$ calculates fairness losses.

called the balanced error rate (BER) (Feldman et al., 2015, Menon and Williamson, 2018), which guarantees small joint error across demographic groups. The BER represents the sum of false positive rate and false negative rate. Therefore, it is equal to minimizing the below two conditional errors. $\text{BER}_{\mathcal{D}}(\hat{Y}\|Y)$ is defined as

$$\text{BER}_{\mathcal{D}}(\hat{Y}\|Y) \propto p(\hat{Y}=1|Y=0) + p(\hat{Y}=0|Y=1). \tag{2.5.6}$$

and $\text{BER}_{\mathcal{D}}(\hat{S}\|S)$ is defined similarly, where $\hat{S}$ is the predicted sensitive random variable. Cfair is optimized based on the following min-max formulation.

$$\mathcal{L}_{Cfair} = \min_{C,E} \max_{h_0,h_1} \quad \left(\text{BER}_{\mathcal{D}}(C(E(X))\|Y) - \gamma \mathcal{L}_{DP}^{Cfair}\right), \tag{2.5.7}$$

where

$$\mathcal{L}_{DP}^{Cfair} = \text{BER}_{\mathcal{D}^{y=0}}\left(h_0(E(X))\|S\right) + \text{BER}_{\mathcal{D}^{y=1}}\left(h_1(E(X))\|S\right). \tag{2.5.8}$$

This approach proposes that using the balanced error rate along with the conditional alignment helps in achieving equalized odds across the groups without impacting demographic parity.

Cfair-EO is a variant of the Cfair model, which considers Cross-Entropy loss instead of BER loss for the classifier C to achieve equalized odds. In case of equal target class distribution, Cfair and Cfair-EO are the same.

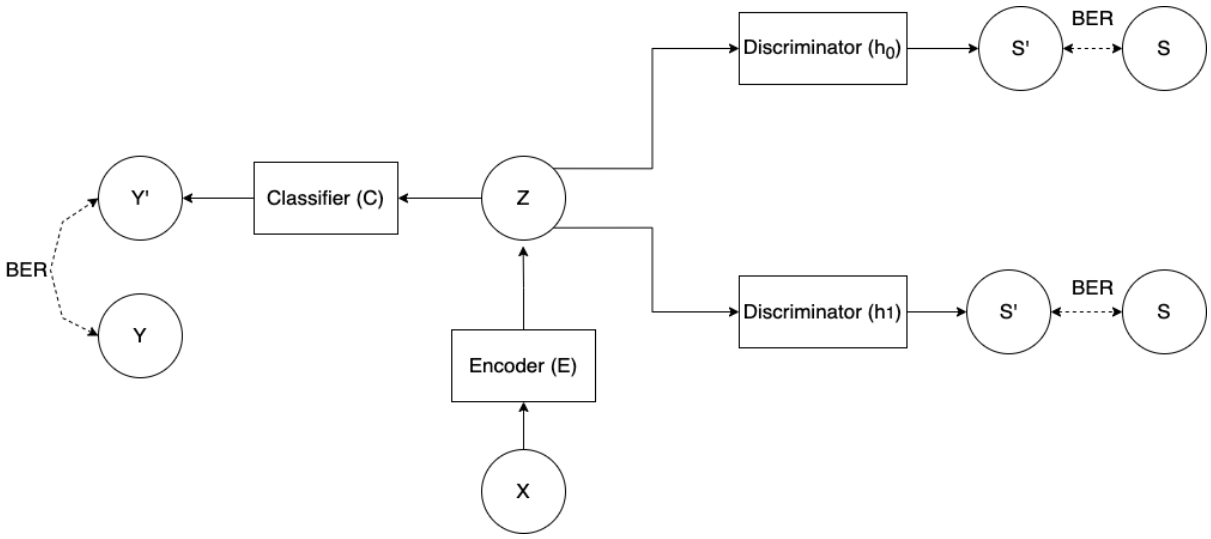


Fig. 2.3. Model for conditional learning of fair representations (Cfair) adapted from (Zhao et al., 2020). The variables are data X, latent representations Z, sensitive attributes S, and labels Y. The Encoder E maps X to Z, the Classifier C predicts Y' from Z, and the Discriminator h_0 and h_1 predicts S' from Z for class labels Y=0 and Y=1 respectively. Loss function BER calculates Balanced Error Rate between predicted and original variables.

Flexibly Fair Representation Learning by Disentanglement (Ffvae). Inspired by FactorVAE (Kim and Mnih, 2018), Ffvae (Creager et al., 2019) performs disentanglement by factorizing latent space as seen in Figure.2.4. It learns a disentangled representation of the inputs, which is flexibly fair because it can be easily modified at test time to achieve demographic parity across various groups.

Given x, $s = (s_1, \ldots, s_N)$, $b = (b_1, \ldots, b_N)$, and z being respectively the input, the sensitive attribute, the sensitive latent, and non-sensitive latent, with N indicating the number of sensitive or non-sensitive features (depending on the dataset), Ffvae trains an encoder $q(z, b|x)$, a decoder $p(x|z, b)$, as well as an adversarial network. The latent representation is disentangled into sensitive b and non sensitive z latent attributes by encouraging both $\mathrm{MI}(b, z)$ and $\mathrm{MI}\left(b_i, s_j\right), \forall i \neq j$ to be low, where MI represents mutual information. Ffvae objective is defined as -

$$\begin{aligned}\mathcal{L}_{\text{Ffvae}}(p, q) = \mathbb{E}_{q(z,b|x)}[\log p(x \mid z, b) + \alpha \log p(s \mid b)]] - \gamma D_{KL}(q(z, b)\|q(z)\prod_j q\left(b_j\right)) \\ - D_{KL}(q(z, b \mid x)\|p(z, b)).\end{aligned} \tag{2.5.9}$$

Eq.(2.5.9) has two terms, the first term consists of a reconstruction term (on left) and a *predictiveness* term $p(s \mid b)$, which aligns sensitive attributes to its respective sensitive latents, the second term is

the *disentanglement* term which decorrelates the sensitive latent representation b from z using an adversarial network.

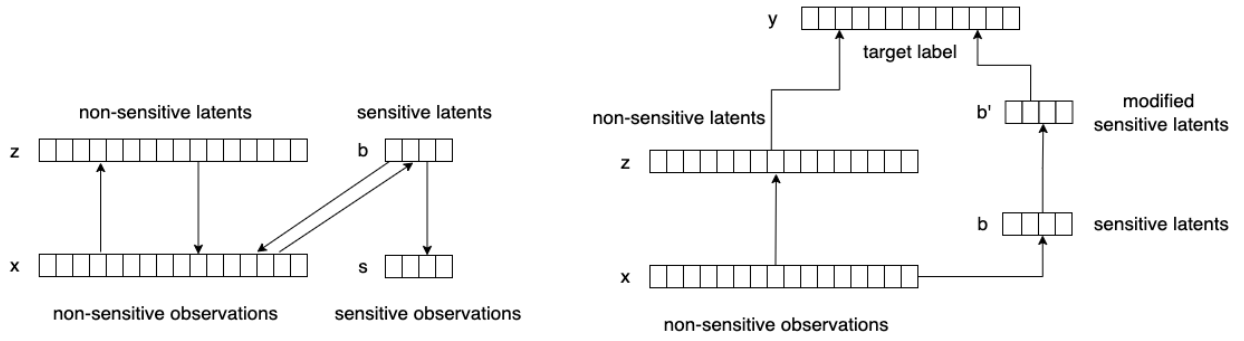


Fig. 2.4. Model train and test setups for Flexibly Fair VAE (Ffvae) model adapted from (Creager et al., 2019). The variables are data x, latent representations x, sensitive attributes s, sensitive latents b, modified latents b' and labels y. Ffvae learns the encoder distribution $q(z, b \mid x)$ and decoder distributions $p(x \mid z, b)$, $p(s \mid b)$ from inputs x and multiple sensitive attributes s. The disentanglement prior structures the latent space by encouraging low $MI(b_i, s_j) \forall i \neq j$ and low $MI(b, z)$ where $MI(.)$ denotes mutual information. The Ffvae latent code $[z, b]$ can be modified by discarding or noising out sensitive dimensions b_j, which yields a latent code $[z, b']$ independent of groups and subgroups derived from sensitive attributes s_j. A held out label y can then be predicted with subgroup demographic parity.

In addition to the models mentioned above, we also did experiments with Mitigating Unwanted Biases with Adversarial Learning (Mubal) model (Zhang et al., 2018) (based on the code released by authors), however, the model was very unstable on our dataset configurations even after extensive hyper-parameter search. We hypothesize that this is due to applying adversarial training directly to the class labels, which makes the model unstable, as indicated by the authors in Figure. 2.5. Due to unstable results, we dropped this model from our evaluation.

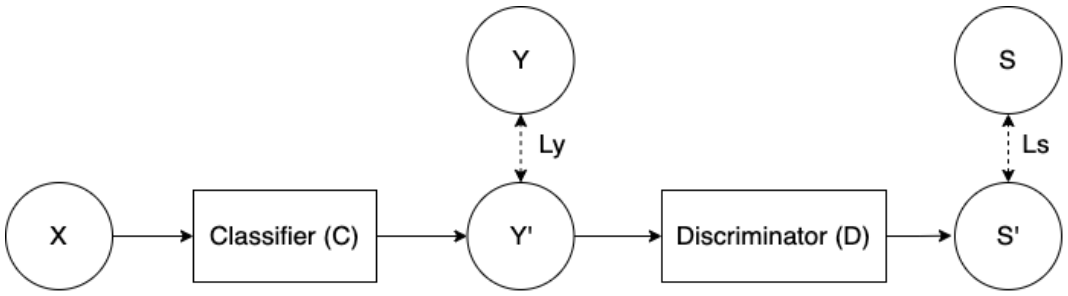


Fig. 2.5. Model for Mitigating Unwanted Biases with Adversarial Learning (Mubal) adapted from (Zhang et al., 2018). The variables are data X, latent representations Z, sensitive attributes S, and labels Y. The Classifier C predicts Y' from Z, and the Discriminator D predicts S' from Y. And Loss functions $\mathcal{L}_y$ calculates classification and $\mathcal{L}_S$ calculates fairness losses.

Chapter 3

Proposed Benchmarking

In this work, we examine the performance of two baselines without a bias-mitigation learning criteria: a multi-layer perceptron (MLP) and a convolutional neural network (CNN). These baselines are compared to,

- Four variants of Learning adversarially fair and transferable representations (Laftr) (Madras et al., 2018) i.e., Laftr-DP, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-EqOdd, which use adversarial learning to achieve group fairness by leveraging different fairness criteria,
- Two variants of Conditional learning of fair representations (Cfair) (Zhao et al., 2020) i.e., Cfair and Cfair-EO, which perform conditional alignment of representations to achieve accuracy-fairness trade-off, and lastly,
- Flexibly fair representation learning by disentanglement (Ffvae) (Creager et al., 2019), which disentangles latent representations into sensitive and non-sensitive features.

For simplicity, we introduce the following shorthand notation. We denote by $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{0,1\}$ the set of inputs and outputs, respectively. Two sets of random variables X and Y take associated values $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Variable y determines eligibility, and we denote the sensitive binary variable by S taking values $s \in \{0,1\}$. Moreover, p indicates the output probability of the classifier as a real number and $\hat{Y}$ is the class predicted random variable based on p, which takes values $\hat{y} \in \{0,1\}$ using a threshold of 0.5. Furthermore, the term $\mathcal{D}_s^y$ is the conditional distribution of the joint distribution $\mathcal{D}$ over $X \times Y \times S$, given $Y = y$, $S = s$.

3.1. Datasets

In our evaluations and tests, we employ two datasets: a synthetic dataset to allow for controlled data modification and generation, and a real dataset to ensure that the results gained are applicable to real data.

CI-MNIST We created a variation of the MNIST dataset (LeCun and Cortes, 2010) in order to test bias-mitigation algorithms in difficult settings and to be able to manage multiple dataset

configurations, called *Correlated and Imbalanced MNIST* or in short CI-MNIST , where we introduce different types of correlations between attributes, dataset features, and an artificial eligibility criterion. For an input image x, the label $y \in \{1,0\}$ indicates eligibility or ineligibility, respectively, given that x is even or odd. We define the background colors (bck) as the protected or sensitive attribute $s \in \{0,1\}$, where blue denotes the unprivileged group and red denotes the privileged group.

The primary aim for employing this dataset is to replicate more connected and imbalanced datasets by controlling different aspects of the data production process. We can analyse models in setups with varying amounts of bias and measure their resilience this way.

To construct such difficult scenarios, we leverage the following dataset components in particular.

- **clr-ratio**: We represent it as (b_e, b_o) pair and it is used to refer to the amount of images in the dataset with blue backgrounds in (even, odd) classes as a percentage. When background is used as a sensitive attribute, the pair (b_e, b_o) denotes the percentage of unprivileged population in (eligible, ineligible) groups, as blue is used for unprivileged group and even and odd indicate respectively eligibility and ineligibility. The rest of the digits in each group have red backgrounds. Using this feature, one can control the correlation between the sensitive attribute (background color) and the eligibility (being even or odd) or create imbalanced datasets between under-represented and over-represented groups.
- **pos-ratio**: Denoted as (l_e, l_o) pair where l_e (l_o) refers to the percentage of even (odd) digits in the dataset with a small box in the top-left half of the image. The rest of the digits have the box in the top-right half. Figure 3.1 depicts image samples with such small boxes.

Note that when pos-ratio is (1, 0) the small box is always on the top-left for even digits and always on top-right for odd digits, so its location indicates the eligibility of the sample (being even or odd), yielding the completely correlated setup. When the pos-ratio is (0.5, 0.5), the little box is on the top-left of the image in 50% of the samples and on the top-right of the image in the remaining 50% of the samples, resulting in a perfectly decorrelated configuration. In order to generate correlated setups, one can manage the degree of correlation between eligibility and the position of the box as a non-sensitive or sensitive property by transitioning between these two extreme examples.We don't use a correlation between the location of little boxes and the colour of the background.

While training sets can be imbalanced, we use a balanced test set, meaning clr-ratio=(0.5, 0.5) and pos-ratio=(0.5, 0.5) in order to evaluate results on balanced dataset sub-groups and ease comparison between different setups.

Adult dataset. We use the Adult database (Dua and Graff, 2017) because of its widespread use in the community to verify that the configurations examined on the proposed synthetic dataset

carry over to more actual instances. (Madras et al., 2018, Zhao et al., 2020, Zhang et al., 2018). The Adult dataset consists of 30,162 training and 15,060 testing samples of individuals with 112 features such as gender, age, and nationality. Given the features of an adult, the prediction task is to determine whether the person makes over 50,000 (eligible) or not (ineligible). The binary eligibility value y is defined by salary being over or less than 50,000.

To generate configurations on the Adult dataset with varying complexity, we consider various thresholds to binarize the sensitive attribute age, which is a multi-valued sensitive attribute and is often binarized (Zafar et al., 2017a, Kamiran and Calders, 2009, Zemel et al., 2013). Note that we evaluate the aforementioned bias-mitigation approaches on a new Adult test set, which is a subset of the Adult test set but balanced in terms of eligibility and sensitive attributes, similar to our CI-MNIST . Our goal is to assess different possible configurations of a real-world dataset in similar setups as the ones used in CI-MNIST .

- **age-ratio**: Finding settings in a real dataset that resembles the ones chosen in a synthetic dataset is not trivial since, contrary to synthetic datasets, the features of a real dataset cannot be controlled for data generation. We binarize age by thresholding on it to find correlations between this sensitive attribute and the eligibility in the Adult dataset that are closest to the ones we used in the CI-MNIST dataset. We call this feature age-ratio and show it in pair (a_e, a_i), which refers to the ratio of people who are unprivileged in (eligible, ineligible) classes. The rest of the people in each class are considered as privileged.

3.2. Experimental Setup

3.2.1. CI-MNIST dataset

Figure 3.2 shows samples of the dataset used in our experiments. Unless otherwise stated, we used 50,000 images for the training set, 10,000 images for each validation and test sets. In CI-MNIST experiments, the eligible and ineligible groups represent each 50% of the training data in both train and test sets. However, while the train set can be imbalanced with respect to sensitive attributes, the test set is always balanced. We initially pad the input image of size 28x28 on the top and sides to give a 32x32 image for the dataset creation. Blue and red colors with a 10% gaussian noise are used as background colors. For the small box, we used a 4x4 sized gray-colored box in the center of the top-left half or top-right half of the padded region of the image. We have experimented with multiple background colors, box colors and box sizes to understand the impact of colors, positions, sizes of the features on our models. Our motivation in choosing the current features is that the models can easily notice these features but find them difficult to remove.

CI-MNIST currently supports multiple sensitive features (multiple background colors and positions of boxes). For more details refer to the released codebase.

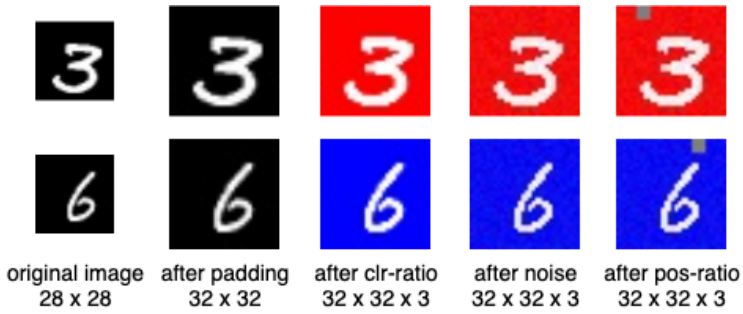


Fig. 3.1. The conversion process used to generate samples. Input image is first padded to become 32x32. The attributes *clr-ratio* is then applied, which decides the background color (blue or red). Noise is then added to the background color. Finally, *pos-ratio* affects the positioning of the box (top row, left half or top row, right half).

3.2.2. Adult dataset

For training on the Adult dataset, we used the full Adult training set consisting of a total of 30,162 samples. We used 20% of the training set as the validation set. In the Adult dataset, the eligible and ineligible groups represent each 25%, 75% of the training data. Hence the data is imbalanced with respect to target label with a skew towards ineligible lower-income class (<=50k). We use age as sensitive attribute and threshold on it in the training sets (as indicated in Table 3.1) to create various *age-ratios*. Note that in all *age-ratios* the size of the training dataset does not change, and the eligible and ineligible groups remain at 25% and 75%. However, through thresholding on age as the sensitive attribute, we change the number of people in privileged and unprivileged groups. We find age-thresholds that closely resemble the dataset ratios used in CI-MNIST , in terms of the percentage of unprivileged individuals in eligible and ineligible groups.

As our original test set is modified to be balanced in terms of both sensitive attributes (*age-ratio*) and target classes, the number of testing samples vary for each *age-ratio*. This is because we use age for the sensitive attribute, and when we change its threshold, the number of examples belonging to unprivileged and privileged group changes. We drop the minimum number of samples from the bigger subgroup of sensitive attribute and target class to make the dataset balanced. In Table 3.1, we mention the age thresholds used for the unprivileged group to achieve our desired *age-ratios* and also indicate the test-set size in each case. The remaining ages in either eligible and ineligible groups are considered privileged.

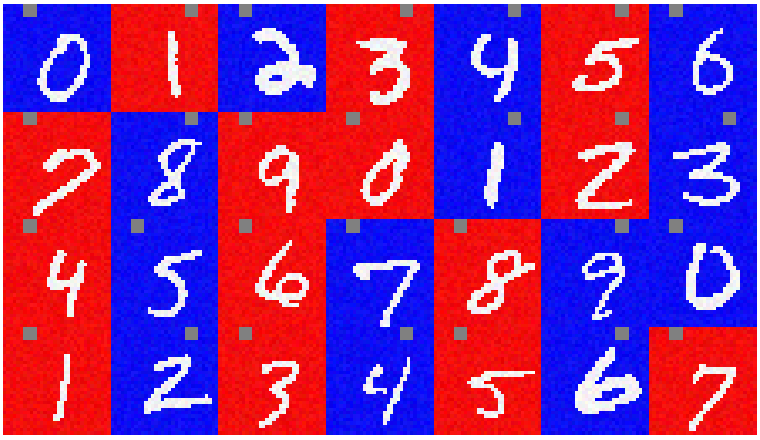


Fig. 3.2. Sampled images from our dataset.

age-ratio	unprivileged age threshold	test set size
(0.5, 0.5)	25 <= age < 44	7,336
(0.1, 0.1)	32 <= age < 36	1,708
(0.01, 0.01)	71 <= age < 75	208
(0.66, 0.33)	38 <= age < 60	5,480
(0.06, 0.36)	0 <= age < 30	908

Table 3.1. Thresholds of age and test set size used for various *age-ratios*. The test set is balanced in terms of both sensitive attribute and target class, while the train set is imbalanced and is of fixed size 30,162.

3.2.3. Privacy and author consent

CI-MNIST : This data is an extension of the publicly available MNIST dataset (LeCun and Cortes, 2010), which does not contain any personal data. Yann LeCun and Corinna Cortes hold the copyright of MNIST dataset, which is a derivative work from original NIST datasets. MNIST dataset is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license (CC BY-SA 3.0).

Adult: The Adult dataset was originally extracted by Barry Becker from the 1994 Census bureau database and the data was first cited in (Kohavi, 1996). It was donated by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics) and publicly released to the community on the UCI Data Repository (Dua and Graff, 2017). It is licensed under Creative Commons Public Domain (CC0).

We do not put these datasets in our repository; instead, we provide code and guidelines on processing the original dataset to obtain the dataset variants used in our experiments.

3.2.4. Architecture

In all models, we used three fully connected layers for discriminator and classifier networks. In the encoder network, we used three fully connected layers for all models except Ffvae, in which we used a convolutional encoder and decoder networks for increased training stability (Kim and Mnih, 2018).

Baseline MLP Setup: The baseline MLP model consists of an encoder for the input image and a classifier for eligibility prediction. Cross entropy loss is used for optimization.

Baseline CNN Setup: The baseline CNN model consists of a CNN encoder for the input image and an MLP classifier for eligibility prediction. Cross entropy loss is used for optimization.

Laftr Setup: The model consists of an encoder, a classifier, and a discriminator. We used an adapted PyTorch version of the original codebase released by the authors of the original paper (Madras et al., 2018). Following the original code's training method, we train the encoder, classifier and train the discriminator in alternate steps. We used two discriminator iterations per encoder-classifier iteration and applied cross-entropy loss for optimization of both the classifier and discriminator. We used the default classification coefficient of 1.0 and used five values of adversarial coefficient $\gamma \in [0.1, 0.5, 1, 2, 4]$, as proposed in the original paper.

Cfair Setup: The model consists of an encoder, a classifier, and two discriminators (one for each eligibility class label). We used the code provided by the authors to run the experiments. We experimented with five values of adversarial coefficient $\gamma \in [0.1, 1, 10, 100, 1000]$, as proposed in the original paper. The binary loss (0-1 loss) in Eq.2.5.7 is NP-hard to optimize directly (Feldman et al., 2009, Ben-David et al., 2003), hence the model uses a convex relaxation of the binary loss, which is a weighted cross-entropy loss as shown below.

$$\begin{aligned} \mathcal{D}(\widehat{Y} \neq y \mid Y = y) =& \frac{\mathcal{D}(\widehat{Y} \neq y, Y = y)}{\mathcal{D}(Y = y)} \\ \leq& \frac{\mathrm{CE}_{\mathcal{D}^y}(\widehat{Y} \| Y)}{\mathcal{D}(Y = y)} \end{aligned} \tag{3.2.1}$$

Hence we can relax the optimization problem to a cost-sensitive cross-entropy loss minimization problem, where the weight for each class is given by the inverse marginal probability of the corresponding class. This allows us to equivalently optimize the objective function without explicitly computing the conditional distributions.

Ffvae Setup: The model consists of a convolutional encoder, a convolutional decoder, a fully connected classifier, and a fully connected discriminator. We used the code provided by authors to run the experiments. We applied adversarial coefficient $\gamma \in [10, 50, 100]$ and the alignment

coefficient $\alpha \in [10, 100, 1000]$. We observed that the training of Ffvae becomes unstable for higher values of γ. This is due to the fact that the stability between *predictiveness* and *disentanglement* gets harder to achieve as they work against each other when the sensitive attribute and the eligibility are correlated. Ffvae model takes ELBO loss for the VAE and approximates the *disentanglement* term using the mean error difference between discriminator logits (Kim and Mnih, 2018). The model uses cross-entropy loss for the *predictiveness* term and the discriminator network.

In CI-MNIST experiments, we kept the widths of encoder, decoder, discriminator constant at 32, and the encoded latent representation size is 16 for all models. We experimented with two values of classifier widths 32, 64. However, we were unable to observe the trend that is recently highlighted by (Sagawa et al., 2020) that increasing model capacities may cause unfairness towards minorities while increasing accuracy. However, this needs to be further investigated. In Adult experiments, we kept the widths, latent representation sizes the same as their respective original papers. Tables 3.2, 3.3 show the architectural details for CI-MNIST and Adult experiments.

Table 3.2. Architectures used for Baseline MLP, Baseline CNN, Laftr, Cfair, Ffvae models for CI-MNIST dataset.

MLP Encoder	MLP Classifier/Discriminator	CNN Encoder
Input $\in \mathbb{R}^{3072}$	Input $\in \mathbb{R}^{16}$	Input $32 \times 32 \times 3$ image
FC. 32 LReLU	FC. 32 LReLU	4×4 conv. 32 LReLU. stride 2, padding 1
FC. 32 LReLU	FC. 32 LReLU	4×4 conv. 64 LReLU. stride 2, padding 1
FC. 16 LReLU	FC. 2 LReLU	4×4 conv. 64 LReLU. stride 2, padding 1
		4×4 conv. 256 LReLU. stride 1
		1×1 conv. 16 LReLU.

Ffvae Encoder	Ffvae Decoder
Input: $32 \times 32 \times 3$ image	Input $\in \mathbb{R}^{16}$
4×4 conv. 32 LReLU. stride 2, padding 1	FC. 128 LReLU
4×4 conv. 64 LReLU. stride 2, padding 1	FC. 1024 LReLU, Resize $64 \times 4 \times 4$
4×4 conv. 64 LReLU. stride 2, padding 1	4×4 upconv. 64 LReLU. stride 2, padding 1
Flatten 1024, FC. 128 LRELU	4×4 upconv. 32 LReLU. stride 2, padding 1
FC. 2×16	4×4 upconv. 3 LReLU. stride 2, padding 1

For sensitive information removal experiments in Section 4.1, sensitive features are predicted from latent representations from model-specific encoders. We use the same architecture as MLP Classifier in Table 3.3 for these experiments.

Table 3.3. Architectures used for Baseline MLP, Laftr, Cfair, Ffvae models for Adult dataset.

MLP Encoder	MLP Classifier	Ffvae Encoder	Ffvae Classifier/ Discriminator
Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{16}$	Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{60}$
FC. 32 LReLU	FC. 32 LReLU	FC. 200 LReLU	FC. 200 LReLU
FC. 32 LReLU	FC. 32 LReLU	FC. 60 LReLU	FC. 2 LReLU
FC. 16 LReLU	FC. 2 LReLU		

Laftr Encoder	Laftr Classifier/Discriminator	Cfair Encoder	Cfair Classifier/Discriminator
Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{8}$	Input $\in \mathbb{R}^{112}$	Input $\in \mathbb{R}^{60}$
FC. 8 LReLU	FC. 2 LReLU	FC. 60 LReLU	FC. 0/50 LReLU
			FC. 2 LReLU

3.2.5. Hyperparameter details

Laftr: We use adversarial coefficient $\gamma \in [0.1, 0.5, 1, 2, 4]$ as hyperparameter as proposed in the original paper and we use two discriminator iterations per encoder-classifier iteration.

Cfair: We use adversarial coefficient $\gamma \in [0.1, 1, 10, 100, 1000]$ as hyperparameter as proposed in the original paper.

Ffvae: We use adversarial coefficient $\gamma \in [10, 50, 100]$ and the alignment coefficient $\alpha \in [10, 100, 1000]$ as hyperparameters as proposed in the original paper. We also use patience epochs 5 for early stopping in VAE training as a hyperparameter.

Other general hyperparameters considered for all the models include classifier, encoder, and discriminator widths, number of layers, and latent representation size with values mentioned in Tables 3.2 and 3.3. Leaky ReLU is used for all activation functions, and Glorot (Glorot and Bengio, 2010) is used to initialize all weights. The models are trained using Adam optimizer with a learning rate of 1e-3. Models are trained for 500 epochs, with early stopping of 5 epochs patience on the validation set's loss to find the best model. Please check our repository for a complete set of hyper-parameters and training setups.

We leverage the datasets introduced in section 3.1 to create challenging scenarios in which,

- We alter the balance of the under-represented group,
- Correlate sensitive attribute and eligibility, and
- Include scenarios in which small features in the image, considered either sensitive or non-sensitive attributes, are associated with eligibility.

We change the dataset from a balanced setup to varying imbalanced setups at training time in all scenarios, but we always evaluate the models in a balanced setup at test time. This allows us to report on the model's performance over a diverse variety of sub-groups. The ratio of unprivileged (blue) to privileged (red) background on CI-MNIST in the balanced setup is 50% for both eligible and ineligible groups, providing a *clr-ratio* of $(0.5, 0.5)$. Furthermore, with a *pos-ratio* of $(0.5, 0.5)$,

the small-box location and eligibility are unrelated. In the balanced configuration, the value of *age-ratios* on Adults is $(0.5, 0.5)$.

To provide results, we first average metrics across three random seeds, then report the best value over hyper-parameters for each metric on the test set for each metric. As a result, we can report the best possible results on each metric without having to compromise between accuracy and fairness (which usually have a tradeoff and improving on one metric deteriorates the other, so we show best possible outcome on each metric). This would allow us to show bias even when the best model variation is used for each metric. This is in the models' favour because the best possible outcome is provided. Hence each column can correspond to a different hyper-parameter. Next, we present results in each setting.

Notes on all settings: In settings 3 and 4, we could not emulate similar scenarios on the Adults dataset, as each feature (such as age) is one out of 112 components; hence they all occupy a single dimension. Moreover, in settings 1 and 2, we could not emulate the extreme scenarios of CI-MNIST dataset in the Adult dataset. This highlights the difficulty of emulating all scenarios in real datasets and motivates the usage of the proposed synthetic data in conjunction with real datasets.

3.3. Setting 1: Impact of reducing the representation of the unprivileged group.

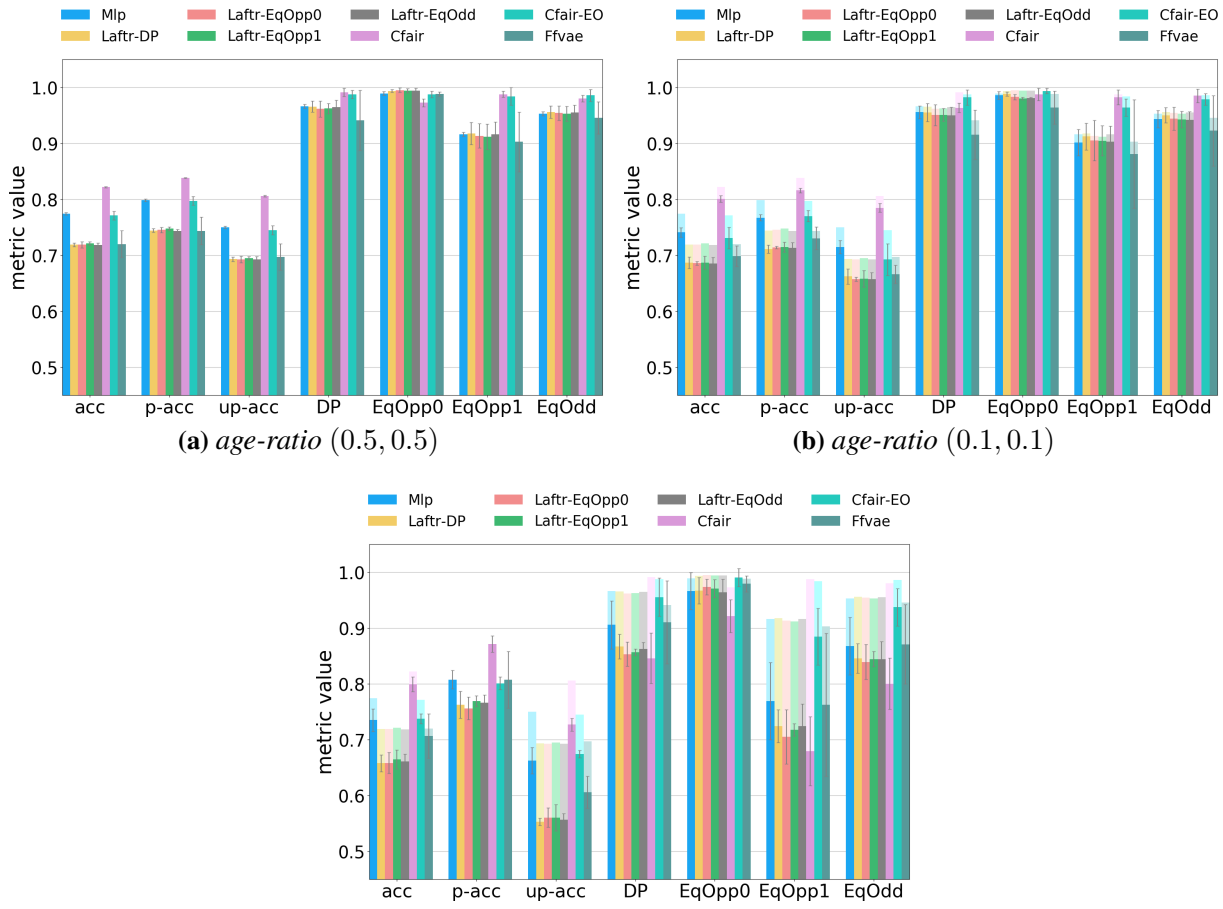
Experiment setup: We first evaluate the impact of the representation percentage of the unprivileged group in both eligible and ineligible groups. The question we want to ask is whether the models are robust to the small presence of under-represented groups. To obtain such a setup, in CI-MNIST dataset we change *clr-ratio* from the balanced setup of $(0.5, 0.5)$ to imbalanced cases of $(0.1, 0.1)$, $(0.01, 0.01)$, and $(0.001, 0.001)$. In the Adult dataset, we change *age-ratio* from $(0.5, 0.5)$ to the imbalanced cases of $(0.1, 0.1)$ and $(0.01, 0.01)$.

Model evaluation and discussion: Figures 3.3 and Figures 3.4 compare models trained on balanced and reduced representation data, both on Adult and CI-MNIST datasets. We find that when switching from balanced to imbalanced configurations, both unprivileged group accuracy and fairness measures decrease for all models. This suggests that even bias-mitigation algorithms are prone to under-representation of disadvantaged groups. We see the same trend on both CI-MNIST and Adults, with bias increasing as the datasets grow more unbalanced.

Due to the small ratio of the unprivileged group to the privileged group, this scenario generates bias. In this scenario, the bias could be attributable to two factors: one, the imbalance between the two groups, and the other, the lack of data from the underprivileged group. The *clr-ratios* of $(0.1, 0.1)$, $(0.01, 0.01)$, and $(0.001, 0.001)$, in particular, are all affected by group imbalances.

In the latter two situations, however, there are 250 and 25 samples from the unprivileged group in CI-MNIST , respectively, compared to 24,750 and 24,975 samples from the privileged group. Here, the models must address data scarcity along with the group imbalance.

We report the complete set of results for debiasing models of MLP, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, in Tables 6.1 to 6.16.

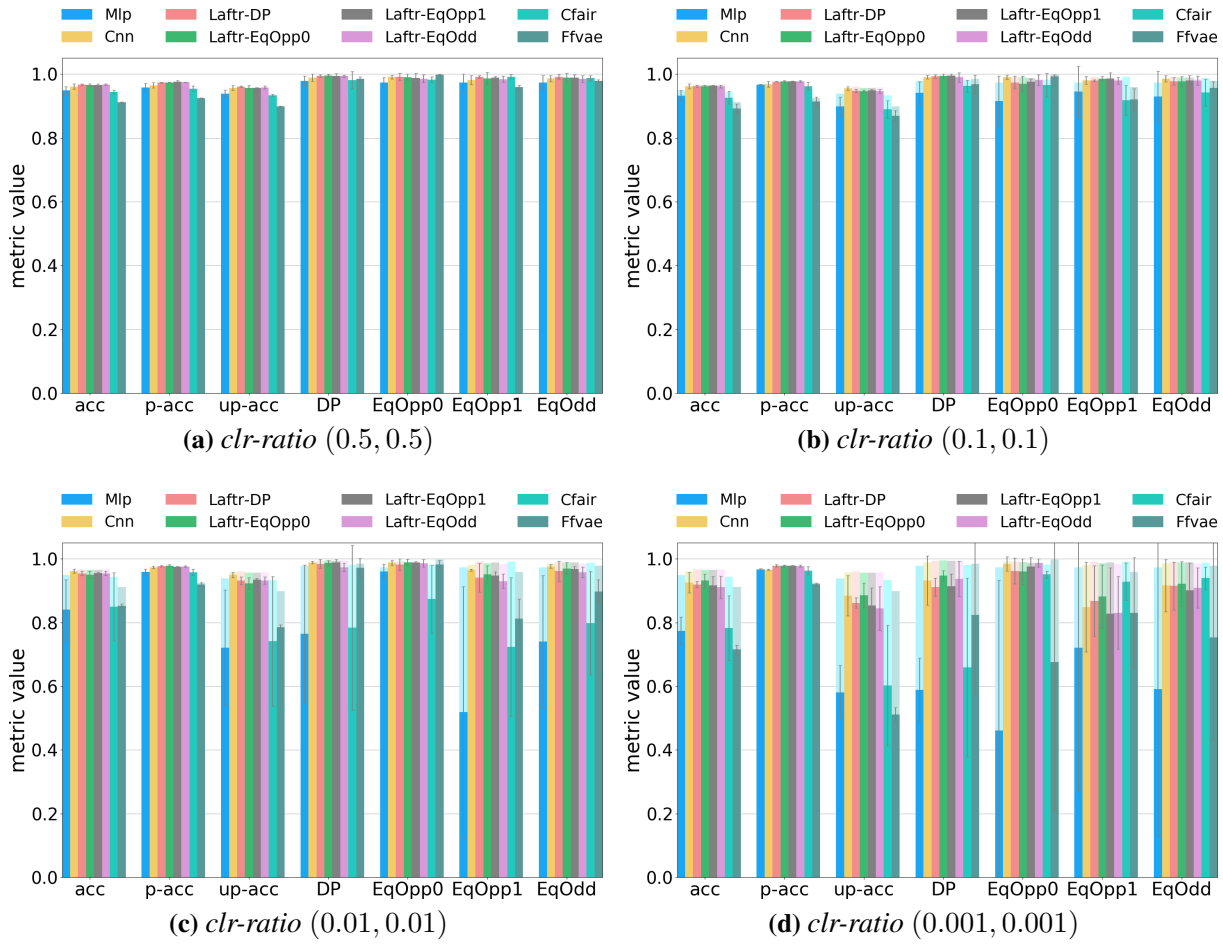


(a) *age-ratio* $(0.5, 0.5)$

(b) *age-ratio* $(0.1, 0.1)$

(c) *age-ratio* $(0.01, 0.01)$

Fig. 3.3. Comparing different models while decreasing minority representation for Adult dataset. Sub-figure 3.3a shows the balanced case. Sub-figures 3.3b and 3.3c when compared with 3.3a shows the impact of reducing unprivileged group in Setting 1. In sub-figures 3.3b and 3.3c the pale colors show the decrease in performance compared to the balanced case in 3.3a. Note that in 3.3a the models are not perfect as the performance is not always 1. Higher is better for all metrics. For each bar, the standard deviation is also shown. Best viewed in colors.



(a) *clr-ratio* $(0.5, 0.5)$ **(b)** *clr-ratio* $(0.1, 0.1)$

(c) *clr-ratio* $(0.01, 0.01)$ **(d)** *clr-ratio* $(0.001, 0.001)$

Fig. 3.4. Comparing different models while decreasing minority representation for CI-MNIST dataset. Sub-figure 3.4a shows the balanced case. Sub-figures 3.4b, 3.4c, and 3.4d when compared with 3.4a shows the impact of reducing unprivileged group in Setting 1. In sub-figures 3.4b, 3.4c, and 3.4d the pale colors show the decrease in performance compared to the balanced case in 3.4a. Note that in 3.4a the models are not perfect as the performance is not always 1. Higher is better for all metrics. For each bar, the standard deviation is also shown. Best viewed in colors.

3.4. Setting 2: Impact of correlation of sensitive attribute with eligibility.

Experiment setup: In this setting, we wish to see how resistant models are to the sensitive attribute and eligibility correlation. The objective is to see if models perform properly when such correlations are present. To that purpose, we use a training set with sensitive attribute and eligibility correlations to train models.

To provide such a scenario, in CI-MNIST we change *clr-ratio* from the balanced case of $(0.5, 0.5)$ to correlated cases of $(0.1, 0.9)$ and $(0.01, 0.99)$, and for Adult dataset we change *age-ratio* from $(0.5, 0.5)$ to $(0.06, 0.36)$, where the unprivileged group becomes under-represented in the eligible group and over-represented in the ineligible group, hence correlating eligibility and sensitive features. Note that the setting selected for the Adult dataset is the closest to the selected CI-MNIST setting that we could emulate with this data.

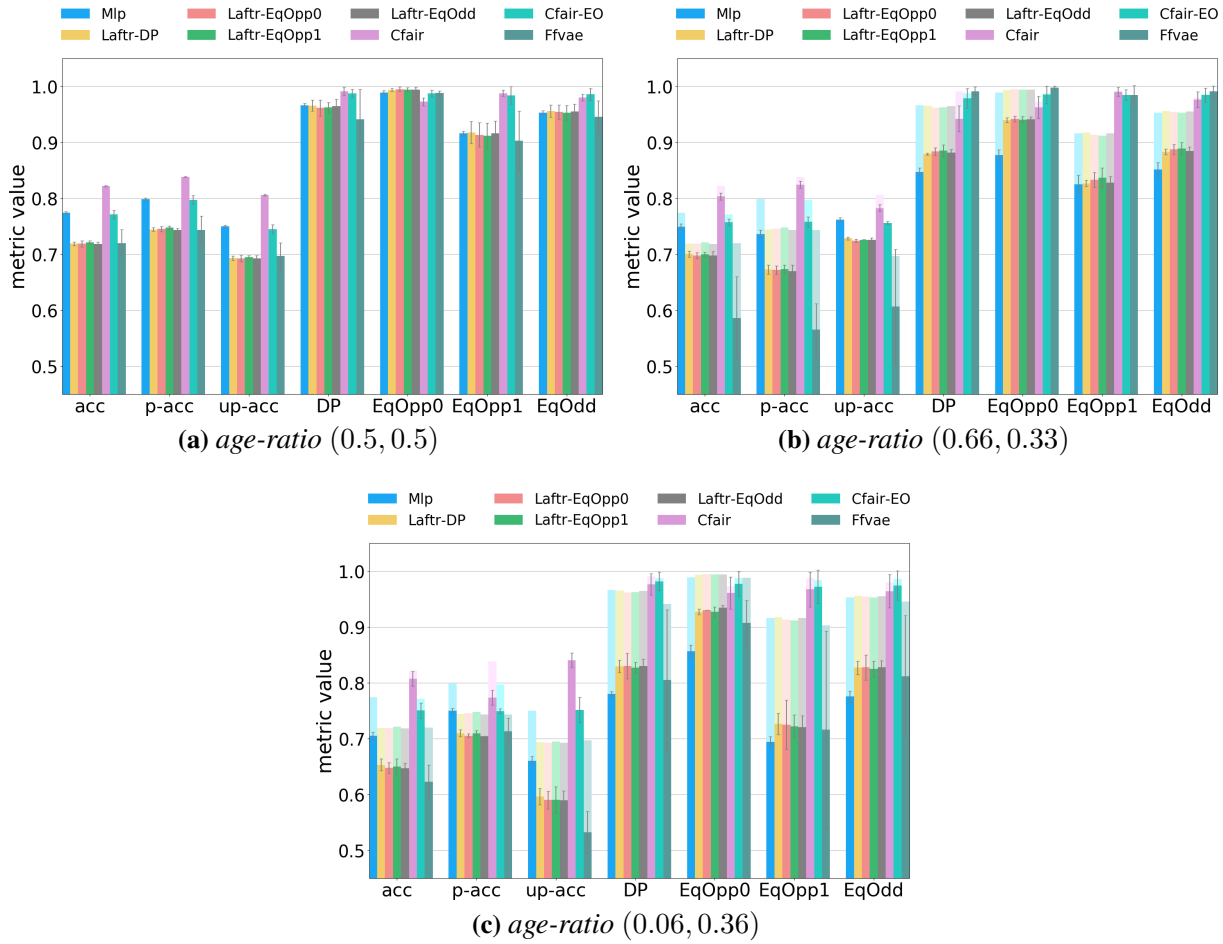
Model evaluation and discussion: Figures 3.5 and Figures 3.6 compare model performances under different metrics for the Adult and CI-MNIST datasets. For all models, we observe a large drop in both accuracy and fairness metrics on both Adult and CI-MNIST . Note that the bias increases when the correlation level increases from 3.6b to 3.6c. The drop is greater than in the preceding evaluation scenario, in which unprivileged representation was lowered. This shows that models are even more vulnerable to eligibility and sensitive attribute correlations. It's also worth noting that we see the same trends in CI-MNIST in Settings 1 and 2 as we do in Adults, demonstrating the data's transferability.

We report the complete set of results for debiasing models of MLP, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, in Tables 6.17 to 6.32.

3.5. Setting 3: Impact of correlation of non-sensitive attribute with eligibility.

Experiment setup: In this setting, we want to see how well the models exploit the relationship between non-sensitive and non-predominant attributes and eligibility. To evaluate this setting on CI-MNIST , we keep *clr-ratio* at $(0.5, 0.5)$; however, we change *pos-ratio* from the balanced setting of $(0.5, 0.5)$ to the imbalanced settings of $(0.9, 0.1)$. Note that this would make the position of box highly correlated with eligibility.

Model evaluation and discussion: Figures 3.7a and 3.7c compare different models under this setting. In this case, since *pos-ratio* is a non-predominant non-sensitive feature compared to *clr-ratio*. We notice a slight non-consistent drop in fairness measurements among models indicating that the models are only slightly biased, but we do see a consistent decrease in accuracy metrics. Fairness measurements are calculated using the sensitive feature, in this case, the backdrop color.



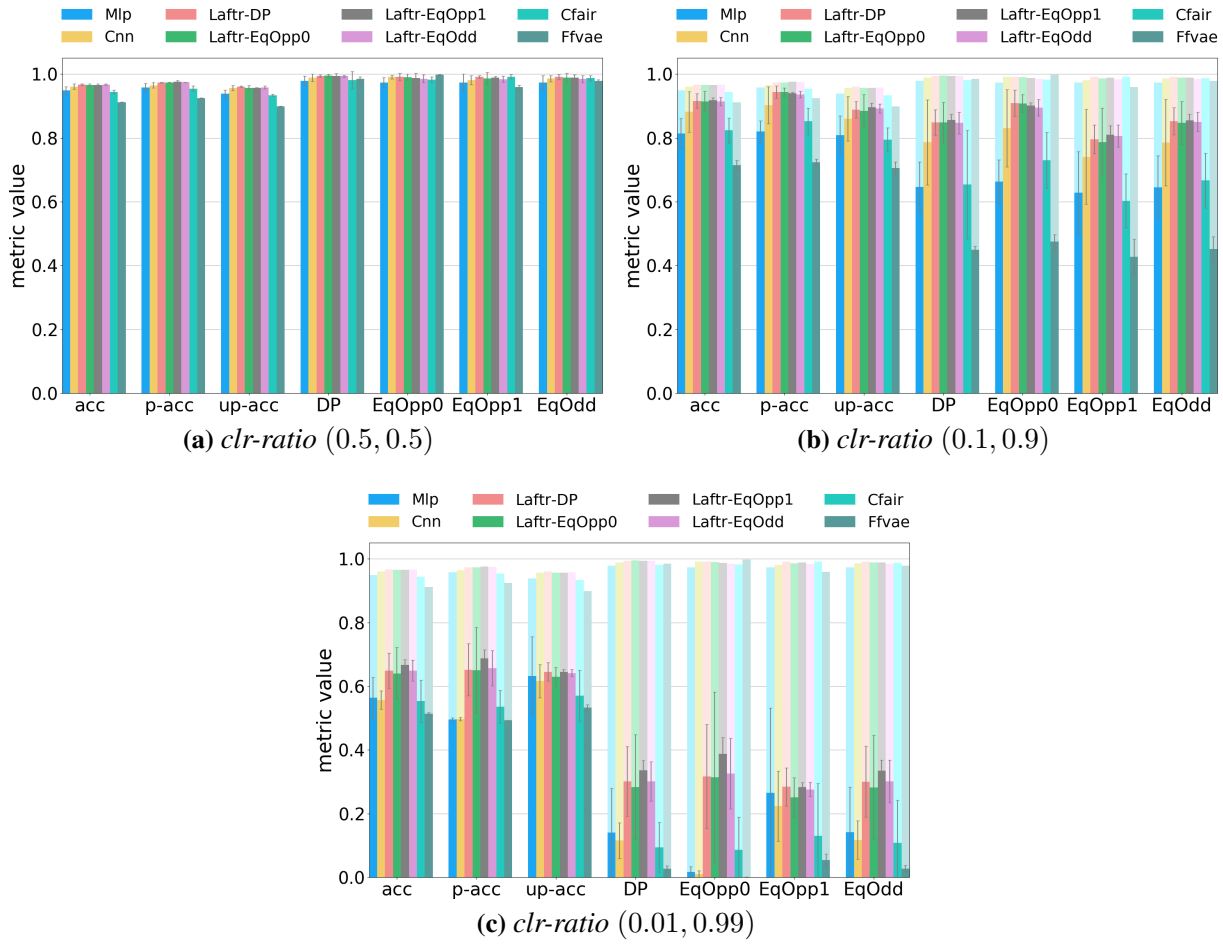
(a) *age-ratio* $(0.5, 0.5)$ **(b)** *age-ratio* $(0.66, 0.33)$

(c) *age-ratio* $(0.06, 0.36)$

Fig. 3.5. Comparing different models while shifting correlation of sensitive attribute (age) with the eligibility for Adult dataset. Sub-figure 3.5a shows the balanced case. Sub-figures 3.5b and 3.5c when compared with 3.5a shows the impact of correlation of sensitive attribute and eligibility in Setting 2. In sub-figures 3.5b and 3.5c the pale colors show the decrease in performance compared to the balanced case in 3.5a. Note that in 3.5a the models are not perfect as the performance is not always 1. Higher is better for all metrics. For each bar, the standard deviation is also shown. Best viewed in colors.

3.6. Setting 4: Impact of correlation of non-predominant features with eligibility

Experiment setup: When a non-predominant, non-sensitive trait was correlated with eligibility in the prior study, we found no significant bias. In this setting, we want to see if models are biased when a non-predominant trait that is correlated with eligibility becomes a sensitive attribute. It's worth noting that bias-mitigation models use a sensitive attribute as an input, try to remove bias by removing the sensitive attribute, and calculate fairness metrics for the sensitive attribute in question.



(a) *clr-ratio* $(0.5, 0.5)$ **(b)** *clr-ratio* $(0.1, 0.9)$

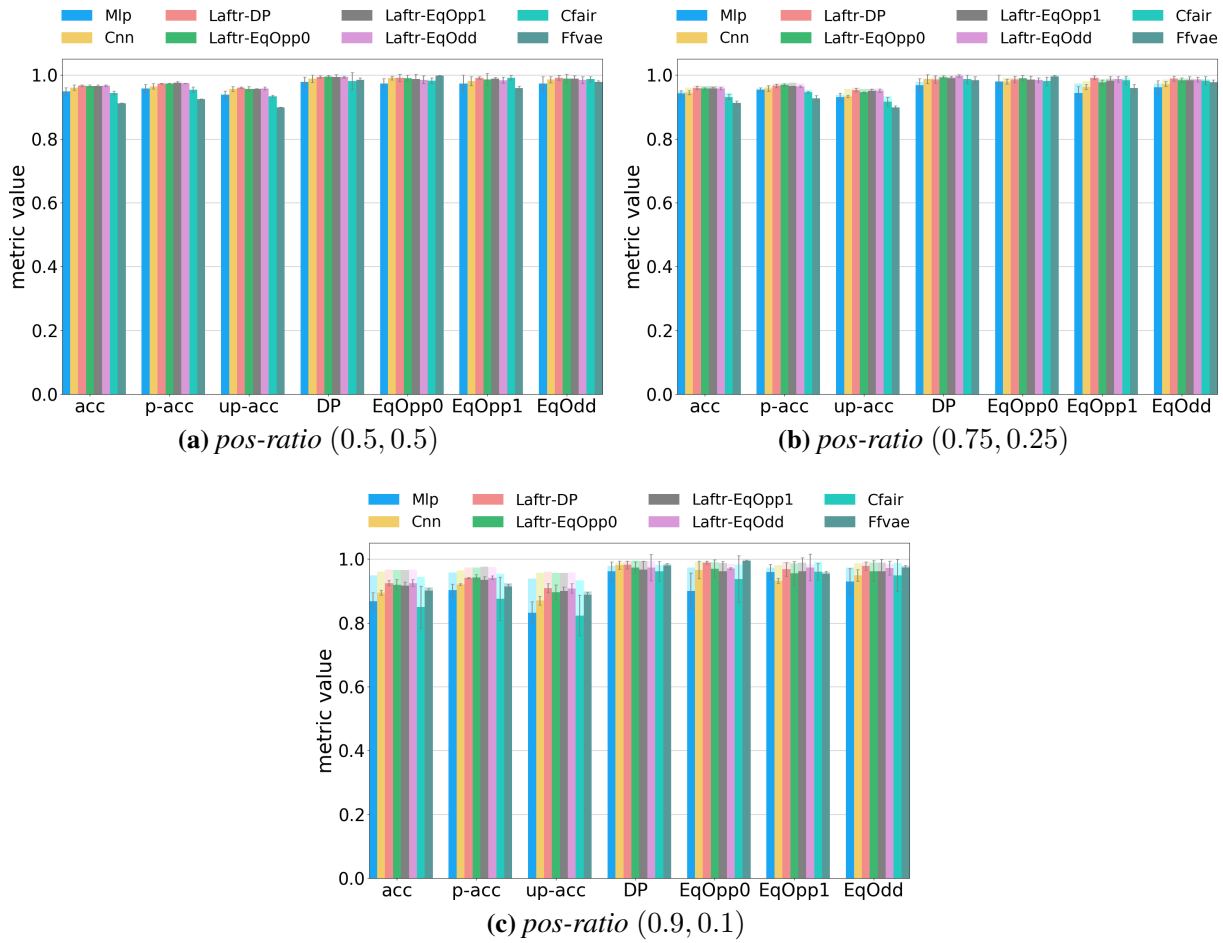
(c) *clr-ratio* $(0.01, 0.99)$

Fig. 3.6. Comparing different models while shifting correlation of sensitive attribute (bck) and the eligibility for CI-MNIST dataset. Sub-figure 3.6a shows the balanced case. Sub-figures 3.6b and 3.6c when compared with 3.6a shows the impact of correlation of sensitive attribute and eligibility in Setting 2. In sub-figures 3.6b and 3.6c the pale colors show the decrease in performance compared to the balanced case in 3.6a. Note that in 3.6a the models are not perfect as the performance is not always 1. Higher is better for all metrics. For each bar, the standard deviation is also shown. Best viewed in colors.

We use the same configuration as in Setting 3, but change the sensitive attribute from background color to the small box location, optimize model and calculate metrics for the new sensitive attribute.

Model evaluation and discussion: Figures 3.8a and 3.8c compare the performance of different models. As we can observe, the performances of most of the models drop, except that of Cfair, which mostly maintains its previous fairness performance. The obtained results indicate that correlation of even non-predominant sensitive features (i.e., small boxes) with eligibility can cause bias, although the impact is lower compared to Setting 2, where the predominant component (background color) was correlated with eligibility. Hence, the level of the predominance of the sensitive attribute can change the degree of bias of the trained models.

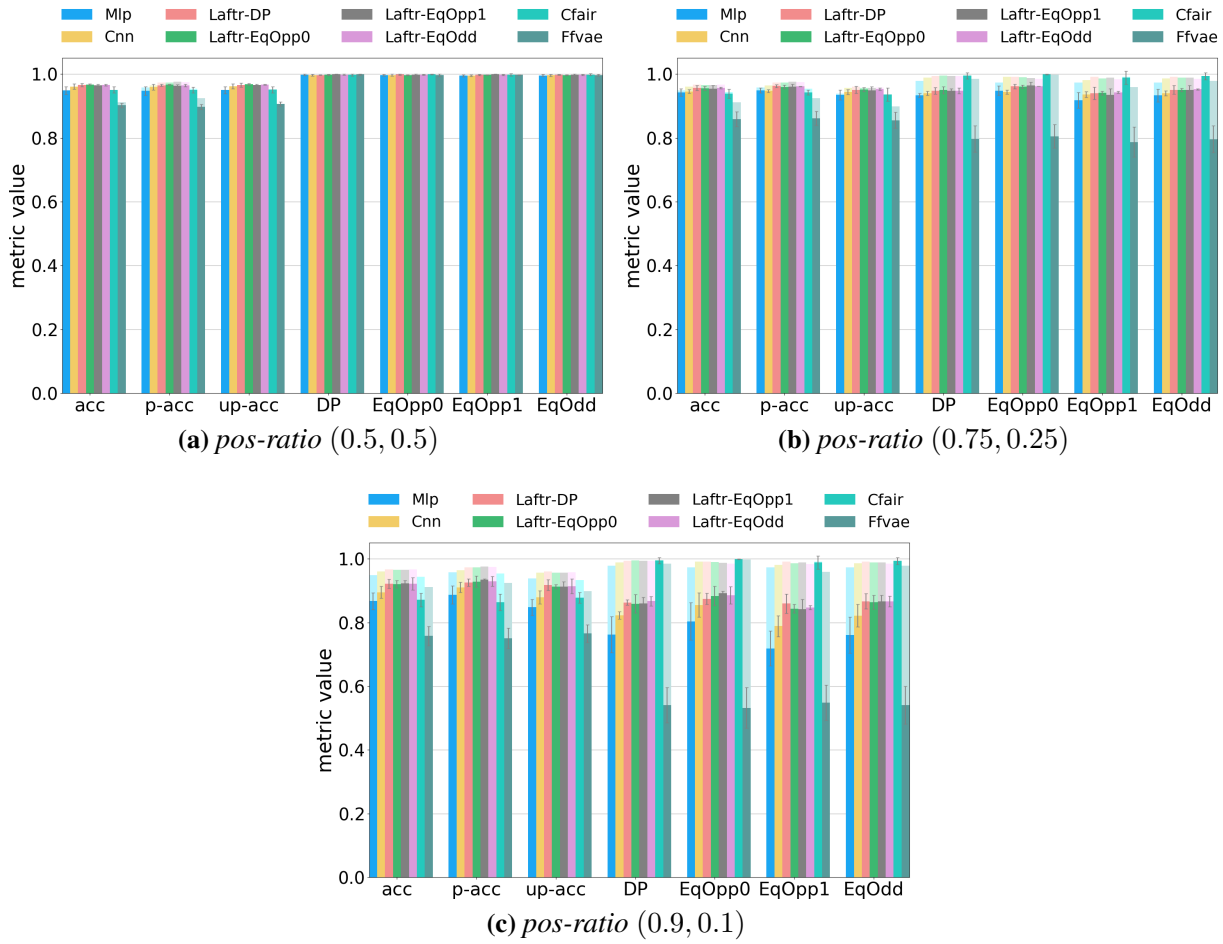
Results are depicted in Figure in Tables 6.41 to 6.48, corresponding to the experimental setup described in Setting 4. The results are obtained by comparing the baseline model with the debiasing models of MLP, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, when the position and a small sized feature of the image correlates with eligibility. Figure 3.8 compares all models side-by-side.



(a) *pos-ratio* $(0.5, 0.5)$ **(b)** *pos-ratio* $(0.75, 0.25)$

(c) *pos-ratio* $(0.9, 0.1)$

Fig. 3.7. Comparing different models while shifting correlation of a non-sensitive attribute and the eligibility for CI-MNIST dataset. Sub-figure 3.7a shows the balanced case. Sub-figures 3.7b and 3.7c when compared with 3.7a shows the impact of shifting correlation of a non-sensitive attribute and the eligibility in Setting 3. In sub-figures 3.7b and 3.7c the pale colors show the decrease in performance compared to the balanced case in 3.7a. Note that in 3.7a the models are not perfect as the performance is not always 1. Higher is better for all metrics. For each bar, the standard deviation is also shown. Best viewed in colors.



(a) *pos-ratio* $(0.5, 0.5)$

(b) *pos-ratio* $(0.75, 0.25)$

(c) *pos-ratio* $(0.9, 0.1)$

Fig. 3.8. Impact of position and small visual components on different models' performance for CI-MNIST dataset. Sub-figure 3.8a shows the balanced case. Sub-figures 3.8b and 3.8c when compared with 3.8a shows the impact of position and small visual components in Setting 4. In sub-figures 3.8b and 3.8c the pale colors show the decrease in performance compared to the balanced case in 3.8a. Note that in 3.8a the models are not perfect as the performance is not always 1. Higher is better for all metrics. For each bar, the standard deviation is also shown. Best viewed in colors.

3.7. Reproducibility

We have released our code at https://github.com/charan223/FairDeepLearning. We have provided the instructions to reproduce all experiments reported. Model, dataset, architectural, and hyper-parameter details are presented in Section 3.2. The detailed results on all experiments are presented in Section 6.1. For each chart, we provide confidence-interval around the mean and run each experiment with three different seeds. We trained about 3000 models on 2 16GB NVIDIA Tesla V100 GPUs for 14 days.

Chapter 4

Discussion

4.1. Model Stability and Performance

4.1.1. Variation due to random seeds

The random seed is a possible source of variation in deep learning models because it is used in initializing model parameters and data sampling during training, which can cause the model to converge to various solutions over time. We trained each model three times using three random seeds to analyse the impact of the random seed. We reported the mean and standard deviation of the results – see Figures 4.1 and 4.2 for all of the settings described in our experiments. We discovered that the seed we chose had a greater impact on some models than others. Almost all models demonstrate instability regarding change of results given different seeds, depending on the experiment; however, Laftr models show better stability throughout all experiments. In our bias-mitigation strategy, we tested with different seeds to assess the model's susceptibility to minor training differences. This is complementary to other studies of stability, such as cross-validation studied in (Friedler et al., 2019). In Figures 4.1 and 4.2 we illustrate the standard deviation of all models for all of the experiments of Adult and CI-MNIST datasets.

4.1.2. Correlation between dataset features and model's prediction.

If a model's prediction is correlated with dataset features, the model could rely on correlations in the data when making predictions and thus be unfair when applied. To verify this, we measured the correlation between dataset features and fairness metrics for each model using Spearman correlation matrices, measured separately for each setting. The results are presented in Figure 4.3. We observe that all models are subject to bias as their predictions correlate with dataset features; however, Cfair is slightly less biased in capturing correlations in some setups.

In Figure 4.3 we present Spearman Correlation plots for each dataset and each setting of the experiments. The correlation plot of Setting 3 illustrates that the models are still picking up on the

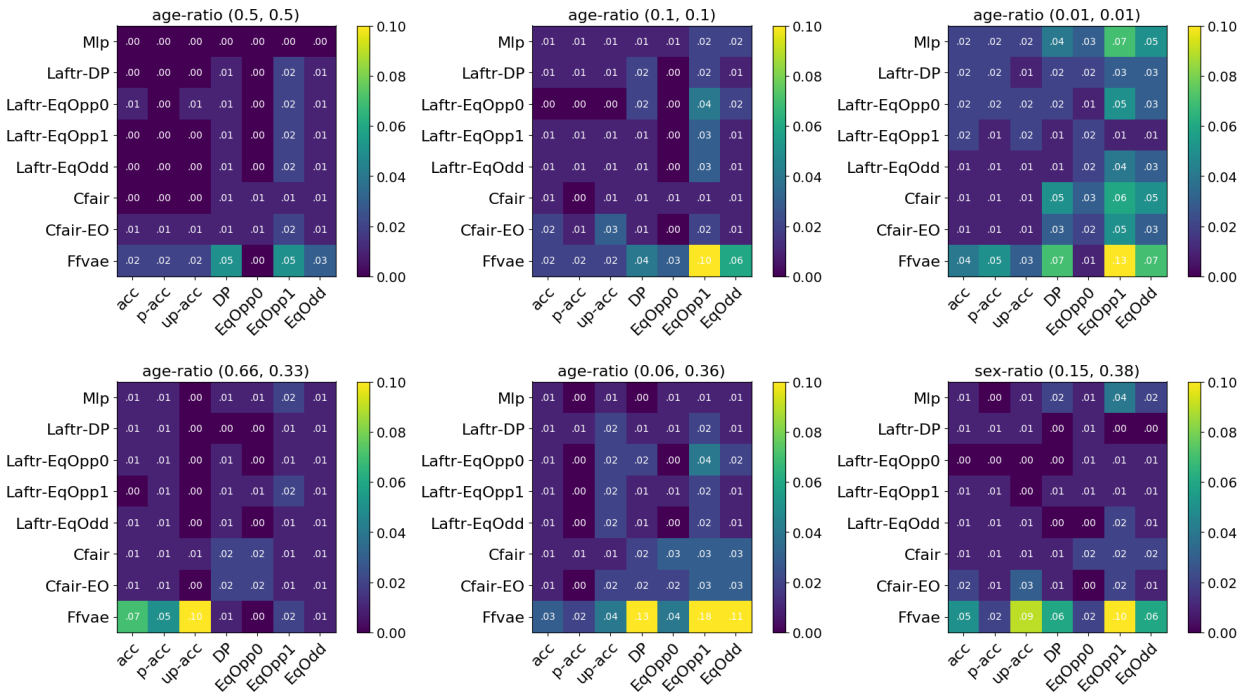


Fig. 4.1. Standard deviation of different fairness metrics (x-axis) in different models (y-axis) over three seeds for Adult dataset. Each plot corresponds to a different experimental setup presented in Section 3.3.

association between the non-sensitive feature and eligibility, even though there is no substantial bias in this setting. This shows that, even though models remain biased towards unobserved but possibly sensitive features, superior fairness results in such circumstances are partly due to how fairness metrics are derived (w.r.t. the selected sensitive features, which in this case was the background attribute).

4.1.3. Sensitive information removal

The researched bias-mitigation algorithms strive to eliminate sensitive data from the latent space. One natural question is: how successful are models at accomplishing this goal? We freeze the model's parameters after training and train a classifier to predict the sensitive attribute class using the model's latent features to see if the sensitive information is still present in the latent representation. For each configuration in Section 3, we train such a classifier on its training set and report the results on the test set, which is shown in sens-acc column in Tables 6.1 to 6.48 of Settings 1 to 4.

We see a few interesting patterns: First when the sensitive attribute is dominant, the latent representation of models contains more information about the sensitive attribute than when the sensitive attribute is less prominent. This may be seen when the sensitive property in CI-MNIST

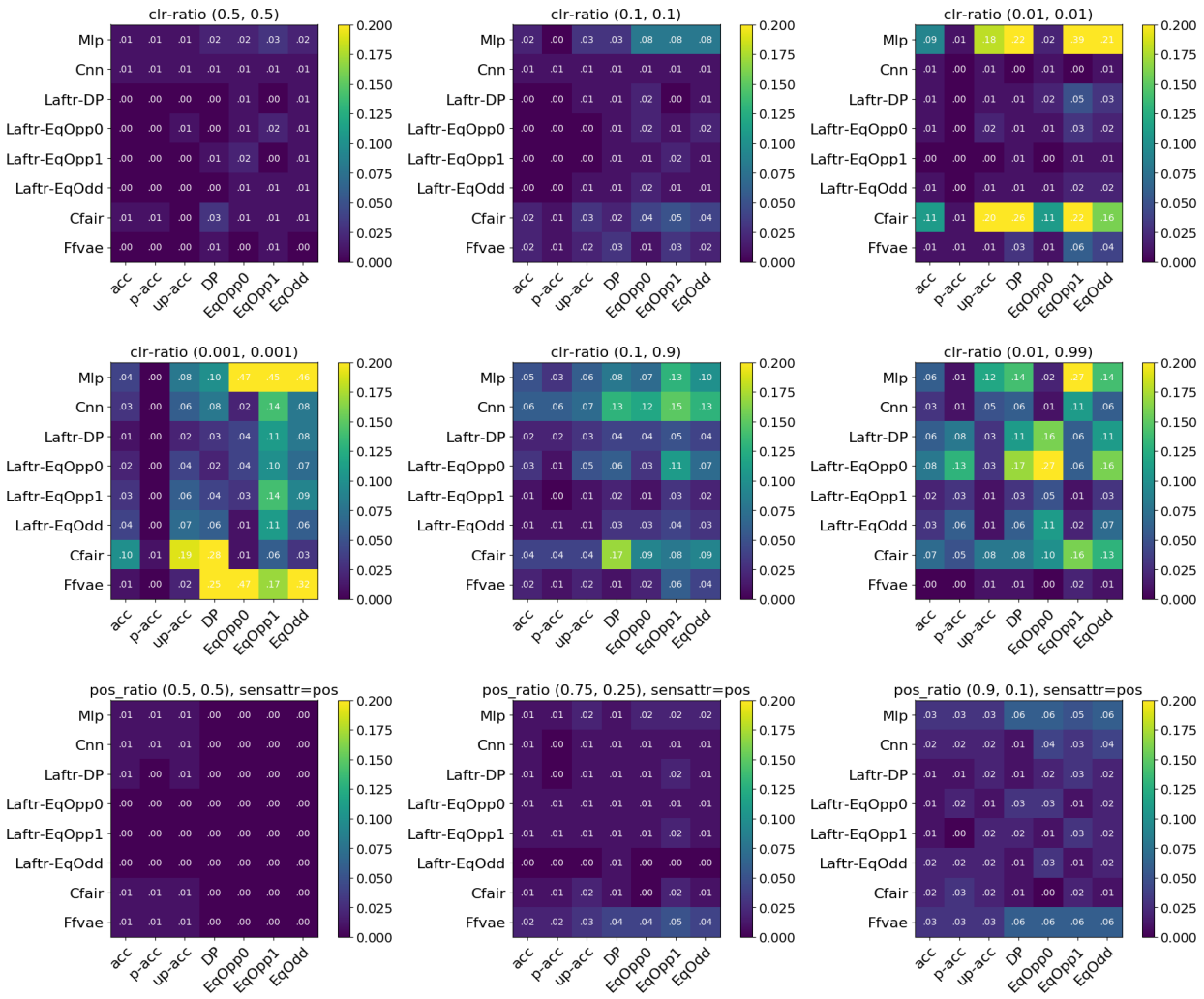
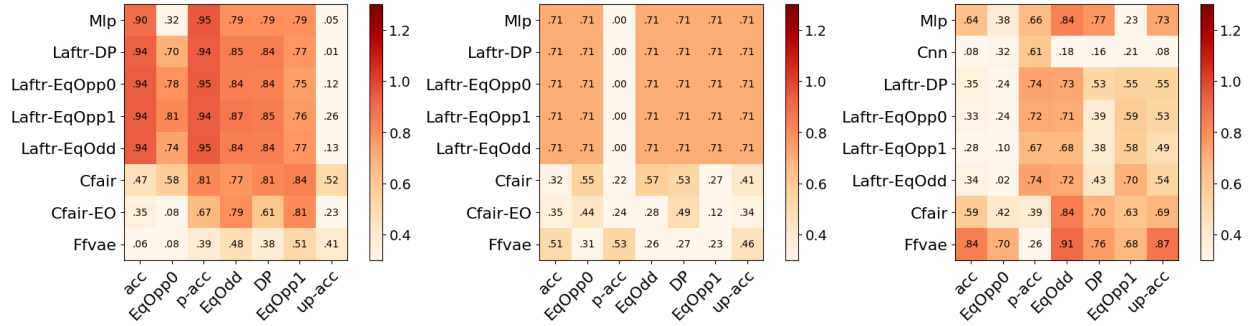


Fig. 4.2. Standard deviation of different fairness metrics (x-axis) in different models (y-axis) over three seeds for CI-MNIST dataset. Each plot corresponds to a different experimental setup presented in Section 3.3.

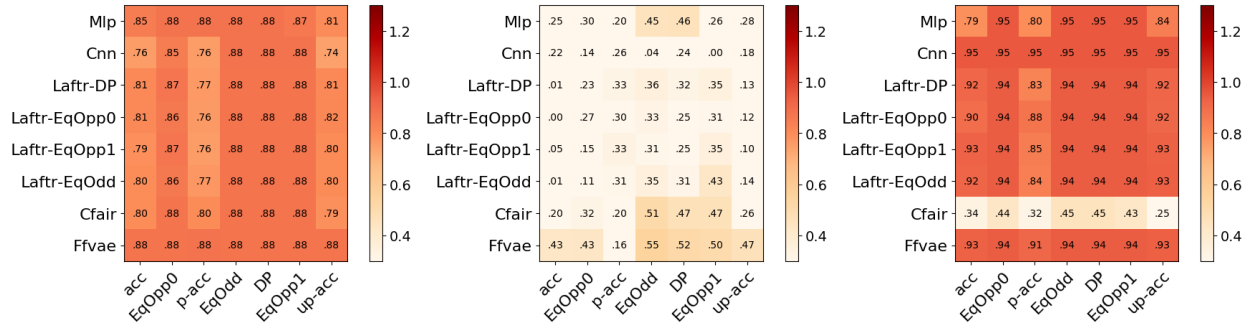
is the backdrop colour in settings 1, 2, and 3, as opposed to the small box in configuration 4. Because age is one of 112 features of the Adult, it was relatively simple for models to remove the sensitive information. Second, even though bias-mitigation approaches outperform the baselines (MLP, CNN) in bias mitigation, we find that sensitive information is more significant in the latent representation of models trained with bias-mitigation algorithms in many cases. We believe that, rather than completely deleting sensitive information from their latent space, the success of these bias-mitigation approaches in generating fairer results is due to the sensitive group balance enforced through their loss functions.



(a) Adult, Setting 1, showing correlation for *age-ratio* attribute with fairness metrics

(b) Adult, Setting 2, showing correlation for *age-ratio* attribute with fairness metrics

(c) CI-MNIST, Setting 1, showing correlation for *clr-ratio* attribute with fairness metrics



(d) CI-MNIST, Setting 2, showing correlation for *clr-ratio* attribute with fairness metrics

(e) CI-MNIST, Setting 3, showing correlation for *pos-ratio* attribute with fairness metrics

(f) CI-MNIST, Setting 4, showing correlation for *pos-ratio* attribute with fairness metrics

Fig. 4.3. Each plot depicts correlation of one dataset attribute with fairness metrics for one setting and one dataset in Section 3. On the Adult dataset, we depict the correlation of *age-ratio* with fairness metrics as this attribute has been the sensitive feature that is changed in the experiments. On CI-MNIST , in Settings 1 and 2, we depict *clr-ratio*, and in Settings 3 and 4, we show *pos-ratio*, hence showing only the feature that is changed from the balanced case. Note that contrary to other cases, in Setting 3 *pos-ratio* is not the sensitive attribute, and background is the sensitive attribute. We plot the absolute Spearman correlation metric, where we use absolute difference of the dataset attribute from the balanced case (0.5) as input to the Spearman function. This is because numbers 1 and 0 have a similar meaning as they are equally away from the balanced case. Finally, we report absolute averaged correlation values over all instances. Values range in $[0, 1]$, where one indicates maximum correlation. Almost all bias-mitigation models suffer from not mitigating the strong correlation between the overall accuracy and the sensitive attribute.

4.1.4. Model Performance

While model performance was generally close, we often observed that Cfair or Laftr performed slightly better than other models, depending on the setting. It's worth noting that, unlike Ffvae, both models apply fairness criteria directly to their learned latent spaces rather than disentangling sensitive and non-sensitive components of their latent representations. We found that disentangling characteristics was less effective in removing sensitive data. This is especially evident in the Adults

dataset's setting 1 and 2 results – see Tables 6.4 and 6.20, respectively, and in setting 4 of the CI-MNIST dataset – see Table 6.44), as we observe sens-acc is higher in the non-sensitive latent space of Ffvae compared to Cfair and Laftr.

Since bias mitigation applied to a unified latent space and explicit latent space disentangling constitute two prominent bias-mitigation strategies, a natural question is whether these strategies can benefit from one another. To verify this, we merged both strategies and investigated the effect on the models' performances. In one setup, we merged Laftr-DP with Ffvae, and in another, we merged Cfair with Ffvae. Check Section 4.1.5 for details of merging models. We evaluated these merged models in settings 1 and 2 of the Adults dataset. Tables 4.1 to 4.4 present results for the merged models.

4.1.5. Merging bias-mitigation algorithms.

Tables 4.1 and 4.2 show results for merging Cfair and Ffvae and Tables 4.3 and 4.4 show results for merging Laftr and Ffvae.

To merge Ffvae with Laftr we added to Ffvae objective in Eq.(2.5.9), the $\mathcal{L}_{DP}^{Laftr}$ term in Eq.(2.5.4), yielding

$$\mathcal{L}_{DP}^{Ffvae-Laftr} = \mathcal{L}_{\text{Ffvae}}(p,q) - \eta\mathcal{L}_{DP}^{Laftr} \tag{4.1.1}$$

Similarly, to merge Ffvae with Cfair we added to Ffvae objective in Eq.(2.5.9), the $\mathcal{L}_{DP}^{Cfair}$ term in Eq.(2.5.8), yielding

$$\mathcal{L}_{DP}^{Ffvae-Cfair} = \mathcal{L}_{\text{Ffvae}}(p,q) - \eta\mathcal{L}_{DP}^{Cfair} \tag{4.1.2}$$

Where η is a hyper-parameter, balancing the two losses. In both cases, the added loss ($\mathcal{L}_{DP}^{Laftr}$ or $\mathcal{L}_{DP}^{Cfair}$) is applied to the non-sensitive latent z of the Ffvae model. Please check Section 4.1 for the discussion on the obtained results.

Table 4.1. Merged Ffvae and Cfair results when decreasing minority representation for Adult dataset, sensitive attribute:age. Added $\mathcal{L}_{DP}^{Cfair}$ to Eq.(2.5.9). Compare with Ffvae Table 6.4 and Cfair Table 6.2 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.77	0.81	0.73	0.99	0.99	0.97	0.98	0.92
(0.1, 0.1)	0.75	0.78	0.72	0.99	1.0	0.99	0.99	0.98
(0.01, 0.01)	0.72	0.83	0.62	0.94	1.0	0.85	0.93	0.79

We observe an improvement of the fairness metrics (DP, EqOpp, EqOdd) in joint models compared to individual models, while accuracy has not changed or deteriorated. In particular, when comparing the results in Tables 4.1 to 4.4 to the tables obtained from individual models, we

Table 4.2. Merged Ffvae and Cfair results on correlation of sensitive attribute (age) and eligibility for Adult dataset. Added $\mathcal{L}_{DP}^{Cfair}$ to Eq.(2.5.9). Compare with Ffvae Table 6.20 and Cfair Table 6.18 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.77	0.81	0.73	0.99	0.99	0.97	0.98	0.92
(0.66, 0.33)	0.74	0.73	0.75	1.0	1.0	1.0	1.0	0.92
(0.06, 0.36)	0.7	0.76	0.64	0.98	0.99	0.96	0.97	0.97

Table 4.3. Merged Ffvae and Laftr-DP results when decreasing minority representation for Adult dataset, sensitive attribute:age. Added $\mathcal{L}_{DP}^{Laftr}$ to Eq.(2.5.9). Compare with Ffvae Table 6.4 and Laftr-DP Table 6.8 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.66	0.71	0.61	0.99	1.0	0.98	0.99	0.93
(0.1, 0.1)	0.6	0.67	0.54	1.0	1.0	1.0	1.0	0.98
(0.01, 0.01)	0.6	0.69	0.52	1.0	1.0	1.0	1.0	0.92

Table 4.4. Merged Ffvae and Laftr-DP results on correlation of sensitive attribute (age) and eligibility for Adult dataset. Added $\mathcal{L}_{DP}^{Laftr}$ to Eq.(2.5.9). Compare with Ffvae Table 6.20 and Laftr-DP Table 6.24 results.

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.66	0.71	0.61	0.99	1.0	0.98	0.99	0.93
(0.66, 0.33)	0.49	0.5	0.49	1.0	1.0	1.0	1.0	0.93
(0.06, 0.36)	0.6	0.69	0.51	0.97	0.99	0.95	0.97	0.98

observe, on average over different cases, the fairness metrics DP, EqOpp0, EqOpp1, and EqOdd have improved in Ffvae+Cfair by 5.83%, 3.41%, 9.15%, 5.98% and in Ffvae+Laftr by 9.63%, 3.4%, 19.4%, 10.53%, while accuracy has improved in Ffvae+Cfair by only 1.15% and dropped in Ffvae+Laftr by 12.66%. This emphasizes that further investigation in merging the seemingly different bias-mitigation strategies can bring improvement *w.r.t* fairness metrics while maintaining model accuracy.

4.2. Sources of Bias

In our experiments, in addition to random seed, we have considered three different sources of bias:

4.2.1. Reduced representation of the unprivileged group.

To further disentangle the cause of bias, we performed an experiment where we kept the ratio of $(0.001, 0.001)$ in all settings but increased the number of total samples by 10, 100, and 1000 times

to alleviate data scarceness, especially in the $1000x$ case where the under-represented group is as numerous as the original dataset size. The results are shown in Tables 4.5 for CNN (without any bias-mitigation strategy) and in 4.6 for Laftr-EqOpp0, which is one of the strongest bias-mitigation models, where clr-ratios is kept at (0.001, 0.001) but the total dataset size has changed from x to $10x$, $100x$ and $1000x$, where x is the original number of samples. Surprisingly, the source of bias in these models differs. In Laftr, data scarcity appears to be a source of bias, and as the number of samples in the under-represented group grows, the model's performance improves. However, it still falls short of the balanced arrangement illustrated in the first row of Table 6.15. In CNN, however, the imbalance in the data distribution seems to be the main cause of bias, as the model does not effectively leverage the increasing number of samples. These results indicate that different models may be susceptible to varying sources of bias, including data imbalance and scarcity.

Table 4.5. CNN results for measuring whether the bias is due to small ratio or small number of samples.

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.001, 0.001) x	0.93	0.97	0.88	0.93	0.98	0.85	0.92	0.5
(0.001, 0.001) 10x	0.96	0.98	0.95	0.99	0.98	0.96	0.97	0.51
(0.001, 0.001) 100x	0.9	0.98	0.82	0.95	0.9	0.79	0.84	0.52
(0.001, 0.001) 1000x	0.9	0.98	0.82	0.95	0.89	0.78	0.83	0.52

Table 4.6. Laftr-EqOpp0 results for measuring whether the bias is due to small ratio or small number of samples.

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.001, 0.001) x	0.94	0.98	0.89	0.95	0.96	0.88	0.92	0.67
(0.001, 0.001) 10x	0.91	0.98	0.83	0.94	0.81	0.9	0.85	0.5
(0.001, 0.001) 100x	0.94	0.98	0.89	0.99	0.91	0.9	0.91	0.54
(0.001, 0.001) 1000x	0.96	0.99	0.93	0.99	0.95	0.93	0.94	0.54

4.2.2. Correlation of a feature with eligibility.

In some datasets, the unprivileged group, might exhibit lower eligibility, *e.g.,* due to historical reasons like lack of access to facilities or limited resources compared to a privileged group (Kamishima et al., 2012).

As seen in settings 2 and 4, the model can quickly pick up on correlations between features and eligibility in the data, generating bias at test time. However, under such circumstances, we discover that if the sensitive attributes/features dominate, i.e., occupy a larger percentage of the input data, the bias grows stronger, leading the fairness metrics to decrease even more, as seen

when comparing setting 2 to setting 4. Furthermore, we observe that the degree of bias in models augments when the correlation grows in both settings.

4.2.3. Impact of non-predominant correlated features.

The bias-mitigation techniques assume that the sensitive or biased attributes are known *a priori* and can thus be addressed by eliminating these features from the learned representations. The model or a possible annotator may overlook some properties, mainly minor (non-predominant) ones. Wearing spectacles, for example, is linked to ageing. If this sensitive attribute is not known *a priori* to the model, the bias-mitigation algorithm will not be able to handle it appropriately. Our results in setting 3 and the Spearman correlation plot in this setting indicate that such non-predominant but correlated features can be a source of bias and the models still capture the correlation in the data.

Chapter 5

Conclusion

With the increasing use of representation learning algorithms in automatic decision-making tools (Cappelli et al., 2019, Franz et al., 2020, Dzyabura et al., 2019, Shrestha et al., 2020), the rigorous benchmarking of bias-mitigation algorithms has become imperative. We established a framework in this paper to benchmark bias-mitigation algorithms in various circumstances to test their boundaries. Using the suggested benchmark on challenging versions of the given synthetic CI-MNIST and Adults datasets, we comprehensively evaluated and examined a variety of deep learning-based bias-mitigation models, where we regulate the correlation and balance between distinct dataset subgroups. We demonstrated that we could purposely push the models under investigation to their breaking point, even though they have proven their usefulness to the community and were able to perform sufficiently on some of our settings and dataset variants.

5.1. Research Conclusion

Throughout our research, we discovered that these models could exploit sensitive features and produce biases when: 1) the underprivileged group is under-represented, which might be due to population imbalance or scarcity; and 2) the sensitive trait and eligibility have an association. Both scenarios show a more significant bias when the imbalance or correlation rises. According to our findings, the degree of bias in the model was similarly affected by the dimensionality/predominance of the biased input features. Furthermore, we discovered that the sensitive information is present in the latent representation of models trained with bias-mitigation strategies, implying that employing representation from these models for downstream tasks can still lead to a biased treatment. Our results also showed that some models are more susceptible to the random seed than others; hence the variation of the random seed can be a source of bias. As a result, not all models are well-suited to all contexts, and utilizing models without a thorough awareness of their limits may result in adverse outcomes. We also observe that, given the variety of fairness indicators, model selection should be improved, as we cannot identify a single model that works well across all metrics. We hope our benchmark serves as a starting point to verify the robustness of bias-mitigation models.

5.2. Limitations

Some of the limitations of the dataset creation process in this work include,

- Focusing only on single binary target labels due to the extensiveness of evaluating all the cases. Our codebase provides code for creating non-binary target labels for use in future experimentation.
- Using binary-sensitive attributes only for experiments. It would be interesting to extend the present work to non-binary sensitive features and observe the impact of bias-mitigation methods on multiple sensitive features.
- Not evaluating the correlation of input features. Just mitigating the effect of sensitive features might not be enough, as proxy features can partially correlate with the sensitive features (Corbett-Davies and Goel, 2018) affecting the bias mitigation.
- Lower sample size when creating sub-groups in datasets other than MNIST due to non-uniform distributions of sensitive features and target labels. These smaller sample sizes can cause inconsistent model experimentation and give inaccurate results.

On a separate note, we have evaluated some of the recent promising bias-mitigation algorithms out of many proposed models. This field is expanding rapidly, and we could not evaluate all possible models. It would be interesting to evaluate other promising models in the proposed setups and try the proposed settings on more datasets to observe any potential change in performance due to differences in the data modality.

5.3. Potential negative societal impacts

While we report results using currently established fairness metrics, there is neither a universal fairness metric nor a universally accepted definition of fairness (Mukherjee et al., 2020). Therefore, although we take steps towards properly benchmarking bias mitigation algorithms, one should still be cautious about the chosen metrics before using them in real-world applications. Group parity metrics like DP, EqOpp0, EqOpp1, and EqOdd can suffer from statistical limitations, as the true underlying protected group distributions might differ regardless of other unprotected features used for prediction. Thus enforcing these metrics, which try to lower group fairness, can lower accuracies and harm the groups designed to protect (Corbett-Davies and Goel, 2018).

Hence before using any specific metric, the true data distribution should be studied well by designing suitable interventions. Real-world assessments and understanding consequences also help in achieving equitable metrics. Today with representation learning methods used in automatic decision-making applications (Cappelli et al., 2019, Franz et al., 2020, Dzyabura et al., 2019,

Shrestha et al., 2020), more interpretable and explainable models are necessary to avoid any harmful consequences.

5.4. Future work

While there have been many definitions of and approaches to fairness in the literature, the study in this area is anything but complete. Fairness and algorithmic bias still hold several research opportunities. In this section, we provide pointers to outstanding challenges in fairness research and an overview of opportunities for the development of understudied problems.

5.4.1. Challenges

5.4.1.1. **Synthesizing a definition of fairness.** Several definitions of what would constitute fairness from a machine learning perspective have been proposed in the literature (Dwork et al., 2012, Zafar et al., 2017b, Hardt et al., 2016b). These definitions cover a wide range of use cases and are somewhat disparate in their view of fairness. Because of this, it is nearly impossible to understand how one fairness solution would fare under a different definition of fairness. Synthesizing these definitions into one remains an open research problem since it can evaluate these systems as more unified and comparable. Having a more unified fairness definition and framework can also help with the incompatibility issue of some current fairness definitions.

5.4.1.2. **From Equality to Equity.** The definitions presented in the literature mostly focus on equality, ensuring that each individual or group is given the same amount of resources, attention or outcome. However, little attention has been paid to equity, which is the concept that each individual or group is given the resources they need to succeed (Mehrabi et al., 2020, White and Brumfield, 2014). Operationalizing this definition and studying how it augments or contradicts existing definitions of fairness remains an exciting future direction.

5.4.1.3. **Searching for Unfairness.** Defined fairness, it should be possible to identify instances of this unfairness in a particular dataset. Inroads toward this problem have been made in data bias by detecting instances of Simpson's Paradox in arbitrary datasets (Alipourfard et al., 2018); however, unfairness may require more consideration due to the variety of definitions and the nuances in detecting each one.

References

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Alipourfard, N., Fennell, P. G., and Lerman, K. (2018). Can you trust the trend?: Discovering simpson's paradoxes in social data. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., and Wang, X. (2021). Evaluating fairness of machine learning models under uncertain and incomplete information. *CoRR*, abs/2102.08410.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. (2019). Scalable fair clustering.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bellamy, R., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K., and Zhang, Y. (2018a). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *ArXiv*, abs/1810.01943.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018b). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943.

Ben-David, S., Eiron, N., and Long, P. M. (2003). On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66:496–514.

Bengio, Y., Courville, A. C., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 004912411878253.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Bogen, M. and Rieke, A. (2018). Help wanted: an examination of hiring algorithms, equity, and bias.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, pages 4356–4364.

Brunet, M., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. S. (2018). Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91.

Byanjankar, A., Heikkilä, M., and Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 719–725.

Calders, T. and Verwer, S. (2010a). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Calders, T. and Verwer, S. (2010b). Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K., and Varshney, K. (2017). Optimized pre-processing for discrimination prevention. In *NeurIPS*.

Cappelli, P., Tambe, P., and Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *PsychRN: Psychological Applications of Technology & Media (Topic)*.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620.

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

Cohen, J. P., Luck, M., and Honari, S. (2018). Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and*

computer-assisted intervention, pages 529–536. Springer.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *ArXiv*, abs/1808.00023.

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

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Du, M., Yang, F., Zou, N., and Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. *ArXiv*, abs/1104.3913.

Dzyabura, D., Kihal, S. E., Hauser, J., and Ibragimov, M. (2019). Leveraging the power of images in managing product return rates. *Econometrics: Econometric & Statistical Methods - Special Topics eJournal*.

d' Alessandro, B., O'Neil, C., and LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134.

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

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. (2009). Agnostic learning of monomials by halfspaces is hard. *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 385–394.

Franz, L., Shrestha, Y. R., and Paudel, B. (2020). A deep learning pipeline for patient diagnosis prediction using electronic health records. *ArXiv*, abs/2006.16926.

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

Gajane, P. (2017). On formalizing fairness in prediction with machine learning. *CoRR*, abs/1710.03184.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.

Goel, N., Yaghini, M., and Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 116, New York, NY, USA. Association for Computing Machinery.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.

Hardt, M., Price, E., and Srebro, N. (2016a). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, pages 3323–3331.

Hardt, M., Price, E., and Srebro, N. (2016b). Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

Hardt, M., Price, E., Srebro, N., et al. (2016c). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hoffman, M., Kahn, L., and Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2.

Huang, L. and Vishnoi, N. K. (2020). Stable and fair classification.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2019). Wasserstein fair classification. In *UAI*.

Jones, G., Hickey, J. M., Stefano, P. G. D., Dhanjal, C., Stoddart, L. C., and Vasileiou, V. (2020). Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *ArXiv*, abs/2010.03986.

Kamiran, F. and Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6.

Kamiran, F. and Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer.

Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828.

Kelley, H. J. (1960). Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954.

Kim, H. and Mnih, A. (2018). Disentangling by factorising. In *ICML*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

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

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, volume 97, pages 4114–4124. PMLR.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2016). The variational fair autoencoder. *CoRR*, abs/1511.00830.

Louppe, G., Kagan, M., and Cranmer, K. (2017). Learning to pivot with adversarial networks. In *NeurIPS*.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2020). Gender bias in neural natural language processing. In *Logic, Language, and Security*.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. (2018). Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pages 3381–3390.

Malekipirbazari, M. and Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631.

Marx, C. T., Phillips, R. L., Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2019). Disentangling influence: Using disentangled representations to audit model predictions. In *NeurIPS*.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mehrabi, N., Huang, Y., and Morstatter, F. (2020). Statistical equity: A fairness classification objective. *ArXiv*, abs/2005.07293.

Mehrabi, N., Morstatter, F., Peng, N., and Galstyan, A. (2019a). Debiasing community detection: The importance of lowly-connected nodes.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019b). A survey on bias and fairness in machine learning. *ArXiv*, abs/1908.09635.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019c). A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635.

Menon, A. and Williamson, R. (2018). The cost of fairness in binary classification. In *FAT*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mitchell, T. M. et al. (1997). *Machine learning*. McGraw-hill New York.

Mukherjee, D., Yurochkin, M., Banerjee, M., and Sun, Y. (2020). Two simple ways to learn individual fairness metrics from data. *ArXiv*, abs/2006.11439.

Oneto, L. and Chiappa, S. (2020). Fairness in machine learning. *CoRR*, abs/2012.15816.

Oneto, L., Siri, A., Luria, G., and Anguita, D. (2017). Dropout prediction at university of genoa: a privacy preserving data driven approach.

Papamitsiou, Z. and Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology and Society*, 17(4):49–64.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. (2020). An investigation of why overparameterization exacerbates spurious correlations.

Samadi, S., Tantipongpipat, U. T., Morgenstern, J., Singh, M., and Vempala, S. S. (2018). The price of fair PCA: one extra dimension. *CoRR*, abs/1811.00103.

Shrestha, Y. R., Krishna, V., and Krogh, G. (2020). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Social Science Research Network*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Suresh, H. and Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Sutton, R. S. and Barto, A. G. (2005). Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 16:285–286.

Träuble, F., Creager, E., Kilbertus, N., Goyal, A., Locatello, F., Schölkopf, B., and Bauer, S. (2020). Is independence all you need? on the generalization of representations learned from correlated data.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Verma, S. and Rubin, J. (2018a). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7.

Verma, S. and Rubin, J. (2018b). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

White, H. L. and Brumfield, K. (2014). Susan gooden, race and social equity: A nervous area of government. *Journal of Comparative Policy Analysis: Research and Practice*, 16:191 – 192.

Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G. (2017). Controllable invariance through adversarial feature learning. In *NeurIPS*.

Yu, Z., Chakraborty, J., and Menzies, T. (2021). Fairbalance: Improving machine learning fairness on multiplesensitive attributes with data balancing.

Zafar, M., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. (2017a). Fairness constraints: Mechanisms for fair classification. In *AISTATS*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017b). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of International Conference on World Wide Web*, pages 1171–1180.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. (2020). Conditional learning of fair representations. In *International Conference on Learning Representations*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. In *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15*.

Chapter 6

Appendix

6.1. Experiments and Results

6.1.1. Impact of reducing representation of unprivileged group

We report the complete set of results for debiasing models of MLP, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, in Tables 6.1 to 6.16, corresponding to the experimental setup described in Setting 1 of Section 3.3. Each pair in *clr-ratio* column indicate (b_e, b_o), which is the ratio of images with blue background for (even=eligible, odd=ineligible) data. Figures 3.3, 3.4 compare all models side-by-side. Note that to report results, we initially averaged metrics over three seeds, then for each metric, the best value over the fairness coefficients of the models is reported on the test set.

Table 6.1. MLP results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.97	0.99	0.92	0.96	0.66
(0.1, 0.1)	0.74	0.77	0.71	0.96	0.99	0.9	0.95	0.51
(0.01, 0.01)	0.74	0.81	0.66	0.91	0.97	0.77	0.87	0.54

Table 6.2. Cfair results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.82	0.84	0.81	0.99	0.97	0.99	0.98	0.54
(0.1, 0.1)	0.8	0.82	0.79	0.96	0.99	0.98	0.98	0.51
(0.01, 0.01)	0.8	0.87	0.73	0.85	0.92	0.68	0.8	0.54

6.1.2. Impact of correlation of sensitive attribute with eligibility

Table 6.3. Cfair-EO results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.99	0.99	0.98	0.98	0.52
(0.1, 0.1)	0.73	0.77	0.69	0.98	0.99	0.96	0.97	0.51
(0.01, 0.01)	0.74	0.8	0.67	0.96	0.99	0.88	0.94	0.54

Table 6.4. Ffvae results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.74	0.7	0.94	0.99	0.9	0.95	0.93
(0.1, 0.1)	0.7	0.73	0.67	0.92	0.96	0.88	0.92	0.98
(0.01, 0.01)	0.71	0.81	0.61	0.91	0.98	0.76	0.87	0.92

Table 6.5. Laftr-EqOdd results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.96	0.99	0.92	0.96	0.52
(0.1, 0.1)	0.69	0.71	0.66	0.95	0.98	0.9	0.94	0.52
(0.01, 0.01)	0.67	0.77	0.56	0.86	0.96	0.72	0.84	0.52

Table 6.6. Laftr-EqOpp1 results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.7	0.96	0.99	0.91	0.95	0.53
(0.1, 0.1)	0.69	0.72	0.66	0.95	0.98	0.9	0.94	0.52
(0.01, 0.01)	0.67	0.77	0.56	0.86	0.97	0.72	0.84	0.52

Table 6.7. Laftr-EqOpp0 results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.69	0.96	1.0	0.91	0.96	0.52
(0.1, 0.1)	0.69	0.71	0.66	0.95	0.98	0.91	0.95	0.52
(0.01, 0.01)	0.66	0.76	0.56	0.85	0.97	0.71	0.84	0.52

Table 6.8. Laftr-DP results when decreasing minority representation for Adult dataset, sensitive attribute:age, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.97	0.99	0.92	0.96	0.60
(0.1, 0.1)	0.69	0.71	0.66	0.96	0.99	0.91	0.95	0.52
(0.01, 0.01)	0.66	0.76	0.55	0.87	0.97	0.72	0.84	0.55

Table 6.9. MLP results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.96	0.94	0.98	0.97	0.97	0.97	0.91
(0.1, 0.1)	0.94	0.97	0.9	0.94	0.91	0.94	0.93	0.99
(0.01, 0.01)	0.84	0.96	0.72	0.76	0.96	0.52	0.74	0.9
(0.001, 0.001)	0.77	0.97	0.58	0.59	0.46	0.72	0.59	0.67

Table 6.10. CNN results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	0.99	0.99	0.98	0.98	0.56
(0.1, 0.1)	0.96	0.97	0.95	0.99	0.99	0.98	0.98	0.5
(0.01, 0.01)	0.96	0.97	0.95	0.99	0.99	0.96	0.97	0.5
(0.001, 0.001)	0.93	0.97	0.88	0.93	0.98	0.85	0.92	0.5

Table 6.11. Cfair results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.94	0.95	0.93	0.98	0.98	0.99	0.98	1.0
(0.1, 0.1)	0.93	0.96	0.89	0.96	0.97	0.92	0.95	1.0
(0.01, 0.01)	0.85	0.96	0.74	0.78	0.87	0.72	0.79	1.0
(0.001, 0.001)	0.78	0.96	0.6	0.66	0.95	0.93	0.94	1.0

Table 6.12. Ffvae results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.92	0.9	0.98	1.0	0.96	0.98	1.0
(0.1, 0.1)	0.89	0.91	0.87	0.97	0.99	0.92	0.96	1.0
(0.01, 0.01)	0.85	0.92	0.79	0.97	0.98	0.81	0.9	1.0
(0.001, 0.001)	0.72	0.92	0.51	0.82	0.68	0.83	0.76	1.0

Table 6.13. Laftr-EqOdd results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.98	0.98	1.0
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.98	0.98	0.98	0.96
(0.01, 0.01)	0.96	0.98	0.93	0.97	0.99	0.93	0.96	0.8
(0.001, 0.001)	0.91	0.98	0.84	0.94	0.99	0.83	0.91	0.8

Table 6.14. Laftr-EqOpp1 results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.98	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.98	0.98	0.98	0.95
(0.01, 0.01)	0.95	0.97	0.93	0.99	0.99	0.95	0.97	0.85
(0.001, 0.001)	0.92	0.98	0.85	0.91	0.98	0.83	0.91	0.73

Table 6.15. Laftr-EqOpp0 results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.97	0.99	0.98	0.98
(0.01, 0.01)	0.95	0.98	0.92	0.99	0.99	0.95	0.97	0.78
(0.001, 0.001)	0.94	0.98	0.89	0.95	0.96	0.88	0.92	0.67

Table 6.16. Laftr-DP results when decreasing minority representation for CI-MNIST dataset, sensitive attribute:bck, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	1.0
(0.1, 0.1)	0.96	0.98	0.95	0.99	0.97	0.98	0.97	1.0
(0.01, 0.01)	0.96	0.98	0.93	0.98	0.98	0.94	0.96	0.86
(0.001, 0.001)	0.92	0.98	0.86	0.91	0.96	0.87	0.92	0.69

We report the complete set of results for debiasing models of MLP, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, in Tables 6.17 to 6.32, corresponding to Setting 2 in Section 3.4. Each pair in *clr-ratio* column indicate (b_e, b_o), which is the ratio of images with blue background for (even=qualified, odd=unqualified) data. Figures 3.5, 3.6 compare all models side-by-side.

Table 6.17. MLP results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.97	0.99	0.92	0.96	0.66
(0.66, 0.33)	0.75	0.74	0.76	0.85	0.88	0.83	0.85	0.65
(0.06, 0.36)	0.71	0.75	0.66	0.78	0.86	0.69	0.77	0.65

Table 6.18. Cfair results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.82	0.84	0.81	0.99	0.97	0.99	0.98	0.54
(0.66, 0.33)	0.8	0.82	0.78	0.94	0.96	0.99	0.97	0.64
(0.06, 0.36)	0.8	0.77	0.84	0.98	0.96	0.97	0.96	0.54

Table 6.19. Cfair-EO results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.78	0.8	0.75	0.99	0.99	0.98	0.98	0.52
(0.66, 0.33)	0.76	0.76	0.76	0.98	0.99	0.98	0.98	0.54
(0.06, 0.36)	0.75	0.75	0.75	0.98	0.98	0.97	0.97	0.53

Table 6.20. Ffvae results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.74	0.7	0.94	0.99	0.9	0.95	0.93
(0.66, 0.33)	0.59	0.57	0.61	0.99	1.0	0.98	0.99	0.92
(0.06, 0.36)	0.62	0.71	0.53	0.81	0.91	0.72	0.81	0.98

Table 6.21. Laftr-EqOdd results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.96	0.99	0.92	0.96	0.52
(0.66, 0.33)	0.7	0.67	0.73	0.88	0.94	0.83	0.89	0.54
(0.06, 0.36)	0.65	0.7	0.59	0.83	0.93	0.72	0.82	0.47

Table 6.22. Laftr-EqOpp1 results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.7	0.96	0.99	0.91	0.95	0.53
(0.66, 0.33)	0.7	0.67	0.73	0.89	0.94	0.84	0.89	0.55
(0.06, 0.36)	0.65	0.71	0.59	0.83	0.93	0.72	0.82	0.47

Table 6.23. Laftr-EqOpp0 results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.72	0.75	0.69	0.96	1.0	0.91	0.96	0.52
(0.66, 0.33)	0.7	0.67	0.72	0.88	0.94	0.83	0.89	0.55
(0.06, 0.36)	0.65	0.71	0.59	0.83	0.93	0.73	0.83	0.48

Table 6.24. Laftr-DP results on correlation of sensitive attribute (age) and eligibility for Adult dataset, selected best result per attribute

(u-elg, u-inelg)	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.71	0.74	0.69	0.97	0.99	0.92	0.96	0.6
(0.66, 0.33)	0.7	0.67	0.73	0.88	0.94	0.83	0.89	0.57
(0.06, 0.36)	0.66	0.71	0.6	0.83	0.93	0.73	0.83	0.55

Table 6.25. MLP results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.96	0.94	0.98	0.97	0.97	0.97	0.91
(0.1, 0.9)	0.81	0.82	0.81	0.65	0.66	0.63	0.65	0.96
(0.01, 0.99)	0.56	0.5	0.63	0.14	0.02	0.27	0.15	0.98

Table 6.26. CNN results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	0.99	0.99	0.98	0.98	0.56
(0.1, 0.9)	0.88	0.9	0.86	0.79	0.83	0.74	0.78	0.85
(0.01, 0.99)	0.56	0.5	0.62	0.12	0.01	0.22	0.12	1.0

Table 6.27. Cfair results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.94	0.95	0.93	0.98	0.98	0.99	0.98	1.0
(0.1, 0.9)	0.82	0.85	0.79	0.65	0.73	0.6	0.67	1.0
(0.01, 0.99)	0.55	0.54	0.57	0.09	0.09	0.13	0.11	1.0

Table 6.28. Ffvae results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.92	0.9	0.98	1.0	0.96	0.98	1.0
(0.1, 0.9)	0.71	0.72	0.71	0.45	0.48	0.43	0.45	1.0
(0.01, 0.99)	0.51	0.49	0.53	0.03	0.0	0.05	0.03	1.0

Table 6.29. Laftr-EqOdd results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.98	0.98	1.0
(0.1, 0.9)	0.92	0.94	0.89	0.85	0.89	0.81	0.85	0.98
(0.01, 0.99)	0.65	0.66	0.64	0.3	0.33	0.28	0.31	0.97

Table 6.30. Laftr-EqOpp1 results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.98	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.9)	0.92	0.94	0.9	0.86	0.9	0.81	0.85	0.97
(0.01, 0.99)	0.67	0.69	0.65	0.34	0.39	0.28	0.34	0.97

6.1.3. Impact of correlation of non-sensitive attribute with eligibility

We report the complete set of results for debiasing models of MLP, CNN, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, in Tables 6.33 to 6.40, corresponding to the

Table 6.31. Laftr-EqOpp0 results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	0.99
(0.1, 0.9)	0.91	0.94	0.88	0.85	0.91	0.79	0.85	0.94
(0.01, 0.99)	0.64	0.65	0.63	0.28	0.31	0.25	0.28	0.96

Table 6.32. Laftr-DP results on correlation of sensitive attribute (bck) and eligibility for CI-MNIST dataset, selected best result per attribute

clr-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	1.0
(0.1, 0.9)	0.92	0.94	0.89	0.85	0.91	0.8	0.85	1.0
(0.01, 0.99)	0.65	0.65	0.65	0.3	0.32	0.28	0.3	0.99

experimental setup described in Setting 3 of Secton 3.5. Each pair in *pos-ratio* column indicate (l_e, l_o), which specifies the ratio of images with box on left side for (even=eligible, odd=ineligible). Figure 3.7 compare all models side-by-side.

Table 6.33. MLP results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.96	0.94	0.98	0.97	0.97	0.97	0.91
(0.75, 0.25)	0.94	0.95	0.93	0.97	0.98	0.94	0.96	0.98
(0.9, 0.1)	0.86	0.9	0.83	0.96	0.9	0.96	0.93	1.0

Table 6.34. CNN results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	0.99	0.99	0.98	0.98	0.56
(0.75, 0.25)	0.95	0.96	0.93	0.99	0.98	0.96	0.97	0.57
(0.9, 0.1)	0.9	0.92	0.87	0.98	0.97	0.93	0.95	0.63

6.1.4. Impact of position and small features in the input images

Comparing baseline model with debiasing models of MLP, Cfair, Ffvae, Laftr-EqOdd, Laftr-EqOpp1, Laftr-EqOpp0, and Laftr-DP, when position and a small feature of the image correlates with

Table 6.35. Cfair results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.94	0.95	0.93	0.98	0.98	0.99	0.98	1.0
(0.75, 0.25)	0.94	0.95	0.92	0.99	0.98	0.98	0.98	1.0
(0.9, 0.1)	0.85	0.88	0.82	0.96	0.94	0.96	0.95	1.0

Table 6.36. Ffvae results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.92	0.9	0.98	1.0	0.96	0.98	1.0
(0.75, 0.25)	0.92	0.93	0.9	0.98	1.0	0.96	0.98	1.0
(0.9, 0.1)	0.9	0.91	0.89	0.98	0.99	0.95	0.97	1.0

Table 6.37. Laftr-EqOdd results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.98	0.98	1.0
(0.75, 0.25)	0.95	0.96	0.95	1.0	0.98	0.99	0.98	1.0
(0.9, 0.1)	0.93	0.94	0.91	0.97	0.97	0.97	0.97	1.0

Table 6.38. Laftr-EqOpp1 results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.98	0.96	0.99	0.99	0.99	0.99	0.99
(0.75, 0.25)	0.96	0.97	0.95	0.99	0.98	0.98	0.98	0.99
(0.9, 0.1)	0.92	0.93	0.9	0.97	0.96	0.96	0.96	0.99

Table 6.39. Laftr-EqOpp0 results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	0.99
(0.75, 0.25)	0.96	0.97	0.95	0.99	0.99	0.98	0.98	0.99
(0.9, 0.1)	0.92	0.94	0.9	0.97	0.97	0.95	0.96	0.99

eligibility. Results are depicted in Figure in Tables 6.41 to 6.48, corresponding to the experimental

Table 6.40. Laftr-DP results on correlation of non-sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.97	0.96	0.99	0.99	0.99	0.99	1.0
(0.75, 0.25)	0.96	0.97	0.95	0.99	0.99	0.99	0.99	1.0
(0.9, 0.1)	0.93	0.94	0.91	0.98	0.99	0.97	0.98	1.0

setup described in Setting 4 of Section 3.6. Each pair in *pos-ratio* column indicate (l_e, l_o), which specifies the ratio of images with box on left side for (even=eligible, odd=ineligible). Figure 3.8 compare all models side-by-side.

Table 6.41. MLP results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.95	0.95	1.0	1.0	0.99	0.99	0.51
(0.75, 0.25)	0.94	0.95	0.93	0.93	0.95	0.92	0.94	0.59
(0.9, 0.1)	0.87	0.89	0.85	0.76	0.8	0.72	0.76	0.67

Table 6.42. CNN results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	1.0	1.0	0.99	0.99	0.56
(0.75, 0.25)	0.94	0.95	0.94	0.94	0.94	0.94	0.94	0.79
(0.9, 0.1)	0.9	0.91	0.88	0.82	0.86	0.79	0.82	0.88

Table 6.43. Cfair results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.95	0.95	0.95	1.0	1.0	1.0	1.0	0.82
(0.75, 0.25)	0.94	0.94	0.94	0.99	1.0	0.99	0.99	0.88
(0.9, 0.1)	0.87	0.86	0.88	0.99	1.0	0.99	0.99	0.92

Table 6.44. Ffvae results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.91	0.9	0.91	1.0	1.0	1.0	1.0	1.0
(0.75, 0.25)	0.86	0.86	0.86	0.8	0.81	0.79	0.8	1.0
(0.9, 0.1)	0.76	0.75	0.77	0.54	0.53	0.55	0.54	1.0

Table 6.45. Laftr-EqOdd results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.97	1.0	1.0	1.0	1.0	0.99
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.94	0.95	0.64
(0.9, 0.1)	0.92	0.93	0.91	0.87	0.89	0.85	0.87	0.72

Table 6.46. Laftr-EqOpp1 results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.96	1.0	1.0	1.0	1.0	0.93
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.93	0.95	0.65
(0.9, 0.1)	0.92	0.93	0.91	0.86	0.89	0.84	0.86	0.75

Table 6.47. Laftr-EqOpp0 results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.97	0.97	0.97	1.0	1.0	1.0	1.0	0.74
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.94	0.95	0.61
(0.9, 0.1)	0.92	0.93	0.91	0.86	0.88	0.84	0.86	0.74

Table 6.48. Laftr-DP results on correlation of sensitive attribute (pos) and eligibility for CI-MNIST dataset, selected best result per attribute

pos-ratio	acc	p-acc	up-acc	DP	EqOpp0	EqOpp1	EqOdd	sens-acc
(0.5, 0.5)	0.96	0.96	0.97	1.0	1.0	1.0	1.0	1.0
(0.75, 0.25)	0.95	0.96	0.95	0.95	0.96	0.94	0.95	0.95
(0.9, 0.1)	0.93	0.93	0.92	0.86	0.87	0.86	0.86	0.97