Received ; | Revised ; | Accepted

DOI: xxx/xxxx

REVIEW ARTICLE

How do the Existing Fairness Metrics and Unfairness Mitigation Algorithms contribute to Ethical Learning Analytics?

Deho Oscar Blessed* | Chen Zhan | Jiuyong Li | Jixue Liu | Lin Liu | Thuc Duy Le

[1]UniSA STEM, University of South Australia, South Australia, Australia

Correspondence
*Deho Oscar Blessed, Email: oscar.deho@mymail.unisa.edu.au

With the widespread use of learning analytics (LA), ethical concerns about fairness have been raised. Research shows that LA models may be biased against students of certain demographic groups. Although fairness has gained significant attention in the broader machine learning (ML) community in the last decade, it is only recently that attention has been paid to fairness in LA. Furthermore, the decision on which unfairness mitigation algorithm or metric to use in a particular context remains largely unknown. On this premise, we performed a comparative evaluation of some selected unfairness mitigation algorithms regarded in the fair ML community to have shown promising results. Using a 3-year program dropout data from an Australian university, we comparatively evaluated how the unfairness mitigation algorithms contribute to ethical LA by testing for some hypotheses across fairness and performance metrics. Interestingly, our results show how data bias does not always necessarily result in predictive bias. Perhaps not surprisingly, our test for fairness-utility tradeoff shows how ensuring fairness does not always lead to drop in utility. Indeed, our results show that ensuring fairness might lead to enhanced utility under specific circumstance. Our findings may to some extent, guide fairness algorithm and metric selection for a given context.

KEYWORDS:
Fairness, Learning Analytics, Ethical LA, Predictive Modelling, Virtual Learning Environment

1 | PRACTITIONER NOTES

1.0.1 | What is already known about this topic

- LA is increasingly being used to leverage actionable insights about students and drive student success
- LA models have been found to make discriminatory decisions against certain student demographics — therefore, raising ethical concerns.
- Fairness in education is nascent. Only a few works have examined fairness in LA and consequently followed up with ensuring fair LA models.

1.0.2 | What this paper adds

- A juxtaposition of unfairness mitigation algorithms across the entire LA pipeline showing how they compare and how each of them contributes to fair LA
- Ensuring ethical LA does not always lead to a dip in performance. Sometimes, it actually improves performance as well.

- Fairness in LA has only focused on some form of outcome equality, however equality of outcome may be possible only when the playing field is levelled.

1.0.3 | Implications for practice and/or policy

- Based on desired notion of fairness and which segment of the LA pipeline is accessible, a fairness-minded decision maker may be able to decide which algorithm to use in order to achieve their ethical goals.
- LA practitioners can carefully aim for more ethical LA models without trading significant utility by selecting algorithms that find the right balance between the two objectives.
- Fairness enhancing technologies should be cautiously used as guides — not final decision makers. Human domain experts must be kept in the loop to handle the dynamics of transcending fair LA beyond equality to equitable LA.

Statements on open data, ethics and conflicts of interest

All ethical requirements have been considered prior to conducting the analysis. The study was approved by the governing ethical board. (Ethics Protocol Application ID: 204198).
The authors declare no conflict of interests.
Due to privacy concerns, we cannot share the dropout dataset. Access to other data and analysis associated with this publication is available at this link

2 | INTRODUCTION

The availability of powerful data infrastructure and futuristic visions of many educational institutions worldwide has resulted in increased deployment of LA[1] technologies to drive student and institutional success as well as optimal resource allocation Dawson, Jovanovic, Gašević, and Pardo (2017). Given the current COVID-19 pandemic, most educational institutions have transitioned learning activities to online instruction. This has further increased the amount of learner-generated data almost exponentially. As a result, there has been an increased interest in LA as field of research and practice. Since its inception, the prime focus of LA has been utilization of data from various learning environments (e.g., learning management systems (LMS), massive open online courses (MOOCS) or student information services) and/or multimodal sensory (e.g., visual, auditory, reading and writing, and kinesthetic) data to leverage insights on students Joksimović, Kovanović, and Dawson (2019). For instance, using LA, we can predict: students at risk of failing a course Hlosta, Zdrahal, and Zendulka (2017) or learning outcomes Käser, Hallinen, and Schwartz (2017).

Despite the many benefits that LA provides, ethical concerns about their fairness have been raised Baker and Hawn (2021). A significant number of fairness metrics Narayanan (2018) and algorithms satisfying such metrics Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2019) have been developed in the broader ML community over the last decade. Despite the considerable number of research on fairness, we have not seen these fairness-aware algorithms deployed in relevant real-life domains. This is probably because it is non-trivial to find the "best" algorithm or fairness metric for each situation.

Furthermore, there is an ongoing debate on what is termed as the "impossibility theorem", to wit, it is not possible to satisfy all fairness measures simultaneously. Kleinberg, Mullainathan, and Raghavan (2016) analysed the relationship between *calibration within groups*, *balance of negative class* and *balance for positive class* and proved that except in highly constrained special cases, it is impossible to satisfy all three measures simultaneously. Similarly, Chouldechova (2017) showed that in the event of unequal base rates across groups, a recidivism prediction instrument satisfying predictive parity may still result in disparate impact. Berk, Heidari, Jabbari, Kearns, and Roth (2018) demonstrated the incompatibility between six fairness measures and the impossibility to simultaneously maximize accuracy and fairness.

Several surveys have been done on different aspects of fair ML. We categorize them as follows: ones that provide a general literature overview and those that experimentally evaluate fairness measures and unfairness mitigation algorithms. In the first category, Romei and Ruggieri (2011) focuses on areas of application of fairness, methods and approaches to data collection and data analysis. The survey by Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2019) covers the sources bias and how (un)fairness manifests in a different families of ML approaches. In the second category, Žliobaitė (2017) computationally evaluated some fairness measures and shed light on the implications of using a particular measure. Closely related to ours is the work of Friedler et al. (2019) where the authors experimentally compare some pre- and in-processing algorithms to determine how the algorithms relatively perform.

[1] unless otherwise stated, in this work, we use LA to loosely represent ML in the educational settings

In the aforementioned works and more generally fair ML research, little attention has been paid to the domain of education. In the education context, while Yu, Lee, and Kizilcec (2021a) focused on theoretical approaches and recommendations to reducing bias in education, Baker and Hawn (2021) shed light on causes of bias in education and which protected groups are mostly affected. A similar study to ours was done by Riazy, Simbeck, and Schreck (2020). The authors compared Kamishima, Akaho, Asoh, and Sakuma (2012)'s Prejudice Remover and Zafar, Valera, Rogriguez, and Gummadi (2015)'s Margin-based classifier with four baseline classifiers across four fairness metrics. However, the breadth of the algorithms they considered is relatively small. To the best of our knowledge, there is no work in LA that evaluates the unfairness mitigation algorithms across the entire LA pipeline. A critical evaluation of how discrimination occur and are mitigated at various segments of the LA pipeline would guide LA practitioners on which algorithm to use in light of the segment accessible to them.

On that premise, we evaluated a relatively larger collection of unfairness mitigation algorithms spanning the entire LA pipeline (i.e. pre-processing, in-processing and post-processing) across a relatively wide collection of fairness and performance metrics. We selected works regarded to have shown promising results in the fair ML community. We tested for how they relatively compare and contribute to ensuring ethical LA based on some hypotheses detailed in Section 5. We further investigated if our findings would be consistent with domains other than LA by performing same evaluations using datasets other than educational dataset. The rest of our work is structured as follows: in Section 3, we perform literature review. Section 4 captures information on datasets used, the algorithms evaluated and metrics used. Experiments and discussions are made in Section 5, and then conclusions in Section 6.

3 | LITERATURE REVIEW

In this section, we first discuss works on fairness in the broader ML community and then we zero in on fairness in LA. To help our discussion, we made the following denotations. We represent non-sensitive attributes by X, sensitive or protected attribute by A, actual outcome by Y and predicted outcome by $\hat{Y}$. We denote the privileged group and favourable outcome by $A = 1$ and $Y = 1$ (or $\hat{Y} = 1$) respectively. Conversely, we represent unprivileged and unfavourable outcome by $A = 0$ and $Y = 0$ (or $\hat{Y} = 0$) respectively. We structured and reviewed literature as follows: (a) (un)fairness discovery or measurement; (b) unfairness mitigation; and (c) fairness in LA

3.1 | (Un)fairness Discovery or Measurement

According to Saxena et al. (2019), fairness may regarded as the absence of prejudice or any form of favouritism towards an individual or a group based on their sensitive attributes. There are two broad approaches by which fairness is measured in literature, namely (a) correlation-based approaches (b) causality-based approaches. Most of the fairness measures that we discussed in this work and more generally in the fair ML community are correlation-based (i.e., the correlation between the sensitive attribute and the target label is used to imply the existence of unfairness). Kusner, Loftus, Russell, and Silva (2017) among others, argue that correlation-based fairness notions usually suffer from statistical anomalies. More so, in a legal setting, a more admissible evidence of unfairness is based on counterfactual reasoning. For example, all other factors held constant, would a male student that was predicted to dropout receive the same prediction had he been a female student? These questions are best answered using causality. Both correlation-based and causality-based approaches to fairness measurements coalesce around two underlying notions of fairness, namely group fairness and individual fairness. We discuss them in detail as follows:

3.1.1 | Group Fairness (GF)

There are many variants of group fairness notion, however, the overarching idea is some form of statistical or predictive parity across demographic groups. We put the group fairness measures that we rewiewed into three categories according to the criteria that their computations are predicated on.

*Measurements based on actual **or** predicted outcome*: These measurements are the most basic and intuitive fairness measures. However, they are faced with lots of limitations. Some common examples of these measurements are statistical parity (SP) Dwork, Hardt, Pitassi, Reingold, and Zemel (2012) and disparate impact (DI) Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian (2015). SP ensures that proportions of individuals getting a favourable outcome are equal across demographics. The DI measure is based on the four-fifths rule. SP and DI are computed by finding the difference and ratio respectively in proportions of each demographic group getting a favourable outcome.

*Measurements based on actual **and** predicted outcome*: These measures ensure that predictions are more faithful to the ground-truth. Hardt, Price, and Srebro (2016) introduced the equal odds (EO) which requires that the true positive and false positive rates are equal for both protected and unprotected groups. For instance in a dropout prediction, EO ensures that male and female students are accurately and falsely predicted to dropout at similar rates. Equal opportunity (EOP) which is a relaxed version of EO ensures that the true positive rates for both protected and unprotected groups are equal. Furthermore, some other error metrics that are computed from a confusion matrix can be group-conditioned to measure fairness.

A difference in these group-conditioned error measures can be regarded as discrimination. Verma and Rubin (2018) lists a collection of such group-conditioned error measures, namely positive predictive value difference (PPV-diff), false discovery rate difference (FDR-diff), negative predictive value difference (NPV-diff) and false omission rate difference (FOR-diff).

Measurement based on Generalised Entropy Indices: This measure is borrowed from an economic principle used to measure inequality of income among a population. This measure quantifies unfairness by measuring how the outcome of an algorithm **benefits** different individuals or groups unequally. The benefit for an individual i can be computed using a benefit function b_i. The choice of benefit function is dependent on the domain of application. For simplicity, we use the binary benefit function defined in Speicher et al. (2018) which is computed as $b_i = \hat{y}_i - y_i + 1$. The benefit of a group g, is the mean of the benefits received by individuals in that group ; $\mu_g = \frac{1}{|g|}\sum_{i \in g} b_i$. For a constant $\alpha \notin \{0, 1\}$, the generalized entropy of benefits $b_1, b_2, ...b_n$ with mean benefit μ is computed as:

$$\varepsilon^{\alpha}(b_1, b_2, ..., b_n) = \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^{n} [(\frac{b_i}{\mu})^{\alpha} - 1] \tag{1}$$

The generalized entropy index has the subgroup decomposability characteristic. It can be decomposed into a group fairness (between-group entropy index (BGEI)) and individual fairness (within-group entropy index (WGEI)) components i.e., $\varepsilon^{\alpha}(b_1, b_2, ..., b_n) = \varepsilon^{\alpha}_{\beta}(b) + \varepsilon^{\alpha}_{\omega}(b)$. The BGEI and WGEI are denoted by $\varepsilon^{\alpha}_{\beta}(b)$ and $\varepsilon^{\alpha}_{\omega}(b)$ respectively. The theil index is special case of generalized entropy index with $\alpha = 1$. The theil index also has a group fairness (between-group theil index (BGTI)) and individual fairness (within-group theil index (WGTI)) components.
Although group fairness measures are able to measure group-based discrimination, individual-level unfairness are not properly detected. Individual fairness was thus introduced.

3.1.2 | Individual Fairness (IF)

Individual fairness is based on the idea that similar individuals should be treated similarly. There are fewer individual fairness measurements compared to group fairness. We put the individual fairness measures that we rewiewed into three categories based on the following criteria:

Measurements based on similarity metric: A predictor satisfies individual fairness if individuals i and j who are similar with respect to a given similarity metric; such as a distance function (i.e. $d(i, j) < \epsilon$) defined for a particular task are given similar outcomes. IF was introduced by Dwork, Hardt, Pitassi, Reingold, and Zemel (2012). The similarity metric in their work, however, is assumed to be given which may be unrealistic in certain instances as was stated in their paper as the most challenging aspect of their work. Based on this idea, Zemel, Wu, Swersky, Pitassi, and Dwork (2013) introduced the **Consistency (IF)**[2] measure. This measure compares the outcome of an individual i to those of its k-nearest neighbours.

Measurements based Generalised Entropy Indices (GEI): As we already discussed under the group fairness, the GEI can be decomposed into group and individual fairness components, the WGEI and WGTI components are used for measuring individual-level unfairness.

Measurements based Counterfactual Reasoning: We discuss two of the highly cited causality-based measures. Kusner, Loftus, Russell, and Silva (2017) introduced the *counterfactual fairness* measure which compares the same individual with an "imagined" version of themself. A predictor is counterfactually fair if a male individual in the real world will have the same outcome in the counterfactual world where his gender is flipped to female. One limitation of Kusner, Loftus, Russell, and Silva (2017)'s measure is that it assumes the entire effect of the sensitive attribute on the target label to be "discriminatory". To tackle this limitation, Chiappa (2019) introduced the *"path-specific counterfactual fairness"*. This measure checks the causal effect of the sensitive attribute on the target label along fair and unfair causal pathways. Causal discovery of unfairness is still nascent and has a great potential of improving fair ML. To the best of our knowledge, results from existing works show that algorithmic discrimination may be relatively "better" discovered and mitigated when tackled from the causal perspective. However the tools available for our comparative evaluations are all of the correlation-based approaches. Thus, we leave evaluation of causality-based approaches for future work.

3.2 | Unfairness Mitigation

A naïve way of mitigating unfairness would be simply deleting all the sensitive features in the dataset that serve as the basis for discrimination. This however, is far from being adequate as non-sensitive attributes may act as proxies for sensitive attributes (usually referred to as *redlining*) Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian (2015). Researchers have come up with unfairness mitigation algorithms that tackle the various segments of a ML pipeline. There are those that that remove biases in the training data (pre-processing), the algorithm itself (in-processing) or the predicted outcomes (post-processing).

[2]For the sake of clarity, **consistency** as a fairness measure is emboldened to differentiate it from the noun consistency

3.2.1 | Pre-processing

Pre-processing methods work on the motivation that the quality of any model depends on the quality of the training data, hence a fair dataset would result in fair outcomes. Pre-processing approaches modify biased historical data to remove discriminatory patterns. The Learning Fair Representation (LFR) algorithm of Zemel, Wu, Swersky, Pitassi, and Dwork (2013) learns a latent representation that encodes the essential information of the data while obfuscating information related to sensitive attributes in order to satisfy **consistency** and SP. Similarly, in order to satisfy both group and individual fairness, Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney (2017)'s Optimized Preprocessing (OptimPreProc) used the idea of probabilistic transformation of the training dataset formulated as a convex optimization problem with three constraints, namely minimizing discrimination, preserving utility and reducing distortion in the individual data samples. Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian (2015)'s Disparate Impact Remover (DIR) ensures group fairness while preserving rank-ordering by editing feature values. Although the features are changed to ensure fairness, labels remain unchanged, thus label biases may still be present in DIR's fair data. Kamiran and Calders (2012) developed ways to modify a biased data by *massaging*, where the labels of some individuals in the dataset are changed, *reweighing algorithm (RW)*, where weights are systematically assigned to individuals to achieve group fairness and by *sampling*, where sample sizes of the subgroups are changed to make the dataset fair. Pre-processing approach is used for algorithms that have access to the training data and can modify it. The fair pre-processed data can thereafter be used to learn any ML algorithm to make fair decisions.

3.2.2 | In-processing

In-processing methods achieve fairness by explicitly introducing extra fairness constraints in the training algorithm. The authors of Kamiran, Calders, and Pechenizkiy (2010) modified the splitting criterion of a decision tree classifier to ensure fairness. Zhang, Lemoine, and Mitchell (2018) used Adversarial Debiasing (AdDeb) to learn classifiers that maximize accuracy while removing influence of sensitive attributes on predictions. The inability of the adversary to predict the sensitive group based on the predictions made by the predictor shows that the predictions are fair since they are not dependent on the protected attributes. A discrimination-aware regularization term is added to the objective function of a logistic regression model in Kamishima, Akaho, Asoh, and Sakuma (2012)'s Prejudice Remover (PR). The prejudice remover can be applied to ensure fairness in any predictive algorithm with probabilistic discriminative model. Three approaches were used by Calders and Verwer (2010) for a naive bayes classifier viz. (a) modification of conditional probability distribution, (b) learning group-specific models, (c) adding a latent variable to the bayesian model. The work in Celis, Huang, Keswani, and Vishnoi (2018) introduces what they call a meta-algorithm (Meta) for classification which takes a generic collection of fairness constraints as a unifying framework to satisfy some of the existing fairness metrics. In-processing approaches are used in scenarios where access to and modification of the internals of the ML algorithm is possible.

3.2.3 | Post-processing

Post-processing approaches work by modifying the results of a previously trained model to achieve desired measures of fairness with respect to different groups. Post-processing may be regarded as shifting the decision boundary of baseline classifiers for different groups to achieve some fairness notion. Kamiran, Calders, and Pechenizkiy (2010) relabelled the leave nodes of a decision tree classifier to achieve group fairness. Hardt, Price, and Srebro (2016) used linear programming to find probabilities with which output labels were changed to optimize equal odds and equal opportunity. Similar to equal odds algorithm is another work by Pleiss, Raghavan, Wu, Kleinberg, and Weinberger (2017) called Calibrated Equal Odds (CalEqOdds) with a single error constraint (such as false positive parity or false negative parity). CalEqOdds optimizes over a calibrated classifier score outputs to determine the probabilities with which outputs are post-processed to satisfy equal odds constraint. Post-processing approaches are useful for blackbox systems where there is only access to the predicted outcomes.

3.3 | Fairness in LA

LA is increasingly being used by educational institutions to leverage actionable insights such as: predicting students at risk of failing a course Hlosta, Zdrahal, and Zendulka (2017) or learning outcomes Käser, Hallinen, and Schwartz (2017) for the necessary interventions to be made. Aside the obvious benefits that LA provides, their increased adoption and deployment has raised ethical concerns about fairness. We briefly overview literature on fairness in LA. For a comprehensive literature review of fairness in LA and LA itself in general, we refer readers to Baker and Hawn (2021); Yu, Lee, and Kizilcec (2021a) and Gardner and Brooks (2018); Lang, Siemens, Wise, and Gasevic (2017) respectively as that is beyond the scope of this paper. Although fairness has been quite extensively explored in other domains such as criminal justice, it is nascent in the context of LA. The limited examinations of fairness does not in effect imply limited evidence of unfairness in LA Blanchard (2012); Hu and Rangwala (2020); Ocumpaugh, Baker, Gowda, Heffernan, and Heffernan (2014); Yu, Li, Fischer, Doroudi, and Xu (2020). For example, some models tend to generalize well for urban and suburban students but not rural students Baker and Gowda (2010); Ocumpaugh, Baker, Gowda, Heffernan, and Heffernan (2014). Blanchard (2012) showed how students from WEIRD (white, educated, industrialized, rich and democratic) were oversampled

by an intelligent tutoring systems (ITS) compared to those from non-WEIRD countries. Also, models for predicting course failure were found by Hu and Rangwala (2020) to discriminate against African-Americans.

There have been a few works advancing the ethical LA agenda by designing fair LA models. In a "naive" attempt at unfairness mitigation,Yu, Lee, and Kizilcec (2021b) carried out a study to ascertain whether the inclusion or exclusion of protected attributes had an effect on fairness and performance of an LA model. They found no significant difference in performance and fairness between the protected attribute-aware and protected attribute-blind models. In order to correct the racial and gender bias in their LA model, Lee and Kizilcec (2020) post-processed a Random Forest model by setting protected group-specific classification thresholds to achieve equal opportunity. Similarly, Hu and Rangwala (2020), built an LA model to ensure fairness with respect to race and gender using metric free individual fairness. Their results interestingly suggest that biased data may not always necessarily lead to biased predictions. Given that there are many different options of LA models with varying levels of fairness and performance for a particular task, the decision on which one to choose becomes difficult. In light of that, Sha et al. (2021) compared four traditional LA models and two deep learning (DL) LA models. They found the traditional models to be relatively more fairer but less accurate compared to the DL models and vice versa. They also reported how simple equal sampling of protected groups improved fairness. Similar in result with respect to fairness but contrasting performance results,Kung and Yu (2020) also reported how interpretable models such as Logistic regression were equal or less unfair with interestingly no compromise in accuracy compared to complex models and thus advocating for interpretable LA models. Gardner, Brooks, and Baker (2019) also showed in their work no strict evidence of fairness-accuracy tradeoff. The authors performed a replication study on five Massive Open Online Course (MOOC) dropout prediction models and measured fairness using a metric they refer to as Absolute Between-ROC Area (ABROCA). Gardner, Brooks, and Baker (2019) found unfairness to be mostly associated with course gender imbalance. Given that there are existing algorithms in the fair ML community dedicated to unfairness mitigation, it is crucial to investigate how these algorithms can be leveraged to build fair LA models. To that effect, Riazy, Simbeck, and Schreck (2020) compared Kamishima, Akaho, Asoh, and Sakuma (2012)'s Prejudice Remover and Zafar, Valera, Rogriguez, and Gummadi (2015)'s Margin-based classifier with four baseline LA models. As expected, the dedicated unfairness mitigation algorithms improved the fairness compared to the baseline LA models. However, the breadth of unfairness mitigation algorithms in Riazy, Simbeck, and Schreck (2020)'s work is relatively small (i.e., two in-processing algorithms). To further advance their work, we comparatively evaluated some selected unfairness mitigation algorithms spanning the entire LA pipeline by testing them on some fairness and performance related hypotheses detailed in Section 5.

4 | DATA SETS, ALGORITHMS AND METRICS

4.1 | Data sets

We performed the experiments on five real world datasets. Four of them: (German credit Lichman (2013), Law School Kusner, Loftus, Russell, and Silva (2017), Compas Angwin, Larson, Mattu, and Kirchner (2016) and Voilent Crime Angwin, Larson, Mattu, and Kirchner (2016)) are commonly used in the fair ML community . We also used first semester records of 3 cohorts of students from 2015-2017 for a particular program in a large public Australian University to predict if they would dropout of the program within their 3-year duration of study. Dropout prediction is one of the most common LA tasks, as such, most of our analysis would be based on the dropout dataset. Unless otherwise stated, all datasets used are the preprocessed version from Friedler et al. (2019). For the Law school data, we used the preprocessed version used in Kusner et al. (2017). We binarized all categorical features in the dropout data. For non-binary categorical features, binarization was done using an indicator function. Some courses required more online engagement than others, thus to make them comparable, we "z-normalized" the online engagements for each specific course. Data details are summarised in Table 1 .

TABLE 1 Details about dataset used for experimental evaluation

Data	Domain	Samples	Protected Attribute	Prediction
Dropout	Education	696	Home Language	dropout
Law School	Education	21,790	Race	first year average
German Credit	Finance	1,000	Age	loan default
Compas	Criminal Justice	6,167	Race	recidivism
Violent Crime	Criminal Justice	4,010	Race	recidivism

4.2 | Measures

The fairness measurements for the experimental evaluation have already been discussed in Subsection 3.1. First of all, we chose these fairness measures because they are the most commonly used measures in the fair ML community Mehrabi et al. (2019); Verma and Rubin (2018). More so, a recent experiment by Srivastava, Heidari, and Krause (2019) examined most of the fairness measures that we considered in this work . In Srivastava, Heidari, and Krause (2019)'s experiment, the participants were each presented with 20 hypothetical questions w.r.t fairness across different contexts and were asked to choose an operationalisation of fairness that they deemed fair w.r.t each question and explain the reason behind their choice. Their results showed how these fairness measures (considered in this work) aligns with the *lay person's* perception of fairness. We would like to point out that although the fairness measures we considered are not exhaustive of *all* contexts, the collection is relatively large enough to generalise for most contexts. We measured predictive performance w.r.t accuracy, precision (aka PPV) and NPV. Using the dropout data as a case study, the accuracy indicates how often the model correctly predicts dropout and non-dropouts on the overall level. Zeroing in on each respective class, we use precision to measure how often students that were predicted to dropout actually dropped out. Conversely, we used NPV to measure how often students predicted not to dropout actually did not dropout. The predictive performance and fairness measures used in this study are summarized in Table 2

TABLE 2 Performance and fairness measures (GF= group fairness, IF= individual fairness). The full meaning of the performance and fairness measures can be found Subsection 3.1. Also, see the same for the full formulae for BGEI, BGTI, WGEI, and WGTI

Measure	Formula	Type
Accuracy	$TP + TN/(TP + TN + FP + FN)$	Performance
Precision	$TP/(TP + FP)$	Performance
NPV	$TN/(TN + FN)$	Performance
DI	$(P(\hat{Y} = 1\|A = 0))/P(\hat{Y} = 1\|A = 1) \geq \tau = 0.8$	GF
SP	$P(\hat{Y} = 1\|A = 0) - P(\hat{Y} = 1\|A = 1)$	GF
EO	$P(\hat{Y} = 1\|A = 0, Y = y) = P(\hat{Y} = 1\|A = 1, Y = y), y \in \{0, 1\}$	GF
EOP	$P(\hat{Y} = 1\|A = 0, Y = 1) - P(\hat{Y} = 1\|A = 1, Y = 1)$	GF
PPV-diff	$(TP/(TP + FP)_{A=0}) - (TP/(TP + FP)_{A=1})$	GF
FDR-diff	$(FP/(TP + FP)_{A=0}) - (FP/(TP + FP)_{A=1})$	GF
NPV-diff	$(TN/(TN + FN)_{A=0}) - (TN/(TN + FN)_{A=1})$	GF
FOR-diff	$(FN/(TN + FN)_{A=0}) - (FN/(TN + FN)_{A=1})$	GF
BGEI	$\varepsilon^{\alpha}_{\beta}(b), \alpha \notin \{0, 1\}$	GF
BGTI	$\varepsilon^{\alpha}_{\beta}(b), \alpha = 1$	GF
WGEI	$\varepsilon^{\alpha}_{\omega}(b), \alpha \notin \{0, 1\}$	IF
WGTI	$\varepsilon^{\alpha}_{\omega}(b), \alpha = 1$	IF
Consistency	$Cons_i = 1 - \frac{1}{Nk} \sum_{i=1}^{N} \sum_{j \in kNN(x_i)} \left\lvert \hat{y}_i - \hat{y}_j \right\rvert$	IF

4.3 | Algorithms

For our baseline models[3], we did model selection using H2O.ai's AutoML and found Gradient Boosted Machines (GBMs) had the best performance ahead of XGBoost, Random Forest and Extremely Randomised Trees. The AutoML module auto handles the hyperparameter tuning, we evaluated the hyperparameters by manually setting the nfolds argument to 10 for 10-fold cross validation. In addition to the GBM, we also used a Logistic Regression (LR) model as our baseline partly because of their interpretability and also Kung and Yu (2020)'s research showed they tend to be relatively fair compared to complex models without compromising accuracy. The unfairness mitigation algorithms that we evaluated in this work

[3]We refer to models without any fairness constraint (assumed to be biased) as baseline models. Subsequently, such models produce baseline predictions

are among those considered in the fair ML community to have shown promising results in unfairness mitigation. These algorithms have gained a lot of attention (see citation count in Table 3) and have undergone rigorous review process before being published in top-tier ML conferences such as NIPS and ICML. We opine that these algorithms are prime candidates for real-world adoption if there be such a situation. The details of the unfairness mitigation algorithms are summarised in Table 3 .

TABLE 3 Selected unfairness mitigation algorithms

Algorithm	Citations (October, 2021)	Type	Modifies
DIR	1006	Pre-processing	Biased Data
RW	526	Pre-processing	Biased Data
LFR	1063	Pre-processing	Biased Data
PR	466	In-processing	Biased Model
AdDeb	487	In-processing	Biased Model
Meta	128	In-processing	Biased Model
EqOdds	1856	Post-processing	Biased Predictions
CalEqOdds	392	Post-processing	Biased Predictions

5 | EXPERIMENTS, RESULTS AND DISCUSSION

Although we used many datasets for our evaluations, our analysis would be mostly based on the student dropout dataset. The other datasets are for benchmarking and test of generalizability of findings. The favorability of our prediction outcome (i.e. being predicted to dropout) may be regarded as a two-sided coin. For instance, if being predicted to dropout comes with some form of intervention package to help the "at-risk" student, then we consider that to be a favourable outcome. However, if being predicted to dropout negatively affects the reputation of the student, then it becomes unfavourable. In this work, we consider being predicted to dropout as favourable. English as home language students are considered as privileged group and all others unprivileged.

5.1 | Experiments

We randomly shuffled each dataset into ten shuffles and performed five-fold cross validation on each shuffle. We thus have 50 results for each performance and fairness metric for every algorithm. For the baseline algorithms, we used the H2O.ai python implementation. We have already discussed in Subsection 4.3 how the internal tunings were done. For the unfairness mitigation algorithms, we used implementations in IBM's AI Fairness 360 (AIF360) package Bellamy et al. (2018). DIR has a hyper-parameter λ for tuning the tradeoff between fairness and accuracy. $\lambda = 0$ signifies no-fairness while $\lambda = 1$ ensures maximum fairness. We experimented on values of λ in $[0, 1]$ space. PR also has a hyper-parameter η. Just like the authors of the algorithm, we explored the influence of the η values between 0 and 100. We do same for Meta's τ hyper-parameter at increments of 0.2 in $[0, 1]$.

The operation of the pre-processing and post-processing algorithms are not straightforward like the in-processing algorithms. For the pre-processing, we first transform (i.e., pre-process) the "biased data" into a "fair data" and then train the GBM and LR on the "fair data". The predictions of the now fair GBM and LR models on the test set are then evaluated for fairness and performance. In the case of the post-processing, we first train the *supposedly biased* GBM and LR models, and then post-process their predictions on the the test set based on a fairness constraint. We compared all algorithms using the average statistic of all performance and fairness measures for all shuffles. The standard deviations are shown by error bars in plots. We would like to state that since some algorithms are designed specifically to satisfy some specific fairness measures, those algorithms may perform better than others when all the algorithms are compared using that specific measure. We only included figures that are relatively sufficient to explain the our findings. Please refer to the supplementary material for the remaining figures.

We formulate our results and discussions as hypothetical questions that we seek to answer. Although some of the questions have already been asked in literature, just as Gardner, Brooks, and Baker (2019) puts it: replication studies helps to either solidify the substantiality of discoveries in prior works, or report new discoveries. More so, it has been shown that up to date, there is limited replication studies in LA: estimated at approximately 0.13%.Gardner, Brooks, and Baker (2019). Our evaluation questions are as follows:

5.2 | Do models trained without fairness constraint always replicate biases in data?

Most of the existing research in fair ML show that models trained without some fairness constraint leads to unfair outcomes. Our aim is to evaluate if data bias "necessarily" result in predictive bias. The only fairness metrics we could compute from the the "ground truth" in the data itself are disparate impact (DI), statistical parity (SP) and **consistency**. We computed these metrics from the biased test data and compared them with those computed from the predictions of the baseline models and the fairness-aware models. For some metrics, the baseline models as well as some fairness-aware models replicated (or exacerbated) the biases in the data. From Figure 1 , we observed that unfairness w.r.t SP was either replicated or marginally exacerbated by the baseline models across all datasets. With our student dropout prediction, this implies that in the situation where such models are deployed, some Non-English as home language students likely to drop-out may end up not being identified for the necessary interventions to be made. This is consistent with the unequal dropout distributions (50.33% for Non-English and 64.95% for English as home language students) in the dataset — perhaps, another reason why equal sampling should be encouraged in order to ensure fair LA models. Contrary but not entirely surprising, in Hu and Rangwala (2020)'s work, an unconstrained Logistic regression model made fair predictions even though there was evidence data bias — of course the dataset used and the how it was preprocessed before model training plays a huge role. Similarly, our GBM model had marginally better fairness in terms of DI on the law school and dropout data compared to the DI in the bias data. We observed that some unfairness mitigation algorithms had relatively less fairer results w.rt. to some fairness metrics compared to the baseline. Hu and Rangwala (2020) also made similar findings. This however is not entirely surprising due to the "impossibility theorem" stated earlier. From Figure 1 , the value for **consistency** metric in the biased data remained unchanged in the baseline predictions and most of the fairness-aware models. This however is expected since those algorithms are designed to satisfy group fairness. The LFR algorithm was introduced with **consistency**, thus, its not surprising that is satisfies **consistency** better than other algorithms. Intuitively, this suggests that it would be prudent for a decision maker wanting to use some unfairness mitigation algorithm to firstly use the fairness metric that came bundled with the algorithm except in cases of "metric-agnostic" algorithms. Overall, we find that data bias does not necessarily result in predictive bias; sometimes (rarely), even the baseline models may be relatively fairer compared to some fairness-aware models w.r.t some metrics.

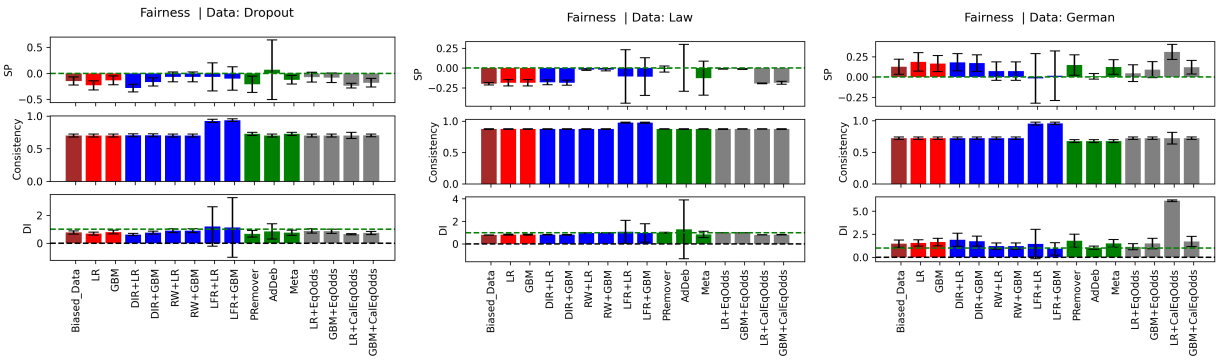

FIGURE 1 Comparing unfairness in biased data (data) with unfairness in predictions by a (fairness-aware or fairness-unaware (baseline)) model. Ideal fairness value is indicated by green short-dashed line. For consistency, ideal value= 1. Color code: Biased data= brown, baseline= red, pre-processing= blue, in-processing= green, post-processing= gray.

5.3 | Do constraints always adversely affect predictive performance?

We also evaluated for a simple yet important question as to whether models trained without any fairness constraint performed better than those with fairness constraints. We compared the predictive performance of the baseline models with fairness-aware models across 3 performance metrics. From Figure 2 , for the dropout dataset, we observed marginal differences in predictive performance between the fairness-aware models and fairness-unaware models across all three performance metrics — accuracy, precision and NPV. We observed that, the LFR algorithm actually improves predictive performance compared to the baseline model for the dropout data. We think this is probably because the latent representation of the biased data learnt by the LFR algorithm was "richer" than the biased data. This performance of the LFR algorithm is however not consistent across all datasets. On average, we find that the fairness-aware models had slight dip in performance metrics across datasets. This suggests that aiming for ethical LA may sometimes come at a price. We further discuss this fairness-utility tradeoff in Subsection 5.4.

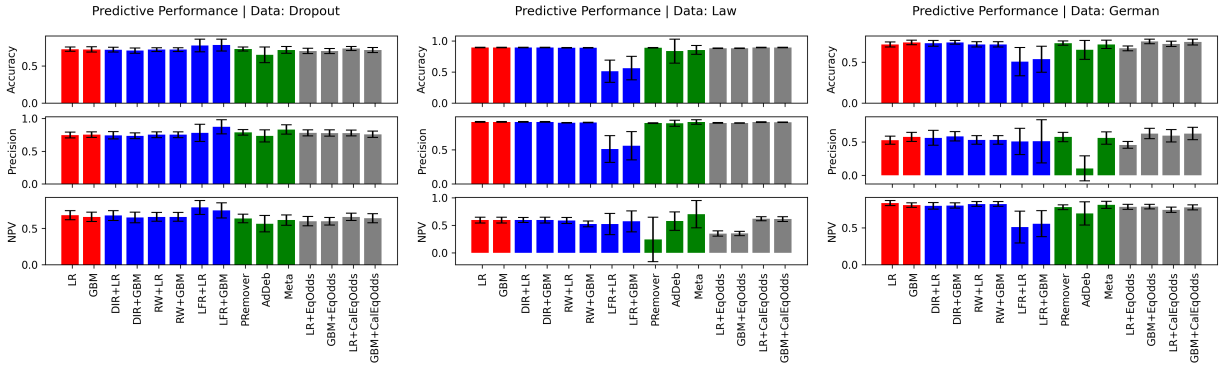

FIGURE 2 Comparing predictive performance of baseline models with fairness-aware models. Color code: baseline= red, pre-processing= blue, in-processing= green, post-processing= gray.

5.4 | Does more fairness necessarily imply less predictive performance and vice versa?

Which unfairness mitigation algorithm is able to ensure much fairness without compromising a lot of utility [4]? In this section we compare the baseline and the various unfairness mitigation algorithms. For some algorithms, we observed that improving fairness caused a slight dip in performance and vice versa. From Figure 3 , we observed that there is no single "winner" algorithm across all metrics and datasets. For example, for the dropout dataset, the LFR was able to find a relatively better balance between predictive performance and fairness, however it performed worse compared to the other algorithms on the law school and german credit data. Conversely, the AdDeb performed badly compared to the other algorithms on the dropout data but it was able to find a relatively better fairness-utility tradeoff on the law school data. Perhaps not surprisingly, we observed that majority of the algorithms had similar individual fairness-utility tradeoff and varied group fairness-utility tradeoff. Although no algorithm is consistent in finding an **optimal** balance between utility and the various fairness measurements, on the overall level, in a real-world (fair dropout prediction) deployment situation, our results show that the LFR might likely be the first choice as it finds a relatively better balance between performance and fairness (both group and individual) compared to the other algorithms.

Also for algorithms that had hyper-parameters, for determining fairness-utility tradeoff, we evaluate this tradeoffs at different hyper-parameter levels. From Figure 4 , for the prejudice remover, we observed that increasing the η value had very marginal effect on utility. However, there was consequent improvement in SP and EO across datasets. Increasing the η value barely had an effect on the BGEI, this is not surprising as the algorithm was not designed to satisfy BGEI, moreover, the BGEI had already been satisfied even at $\eta = 0$. We found the Meta algorithm to be very sensitive to variations in its hyper-parameter τ. For the dropout data, we observed that increasing τ resulted in an increase in fairness w.r.t SP, WGEI and EO and a consequent decrease in all performance metrics. This goes on till at $\tau = 0.8$ where a further increase in τ actually worsens fairness and rather improves accuracy and precision. We observed similar phenomena for the violent crimes and the german credit data. This we find quite surprising because the higher the τ value, the more we expected the fairness measure to approach 0. This suggests that sometimes, sacrificing utility does not necessarily always improve fairness. Therefore, a careful hyper-parameter optimization is very crucial.

For the DIR, we generated fair datasets at different repair levels and trained and tested the GBM and LR models at each repair level. From Figure 5 , we observed very different results for the two models trained on the fair data produced by the DIR. This was an interesting observation, such that given two different models trained on the same fair data, the internals of each respective model may lead to different fairness results. For the logistic regression model, we also found that for every increase in repair level, there was a corresponding increase in accuracy till at $\lambda = 0.5$ for the dropout data and $\lambda = 0.7$ for the violent crime and german credit datasets. At these respective repair levels, a further increase in repair led to a fall in accuracy. Again, careful hyper-parameter selection is key. The increase in accuracy of the GBM and LR models with increase in repair level is consistent with what we found for the LFR using the dropout dataset. This suggests that sometimes, the new forms of data learnt by the pre-processing algorithms are "richer" for training LA models than the biased data.

5.5 | What is the consistency between group fairness and individual fairness

We do not dispute the existence of incompatibilities that have been shown by the earlier cited works, however, unlike the "impossibility theorem", we introduce what we term "consistency theorem". In this work, we investigate how consistent some correlated group and individual fairness

[4] we use performance and utility interchangeably

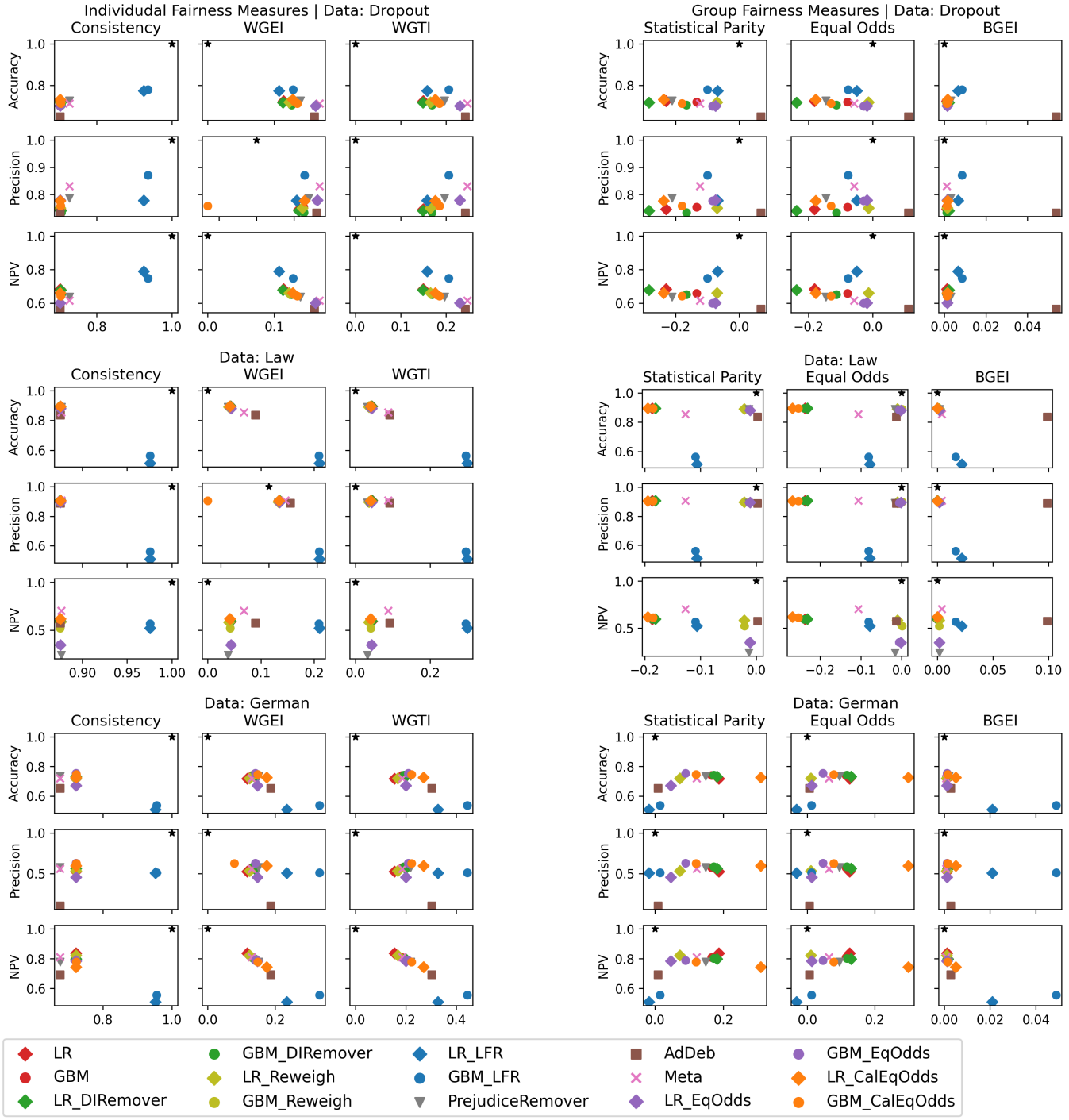

FIGURE 3 Utility-Fairness tradeoff of algorithms. Ideal fairness-utility tradeoff values are indicated by means of a black star on the plots. Different shapes refer to different algorithms while different colors refer to different "fairness variant" of a particular algorithm. For instance, red dots (o) refer to a baseline GBM while blue dots (o) refer to a GBM model trained on a fair data from the LFR algorithm. First three columns are individual fairness-utility tradeoff and last 3 columns are group fairness-utility tradeoff

measures are. More formally, assuming there is a decision maker wanting to achieve consistency of two fairness measures out of a pool of correlated group fairness measures $gf_1, gf_2, gf_3 ... gf_k$ and individual fairness measures $if_1, if_2, if_3 ... if_k$, our work is to show whether any two measures for example, gf_1 and if_1 are consistent and the algorithms that satisfy these consistencies. The knowledge of these consistencies would help the decision maker in deciding which unfairness mitigation algorithm to choose in order to consistently achieve the desired fairness measures simultaneously.

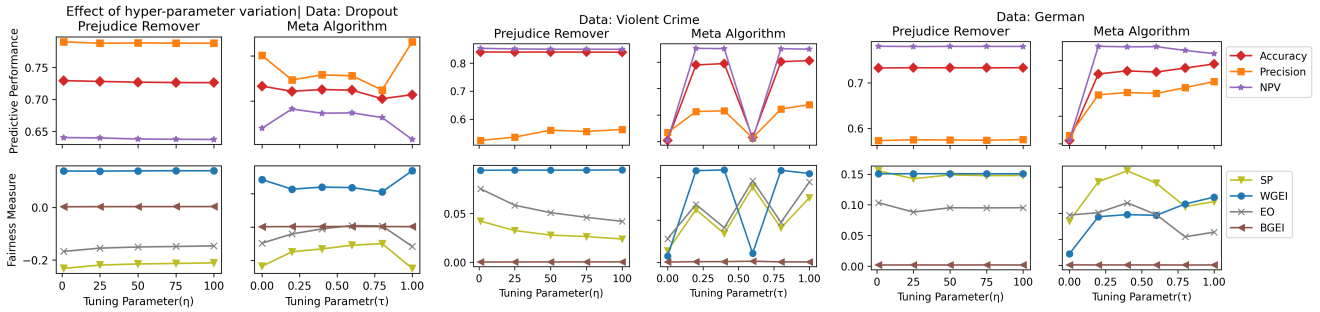

FIGURE 4 Effect of hyper-parameter variations of Prejudice remover, Meta Classifier on performance and fairness. Ideal value for fairness measures is 0 and that of performance measures is 1. Note: Hyper-parameters can be thought of as "sliders", i.e., increasing hyper-parameter value should increase fairness with (a probable) drop in performance and vice versa

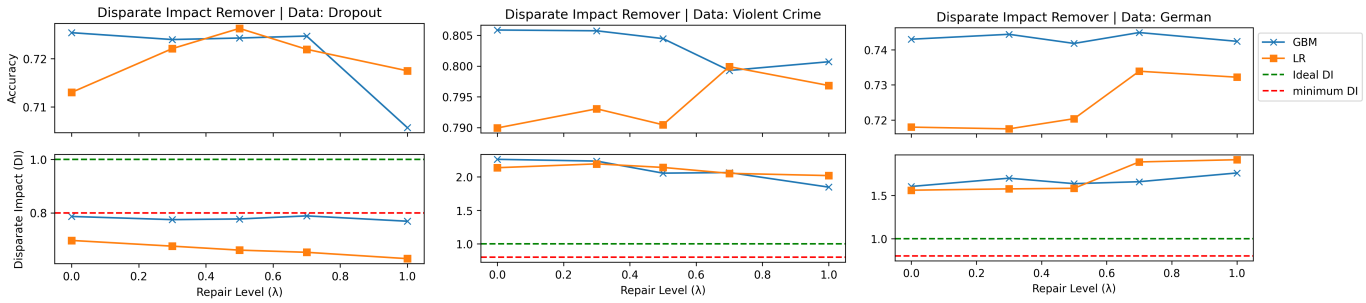

FIGURE 5 Effect of DIR repair level variations on accuracy and disparate impact of GBM and LR models

From Figure 6 ,we observed that all algorithms did well in satisfying the BGEI. As reported by Speicher et al. (2018), we observed that increasing the between-group fairness in most cases may relatively worsen the within-group fairness. We opine that "bounded consistency" can be achieved if the fairness measurements are chosen to lie within a lower bound. For example, consider the WGEI vs BGEI for the dropout dataset in Figure 6 , if we set the WGEI to lie within [0-0.13] and BGEI to lie within [0-0.1], majority of the algorithms would satisfy both measures within these bounds. For our student dropout prediction, this bounded consistency implies that most of the algorithms (LFR, DIR, CalEqOdds, Reweighing and even sometimes the baseline) would assign favourable outcomes to both within and between student groups whose home language are English and not English. Depending on how high the stakes involved are, a higher threshold value would be ideal. Also, we found the LFR algorithm tends to perform relatively better in satisfying statistical parity and **consistency** across datasets compared to the other algorithms. Again, this is quite expected as the LFR was designed with **consistency** and statistical parity in mind. We observed many correlations between the various fairness metrics. For instance, equal opportunity was correlated with WGEI for most algorithms across datasets. Intuitively, this suggests that when students who dropped out are actually predicted to dropout at equal rates regardless of their protected attribute, it equally benefits individuals within each demographic group. Overall, we observed that no single algorithm is able to consistently satisfy all fairness measures optimally. In order achieve **a near-optimal** consistency, if possible, an algorithm may be designed to optimize for consistency of some group fairness measure and some other individual fairness measure that are not affected by the impossibility theorem.

6 | CONCLUSION

We analysed some selected unfairness mitigation algorithms based on some hypotheses to comparatively evaluate how they contribute to ethical LA. Our results aligns with similar findings of Hu and Rangwala (2020) — although most LA models may pick up biases in training data, data biases do not always necessarily imply predictive bias. We found that there are instances where some unfairness mitigation algorithms perform worse with respect to some fairness metric compared to a no-fairness model. Also, with regard to utility-fairness tradeoff, we find that, optimizing for more fairness does not always lead to reduction on predictive performance. More interestingly, for algorithms that modify the data (specifically LFR and DIR), we found that sometimes, the debiased data is more "richer" for ensuring accurate predictions compared to the biased

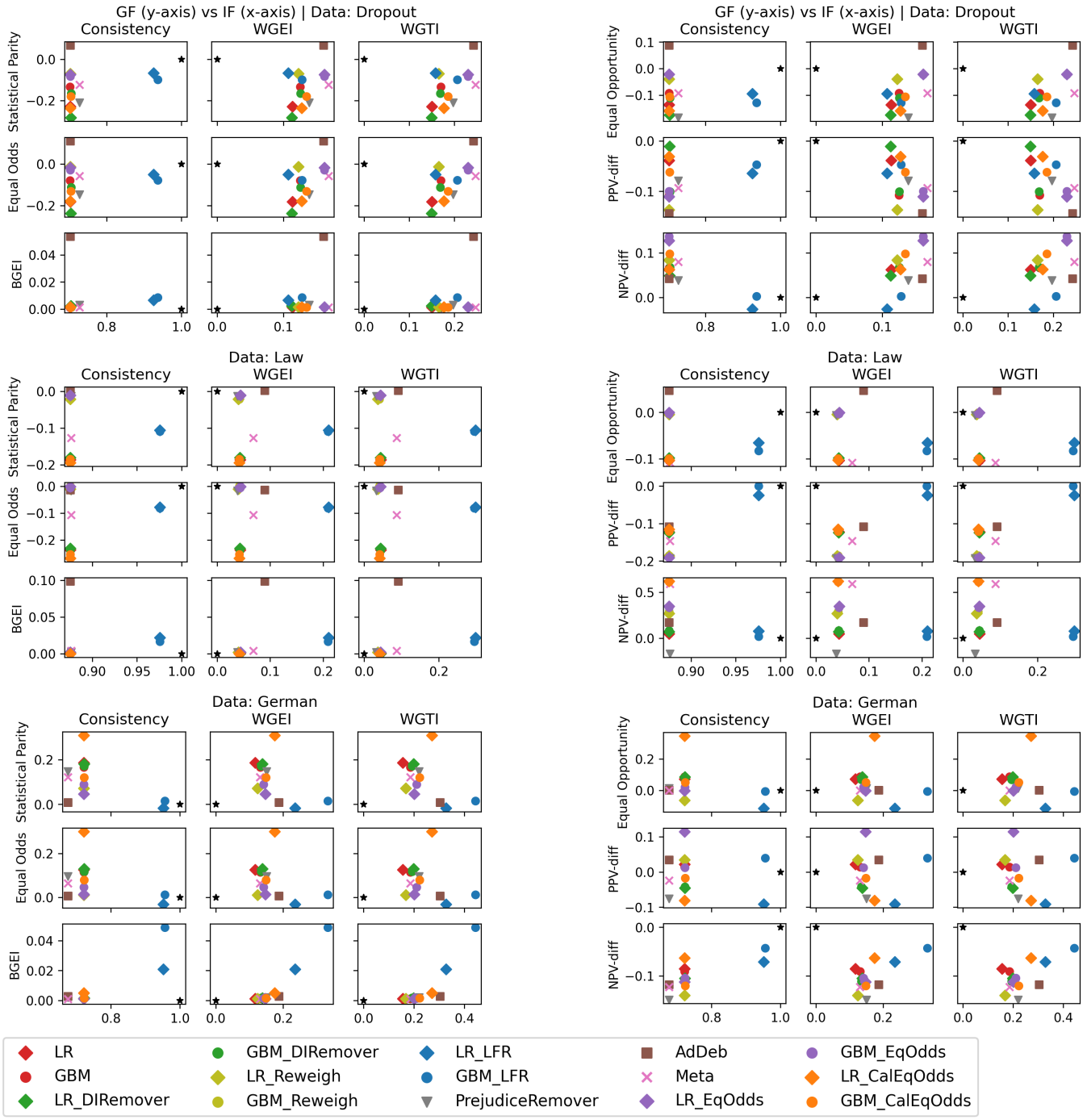

FIGURE 6 Consistency of group and individual fairness across the baseline and the various unfairness mitigation algorithms. Ideal points of consistency are indicated by means of a black star on the plots

data. Furthermore, our results suggest that if reasonable thresholds are decided by policy-makers we can have unfairness mitigation algorithms that satisfy bounded consistency. For algorithms that provide hyper-parameters for fairness-accuracy tradeoff, our results suggest that careful hyper-parameter optimization is the major key to get the right balance between fairness and utility.

We believe that humans are the ones at receiving end of algorithmic decisions. Therefore issues of fairness should not be entirely left to models satisfying some mathematical definition of fairness. There should be a human-in-the-loop who makes final decisions.

All of the algorithms we evaluated are based on the equality of some fairness metric. We suggest its about time fair LA focused a lot more on equity. Equality is feasible only when there is a levelled-playing ground. Thus in LA, the goal should be equitable LA where students are given not

necessarily equal treatment but rather "needed" treatment to ensure their success. How to measure what is "needed" in itself is a research worth investigating.

AUTHORS

Deho Oscar Blessed received his bachelor's degree in Computer Science and Engineering from University of Mines and Technology, Tarkwa, Ghana. He is currently a PhD student in Computer and Information Science at the University of South Australia, Australia. His main research interest are fairness in machine learning, learning analytics and explainable AI. ORCID: 0000-0001-5723-2564

Chen Zhan is a research associate and data scientist at the Centre for Change and Complexity in Learning (C3L) at the University of South Australia (UniSA). He obtained his master degree in software engineering from the University of Science and Technology of China (USTC) in 2016 and completed his PhD in computer and information science from UniSA in 2020. His research centres around data mining and its application in various disciplines, including environmental science, economics, bioinformatics and pharmacoepidemiology. Recently, he focuses on artificial intelligence and learning analytics for the education sector. In particular, he contributes to the research in the development of products and services of privacy-preserving learning analytics for the education technology industry. ORCID: 0000-0002-4794-8339

Jiuyong Li received the PhD degree in computer science from Griffith University, Brisbane, Australia, in 2002. He is currently a Professor at the University of South Australia, Australia. His main research interests include data mining, causal discovery, privacy and fairness, and bioinformatics. His research work has been supported by seven Australian Research Council Discovery projects and he has led several industry and applied projects. He is a member of the Australian Computer Society National Committee for Artificial Intelligence. ORCID:0000-0002-9023-1878

Jixue Liu got his PhD in computer science from the University of South Australia. He is an Associate Professor in the University. He has published widely in databases and artificial intelligence. His work covers the topics of integrity constraint discovery, data analytics in texts and time series, entity linking, fairness computing, privacy in data, XML functional dependencies and data integration and transformation. ORCID:0000-0002-0794-0404

Lin Liu received her Bachelor and Master degrees in electronic engineering from Xidian University, Xi'an, China, and her PhD in computer systems engineering from the University of South Australia, Australia. She is currently an Associate Professor at the University of South Australia. Her research interests include data mining, machine learning, causal inference, and bioinformatics. ORCID: 0000-0003-2843-5738

Thuc Duy Le is an associate professor at the University of South Australia. His research focuses on the development of causal discovery methods and their applications in bioinformatics. He is currently a DECRA fellow and was also a National Health & Medical Research Council (NHMRC) ECR Fellow (2017-2019). He has served as an academic editor and a reviewer for conferences in data mining and journals in bioinformatics. ORCID: 0000-0002-9732-4313

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals [Journal Article]. *And it's biased against blacks. ProPublica*, *23*.

Baker, R. S., & Gowda, S. M. (2010). An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools [Conference Proceedings]. In *Educational data mining 2010*. ERIC.

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education [Journal Article].

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., … Mojsilovic, A. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias [Journal Article]. *arXiv preprint arXiv:1810.01943*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art [Journal Article]. *Sociological Methods and Research*, 0049124118782533.

Blanchard, E. G. (2012). On the weird nature of its/aied conferences [Conference Proceedings]. In *International conference on intelligent tutoring systems* (p. 280-285). Springer.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification [Journal Article]. *Data Mining and Knowledge Discovery*, *21*(2), 277-292.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention [Conference Proceedings]. In *Advances in neural information processing systems* (p. 3992-4001).

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2018). Classification with fairness constraints: A meta-algorithm with provable guarantees [Conference Proceedings]. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 319-328).

Chiappa, S. (2019). Path-specific counterfactual fairness [Conference Proceedings]. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, p. 7801-7808).

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

Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017). From prediction to impact: Evaluation of a learning analytics retention program [Conference Proceedings]. In *Proceedings of the seventh international learning analytics & knowledge conference* (p. 474-478).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness [Conference Proceedings]. In *Proceedings of the 3rd innovations in theoretical computer science conference* (p. 214-226).

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact [Conference Proceedings]. In *proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (p. 259-268).

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning [Conference Proceedings]. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 329-338).

Gardner, J., & Brooks, C. (2018). Student success prediction in moocs [Journal Article]. *User Modeling and User-Adapted Interaction*, *28*(2), 127-203.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis [Conference Proceedings]. In *Proceedings of the 9th international conference on learning analytics & knowledge* (p. 225-234).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning [Conference Proceedings]. In *Advances in neural information processing systems* (p. 3315-3323).

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data [Conference Proceedings]. In *Proceedings of the seventh international learning analytics & knowledge conference* (p. 6-15). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3027385.3027449` doi: 10.1145/3027385.3027449

Hu, Q., & Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students [Conference Proceedings]. In *Proceedings of the 13th international conference on educational data mining (edm 2020)* (p. 431-437).

Joksimović, S., Kovanović, V., & Dawson, S. (2019). The journey of learning analytics [Journal Article]. *HERDSA Review of Higher Education*, *6*, 27-63.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination [Journal Article]. *Knowledge and Information Systems*, *33*(1), 1-33.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning [Conference Proceedings]. In *2010 ieee international conference on data mining* (p. 869-874). IEEE.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer [Conference Proceedings]. In *Joint european conference on machine learning and knowledge discovery in databases* (p. 35-50). Springer.

Käser, T., Hallinen, N. R., & Schwartz, D. L. (2017). Modeling exploration strategies to predict student performance within a learning environment and beyond. In *Proceedings of the seventh international learning analytics & knowledge conference* (p. 31–40). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3027385.3027422` doi: 10.1145/3027385.3027422

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores [Journal Article]. *arXiv preprint arXiv:1609.05807*.

Kung, C., & Yu, R. (2020). Interpretable models do not compromise accuracy or fairness in predicting college success [Conference Proceedings]. In *Proceedings of the seventh acm conference on learning@ scale* (p. 413-416).

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness [Conference Proceedings]. In *Advances in neural information processing systems* (p. 4066-4076).

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (2017). *Handbook of learning analytics* [Book]. SOLAR, Society for Learning Analytics and Research New York, NY, USA.

Lee, H., & Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success [Journal Article]. *arXiv preprint arXiv:2007.00088*.

Lichman, M. (2013). Uci machine learning repository. *http://archive.ics.uci.edu/ml*.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making [Journal Article]. *Data Mining and Knowledge Discovery*, *31*(4), 1060-1089.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning [Journal Article]. *arXiv preprint arXiv:1908.09635*.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics [Conference Proceedings]. In *Proc. conf. fairness accountability transp., new york, usa* (Vol. 1170).

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection [Journal Article]. *British Journal of Educational Technology*, *45*(3), 487-501.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration [Conference Proceedings]. In *Advances in neural information processing systems* (p. 5680-5689).

Riazy, S., Simbeck, K., & Schreck, V. (2020). Systematic literature review of fairness in learning analytics and application of insights in a case study [Conference Proceedings]. In *International conference on computer supported education* (p. 430-449). Springer.

Romei, A., & Ruggieri, S. (2011). *A multidisciplinary survey on discrimination analysis* [Generic]. Università di Pisa.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness [Conference Proceedings]. In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (p. 99-106).

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D., & Chen, G. (2021). Assessing algorithmic fairness in automatic classifiers of educational forum posts [Conference Proceedings]. In *International conference on artificial intelligence in education* (p. 381-394). Springer.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices [Conference Proceedings]. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (p. 2239-2248).

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning [Conference Proceedings]. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (p. 2459-2468).

Verma, S., & Rubin, J. (2018). Fairness definitions explained [Conference Proceedings]. In *2018 ieee/acm international workshop on software fairness (fairware)* (p. 1-7). IEEE.

Yu, R., Lee, H., & Kizilcec, R. F. (2021a). Bias in education [Journal Article]. *arXiv preprint arXiv:2103.15237*.

Yu, R., Lee, H., & Kizilcec, R. F. (2021b). Should college dropout prediction models include protected attributes? [Journal Article]. *arXiv preprint arXiv:2103.15237*.

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards accurate and fair prediction of college success: evaluating different sources of student data [Conference Proceedings]. In *Proceedings of the 13th international conference on educational data mining (edm 2020)* (p. 292-301). ERIC.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification [Conference Proceedings]. In *Artificial intelligence and statistics* (p. 962-970). PMLR.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations [Conference Proceedings]. In *International conference on machine learning* (p. 325-333).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning [Conference Proceedings]. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (p. 335-340).