A Comprehensive Strategy to Bias and Mitigation in Human Resource Decision Systems

Silvia D'Amicantonio[1,2,†], Mishal Kizhakkam Kulangara[1,†], Het Darshan Mehta[1,†], Shalini Pal[1,†], Marco Levantesi[1,3,*], Marco Polignano[4], Erasmo Purificato[5,‡] and Ernesto William De Luca[1,3]

[1]Otto von Guericke University Magdeburg, Magdeburg, Germany
[2]Polytechnic University of Milan, Milan, Italy
[3]Leibniz Institute for Educational Media | Georg Eckert Institute, Brunswick, Germany
[4]University of Bari Aldo Moro, Bari, Italy
[5]Joint Research Centre, European Commission, Ispra, Italy

Abstract

In recent years, Machine Learning (ML) and Artificial Intelligence (AI) models have become integral to various business operations, especially within Human Resource (HR) systems. These models are primarily used to automate decision-making processes in recruitment, performance assessment, and employee management, enhancing efficiency and streamlining tasks. However, the increasing use of these automated systems has raised significant concerns about the presence of bias, which can lead to discriminatory practices. Such biases may exclude qualified candidates and diminish opportunities, while also posing substantial risks to a company's reputation, with potential legal and ethical consequences. This paper addresses these challenges by exploring the root causes of bias in HR-related ML models and proposing best practices for mitigation. It presents a thorough examination of fairness concepts and definitions within the context of HR decision-making, emphasizing the complex nature of selecting appropriate mitigation techniques based on the specific models and datasets used. Through an empirical evaluation of various mitigation strategies, the study reveals that no single approach can fully satisfy all fairness metrics, highlighting the inherent trade-offs between accuracy and fairness. The findings offer valuable insights into optimizing these trade-offs and provide actionable recommendations for achieving fairer, unbiased outcomes in automated HR systems. Additionally, this research underscores the ongoing need for further study and discussion to enhance transparency and fairness in ML models, contributing to a more equitable HR landscape.

Keywords

Machine Learning, Biases and Fairness, Human Resource Decision-Making, Mitigation Strategies

1. Introduction

The rise of Artificial Intelligence (AI) and Machine Learning (ML) has revolutionized numerous industries, with Human Resources (HR) being one of the most significantly impacted [2, 3, 4]. Globally, companies are increasingly adopting these technologies to enhance decision-making capabilities and boost efficiency [5]. A specific class of ML technologies, commonly referred to as Black Box mod-

XAI.it - 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024 [1]

*Corresponding author.
†These authors contributed equally and are listed in alphabetical order.
‡The author contributed to this work while affiliated with Otto von Guericke University Magdeburg, Germany. The view expressed in this paper is purely that of the author and may not, under any circumstances, be regarded as an official position of the European Commission.
silvia.damicantonio@mail.polimi.it (S. D'Amicantonio); mishal.kizhakkam@st.ovgu.de (M. K. Kulangara); het.mehta@st.ovgu.de (H. D. Mehta); shalini.pal@st.ovgu.de (S. Pal); marco.levantesi@ovgu.de (M. Levantesi); marco.polignano@uniba.it (M. Polignano); erasmo.purificato@acm.org (E. Purificato); deluca@ovgu.de (E. W. De Luca)
https://hcai.ovgu.de/Staff/Ph_D+Students/Marco+Levantesi.html (M. Levantesi); https://marcopoli.github.io/ (M. Polignano); https://erasmopurif.com/ (E. Purificato); https://ernestodeluca.eu/ (E. W. De Luca)
0009-0001-3740-7539 (M. Levantesi); 0000-0002-3939-0136 (M. Polignano); 0000-0002-5506-3020 (E. Purificato); 0000-0003-3621-4118 (E. W. De Luca)

els [6, 7, 8], is characterized by the opacity of their internal workings as they take inputs and produce output, but the decision-making process remains hidden. Major corporations such as Google, IBM, SAP, and Microsoft are already utilizing these algorithmic systems for automated HR management [9]. Black Box models in HR streamline key functions such as recruitment and performance evaluation. The main drivers for their adoption include cost and time savings, increased productivity, and enhanced certainty in decision-making [10, 11]. AI is widely used to evaluate employee engagement and retention by analyzing feedback surveys and performance data. These insights are then applied to monitor achievement, recommend personalized job opportunities, and set objectives. Additionally, AI tools can assist in corrective actions for underperformance and inappropriate behaviours, and even support training by identifying employees likely to make errors and suggesting relevant skill-improvement programs. In recruitment, AI primarily contributes by screening resumes, identifying key terms in job applications, and analyzing video interviews to evaluate job fit and match candidates to open positions.

Although AI is often viewed as providing fairer, more impartial decision-making than humans, recent studies reveal a high risk of bias and discrimination in these systems [9, 12, 13, 14, 15]. Bias can manifest in several ways during the implementation of decision-making algorithms. For instance, historical data used for training models may reflect past societal imbalances, resulting in these biases being reproduced in AI-driven decisions [16]. The opaque nature of Black Box models exacerbates this issue, making bias identification and mitigation particularly challenging. The complexity of the underlying algorithms and deep learning techniques makes these models difficult to interpret. This lack of trasparency poses ethical and legal risks, potentially leading to discriminatory hiring practices and damaging a company's reputation [9]. In response, researchers and developers have proposed various strategies to address these biases, including using more diverse training datasets, implementing fariness-aware algorithms, and ensuring greater transparency and accountability in AI systems [17, 18]. While AI offers tremendous potential to enhance HR processes, it is essential to recognize and mitigate the biases these systems may introduce. Achieving fair and unbiased AI in HR requires a combination of better data practices, increased transparency, regulation, and continuous scrutiny and adjustment of AI models. This paper explores the inherent biases in HR-related Black Box models and outlines strategies for mitigating these biases to ensure fair and equitable decision-making.

2. Related Work

2.1. Understanding Biases and Mitigation Techniques

To fully grasp the reasons behind bias algorithms, it is essential to first review the concept of bias. We refer to *Cognitive Bias* as the type of bias that can be introduced in hiring processes supported by AI [19]. When the latter is used in hiring, the lack of transparency and accountability can heighten the risk of replicating social discrimination. The following subsections explore potential causes of bias and propose strategies to mitigate them.

2.1.1. Source and Implication of Bias

Cognitive bias, a well-documented phenomenon in human decision-making, can also affect AI-driven recruitment. Soleimani *et al.* [20] identify two primary sources of bias in AI: the **training dataset** and the **algorithm itself**. Training datasets often contain historical data, which may include underrepresented or overrepresented groups. Furthermore, these datasets may encode biases related to sensitive attributes due to mislabeled data. This can result in the exclusion of highly qualified candidates or even lead to legal issues as a consequence of violating anti-discrimination laws [12] [21]. Algorithmic biases can arise when developers make subjective assumptions or use inappropriate selection criteria. For example, including ethnicity, culture or gender in an algorithm can lead to wrong correlations between these attributes and the target variable [22]. Finally, algorithms could fail to account for job-specific requirements and produce decisions that are misaligned with the actual needs of the position [12].

	Source of Bias	Mitigation Best Practices
Dataset	Training dataset non-representative Training dataset out of date	- Expand datasets sources - Keep datasets up to date - Blind recruitment
Algorithm	Unable to formulate assumptions Unable to account for context-specific requirements	- Knowledge Sharing - Third parties Audits

Table 1
Sources of Bias and Mitigation techniques identified in our study. The sources are divided into arising from polluted training dataset or improper algorithm [20]. The mitigation techniques are classified accordingly.

2.1.2. Mitigation Strategies

Bias mitigation can be addressed at multiple stages of AI tool development. First, ensuring that the training dataset is representative of the population is crucial. Data should be sourced from diverse demographic groups and regularly updated to prevent the perpetuation of historical biases. Vivek [23] suggests that **blind recruitment** is an effective method for reducing unconscious bias. In the context of AI, blind hiring involves masking potentially bias-inducing variables from resumes in order to let the algorithm focus purely on skills and experience of candidates. Another key strategy for mitigating algorithmic bias is **knowledge sharing** between AI developers and HR professionals. Soleimani *et al.* [20] demonstrate that exchanging information at different stages of development improves recruitment model performance. Finally, **independent audits** and periodic assessments are vital for detecting biases and ensuring that the algorithm remains fair over time [12]. It is also suggested to release audit results since it can build trust with consumers and ensure transparency. In Table 1 the potential sources of bias as well as their mitigation techniques are summarized.

2.1.3. Regulatory and Ethical Consideration

The rapid expansion of automated decision-making systems has highlighted the need for government regulation to ensure fairness for all individuals. Many countries have enacted laws to prevent discrimination based on ethnicity, gender, religion or nationality [24]. In the context of employment, the EU AI Act (Annex III: Article 6(2)) classifies AI systems used in recruitment, employee management, and termination as high-risk. As per the law, systems must be regulated to ensure fairness, transparency, and non-discrimination in hiring and workplace decisions, thus minimizing bias to protect individual's rights [25]. Additionally, the U.S. Equal Employment Opportunity Commission (EEOC) has established guidelines, like the usage of the *four-fifths rule* (Table 2), to promote equal employment opportunities and prevent bias during the hiring process [26]. Generally speaking, three main theories of discrimination are often used to analyze bias:

- **Disparate Treatment**: Refers to intentional discrimination based on protected characteristics [27, 26]. Using sensitive attributes to build the model can prevent unfairness but it could also violate anti-discrimination laws and produce disparate treatment [27].
- **Disparate Impact**: Addresses unintentional discrimination, where proxy (not explicitly sensitive) attributes lead to disproportionate negative outcomes for a protected group [27].
- **Disparate Mistreatment**: Focuses on differences in misclassification rates between groups based on sensitive attributes, considering false positive and false negative rates when evaluating fairness [28].

The distinction between **disparate impact** and **disparate mistreatment** is important. In cases where ground truth data is unavailable and historical data is unreliable, disparate mistreatment may not be suitable due to difficulty in distinguishing between correct and incorrect classifications. On the other hand, when ground truth data is available, focusing on disparate impact may lead to reverse discrimination [28].

Applicants	Hired	Selection Rate	Percent Hired
80 White	48	48/80	60%
40 Black	12	12/40	30%

Table 2
Example from EEOC guidelines [26]. The four-fifths rule requires that the selection rate for any protected group should be at least 80% of the highest selection rate among groups. In this case, the highest selection rate is 60%, hence for the other group, i.e. *Black*, the selection rate should be at least 48%.

2.2. Fairness Metrics

Several fairness metrics have been proposed to assess the fairness of decision-making systems. This section highlights some of the most widely discussed metrics, which are generally categorized into two main types: **Individual Fairness** and **Group Fairness** [29, 30]. *Individual Fairness* refers to ensuring that predictions are fair for each individual, whereas *Group Fairness* focuses on equal treatment of groups with different values for sensitive attributes.
To define these metrics, the following notation is introduced:

- X : Input feature vector of applicants, excluding sensitive attributes.
- A : Sensitive attributes (e.g., race, gender).
- C : Binary classifier mapping X and A to a prediction C.
- Y : The actual outcome of the model.

Hence, the probability to observe an event E given that the attribute A has assumed value a is:

$$P_a(E) = P(E|A = a) \tag{1}$$

2.2.1. Individual Fairness

Fairness through unawareness: An algorithm can produce fair outcomes by excluding all sensitive attributes from the input feature vector, preventing the system from relying on these attributes to make a decision [17, 18]. Thus, the final outcome can be defined as

$$C = C(X, A) = C(X) \tag{2}$$

potential issue $\longrightarrow$ Attributes that are correlated with sensitive information (proxies) may still lead to biased outcomes.
Fairness through awareness: In this approach, an algorithm is considered fair if it produces similar outcomes for similar individuals. Specifically, if two applicants have similar feature vectors, the probability distributions of their predicted outcomes should also be similar, assuming a small similarity metric $d(i, j)$ [18, 30]:

$$C(X^i, A^i) \approx C(X^j, A^j) \tag{3}$$

Where:

- X^i and A^i are the feature vectors of applicant i.
- X^j and A^j are the feature vectors of applicant j.

Counterfactual Fairness: A model is counterfactually fair if the prediction for an individual remains the same in both the real world and in a counterfactual world where the individual belongs to a different demographic group [30]. The causal relationship between X and A must be such that, if A changes from a to a' then X changes from x to x'. The model is counterfactually fair if:

$$P(C(x, a) = c|X = x, A = a) = P(C(x, a') = c|X = x, A = a') \tag{4}$$

for all c and any value of a' attainable by A [31].
Figure 1 illustrates a causal graph in a hiring scenario. The sensitive attribute, Gender (G), is derived from Years of Experience (proxy), which directly influences the outcome (Hired/Not Hired). This setup would not be counterfactually fair, as the proxy influences the outcome [32]. To avoid proxy discrimination, there should be no proxy connections between the sensitive attribute and the outcome [33].

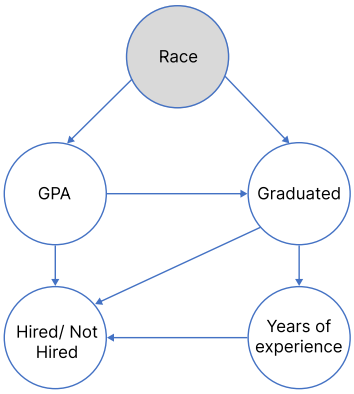


Figure 1: A causal graph representing employee recruitment with G as a sensitive attribute, Years of Experience as a proxy attribute and GPA as a resolving attribute.

2.2.2. Group Fairness

Demographic parity: Also called statistical parity, this metric ensures that the acceptance probability is the same (or within a given percentage) across groups [26, 32]. For some tolerance $\epsilon = \frac{p}{100} \in [0, 1]$,

$$|P_a(C = 1) - P_b(C = 1)| \leq \epsilon \tag{5}$$

Equalized Odds: This metric requires that both protected and unprotected groups have equal True Positive and False Positive rates [29, 34]:

$$P(C = 1 \mid A = 0, Y = y) = P(C = 1 \mid A = 1, Y = y), \quad y \in \{0, 1\} \tag{6}$$

Here, C and A are independent conditional on Y.

Equal Opportunity: This requires that protected and unprotected groups have equal True Positive Rates, focusing on fair positive outcomes [29, 34]:

$$P_a(C = 1 \mid Y = 1) = P_b(C = 1 \mid Y = 1) \tag{7}$$

Overall Accuracy Equality: This metric ensures that the prediction accuracy is the same across groups. In an HR context, it ensures that highly qualified and underqualified applicants are treated equally in both protected and unprotected groups [32, 34]:

$$P_a(C = Y) = P_b(C = Y) \tag{8}$$

Predictive Rate Parity (Sufficiency): This condition is met when the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are equal for both protected and unprotected groups. It helps prevent disparate mistreatment and promotes fairness [28]. Specifically, it ensures:

$$P_a(C = 1|Y = 1) = P_b(C = 1|Y = 1) \quad (1) \tag{9}$$

And for Negative Predictive Value:

$$P_a(C = 1|Y = 0) = P_b(C = 1|Y = 0) \quad (2) \tag{10}$$

A classifier satisfies *Predictive Rate Parity* if both conditions (1) and (2) are met [26].

Treatment Equality: A classifier satisfies this condition when the ratio of False Positives and False Negatives is equal across groups [32]:

$$\frac{P_a(C = 1 \mid Y = 0)}{P_a(C = 0 \mid Y = 1)} = \frac{P_b(C = 1 \mid Y = 0)}{P_b(C = 0 \mid Y = 1)} \tag{11}$$

Fairness Metrics	Advantages	Disadvantages
Fairness through unawareness	Straightforward solution by avoiding explicit use of sensitive attributes	Does not consider the correlation between sensitive and non-sensitive attributes
Fairness through awareness	Considers both the similarity of individuals and the similarity of outcome distributions. Flexible similarity definition for different scenarios.	Choice of distance metrics can impact results and may require fine-tuning. Sensitivity to the definition of similarity, which can vary across scenarios.
Counterfactual fairness	Considers the impact of changes in sensitive attributes on both non-sensitive features and predicted outcomes.	Requires prior knowledge of causal relationships between sensitive and non-sensitive attributes. Practical implementation may be challenging,when causal relationships are complex.
Demographic Parity	Promotes fair representation of all demographic groups.	Rules out an accurate classifier $(C = Y)$, considering it unfair when the base rates of the two groups are significantly different. $P_a(C = 1) \neq P_b(C = 1)$.
Overall Accuracy Equality	Ensures overall accuracy is consistent across different groups. Easy to implement.	Heavily dependent on the error type. It allows you to make up for rejecting qualified members of one group by accepting unqualified members of another group.
Equalized odds	Considers both positive and negative predictive performance. Addresses potential disparities in error rates between groups.	May be sensitive to class imbalances and prevalence differences.
Equal opportunity	Emphasizes equal opportunities for positive outcomes.	Does not consider false positive rates, potentially overlooking negative consequences. Similar to equalized odds, may face challenges in practical implementation.

Table 3
Advantages and Disadvantages of Different Fairness Definitions [18, 29]

2.3. Mitigation Mechanisms

Selecting the appropriate fairness metric for a model requires careful consideration of the legal, ethical, and social implications [35]. As discussed earlier, different fairness metrics offer distinct advantages and disadvantages (see Table 3). A recent research has demonstrated that it is impossible to satisfy multiple fairness notions simultaneously, creating a challenge in achieving balanced outcomes [24]. A key issue is the trade-off between fairness and accuracy. Incorporating fairness as an objective can reduce accuracy, as the focus shifts from purely optimizing prediction accuracy to balancing it with fairness concerns [34]. This creates a need for an established trade-off. Bias mitigation algorithms are designed to balance the dual objectives of maintaining model accuracy and ensuring fairness. These strategies can be applied at different stages of model development. Several methods for bias mitigation have been examined, including the work of Calders and Verwer (2010) [36], Chouldechova (2017) [37], Feldman et al. (2015) [38], Hardy et al. (2016) [39], Kamiran and Calders (2012) [40], Zafar et al. (2017) [41], and Zhang et al. (2018) [42]. These approaches can generally be classified into three categories: *pre-processing*, *in-processing*, and *post-processing techniques*.

Pre-Processing Mechanisms: Pre-processing techniques involve altering the input data to eliminate bias before training the classifier. This approach is useful when the algorithm is allowed to modify the training data [29]. Strategies include removing sensitive attributes, adjusting the labels of instances near decision boundaries (as these are more susceptible to discrimination), and applying reweighing techniques to correct imbalances. Recent approaches suggest altering feature representations in a way that reduces bias without altering the core model [35].

In-Processing Mechanisms: In-processing techniques aim to modify the learning algorithm itself to

reduce discrimination during model training, while keeping the original training data unchanged [29]. This can be achieved by introducing a regularization term to the objective function, which penalizes the mutual information between sensitive attributes and predicted outcomes. Alternatively, constraints can be added to ensure that the model satisfies fairness metrics like equalized odds or reduces disparate impact [28, 35].

Post-Processing Mechanisms: Post-processing methods adjust the predictions of a trained model to meet fairness criteria, without modifying the model or the training data [29]. These approaches are useful when the algorithm can only manipulate the learned model. For instance, some methods adjust the labels predicted by the black-box model using a fairness-driven function. Various studies propose techniques that improve equalized odds or equal opportunity by modifying the outcomes after training [34]. Additionally, it is often suggested to set different thresholds for different groups in a way that both maximizes accuracy and minimizes demographic disparities [35, 43].

3. Study Methodology

In this section, *we address the bias methodologies used to answer our study's research aim.* Our goal is to represent an HR decision system as a graph architecture to empirically evaluate the fairness notions of the predicted outcomes. Additionally, we explore the different standard mitigation techniques to generate results with optimum fairness and accuracy.

In the HR systems, decision-making is inherently comparative, requiring the evaluation of multiple candidates to identify the best fit. Given this complexity, **Graph Neural Network** (GNN) models are well-suited for such tasks. GNNs excel in scenarios where features exhibit intricate relationships and varying degrees of correlation, which significantly influence prediction outcomes. Indeed, our method involved the evaluation of the effectiveness of three GNN architectures: *Graph Convolutional Networks* (GCN), *Graph Attention Networks* (GAT) and *Graph Isomorphism Networks* (GIN).

Each GNN model underwent all three phases of bias mitigation: pre-processing techniques, in-processing, and post-processing. Bias was mitigated using appropriate algorithms at each stage. The model outcomes were evaluated against four key fairness metrics: *statistical parity difference* (SPD), *equal opportunity difference* (EOD), *overall accuracy equality difference* (OAED), and *treatment equality difference* (TED)[1]. By applying these methods to a relevant dataset, we aim to address key experimental questions - how different GNN architectures perform in reducing bias while maintaining predictive accuracy, and what trade-offs arise between fairness and accuracy when various bias mitigation strategies are applied. The findings are discussed in the next section and they disclose how various GNN designs manage and influence such trade-off.

3.1. Data Collection and Pre-Processing

For this study, the *Adult dataset* [44] from the UCI repository has been used. It predicts whether a person's annual income exceeds $50,000. It is a widely recognised dataset containing around **40,000 instances** and **16 attributes**, plus a target variable (income), collected on the 1994 United States Census. The sensitive attribute in this study was *gender*, and the target variable was *annual income.* The income data was converted into binary values as follows:

$$income > 50K \longrightarrow income = 1$$

$$income \leq 50K \longrightarrow income = 0$$

The data was sourced from the Center for Machine Learning and Intelligent Systems at the University of California, Irvine, and is available as a comma-separated values CSV file.

Since GNNs require data to be in graph form, the **K-Nearest Neighbors Graph** (K-NNG) method was employed to convert the dataset into a graph structure. K-NNG connects each entity with its k most

[1]Here the GIT repository with the source code: https://github.com/het28/Bias

similar neighbors based on a similarity metric. This technique was chosen due to the high density of connections in the dataset, allowing the K-NNG to produce sparse graphs with fewer edges, thereby improving computational efficiency compared to fully connected graphs [45].

3.2. Mitigation Mechanisms: A Comparative Analysis

As previously mentioned, bias mitigation can occur at various stages in model development. These stages include pre-processing (modifying data before training), in-processing (adjusting the training process), and post-processing (modifying outcomes post-training). Each mitigation strategy presents distinct advantages and disadvantages, which are summarized in Table 4. In our experimental protocol

Mechanism	Advantages	Disadvantages
Pre-Processing Mechanisms	Pre-processed data can be used for any downstream task.No need to modify classifier. No need to access sensitive attributes at test time.	Mostly used for optimizing before training.May not be able to support all fairness metrics (Statistical Parity or Individual Fairness) due to unavailability of label Y. Compared to the other two methods does not perform well on accuracy and fairness measures.
In-Processing Mechanisms	Good performance on accuracy and fairness measures. Higher flexibility to choose the trade-off between accuracy and fairness measures (depends on specific algorithm). No need to access sensitive attributes at test time.	Methods are task-specific. Do not generalize well across scenarios. Modification of the classifier might not always be feasible.
Post-Processing Mechanisms	Highly adaptable as can be applied after training the classifier. Results in relatively good performance on fairness measures. No need to modify classifier, simplifying implementation.	Need to access protected attributes during the testing phase. Lack the flexibility of picking any accuracy-fairness trade-off.

Table 4
Beside selecting the criterion to measure fairness, it is also need to choose the step in the workflow of a machine learning process in which to apply bias mitigation algorithms. In the table fairness mechanisms are classified conventionally into three categories pre-processing, in-processing and post-processing. Their respective advantages and disadvantages are also outlined [35, 46].

we employed various algorithms to mitigate bias, each corresponding to a different mitigation phase as detailed in Table 5.

Mitigation Mechanisms	***Algorithms***
Pre-processing	Reweighing
In-processing	Prejudice Remover Regularizer Rich Subgroup Fairness
Post-processing	Reject Option Classification

Table 5
Algorithms used for each mitigation phase in this study.

3.3. Mitigation Algorithms

Reweighing - Reweighing is a pre-processing technique designed to adjust the weights of instances in the dataset to mitigate bias. It does so without relabeling the data. For example, features where sensitive attribute a is in the positive class receive higher weights than those in the negative class, and vice versa for attribute b [47]. By adjusting the weights, this method seeks to achieve fairness between protected

and unprotected groups [40]. It is the most ideal algorithm for skewed and imbalanced datasets, which is one of the main causes of bias in HR domain. So, Reweighing was an obvious choice to readjust the representation to balance the different groups in the data, increasing the learning opportunity for the models.

Prejudice Remover Regularizer (PRR) - Prejudice Remover Regularizer is an in-processing technique that introduces a regularization term into the log-likelihood loss function of a classifier. This term penalises discrimination based on sensitive attributes [48]. For HR decision systems, along with fairness, accurate decisions are of primordial importance. Thus, this regularisation term is leveraged as a hyperparameter, which is used to control the degree of penalisation, allowing the model to balance accuracy with fairness [49].

Rich Subgroup Fairness (RSF) - Rich Subgroup Fairness aims to go beyond traditional fairness metrics, which may only evaluate fairness across broad categories such as gender. These broader metrics may overlook biases affecting specific subgroups, such as certain gender-ethnicity intersections. RSF mitigates this by considering finer-grained intersections of various attributes, thereby identifying and addressing biases against more specific protected subgroups. Using this algorithm, it is evaluated if the prediction accuracy across all groups is equal, enforcing the matching representation of false positives and false negatives across all groups [50].

Reject Option Classification (ROC) - Reject Option Classification is a post-processing method that works by adjusting predictions in the low-confidence regions of a probabilistic classifier. This approach reduces discrimination by selectively changing the classification of instances from both protected and unprotected groups. The ROC algorithm uses a variety of parameters, including classification thresholds and fairness metrics, to improve fairness [51]. The process involves swapping predictions (e.g., changing false negatives to true positives) to minimise unfair treatment of different demographic groups. It finds the best confidence bound by itself [47]. While it is effective at reducing bias, ROC is computationally expensive due to the complexity of tuning multiple parameters. Moreover, the algorithm can slightly reduce the accuracy of the unprotected group while increasing it for the protected group [47].

4. Experimental Protocol and Discussion of Results

We organize our experimental runs as following:

1. As first step we evaluate GCN, GAT, and GIN neural networks on *Adult dataset before* applying any mitigation technique.
2. Then we evaluate GCN, GAT, and GIN neural networks on *Adult dataset after* applying mitigation techniques (Reweighing, Prejudice Remover Regularizer, Rich Subgroup Fairness, and Reject Option Classifier).
3. Finally we compared the obtained results and we infer some consideration about their efficacy considering four fairness metrics: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Overall Accuracy Equality Difference (OAED), and Treatment Equality Difference (TED).

Baseline Performance (Pre-Mitigation) As it is possible to observe in Table 6, all three GNN models showed high accuracy, with GCN achieving slightly better results (approx 85%) than GAT and GIN (approx 84%). This aligns with existing literature that emphasizes GCN's superior ability to aggregate neighborhood information effectively, thus enhancing predictive accuracy [52, 53].

However, despite the strong overall accuracy, fairness metrics shown a different picture. GIN exhibited the *largest Equal Opportunity Difference (EOD)*, highlighting its significant **bias in terms of true positive rates** across protected and unprotected groups. GCN showed a relatively high *Treatment Equality Difference (TED)*, indicating a bias in **error rates between demographic groups**. These

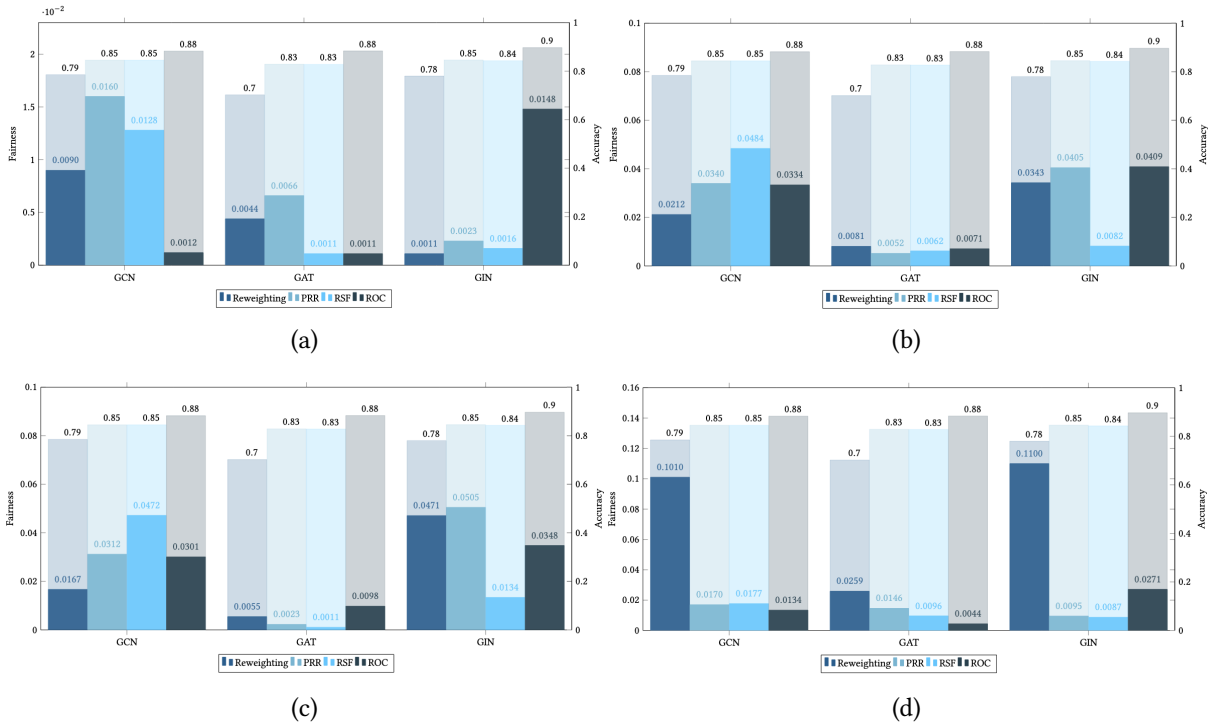
preliminary results reveal a critical insight: *while GNNs perform well in terms of accuracy, they exhibit inherent biases, thus necessitating bias mitigation techniques.*

Model	*Test Accuracy* (↑)	*SPD* (↓)	*EOD* (↓)	*OAED* (↓)	*TED* (↓)
GCN	**0.8449**	0.0098	0.0040	0.0184	0.0303
GAT	0.8293	**0.0020**	**0.0028**	**0.0067**	**0.0064**
GIN	0.8429	0.0079	0.0341	0.0349	0.0083

Table 6
Results before applying Mitigation Techniques. Upward facing arrow (↑) indicates that higher values are better, whereas downward facing arrows (↓) indicate lower values are better.

Model	*Mitigation Technique*	*Test Accuracy* (↑)	*SPD* (↓)	*EOD* (↓)	*OAED* (↓)	*TED* (↓)
GCN	Reweighing	0.7850	0.0090	0.0212	**0.0167**	0.1010
	Prejudice Remover Regularizer	0.8452	0.0160	0.0484	0.0312	0.0170
	Rich Subgroup Fairness	0.8451	0.0128	0.0334	0.0472	0.0177
	Reject Option Classification	**0.8826**	**0.0012**	**0.0081**	0.0301	**0.0134**
GAT	Reweighing	0.7020	0.0044	0.0081	0.0055	0.0259
	Prejudice Remover Regularizer	0.8285	0.0066	**0.0052**	0.0023	0.0146
	Rich Subgroup Fairness	0.8277	0.0050	0.0062	**0.0011**	0.0096
	Reject Option Classification	**0.8833**	**0.0011**	0.0071	0.0098	**0.0044**
GIN	Reweighing	0.7797	**0.0011**	0.0343	0.0471	0.1100
	Prejudice Remover Regularizer	0.8457	0.0023	0.0405	0.0505	0.0095
	Rich Subgroup Fairness	0.8428	0.0016	**0.0082**	**0.0134**	**0.0087**
	Reject Option Classification	**0.8968**	0.0148	0.0409	0.0348	0.0271

Table 7
Results after applying Mitigation Techniques. Upward facing arrow (↑) indicates that higher values are better, whereas downward facing arrows (↓) indicate lower values are better.



(a) (b) (c) (d)

Figure 2: Graph with Left y-axis: fairness evaluation for different models and different techniques. Right y-axis: accuracy of the models. Statistical Parity (2a), Equal Opportunity (2b), Overall Accuracy Equality (2c) and Treatment Equality (2d).

The effectiveness of the bias mitigation strategies, measured in terms of their ability to address fairness concerns without excessively compromising accuracy, is summarized in Table 7 and Figures 2a and 2d. The following paragraphs analyze each technique's impact.

Reweighing This pre-processing method, demonstrated notable improvements in fairness metrics but at the cost of a drop in predictive accuracy. Focusing on the GCN architecture, the model's accuracy fell from approx *85%* to *78.50%*. This reduction is expected, as reweighing tends to penalize the majority class to promote fairness, reducing overall accuracy. On the contrary, SPD seems to be improved fom *0.0098* to *0.0090*, reflecting a more balanced distribution of outcomes across groups, and OAED also decreased from *0.0184* to *0.0167*, suggesting reduced disparities in identifying positive cases. However, as shown in Figure 2c, Treatment Equality Difference (TED) increased to *0.1010*, suggesting a trade-off. This is a typical side-effect of reweighing, where fairness in positive outcomes can lead to greater disparities in misclassification rates, particularly in error rates across demographic groups. Similar consideration can be made for GAT, although in this case SPD registered a higher value after applying the mitigation technique. Finally, for GIN, only SPD showed a decrease from the raw model suggesting that Reweighing might not be the best technique to use with this type of GNN. Overall this result underscores the well-documented tension between fairness and accuracy: *while reweighing can improve outcome parity, it does so at the expense of treatment equality and overall accuracy.*

Prejudice Remover Regularizer (PRR) Prejudice Remover Regularizer, an in-processing technique, showed small variations in accuracy with two models (GCN and GAT) actually performing better, demonstrating its ability to maintain predictive performance. However, every model registered an improvement in only one technique each. For instance, GCN successfully managed to reduce TED (from *0.03* to *0.017*) but failed at satisfying the other metrics. The same holds for GAT and GIN which were able to improve the value of OAED and SPD respectively but didn't succeed at enhancing the rest of the metrics. Although PRR did reduce some of the bias present in the model's predictions, its performance, with respect to the four metrics, strongly depended on the type of GNN used. Moreover, it was not effective in addressing outcome disparities measured by EOD.

Rich Subgroup Fairness (RSF) Rich Subgroup Fairness leads to different impact on model performance and fairness metrics due to its specific approach for handling subgroup disparities. The test accuracy of 84.51 % is nearly identical to the original GCN model's accuracy, which is 84.49 %,same pattern is observed for GAT and GIN maintaining the accuracy at 82.77 % and 84.28 % respectively. It can be observed that SPD increases slightly to *0.0128* for GCN and *0.0050* for GAT, indicating a minor increase in bias regarding the distribution of favorable outcomes across genders. We can observe that RSF aims to be fair across subgroups but not fully able to eliminate all the biases for the GCN and GAT model. On the other hand, RSF worked the better when applied to GIN model when compared with GCN and GAT, since we observed decrease in value for SPD, EOD and OAED, while TED remains almost the same suggesting that RSF has maintained overall predictive performance while addressing fairness.

Reject Option Classification (ROC) The Reject Option Classification emerged as one of the most effective post-processing technique in improving accuracy to almost 89% for all three models. SPD dropped to *0.0012* and *0.0011* for GCN and GAT respectively, indicating near-perfect balance in the distribution of favorable outcomes between groups. TED also improved, decreasing to *0.0134* for GCN and *0.0044* for GAT, hence reflecting more balanced error rates across groups.

Also, OAED was successfully improved for GIN while ROC failed at mitigating bias arising from differences in the true positive rates for protected and unprotected groups.

5. Conclusion

The results presented in this section confirm the hypothesis posed in our research question: bias mitigation techniques do help reduce bias in GNN architectures, but the trade-off between fairness and accuracy is inevitable. Each technique exhibited distinct strengths and weaknesses, depending on the GNN model it was applied to.

The **GAT architecture**, combined with the **ROC algorithm**, produces the best results, offering an optimized balance between accuracy and fairness. This outcome is expected, as it involves prediction swapping and numerous parameters, which also makes it computationally intensive.
While not the top performer, **PRR** and **RSF** consistently maintain accuracy across all GNN models, achieving an effective trade-off, particularly when used with GAT.
Among the fairness metrics, **Treatment Equality** showed the most improvement following the application of mitigation techniques, promoting equal error rates across all groups.
PRR was the least effective in enhancing fairness metrics, indicating that a standard approach like this, which adjusts representation, struggles to improve fairness in complex real-world data.

These findings suggest that no single mitigation technique universally outperforms the others in all fairness metrics, and that careful consideration must be given when selecting the appropriate technique based on the specific fairness requirements and constraints of the task at hand.

Future Directions The growing body of research in fairness and bias mitigation within machine learning underscores the importance of continued investigation, especially as AI systems increasingly influence social and organizational decision-making. Future work should focus on:

- **Exploring more complex and realistic datasets** that encompass multiple sensitive attributes, offering a richer and more representative testing environment.
- **Expanding the use of alternative GNN models**, as variations in model architectures may yield better performance in fairness optimization.
- **Improving model transparency and interpretability**, which will be crucial for building trust in AI-driven HR systems and ensuring these systems are accountable for their decisions.

Such advancements will enable more refined bias mitigation techniques and foster collaboration between researchers and practitioners to create fairer, more equitable machine learning systems for real-world applications.

Acknowledgments

This research is partially funded by PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] M. Polignano, C. Musto, R. Pellungrini, E. Purificato, G. Semeraro, M. Setzu, XAI.it 2024: An Overview on the Future of Explainable AI in the era of Large Language Models, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[2] T. Bogers, D. Graus, M. Kaya, C. Johnson, J.-J. Decorte, Third workshop on recommender systems for human resources (recsys in hr 2023), in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1244–1247. URL: https://doi.org/10.1145/3604915.3608755. doi:10.1145/3604915.3608755.

[3] S. Garg, S. Sinha, A. K. Kar, M. Mani, A review of machine learning applications in human resource management, International Journal of Productivity and Performance Management 71 (2022) 1590–1610.

[4] V. Kakulapati, K. K. Chaitanya, K. V. G. Chaitanya, P. Akshay, Predictive analytics of hr-a machine learning approach, Journal of Statistics and Management Systems 23 (2020) 959–969.

[5] E. Purificato, F. Lorenzo, F. Fallucchi, E. W. D. Luca, The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes, International Journal of Human–Computer Interaction (2023) 1543–1562. URL: https://doi.org/10.1080/10447318.2022.2081284. doi:10.1080/10447318.2022.2081284, publisher: Taylor & Francis.

[6] B. P. Evans, B. Xue, M. Zhang, What's inside the black-box? a genetic programming method for interpreting complex machine learning models, in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1012–1020. URL: https://doi.org/10.1145/3321707.3321726. doi:10.1145/3321707.3321726.

[7] J. Krause, A. Perer, K. Ng, Interacting with predictions: Visual inspection of black-box machine learning models, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 5686–5697. URL: https://doi.org/10.1145/2858036.2858529. doi:10.1145/2858036.2858529.

[8] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 506–519. URL: https://doi.org/10.1145/3052973.3053009. doi:10.1145/3052973.3053009.

[9] A. Köchling, M. C. Wehner, Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development, Business Research 13 (2020) 795–848. URL: https://doi.org/10.1007/s40685-020-00134-w. doi:10.1007/s40685-020-00134-w.

[10] H.-Y. Suen, M. Y.-C. Chen, S.-H. Lu, Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes?, Computers in Human Behavior 98 (2019) 93–101. URL: https://www.sciencedirect.com/science/article/pii/S0747563219301529. doi:https://doi.org/10.1016/j.chb.2019.04.012.

[11] G. Spillo, C. Musto, M. Polignano, P. Lops, M. de Gemmis, G. Semeraro, Combining Graph Neural Networks and Sentence Encoders for Knowledge-aware Recommendations, in: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2023, Limassol, Cyprus, June 26-29, 2023, ACM, 2023, pp. 1–12. URL: https://doi.org/10.1145/3565472.3592965. doi:10.1145/3565472.3592965.

[12] A. Köchling, M. C. Wehner, Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development, Business Research 13 (2020) 795–848. URL: https://doi.org/10.1007/s40685-020-00134-w. doi:10.1007/s40685-020-00134-w.

[13] E. Purificato, L. Boratto, E. W. De Luca, Do Graph Neural Networks Build Fair User Models? Assessing Disparate Impact and Mistreatment in Behavioural User Profiling, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 4399–4403. URL: https://doi.org/10.1145/3511808.3557584. doi:10.1145/3511808.3557584.

[14] E. Purificato, L. Boratto, E. W. De Luca, Toward a Responsible Fairness Analysis: From Binary to Multiclass and Multigroup Assessment in Graph Neural Network-Based User Modeling Tasks, Minds and Machines 34 (2024) 33. URL: https://doi.org/10.1007/s11023-024-09685-x. doi:10.1007/s11023-024-09685-x.

[15] E. Purificato, E. W. De Luca, What Are We Missing in Algorithmic Fairness? Discussing Open Challenges for Fairness Analysis in User Profiling with Graph Neural Networks, in: L. Boratto, S. Faralli, M. Marras, G. Stilo (Eds.), Advances in Bias and Fairness in Information Retrieval,

Communications in Computer and Information Science, Springer Nature Switzerland, Cham, 2023, pp. 169–175. doi:10.1007/978-3-031-37249-0_14.

[16] X. Chang, Gender bias in hiring: An analysis of the impact of amazon's recruiting algorithm, Advances in Economics, Management and Political Sciences 23 (2023) 134–140. doi:10.54254/2754-1169/23/20230367.

[17] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, A. Weller, The case for process fairness in learning: Feature selection for fair decision making, in: NIPS symposium on machine learning and the law, volume 1, Barcelona, Spain, 2016, p. 11.

[18] Z. Chen, J. M. Zhang, M. Hort, M. Harman, F. Sarro, Fairness testing: A comprehensive survey and analysis of trends, ACM Trans. Softw. Eng. Methodol. 33 (2024). URL: https://doi.org/10.1145/3652155. doi:10.1145/3652155.

[19] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (1974) 1124–1131. URL: https://www.science.org/doi/abs/10.1126/science.185.4157.1124. doi:10.1126/science.185.4157.1124. arXiv:https://www.science.org/doi/pdf/10.1126/science.185.4157.1124.

[20] M. Soleimani, A. Intezari, D. Pauleen, Mitigating cognitive biases in developing ai-assisted recruitment systems: A knowledge-sharing approach, International Journal of Knowledge Management 18 (2022). doi:10.4018/IJKM.290022.

[21] D. D. Savage, R. Bales, Video games in job interviews: Using algorithms to minimize discrimination and unconscious bias, ABAJ Lab. & Emp. L. 32 (2016) 211.

[22] M. Polignano, M. de Gemmis, G. Semeraro, Contextualized BERT Sentence Embeddings for Author Profiling: The Cost of Performances, in: O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blecic, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, C. M. Torre, Y. Karaca (Eds.), Computational Science and Its Applications - ICCSA 2020 - 20th International Conference, Cagliari, Italy, July 1-4, 2020, Proceedings, Part IV, volume 12252 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 135–149. URL: https://doi.org/10.1007/978-3-030-58811-3_10. doi:10.1007/978-3-030-58811-3_10.

[23] R. Vivek, Is blind recruitment an effective recruitment method?, International Journal of Applied Research in Business and Management 3 (2022) 56–72. doi:10.51137/ijarbm.2022.3.3.4.

[24] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, K. P. Gummadi, Fairness constraints: A flexible approach for fair classification, Journal of Machine Learning Research 20 (2019) 1–42. URL: http://jmlr.org/papers/v20/18-262.html.

[25] EUAIACT, Annex 3 - euaiact, https://www.euaiact.com/annex/3, 2024. Accessed: 2024-09-27.

[26] D. F. Mujtaba, N. R. Mahapatra, Ethical considerations in ai-based recruitment, in: 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019, pp. 1–7. doi:10.1109/ISTAS48451.2019.8937920.

[27] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev. 104 (2016) 671.

[28] M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, 2017, p. 1171 – 1180. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048347682&doi=10.1145%2f3038912.3052660&partnerID=40&md5=e0ce633abceab72e3214b4c9965d03a8. doi:10.1145/3038912.3052660, cited by: 671; All Open Access, Green Open Access.

[29] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[30] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 214–226. URL: https://doi.org/10.1145/2090236.2090255. doi:10.1145/2090236.2090255.

[31] M. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4069–4079.

[32] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, FairWare '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–7. URL: https://doi.org/10.1145/3194770.3194776. doi:10.1145/3194770.3194776.

[33] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 656–666.

[34] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 3323–3331.

[35] D. Pessach, E. Shmueli, Algorithmic fairness, ArXiv abs/2001.09784 (2020). URL: https://api.semanticscholar.org/CorpusID:210921184.

[36] M. Favier, T. Calders, S. Pinxteren, J. Meyer, How to be fair? a study of label and selection bias, Machine Learning 112 (2010) 5081–5104. doi:https://doi.org/10.1007/s10994-023-06401-1.

[37] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data 5:2 (2017) 153–163. doi:DOI:10.1089/big.2016.0047.

[38] P. S. Hart, L. Feldman, A. Leiserowitz, E. Maibach, Extending the impacts of hostile media perceptions: Influences on discussion and opinion polarization in the context of climate change, Science Communication 37 (2015) 506–532. URL: https://doi.org/10.1177/1075547015592067. doi:10.1177/1075547015592067.

[39] B. Clyne, C. Fitzgerald, A. Quinlan, C. Hardy, R. Galvin, T. Fahey, S. M. Smith, Interventions to address potentially inappropriate prescribing in community-dwelling older adults: A systematic review of randomized controlled trials, Journal of the American Geriatrics Society 64 (2016) 1210–1222. URL: https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/jgs.14133. doi:https://doi.org/10.1111/jgs.14133.

[40] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2012) 1–33. URL: https://doi.org/10.1007/s10115-011-0463-8. doi:10.1007/s10115-011-0463-8.

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

[42] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 335–340. URL: https://doi.org/10.1145/3278721.3278779. doi:10.1145/3278721.3278779.

[43] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 797–806. URL: https://doi.org/10.1145/3097983.3098095. doi:10.1145/3097983.3098095.

[44] B. Becker, R. Kohavi, Adult, UCI Machine Learning Repository, 1996. doi:https://doi.org/10.24432/C5XW20.

[45] A. Boutet, A.-M. Kermarrec, N. Mittal, F. Taiani, Being prepared in a sparse world: The case of knn graph construction, in: 2016 IEEE 32nd International Conference on Data Engineering (ICDE), 2016, pp. 241–252. doi:10.1109/ICDE.2016.7498244.

[46] K. Zanna, A. Sano, Enhancing fairness and performance in machine learning models: A multi-task learning approach with monte-carlo dropout and pareto optimality, ArXiv abs/2404.08230 (2024). URL: https://api.semanticscholar.org/CorpusID:269137478.

[47] P. Janssen, B. M. Sadowski, Bias in algorithms: On the trade-off between accuracy and fairness, 23rd Biennial Conference of the International Telecommunications Society (ITS): "Digital societies

and industrial transformations: Policies, markets, and technologies in a post-Covid world", Online Conference / Gothenburg, Sweden, 21st-23rd June, 2021, International Telecommunications Society (ITS), Calgary, 2021. URL: https://hdl.handle.net/10419/238032.

[48] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: P. A. Flach, T. De Bie, N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 35–50.

[49] B. d'Alessandro, C. O'Neil, T. LaGatta, Conscientious classification: A data scientist's guide to discrimination-aware classification, Big Data 5 (2017) 120–134. URL: https://doi.org/10.1089/big.2016.0048. doi:10.1089/big.2016.0048. arXiv:https://doi.org/10.1089/big.2016.0048, pMID: 28632437.

[50] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html.

[51] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 924–929. doi:10.1109/ICDM.2012.45.

[52] G.-F. Ma, X.-H. Yang, L. Ye, Y.-J. Huang, P. Jiang, Graph convolutional network based on higher-order neighborhood aggregation, in: T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, A. N. Hidayanto (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2021, pp. 334–342.

[53] Y. Liu, K. Ning, Improved graph representation learning based on neighborhood aggregation and interaction fusion, Journal of Intelligent & Fuzzy Systems (2024). URL: https://doi.org/10.3233/JIFS-234086. doi:10.3233/JIFS-234086.