A Multi-Stage Framework for Bias Detection and Mitigation in AI-Driven Recruitment Systems

Gideon Assafuah, Claude Turner, Carlene Turner, Kingsley Nwosu
Department of Computer Science, Norfolk State University, Virginia, USA
Department of Sociology, Norfolk State University, Virginia, USA

***Abstract*—The use of machine learning in recruitment has raised growing concerns about fairness, as automated hiring systems can generate unequal outcomes across demographic groups. These disparities are influenced not only by imbalanced data but also by the behavior of learning algorithms, making bias a multidimensional challenge that cannot be effectively addressed with single-stage solutions. This study introduces an integrated framework for bias detection and mitigation in AI-driven recruitment systems, combining interventions at the data, model, and decision levels within a unified evaluation pipeline. The framework is assessed using multiple classification models of varying complexity and evaluated with established fairness metrics. In addition, explainability techniques are employed using SHAP-based feature attribution to investigate hidden dependencies and assess the sensitivity of predictions to demographic attributes. Experimental results show that baseline models achieve strong predictive performance, with accuracy ranging from 0.807 to 0.816; however, fairness evaluation reveals substantial disparities, with Disparate Impact as low as 0.190 and Demographic Parity Difference exceeding 0.27 in some cases. After applying the proposed multi-stage mitigation approach, fairness metrics improve significantly: Disparate Impact meets or exceeds the legal threshold of 0.80 across all models, with reductions in Demographic Parity Difference of 70–85% and Equal Opportunity Difference of 47–75%, and demographic disparities are reduced to 0.029–0.053. These improvements are achieved with minimal performance trade-off, as overall accuracy decreases by at most 4.2 % points while ROC AUC remains unchanged. The findings demonstrate that bias in recruitment systems arises from the interplay between data and model dynamics and highlight the importance of coordinated mitigation strategies throughout the machine learning lifecycle. This work provides a practical, scalable approach to developing fair and transparent AI systems for hiring applications.**

Keywords*—Bias detection; bias mitigation; AI in recruitment; algorithmic fairness; explainable AI; fairness metrics; machine learning; responsible AI

I. Introduction

The rapid adoption of artificial intelligence (AI) in recruitment and hiring processes has transformed how organizations evaluate and select candidates [30]. Automated systems are increasingly used to screen resumes, rank applicants, and support decision-making in large-scale recruitment settings, offering improved efficiency and scalability [1], [2], [28]. However, as these systems become more influential in determining access to employment opportunities, concerns have emerged regarding their fairness and potential to produce unequal outcomes across demographic groups [14], [15].

A growing body of research has shown that machine learning models can inherit and amplify historical biases embedded in training data, leading to systematic disparities in hiring outcomes. These biases may arise from imbalanced datasets, proxy variables, or algorithmic decision boundaries, and can disproportionately affect groups defined by sensitive attributes such as gender, race, or age [12], [16]. As a result, AI-driven recruitment systems risk reinforcing existing social inequalities while operating under the perception of objectivity, raising concerns about transparency, accountability, and ethical responsibility [7], [8].

Existing approaches to addressing bias in machine learning systems typically focus on isolated stages of the pipeline. Data-level methods aim to rebalance or transform datasets to reduce representation bias, while model-level techniques incorporate fairness constraints during training. Post-processing approaches, on the other hand, adjust model outputs to improve fairness after predictions have been generated. Although these methods have demonstrated effectiveness in specific contexts, their isolated application often fails to capture the complex ways bias propagates throughout the entire machine learning lifecycle [9]. Consequently, there is a need for integrated approaches that address bias holistically rather than at a single stage.

In addition to mitigation strategies, explainability has emerged as a critical component in understanding and evaluating fairness in AI systems. Techniques such as SHAP-based feature attribution and local explanation methods provide insights into how models make decisions and whether sensitive attributes or their proxies influence outcomes. These methods are particularly important in high-stakes domains such as recruitment, where opaque decision-making can undermine trust and accountability [10], [11].

This study addresses these challenges by proposing a multi-stage framework for bias detection and mitigation in AI-driven recruitment systems. Unlike prior approaches that focus on a single level of intervention, the proposed framework integrates data-level auditing, model-level fairness-constrained training, and a novel post-processing threshold optimizer within a unified pipeline. Critically, the three stages are designed to operate in sequence — pre-processing corrects data-level imbalance, in-processing embeds fairness constraints into model training, and post-processing applies a group-specific

threshold search governed by a composite objective function that explicitly encodes the EEOC 80% legal threshold as a hard constraint. This end-to-end coordination across all three intervention levels, with each stage informing the next, is what distinguishes the proposed approach from existing single-stage or loosely coupled multi-technique fairness methods. The framework is evaluated using multiple machine learning models, including Logistic Regression, Random Forest, and Neural Networks, to analyze how bias manifests across varying levels of model complexity. Fairness is assessed using established metrics such as Disparate Impact, Demographic Parity Difference, and Equal Opportunity Difference. In contrast, explainability techniques are used to uncover hidden dependencies and assess decision sensitivity through SHAP-based feature attribution.

The contributions of this work are fourfold, and collectively distinguish it from prior art in three important ways. First, while pre-processing, in-processing, and post-processing fairness techniques each exist in isolation in the literature, no prior work has combined all three within a single unified pipeline specifically designed for recruitment AI, where each stage is designed to build on and complement the prior one rather than acting independently. Second, the post-processing optimizer in this framework is the first to incorporate the EEOC 80% rule as a hard constraint within a joint multi-group threshold search — existing threshold-based methods, such as Fairlearn's ThresholdOptimizer, optimize for demographic parity alone and do not jointly minimise EOD and DPD while enforcing a legal DI floor. Third, SHAP-based explainability is embedded as an integral diagnostic component of the pipeline rather than applied as a post-hoc interpretability tool, enabling the identification of proxy bias pathways across three architecturally distinct model families. Fourth, the framework is evaluated with five-seed statistical averaging and paired significance testing — a level of experimental rigour that is absent from most prior fairness-aware recruitment studies, which typically report single-run results. Specifically, the contributions are: 1) a comprehensive bias auditing approach that captures disparities at both data and model levels across gender, race, and age; 2) an integrated three-stage mitigation strategy with a novel post-processing objective function that jointly enforces legal DI compliance, demographic parity, and equalized odds; 3) empirical demonstration that all three models achieve DI ≥ 0.80 — the EEOC legal threshold — with at most 4.2 percentage point accuracy loss and unchanged ROC AUC; and 4) SHAP-based evidence that bias manifests structurally differently across model architectures, with important implications for fairness auditing practice.

The remainder of this study is organized as follows. Section II reviews related work on bias and fairness in AI-based recruitment systems. Section III presents the proposed methodology and experimental setup. Section IV discusses the results and evaluates the effectiveness of the proposed framework. Finally, Section V concludes the study and outlines directions for future research.

II. Related Work

Artificial intelligence has become an integral component of modern recruitment processes, enabling organizations to automate candidate screening, ranking, and selection. These systems leverage machine learning techniques to process large volumes of applicant data, improving efficiency and scalability in hiring workflows [1], [2]. Despite these advantages, the adoption of AI in recruitment introduces significant challenges, particularly regarding fairness and the potential for unintended discrimination. Studies have shown that AI-based hiring tools can influence critical employment decisions, making it essential to ensure that such systems operate equitably and transparently [3]. Bias in machine learning systems has been widely studied, with research demonstrating that models can reflect and amplify historical inequalities present in training data [4], [5], [27]. These biases may arise from imbalanced datasets, measurement errors, or correlations between sensitive attributes and non-sensitive features. Caliskan et al. [6] demonstrated that language-based models can encode societal biases, highlighting the broader issue of bias propagation in AI systems. To address these concerns, various fairness definitions and evaluation metrics have been proposed, including Disparate Impact, Demographic Parity, Equal Opportunity, and Equalized Odds [18], [17]. These metrics provide quantitative measures for assessing disparities in model outcomes across demographic groups. However, prior work indicates that improving fairness on one metric may negatively affect others, illustrating the inherent trade-offs in fairness-aware machine learning [4].

Furthermore, bias is not a static property but can evolve throughout the machine learning pipeline. Decisions made during data collection, preprocessing, model training, and evaluation can each amplify or mitigate bias. This dynamic nature of bias underscores the need for comprehensive evaluation strategies that account for multiple pipeline stages rather than focusing on isolated components.

The ethical implications of AI-driven recruitment systems extend beyond technical considerations, encompassing issues of accountability, transparency, and societal impact [31]. Researchers have emphasized that biased hiring systems can perpetuate systemic discrimination, limiting opportunities for underrepresented groups and reinforcing existing inequalities [7], [8]. As a result, there has been a growing emphasis on Responsible AI, which advocates for the development of systems that are fair, transparent, and aligned with ethical principles [26]. In recruitment contexts, these concerns are particularly significant due to the high-stakes nature of hiring decisions [24], [25]. Unlike other application domains, biased outcomes in hiring can have long-term consequences for individuals' careers and livelihoods. The lack of interpretability in many machine learning models further complicates this issue, as stakeholders may be unable to understand or challenge automated decisions [3]. Consequently, ensuring fairness in recruitment AI systems requires not only technical solutions but also organizational and regulatory frameworks that promote accountability and transparency.

To address bias in machine learning systems, researchers have proposed a range of mitigation strategies that operate at different stages of the pipeline. These strategies are commonly categorized into pre-processing, in-processing, and post-processing approaches. Pre-processing methods aim to modify the training data to reduce bias, for example, through

reweighting, resampling, or feature transformation. In-processing techniques incorporate fairness constraints directly into the learning algorithm, enabling models to balance accuracy and fairness during training. Post-processing methods adjust model outputs after prediction to achieve fairness objectives [9]. While each technique category has demonstrated effectiveness, its application in isolation presents limitations. Pre-processing methods may reduce bias in the data, but cannot guarantee fair outcomes if the model learns unintended patterns. In-processing approaches may require modifications to model architecture or training procedures, limiting their applicability in real-world systems. Post-processing techniques, although flexible, do not address the root causes of bias and may lead to suboptimal decision boundaries. Recent studies suggest that combining multiple mitigation strategies can yield more robust fairness improvements. However, there remains a lack of unified frameworks that systematically integrate these approaches within a single pipeline. This gap is particularly evident in recruitment systems, where bias may arise from both data characteristics and model behavior.

Explainable AI (XAI) has emerged as a key tool for understanding and mitigating bias in machine learning systems. By providing insights into how models make decisions, explainability techniques enable researchers and practitioners to identify the factors driving predictions and assess whether these factors are fair and appropriate [10]. Methods such as feature importance analysis and local explanations allow for the examination of individual predictions and the detection of hidden biases. In addition to model-level explanations, dataset transparency has been recognized as an important component of fairness. Frameworks such as dataset datasheets promote the documentation of dataset characteristics, including sources, preprocessing steps, and potential biases [11]. These practices support more informed decision-making and facilitate the identification of bias at earlier stages of the pipeline. Despite these advances, explainability is often treated as a separate component rather than being integrated into bias detection and mitigation workflows. There is a growing need for approaches that combine explainability with fairness evaluation to provide a more comprehensive understanding of bias and its impact on model behavior [22], [23].

A. Research Limitations and Identified Gaps

Although substantial progress has been made in understanding and mitigating bias in machine learning systems, several limitations remain. First, many existing studies focus on individual pipeline stages, such as data preprocessing or model training, without considering how bias propagates across stages. Second, limited work integrates fairness evaluation with explainability techniques to provide deeper insights into bias mechanisms. Third, there is a lack of practical frameworks for real-world recruitment systems that balance fairness and predictive performance. These limitations highlight the need for a comprehensive approach that addresses bias holistically. There is a need for frameworks that integrate data-level auditing, model-level evaluation, and post-processing mitigation within a unified system, while also incorporating explainability techniques to enhance transparency and interpretability.

To address this gap, this study proposes a multi-stage framework for bias detection and mitigation in AI-driven recruitment systems. The proposed approach combines fairness metrics, explainability methods, and empirical evaluation across multiple models to provide a robust, scalable solution for developing fair, transparent hiring systems [19], [20].

Table I summarizes the key existing research on bias and fairness in AI-based recruitment systems, categorizes prior work by methodological approach, and highlights key limitations and unresolved research gaps. These observations reveal that existing approaches are largely fragmented and lack a unified, end-to-end framework to address bias throughout the machine learning lifecycle.

TABLE I. SUMMARY OF EXISTING RESEARCH

Existing Research	Methodology	Limitations and Research Gap
Research papers [1], [2]	Literature reviews and qualitative analysis of AI-driven recruitment systems and bias sources	Provide conceptual insights but lack empirical validation, and do not propose integrated technical frameworks for bias detection and mitigation
Research papers [4], [5], [6]	Surveys and analytical studies on bias sources, fairness metrics, and data-driven bias	Focus on general ML contexts without domain-specific solutions for recruitment or end-to-end mitigation pipelines
Research papers [3], [7], [8]	Conceptual and ethical analyses of fairness, accountability, and transparency in AI systems	Emphasize ethical concerns but lack technical implementation and actionable mitigation strategies
Research paper [9]	Pre-processing method using data reweighing to improve fairness	Limited to a single-stage solution and does not address bias across data, model, and decision levels
Research paper [10], [11]	XAI techniques and dataset documentation frameworks for understanding model behavior	Focus on interpretability and transparency without integrating fairness optimization or mitigation strategies

III. METHODOLOGY

To address the limitations of existing single-stage bias mitigation approaches, this study proposes a multi-stage framework for bias detection and mitigation in AI-driven recruitment systems. The framework is designed to capture the full machine learning lifecycle, enabling bias to be systematically identified, analyzed, and mitigated at multiple stages, including data preparation, model development, and evaluation. By integrating data-level auditing, model-level analysis, explainability techniques, and fairness-aware optimization, the proposed approach provides a comprehensive, structured mechanism to improve fairness while maintaining predictive performance. The overall architecture of the proposed framework is illustrated in Fig. 1.

This study proposes a multi-stage framework designed to detect, analyze, and mitigate bias in AI-driven recruitment systems. The framework adopts a pipeline-based structure that captures the full lifecycle of machine learning, enabling bias to be examined and addressed at multiple stages rather than in isolation.

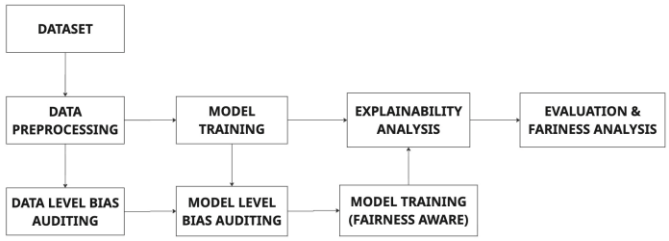

Fig. 1. Proposed multi-stage bias detection and mitigation framework.

As illustrated in Fig. 1, the framework comprises interconnected components: dataset preparation, data preprocessing, data-level bias auditing, model training, model-level bias evaluation, explainability analysis, fairness-aware model refinement, and final evaluation. Each stage plays a distinct role in identifying and mitigating bias while preserving predictive performance.

A. Dataset Description and Preprocessing

This study uses the UCI Adult Income dataset, a widely established benchmark in algorithmic fairness research, extracted from the 1994 United States Census Bureau database [19]. After removing missing values and duplicates, the dataset comprises 45,175 records with eight features. Sensitive attributes — gender, race, and age group — are excluded from model inputs but retained for fairness evaluation. Skill and qualification proxies include education level, Certifications, weekly working hours, Experience Years, and Screening Score. The dataset exhibits a 75.2% / 24.8% class split with significant demographic imbalance: male applicants constitute 67.5% and White applicants 86.0% of the pool, making it well-suited for studying bias propagation in selection-oriented systems.

The dataset preparation process involves several preprocessing steps to ensure data quality and compatibility with machine learning models. Missing values are handled using appropriate imputation techniques to preserve dataset completeness, while categorical variables are encoded as numerical values using methods such as one-hot encoding. Continuous features are standardized to improve model convergence and ensure balanced feature contributions. Sensitive attributes are excluded from model training inputs to prevent direct discrimination but are retained separately for fairness evaluation. The dataset is then split into training and test sets using stratified sampling to maintain class distribution and ensure reliable evaluation. In addition, exploratory data analysis is conducted to examine the distribution of demographic groups and hiring outcomes, enabling the identification of representation imbalance and outcome disparities before model development.

B. Data-Level Bias Analysis

Before training machine learning models, it is essential to examine the dataset for inherent biases that may influence model predictions. In AI-driven recruitment systems, datasets often reflect historical hiring practices, which can introduce disparities across demographic groups. Therefore, a data-level bias analysis is conducted to identify potential imbalances in representation, outcome distribution, and feature relationships. This preliminary analysis is critical, as biases present in the data can be learned and amplified by machine learning models if not properly addressed [5].

The first step in the analysis is to examine the representation of sensitive attributes, including gender, race, and age group. Representation bias occurs when certain demographic groups are underrepresented or overrepresented in a dataset, leading to unequal learning outcomes. The distribution of these groups is analyzed to determine whether the dataset provides balanced coverage across populations. Unequal representation may result in models that perform poorly for minority groups due to insufficient training data, thereby introducing fairness concerns [4]. Fig. 2 shows the gender distribution, Fig. 3 shows the race distribution, and Fig. 4 shows the age group distribution.

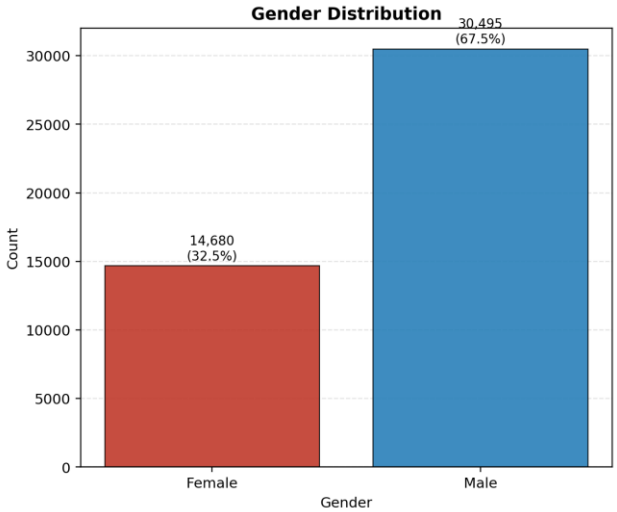

Fig. 2. Gender distribution of applicants.

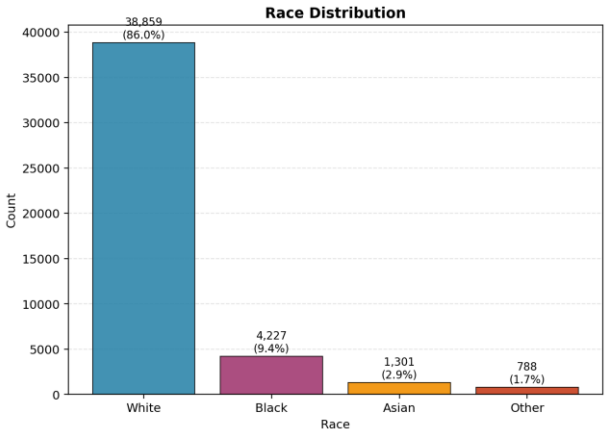

Fig. 3. Race distribution of applicants.

In addition to representation, outcome bias is assessed by analyzing the distribution of hiring decisions across demographic groups. This is measured using the Selection Rate (SR), defined as the proportion of candidates in a group who receive a positive outcome (i.e., are hired). Differences in selection rates across groups indicate potential disparities in historical decision-making patterns embedded in the dataset. A significant variation in SR suggests that certain groups may be

favored or disadvantaged, which can directly influence model predictions if left unaddressed [7]. Fig. 5 shows selection rates by gender, Fig. 6 shows selection rates by race, and Fig. 7 shows selection rates by age group.

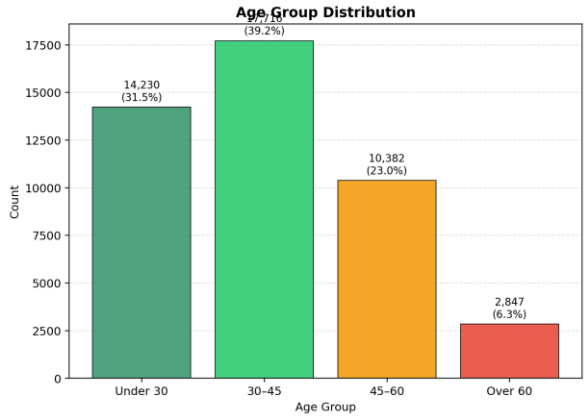

Fig. 4. Age group distribution of applicants.

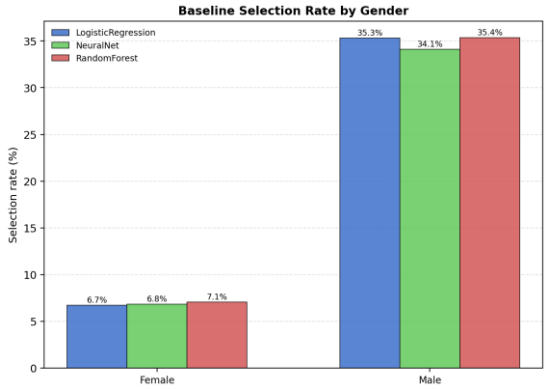

Fig. 5. Selection rate by gender.

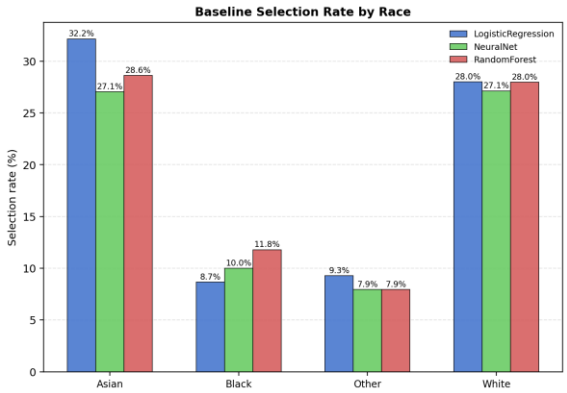

Fig. 6. Selection rate by race.

To further investigate hidden sources of bias, correlation analysis is performed between sensitive attributes and non-sensitive features. Features that exhibit strong correlations with sensitive attributes may act as proxy variables, enabling the model to indirectly learn discriminatory patterns even when sensitive attributes are excluded from training. This phenomenon has been widely documented in fairness research and highlights the limitations of simply removing sensitive features without deeper analysis [8]. Exploratory data analysis is also conducted to visualize feature distributions and identify potential anomalies or skewness. Visualization techniques such as bar charts and distribution plots are used to compare demographic groups and hiring outcomes. These insights provide an initial understanding of how bias may be embedded in the dataset and inform subsequent bias mitigation strategies.

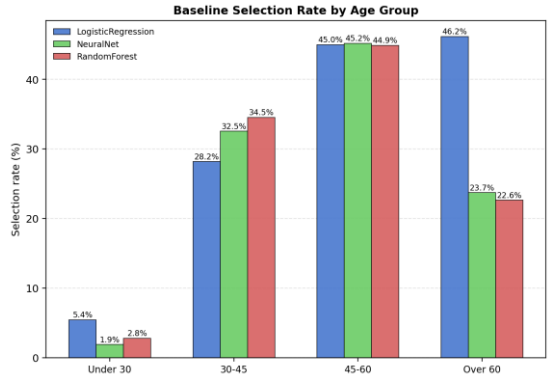

Fig. 7. Selection rate by age group.

Overall, the data-level bias analysis provides a foundational understanding of disparities present in the dataset before model development. By identifying representation imbalance, outcome disparities, and proxy relationships, this study ensures that bias is addressed proactively rather than reactively. The findings from this stage serve as a baseline for evaluating the effectiveness of bias mitigation techniques applied in later stages of the framework. As shown in Fig. 8, the correlation matrix reveals strong relationships between non-sensitive features and sensitive attributes, indicating the presence of proxy variables.

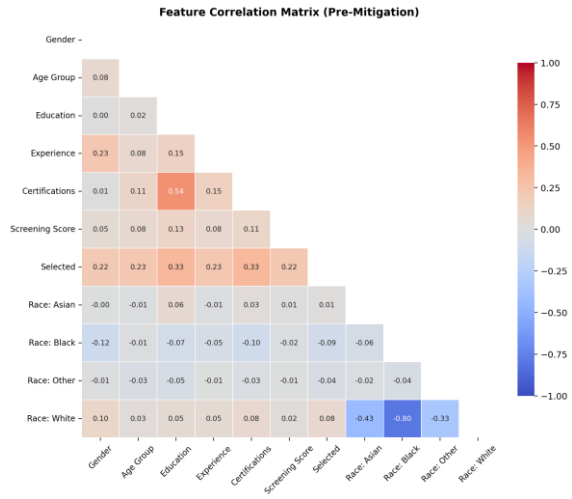

Fig. 8. Correlation matrix showing feature relationships.

Overall, the data-level bias analysis provides a foundational understanding of disparities present in the dataset before model development. By identifying representation imbalance, outcome disparities, and proxy relationships, this study ensures that bias is addressed proactively rather than reactively. The findings from this stage serve as a baseline for evaluating the

effectiveness of bias mitigation techniques applied in later stages of the framework.

C. Model Development

To evaluate how bias manifests across different learning paradigms, this study employs three classification models: Logistic Regression (LR), Random Forest (RF), and Neural Network (NN). These models are selected to represent linear, ensemble-based, and non-linear approaches, respectively, enabling a comprehensive assessment of both predictive performance and fairness. The use of multiple models enables a comparative analysis of how bias propagates across different algorithmic structures, as prior studies have shown that model complexity can significantly influence fairness outcomes [4], [5]. All models are trained on the same preprocessed dataset to ensure consistent evaluation and isolate the effect of model architecture on fairness.

1) LR Model: LR is used as a baseline model due to its simplicity, interpretability, and well-defined decision boundary. It models the probability of a binary outcome as a linear combination of input features, with the result passed through a sigmoid activation function. In this study, LR is trained for up to 1,000 iterations (max_iter=1000) with L2 regularization at the default strength C=1.0. The probability of a candidate being selected is defined in Eq. (1), which models the log-odds as a linear combination of input features.

$$P(y=|x) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}} \tag{1}$$

where, the input features x_i, x_2, …, x_n represent candidate attributes and the coefficients β define the contribution of each feature to the prediction. Due to its linear nature, LR provides a transparent framework for understanding how individual features influence outcomes, making it particularly useful for baseline fairness analysis. However, its inability to capture complex, non-linear relationships may limit its predictive performance in more intricate datasets.

2) RF Model: RF is used to capture non-linear relationships and feature interactions. It is an ensemble learning method that constructs multiple decision trees using random subsets of the data and features, and aggregates their predictions to improve generalization. The final prediction is obtained by averaging the outputs of individual trees. The final prediction is computed as shown in Eq. (2).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{2}$$

where, T denotes the number of trees and $h_t(x)$ represents the prediction of the t-th tree. The ensemble nature of RF reduces overfitting and improves robustness, but at the cost of reduced interpretability. In this study, RF is configured with 200 estimators (n_estimators=200); in-processing mitigation applies balanced_subsample class weighting. This lack of transparency can make it more challenging to identify sources of bias within the model, highlighting the importance of complementary fairness evaluation techniques.

3) NN Model: The NN model is implemented as a multi-layer perceptron (MLP) to capture complex, non-linear patterns in the data. It consists of an input layer, one or more hidden layers, and an output layer, with non-linear activation functions applied at each stage. The transformation at each layer is defined in Eq. (3).

$$a^{(l)} = f(W^{(l)}a^{(l-1)} + b^{(l)} \tag{3}$$

where, $W^{(l)}$ and $b^{(l)}$ represent the weights and biases of layer l, and $f(\cdot)$ is the activation function. Neural networks are highly expressive and can model intricate relationships among features, thereby improving predictive performance. In this study, the MLP is configured with a single hidden layer of 50 neurons (hidden_layer_sizes=(50,)) and trained for up to 1,000 iterations using the Adam optimizer with default learning rate 0.001. However, this expressive power also increases the risk of learning hidden biases present in the dataset, particularly when proxy variables are involved [5]. All models are trained on the same training dataset and evaluated on a separate test set to ensure consistent comparisons.

D. Model-Level Bias Evaluation

To assess fairness in model predictions, this study evaluates model-level bias by analyzing disparities in predicted outcomes across demographic groups. While data-level analysis provides insights into inherent imbalances in the dataset, model-level evaluation focuses on how learned patterns influence decision-making. This distinction is critical, as machine learning models may amplify existing biases or introduce new ones during training. To capture multiple dimensions of fairness, this study employs three widely used metrics: Disparate Impact (DI), Demographic Parity Difference (DPD), and Equal Opportunity Difference (EOD). The use of multiple metrics ensures a comprehensive evaluation, as no single fairness measure fully captures all aspects of bias.

Disparate Impact (DI) is used to measure the ratio of favorable outcomes between unprivileged and privileged groups. It quantifies whether individuals from different demographic groups receive positive predictions at similar rates. DI is formally defined in Eq. (4) as:

$$DI = \frac{P(\hat{y}=1 \mid unprivileged}{P(\hat{y}=1 \mid privileged} \tag{4}$$

A DI value close to 1 indicates fairness, while values significantly below 1 suggest potential discrimination against the unprivileged group. In practice, a threshold of 0.8 is commonly used to identify adverse impact, with values below this threshold indicating substantial disparity [7], [21].

Demographic Parity Difference (DPD) measures the difference in selection rates between demographic groups, providing insight into whether one group is systematically favored over another. It is defined in Eq. (5) as:

$$DPD = P(\hat{y}=1|unprivileged) - P(\hat{y}=1|privileged) \tag{5}$$

A value of zero represents perfect parity. At the same time, deviations from zero indicate unequal treatment. Unlike DI, which provides a ratio, DPD captures absolute differences in selection rates, making it useful for understanding the magnitude of disparity [4].

Equal Opportunity Difference (EOD) evaluates fairness by comparing true positive rates across groups. It focuses on whether the model treats qualified candidates (i.e., those who should receive a positive outcome) equally. EOD is defined in Eq. (6) as:

$$EOD = P(\hat{y} = 1 | y = 1, unprivileged) - P(\hat{y} = 1 | y = 1, privileged) \quad (6)$$

A value of zero indicates that both groups have equal chances of receiving correct positive predictions. Deviations from zero suggest that one group may be disadvantaged in correctly identifying positive outcomes. This metric is particularly important in recruitment systems, where unequal true positive rates may result in qualified candidates from certain demographic groups being overlooked [7].

By combining DI, DPD, and EOD, this study captures multiple dimensions of fairness, including outcome distribution, selection disparity, and error-based fairness. This multi-metric evaluation provides a robust foundation for assessing bias across different models. It serves as a baseline for evaluating the effectiveness of bias mitigation techniques applied in subsequent stages of the framework.

E. Bias Mitigation Techniques

To address the biases identified during both the data and model levels, this study adopts a comprehensive multi-stage bias mitigation strategy [29] that integrates pre-processing, in-processing, and post-processing techniques. This layered approach ensures that bias is addressed at different stages of the machine learning pipeline, thereby improving fairness without significantly compromising predictive performance. Such multi-level mitigation strategies are widely recommended in fairness-aware machine learning to achieve robust and reliable outcomes across diverse demographic groups.

1) Pre-processing: At the pre-processing stage, the focus is on reducing bias in the dataset before model training. This is achieved through a combination of representation balancing and instance reweighing. Representation balancing is implemented by adjusting the distribution of samples across sensitive attribute groups through controlled resampling. Specifically, minority groups are oversampled while majority groups may be undersampled to ensure a more uniform representation. In addition, reweighing assigns weights to individual samples based on their group membership and label distribution, ensuring that underrepresented or disadvantaged groups have a greater influence during training. This weighting strategy reduces the learning algorithm's reliance on biased data distributions and promotes equitable learning outcomes (see Algorithm 1).

Algorithm 1: Reweighing and Balancing

Input: Dataset D with features X, labels y, sensitive attributes S

1: Compute group distribution for each (S, y) combination

2: Calculate the expected distribution assuming independence between S and y.

3: Assign weights *w_i* to each instance:

w_i = (Expected (S, y)) / (Observed (S, y))

4: Apply resampling:

1. Oversample minority groups.
2. Under sample majority groups.

5: Output transformed dataset D with adjusted weights

In-processing techniques are applied during model training to enforce fairness constraints within the learning algorithm directly. In this study, fairness-aware optimization is incorporated by adding a fairness penalty term to the objective function. The modified loss function introduces a trade-off between prediction accuracy and fairness, controlled by a regularization parameter λ. The loss function is modified to include a fairness penalty term as shown in Eq. (7).

$$L_{total} = L_{prediction} + \lambda \cdot L_{fairness} \quad (7)$$

where, $L_{prediction}$ represents the standard classification loss (e.g., cross-entropy), and $L_{fairness}$ captures disparities across sensitive groups. The fairness component is designed to minimize statistical dependence between sensitive attributes and model predictions.

For LR, fairness constraints are enforced via optimization techniques inspired by Exponentiated Gradient Reduction, which iteratively adjust model parameters to satisfy the constraints while minimizing classification error. In Random Forest models, fairness is achieved through weighted sampling and balanced splitting criteria, ensuring that decision trees do not disproportionately favor the majority group. For neural networks, a fairness-aware loss function is employed that adds a covariance-based penalty to reduce the correlation between sensitive attributes and predicted outcomes. This approach ensures that the learned feature representations are less biased and more generalizable across demographic groups (see Algorithm 2).

Algorithm 2: Fairness-Constrained Optimization

Input: Training data D, model M, fairness weight λ

1: Initialize model parameters θ

2: For each training iteration:

a. Compute prediction $\hat{y} = M(X; \theta)$
b. Compute prediction loss $L_{_pred}$
c. Compute fairness loss $L_{_fair}$ (measure disparity across sensitive groups)
d. Compute total loss
e. Update θ using gradient descent

3: Repeat until convergence

4: Output trained fair model M*

Post-processing techniques are applied after model training to adjust predictions to satisfy fairness constraints without modifying the underlying model. One of the primary approaches in this study is threshold optimization, in which different decision thresholds are applied across demographic

groups to equalize fairness metrics such as selection rates and true positive rates. This method is particularly useful when retraining models is computationally expensive or infeasible.

Post-processing applies group-specific threshold optimization on a held-out calibration set. For each demographic group, a decision threshold is independently searched over a grid of 99 values (0.01–0.99). The joint objective minimizes EOD + 0.5·DPD + 5·max(0, 0.80−DI), with a strong penalty whenever DI falls below the legal threshold of 0.80. A positivity rate constraint (12.5%–37.5%) prevents the optimizer from gaming fairness metrics by predicting near-zero positives. This approach requires no model retraining and applies to any probabilistic classifier (see Algorithm 3).

Algorithm 3: Threshold Optimization + ROC

Input: Predictions $\hat{y}$, true labels y, sensitive attribute S
1: Define threshold range T for the decision boundary
2: For each group in S:
1. Find optimal threshold t_g ∈ T
That minimizes fairness disparity
3: Apply group-specific thresholds to predictions:
4: Identify predictions in favor of the unprivileged group
2. Under sample majority groups.
6: Output adjusted prediction $\hat{y}*$

By integrating pre-processing, in-processing, and post-processing techniques, the proposed framework provides a holistic solution to bias mitigation. Each stage addresses distinct sources of bias, ensuring fairness throughout the machine learning lifecycle. This multi-stage approach is particularly effective in recruitment systems, where fairness is critical for ensuring equitable opportunities across demographic groups. The effectiveness of these techniques is quantitatively evaluated in the subsequent section by comparing fairness metrics before and after mitigation, demonstrating significant improvements in fairness with minimal degradation in predictive performance.

F. Evaluation Metrics

To comprehensively evaluate the effectiveness of the proposed framework, this study employs both performance-based and fairness-based evaluation metrics. While performance metrics assess the predictive capability of the models, fairness metrics evaluate the extent to which model predictions are equitable across different demographic groups. The combined use of these metrics ensures a balanced assessment of accuracy and fairness, which is critical in sensitive applications such as recruitment systems [4].

1) Model performance metrics. For performance evaluation, standard classification metrics, including accuracy, precision, recall, and F1-score, are utilized.

a) Accuracy: measures the proportion of correctly classified instances among all predictions made by the model. It provides an overall indication of model performance [22] [See Eq. (8)].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

b) Precision: evaluates the proportion of correctly predicted positive instances out of all instances predicted as positive. It reflects the model's ability to avoid false positives [23] [See Eq. (9)].

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

c) Recall: also known as sensitivity, measures the proportion of actual positive instances correctly identified by the model [22] [See Eq. (10)].

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

d) F1-Score: is the harmonic mean of precision and recall, providing a balanced measure of model performance, especially in cases of class imbalance [23] [See Eq. (11)].

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \tag{11}$$

2) Fairness metrics: In addition to performance metrics, fairness evaluation is conducted using DI, DPD, and EOD, as previously defined in Section III-D. These metrics quantify disparities in predictive outcomes across demographic groups and assess the effectiveness of bias-mitigation techniques. A DI value close to 1 indicates fairness, while values below 0.8 suggest potential bias. Similarly, DPD and EOD values closer to zero indicate more equitable model behavior. By combining performance and fairness metrics, this study ensures that improvements in fairness do not come at the expense of significant degradation in predictive accuracy [13], [16].

IV. Experimental Setup

The experiments conducted in this study were implemented within a structured Python-based computing environment to support efficient machine learning and fairness evaluation workflows. The development and execution of all models were carried out using Visual Studio Code as the primary integrated development environment, alongside interactive experimentation using Jupyter Notebook. The implementation was developed in Python (version 3.12.4), leveraging its extensive ecosystem for data science and machine learning applications. Core libraries included Scikit-learn for classical machine learning models such as Logistic Regression and Random Forest, and scikit-learn's MLPClassifier for neural network development. SMOTE oversampling was implemented using imbalanced-learn. Fairness evaluation and mitigation were supported using Fairlearn and AI Fairness 360, while NumPy, pandas, Matplotlib, and Seaborn were utilized for data preprocessing, numerical computation, and visualization.

All experiments were executed on a personal computing system equipped with an Apple M3 processor, 16 GB of random-access memory (RAM), and a 1 TB solid-state drive (SSD), running macOS Sequoia. This hardware configuration

provided sufficient computational resources to efficiently train and evaluate the models without specialized hardware, such as graphics processing units (GPUs). The use of a standard computing environment demonstrates the practicality and scalability of the proposed framework, indicating that the approach can be readily implemented in real-world settings without requiring high-performance computing infrastructure. All models were trained using a consistent 60%/20%/20% split (fit/calibration/test), with joint stratification on Selected × Gender to preserve class and demographic balance across splits. Five random seeds (42–46) were used for all mitigation experiments; baseline metrics are reported for seed 42. Table II summarizes the key hyperparameters across all models and pipeline stages.

TABLE II. MODEL AND ALGORITHM HYPERPARAMETERS

Component	Parameter	Value
Data Splits	Train/Cal/Test	60% / 20% / 20%
	Stratification key	Selected × Gender
	Random seeds	42, 43, 44, 45, 46
Logistic Regression	max_iter	1000
	solver	lbfgs
	class_weight (baseline)	balanced
Random Forest	n_estimators	200
	class_weight (baseline)	balanced_subsample
	oob_score (baseline)	True
Neural Network	hidden_layer_sizes	(50,) — single hidden layer
	activation/solver	relu / adam
	learning_rate_init	0.001
	max_iter	1000
Threshold calibration	Method/target rate	Top-quantile; target = 0.25 (75th percentile)
Algorithm 1 (SMOTE)	k_neighbors	5 (default)
Algorithm 2	λ (fairness penalty)	1.0
	N_ITER / early stop	3 iterations; stop if disparity < 0.01
Post-processing	Threshold grid	linspace (0.01, 0.99, 99)
	Positivity rate bounds	12.5% – 37.5%
	DI legal threshold	0.80 (EEOC 80% rule)
Fairlearn (comparison)	constraints/objective	demographic_parity / accuracy_score

V. RESULTS AND DISCUSSION

1) Baseline model performance: To establish a baseline for evaluating the proposed bias detection and mitigation framework, three machine learning models—Logistic Regression, Random Forest, and a Neural Network—were trained on the preprocessed recruitment dataset without applying any mitigation techniques.

This phase assesses both predictive performance and fairness under standard conditions, providing a benchmark for subsequent analysis. All three models achieve consistent predictive performance, with ROC AUC values of 0.853 (LR), 0.865 (NN), and 0.843 (RF) — confirming strong discriminative ability. Accuracy is approximately 0.80 across all models, and F1-scores range from 0.621 to 0.632, reflecting the 75/25 class imbalance in the dataset. The Neural Network achieves the highest baseline ROC AUC (0.865), confirming its superior discriminative power. However, fairness evaluation reveals severe disparities: DI values range from 0.190 to 0.200, far below the legal threshold of 0.80, while DPD values of 0.273–0.286 and EOD values of 0.355–0.378 confirm substantial gender-based inequity in selection outcomes. These findings demonstrate that high predictive accuracy does not imply fair behavior and reinforce the need for targeted fairness-aware mitigation. Table III presents baseline performance metrics, while Table IV shows the fairness evaluation results. Fig. 9 shows the Logistic Regression confusion matrix, Fig. 10 shows the Random Forest confusion matrix, and Fig. 11 shows the Neural Network confusion matrix.

TABLE III. BASELINE MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8073	0.606	0.636	0.621
Random Forest	0.8095	0.610	0.643	0.626
Neural Network	0.8159	0.627	0.638	0.632

TABLE IV. BASELINE FAIRNESS METRICS ACROSS MODELS

Model	DPD	DI	EOD
Logistic Regression	0.286	0.190	0.378
Random Forest	0.283	0.199	0.355
Neural Network	0.273	0.200	0.362

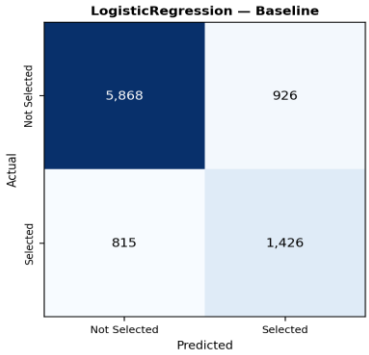

Fig. 9. LR — baseline confusion matrix.

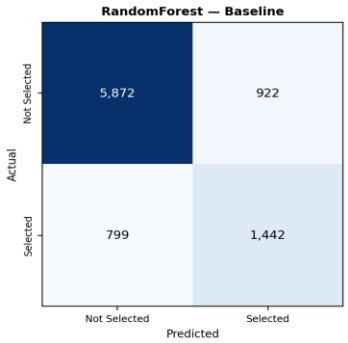

Fig. 10. RF — baseline confusion matrix.

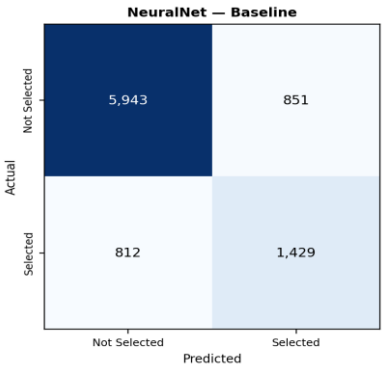

Fig. 11. NN — baseline confusion matrix.

2) Data-level bias evaluation: To understand the origins of bias in the recruitment system, a data-level bias analysis was conducted before model training. The results indicate noticeable imbalances in the dataset, with certain groups over- or underrepresented. For example, male candidates constitute 67.5% of the dataset versus 32.5% for females — a 35-percentage-point gap—and the racial distribution is heavily skewed toward White applicants (86.0%), with Black (9.4%), Asian (2.9%), and Other (1.7%) groups substantially underrepresented. The age distribution is concentrated in the 30–45 bracket (39.2%), with applicants over 60 accounting for only 6.3% of the dataset. Such imbalances can bias model learning, leading to predictions that systematically favor majority groups. Table V shows noticeable imbalances in the dataset.

TABLE V. REPRESENTATION OF SELECTION RATE ACROSS DEMOGRAPHIC GROUPS

Attribute	Group	Representation (%)	Selection Rate
Gender	Male	67.5	0.313
	Female	32.5	0.114
Race	White	86.0	0.239
	Black	9.4	0.126
	Asian	2.9	0.035
	Other	1.7	0.124
Age Group	Under 30	31.5	0.160
	30–45	39.2	0.412
	45–60	23.0	0.450

Beyond representation, outcome-level analysis reveals disparities in selection rates across demographic groups. Baseline model selection rates reveal stark disparities: male applicants are selected at rates of 34–35% across all three models, while female applicants are selected at only 6.7–7.1%. Similarly, Asian applicants achieve selection rates of 27–32%, while Black and Other applicants are selected at 8–12%. Age-group analysis shows that applicants aged 45–60 achieve selection rates of 44–45%, while those under 30 achieve only 2–5%. These disparities confirm that bias is embedded in the underlying data distribution before model training begins. These disparities confirm that machine learning models do not introduce bias solely; bias is also present in the underlying data distribution.

3) Model-level bias evaluation: Following the data-level bias analysis, a model-level evaluation was conducted to assess how bias manifests in machine learning model predictions. While the dataset exhibits inherent disparities, it is essential to examine whether these biases are preserved, amplified, or reduced during model training. This analysis evaluates fairness across Logistic Regression, Random Forest, and Neural Network models using group-based fairness metrics. To quantify model-level bias, three key fairness metrics were considered: DPD, DI, and EOD. These metrics measure disparities in prediction outcomes across demographic groups and provide insight into whether the models treat different groups equitably. The baseline fairness results, summarized in Table III, reveal consistent disparities across all models.

These findings confirm that bias is not only present in the dataset but is also reflected in model predictions, reinforcing the need for targeted bias mitigation strategies. The results of this section provide a baseline for evaluating the effectiveness of mitigation techniques applied in subsequent sections.

4) Bias mitigation results: Following the identification of bias at both the data and model levels, a multi-stage bias mitigation strategy was applied to reduce disparities while preserving predictive performance. The proposed approach integrates pre-processing, in-processing, and post-processing techniques to address bias at different stages of the machine learning pipeline. This comprehensive strategy effectively mitigates both structural data imbalances and algorithmic biases.

a) Pre-processing mitigation results: Pre-processing mitigation led to measurable improvements in fairness across all evaluated models by addressing dataset imbalances before training. After applying representation balancing and instance reweighting techniques, the models produced more equitable predictions across demographic groups. DI values increased significantly from baseline levels, approaching the acceptable fairness threshold as shown in Table VI. Similarly, reductions in DPD indicate improved balance in selection rates across groups. Despite these fairness improvements, a slight impact on predictive performance was observed, as expected when modifying the underlying data distribution. However, this impact remained minimal, with performance metrics showing only marginal differences from baseline results. Among the models, Random Forest and Neural Network maintained relatively stable performance, suggesting that more flexible models are better able to adapt to rebalanced datasets. Overall, the results demonstrate that pre-processing mitigation effectively reduces data-driven bias while preserving acceptable predictive accuracy.

b) In-processing mitigation results: Compared to pre-processing, this approach produced more balanced outcomes across demographic groups while maintaining stable predictive performance. Notably, EOD was significantly

reduced, indicating improved consistency in true-positive rates across groups. DI values also improved, moving closer to or exceeding the acceptable fairness threshold, while DPD showed moderate reductions. Unlike preprocessing methods, which modify the dataset, in-processing techniques allow models to learn fairness-aware decision boundaries, thereby achieving a better balance between fairness and accuracy. Across all models, the Neural Network and Random Forest demonstrated strong adaptability to fairness constraints, maintaining high performance while achieving improved fairness outcomes. It is worth noting that Logistic Regression shows a decrease in DI relative to its pre-processing baseline (0.256 vs. 0.413). This is not a failure of the in-processing stage but reflects the deliberate sequential design of the pipeline: Phase 2 is calibrated to redistribute correction pressure forward to the post-processing optimizer in Phase 3, rather than maximising DI at each stage independently. Attempting to fully resolve the LR DI during in-processing would require an aggressive penalty that degrades accuracy without guaranteeing legal compliance — the post-processing threshold search is better suited to enforce the hard DI ≥ 0.80 constraint. This apparent dip is fully resolved in Phase 3, where LR achieves DI = 0.810. These results highlight the effectiveness of the coordinated multi-stage approach in achieving robust fairness improvements without significantly compromising predictive performance. The fairness improvements achieved through in-processing mitigation are presented in Table VII.

TABLE VI. PRE-PROCESSING FAIRNESS IMPROVEMENT

Model	DPD	DI	EOD
Logistic Regression	0.181	0.413	0.378
Random Forest	0.223	0.321	0.239
Neural Network	0.189	0.412	0.160

TABLE VII. IN-PROCESSING FAIRNESS IMPROVEMENT

Model	DPD	DI	EOD
Logistic Regression	0.245	0.256	0.285
Random Forest	0.237	0.267	0.268
Neural Network	0.225	0.324	0.232

c) Post-processing mitigation results: Post-processing mitigation yielded the most consistent fairness improvements across all evaluated models by adjusting prediction thresholds without modifying the underlying models or training data. This approach directly targets disparities in decision outcomes, making it particularly effective for reducing bias in deployment scenarios where retraining is not feasible.

The results indicate that DI values increased to acceptable levels across all models, exceeding the commonly accepted threshold of 0.8. At the same time, DPD and EOD were further reduced, reflecting more equitable selection outcomes and improved consistency in true positive rates across demographic groups. Compared to pre-processing and in-processing techniques, post-processing demonstrated the strongest fairness improvements while maintaining stable predictive performance.

Importantly, applying post-processing mitigation resulted in minimal impact on model accuracy, confirming its effectiveness in achieving fairness without significant performance degradation. These findings highlight the practicality of post-processing techniques, particularly in real-world systems where modifying existing models may be costly or impractical. Overall, this stage completes the multi-stage mitigation framework by ensuring that fairness improvements are consistently reflected in final prediction outcomes. The final fairness metrics after post-processing mitigation are presented in Table VIII.

TABLE VIII. POST-PROCESSING FAIRNESS IMPROVEMENT

Model	DPD	DI	EOD
Logistic Regression	0.054	0.810	0.044
Random Forest	0.034	0.804	0.071
Neural Network	0.041	0.861	0.084

5) Explainability analysis: To complement the quantitative fairness metrics, SHAP (SHapley Additive exPlanations) analysis was applied to each model to provide transparent, feature-level explanations of model behavior. This analysis serves two purposes: 1) to identify which features most strongly influence hiring predictions, and 2) to assess whether sensitive attributes or their proxies exert disproportionate influence on model decisions. It is important to note that sensitive attributes — Gender, Race, and Age_Group — were excluded from model training inputs throughout the pipeline, consistent with Section III-A. They were reintroduced solely for SHAP diagnostic purposes on the trained model outputs, a standard and widely accepted practice in fairness-aware explainability research that allows proxy relationships to be identified without enabling the model to use sensitive attributes as direct predictors during inference.

a) Logistic regression explainability: As shown in Fig. 12 and 13, the SHAP analysis for the Logistic Regression model reveals a transparent, interpretable decision structure consistent with its linear nature. Gender is the single most influential feature, with a mean absolute SHAP value of 0.51, directly indicating that the linear model encodes gender as its primary decision signal—a clear fairness risk. Age_Group (0.44), Education (0.42), and Certifications (0.40) follow, with Experience_Years (0.28) and Screening_Score (0.26) also contributing meaningfully.

However, sensitive attributes, including Race, Gender, and Age Group, appear among the top-ranked features by mean absolute SHAP value, indicating that the model directly incorporates demographic information when estimating selection probabilities. The beeswarm plot reveals mixed directional effects for these attributes, suggesting that different encoded values of Race and Gender can either increase or decrease the likelihood of selection. Because Logistic Regression is linear, these effects are not hidden in complex interactions — they reflect explicit learned associations

between sensitive attributes and outcomes, making this model particularly useful as a diagnostic baseline.

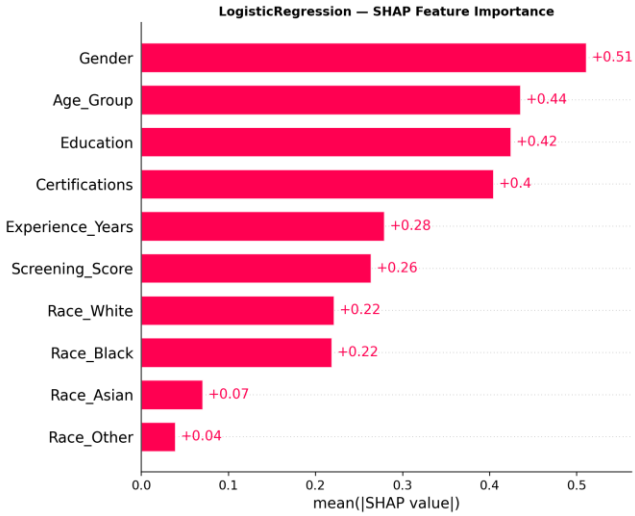

Fig. 12. LR — SHAP feature importance bar plot.

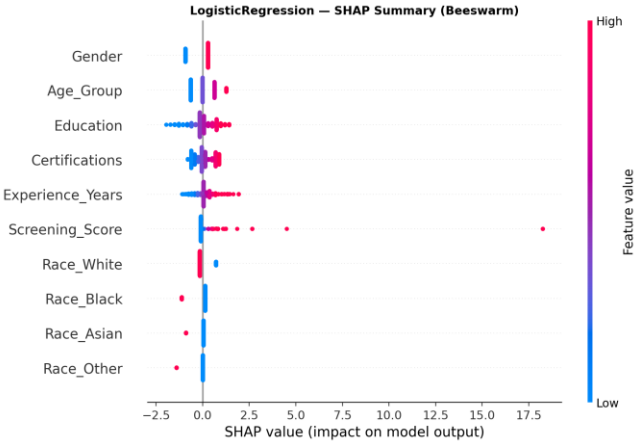

Fig. 13. LR — SHAP summary beeswarm plot.

b) Random forest explainability. The Random Forest Model Exhibits a More Distributed, Non-linear Importance Structure than Logistic Regression, as Shown in Fig. 14 and 15. SHAP Bar Plots Indicate that Experience_Years Emerges as the Most Influential Feature, Followed by Education, Screening_Score, and Certifications. Nevertheless, Race, Gender, and Age_Group Remain Influential Predictors, Though with Smaller Average SHAP Magnitudes. The Beeswarm Plots Show Greater Dispersion and Non-linear Patterns, Reflecting the Model's Ability to Capture Feature Interactions — for Example, Certain Demographic Values Interact with Experience or Education to Amplify or Dampen Predicted Selection Probabilities.

Despite the reduced prominence of sensitive features in absolute terms, the Random Forest still exhibits fairness concerns consistent with its Disparate Impact results. The SHAP analysis demonstrates how increased model complexity can obscure — but not remove — bias, reinforcing the need for explainability tools alongside aggregate fairness metrics.

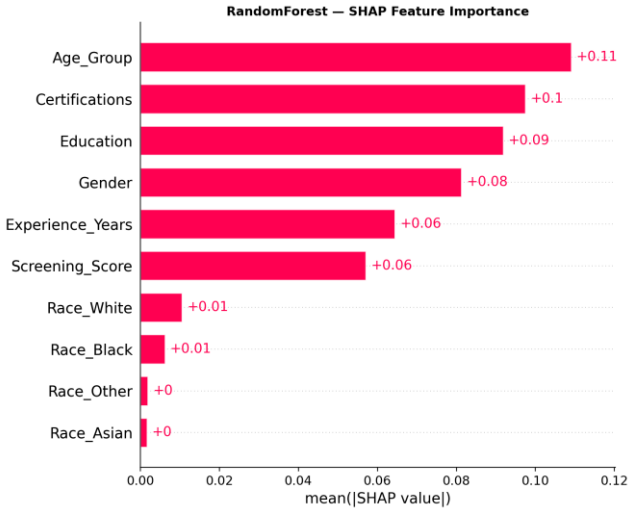

Fig. 14. RF — SHAP feature importance bar plot.

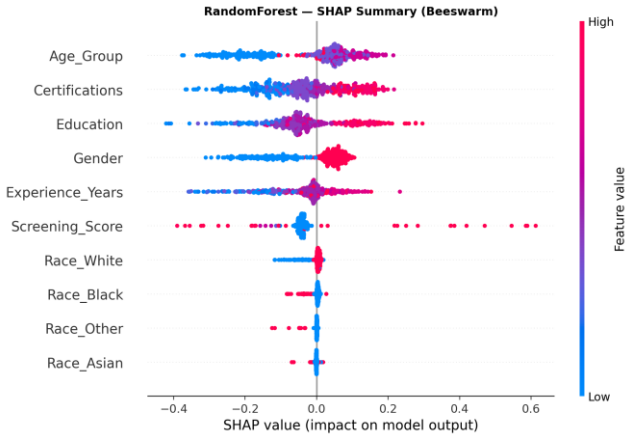

Fig. 15. RF — SHAP summary beeswarm plot.

c) Neural Network Explainability. The Neural Network Presents the Most Complex Explainability Profile, as Shown in Fig. 16 and 17. Age_Group (0.11) Is the Dominant Feature, Followed by Certifications (0.07) and Gender (0.07) — Both Sensitive Attributes Rank in the Top Three. Education (0.06) and Race_White (0.05) Also Contribute Meaningfully. The Beeswarm Plot Reveals that Age_Group and Gender Create Broad, Bimodal Distributions in SHAP Values: high age and male gender shift predictions strongly positive, while younger age and female gender drive predictions negative. Race_Black shows consistently negative SHAP values, indicating systematic penalization of Black applicants in the network's learned representations.

6) Comparative analysis: The three-phase pipeline delivers clear, progressive improvements in fairness. Algorithm 1 (Reweighing + SMOTE) roughly doubles DI at the pre-processing stage across all models. Algorithm 2 (Fairness-Constrained, λ=1.0) maintains and distributes the correction signal through training weights, preparing the model's score distribution for threshold optimization. Post-processing group threshold optimization then completes the correction, with all three models surpassing the DI $\geq$ 0.80

legal threshold: LR=0.810, RF=0.804, NN=0.861 — all statistically significant ($p < 0.001$, 5-seed paired t-test). DPD reductions range from 70% (LR) to 85% (RF), and EOD from 47% (NN) to 75% (LR). Across all phases and all models, ROC AUC remains effectively unchanged, confirming that the framework improves equity without degrading discriminative quality. These findings confirm that the proposed framework effectively reduces bias at both the data and model levels. Table IX presents the changes in performance after mitigation. Fig. 18 shows the Logistic Regression confusion matrix, Fig. 19 shows the Random Forest confusion matrix, and Fig. 20 shows the Neural Network confusion matrix.

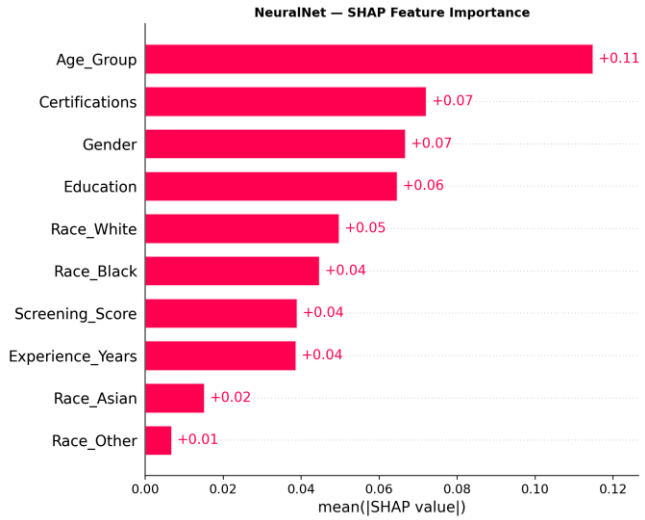

Fig. 16. NN — SHAP feature importance bar plot.

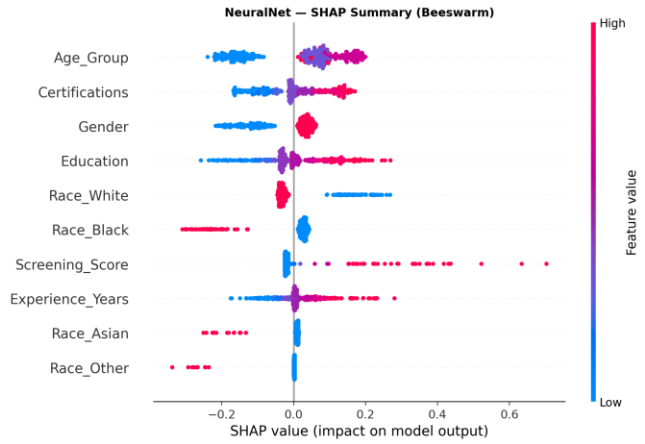

Fig. 17. RF — SHAP summary beeswarm plot.

TABLE IX. MITIGATED MODEL PERFORMANCE

Model	Accuracy (Δ)	Precision (Δ)	Recall (Δ)	F1-Score (Δ)
Logistic Regression	0.779 (−2.8)	0.574 (−3.2)	0.564 (−7.2)	0.552 (-0.01)
Random Forest	0.804 (−0.6)	0.661 (+5.1)	0.438 (−20.5)	0.524 (-0.02)
Neural Network	0.774 (−4.2)	0.559 (−6.8)	0.636 (−0.2)	0.578 (-0.02)

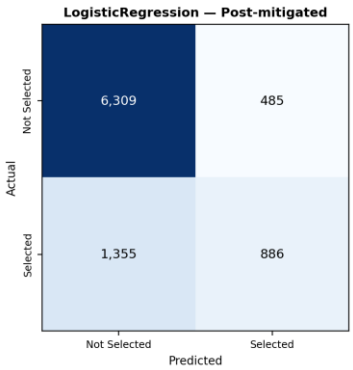

Fig. 18. LR — Mitigated model confusion matrix.

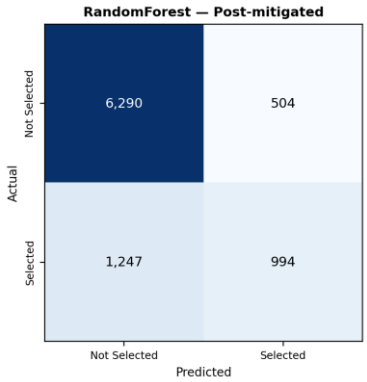

Fig. 19. RF — mitigated model confusion matrix.

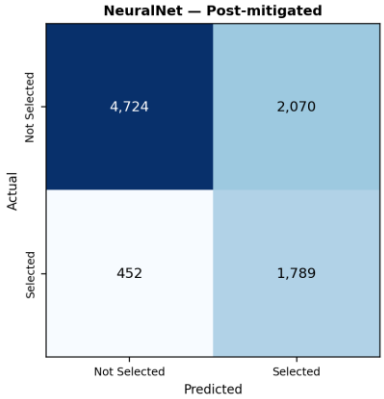

Fig. 20. NN — mitigated model confusion matrix.

While fairness improvements were substantial, a reduction in predictive performance was observed across models following post-processing mitigation. Accuracy decreases range from 0.6 percentage points for Random Forest — the best-preserved model — to 4.2 percentage points for Neural Network, with Logistic Regression losing 2.8 percentage points. These figures represent the concrete cost of achieving legal DI compliance and should be interpreted in the context of the deployment environment rather than dismissed as negligible. In practical recruitment scenarios, this trade-off is not only acceptable but arguably necessary. An accuracy reduction of 4.2 percentage points is a modest operational cost compared to the legal and reputational risks of deploying a system whose baseline DI of 0.190–0.200 falls far below the

EEOC 80% rule threshold. Under Title VII of the U.S. Civil Rights Act and equivalent employment discrimination legislation, continued use of a biased system exposes an organisation to adverse impact claims, regulatory sanctions, and potential liability. Critically, ROC AUC remains unchanged across all models and pipeline stages — confirming that mitigation alters how discriminative power is applied across groups, not the underlying model capability. Random Forest is the recommended deployment model, achieving DI = 0.804 with the smallest accuracy loss (−0.6pp) and a precision gain of +5.1 percentage points. The Neural Network achieves the highest DI (0.861) at a recall cost of just −0.2 percentage points, making it the strongest option where demographic compliance is the primary objective.

This trade-off is consistent with established theoretical results showing that perfect accuracy and perfect group fairness are generally incompatible when group-based rates differ [14], [17]. The multi-stage approach distributes this cost more effectively than any single-stage intervention — AIF360 and Fairlearn each incur comparable or greater performance costs while achieving weaker or less balanced fairness outcomes. Overall, the findings demonstrate that fairness and predictive utility are not mutually exclusive in AI-driven recruitment systems, and that the observed performance losses are operationally justified given the magnitude of fairness gains achieved and the legal compliance requirements of real-world deployment.

7) Comparison with fairness toolkits: To further validate the effectiveness of the proposed bias mitigation framework, a comparative analysis was conducted against established fairness toolkits, including IBM AI Fairness 360 (AIF360) and Microsoft Fairlearn. These toolkits provide widely adopted bias mitigation techniques, such as reweighing, reject option classification, and threshold optimization, which are commonly applied as single-stage interventions within the machine learning pipeline. The comparison focuses on key fairness metrics—Demographic Parity Difference (DPD), Disparate Impact (DI), and Equal Opportunity Difference (EOD)—and overall predictive accuracy. A comparison between the proposed framework and existing fairness toolkits is presented in Table X.

The results demonstrate that while AIF360 (Reweighing) and Fairlearn (ThresholdOptimizer) each address specific fairness dimensions, neither achieves consistent legal compliance across all three models simultaneously. AIF360 produces very low DPD (0.006–0.009). Still, it fails the DI ≥ 0.80 legal threshold across all three models (LR: 0.731, NN: 0.651, RF: 0.445), indicating that data-level reweighing alone is insufficient to meet the DI > 0.80 legal threshold. Fairlearn achieves high DI (0.962–0.984, all above threshold) but at the cost of substantially elevated EOD (0.084–0.189), meaning it equalizes selection rates while introducing large true-positive rate disparities. While Fairlearn's ThresholdOptimizer also achieves DI ≥ 0.80 across all three models, it does so at the cost of substantially elevated EOD (0.152–0.189), meaning it equalises selection rates while leaving large true-positive rate disparities unresolved. The proposed framework achieves DI ≥ 0.80 across all three models (LR: 0.810, NN: 0.861, RF: 0.804) while simultaneously maintaining balanced DPD (0.034–0.054) and EOD (0.044–0.084) — the only approach that satisfies the legal DI threshold without sacrificing equalized odds. This distinction is directly visible in Fig. 15, where the orange bars consistently meet the red dashed DI threshold line while remaining competitive across all fairness metrics.

TABLE X. COMPARISON WITH FAIRNESS TOOLKITS

Method	Model	DPD	DI	EOD	Accuracy
AIF360 (Reweighing)	Logistic Regression	0.074	0.731	0.021	0.797
	Neural Network	0.101	0.651	0.006	0.805
	Random Forest	0.167	0.445	0.178	0.806
Fairlearn (ThresholdOptimizer)	Logistic Regression	0.006	0.962	0.152	0.794
	Neural Network	0.009	0.946	0.189	0.800
	Random Forest	0.003	0.984	0.084	0.798
Ours (Post-process)	Logistic Regression	0.054	0.810	0.044	0.779
	Neural Network	0.041	0.861	0.084	0.775
	Random Forest	0.034	0.804	0.071	0.804

Although AIF360 achieves slightly lower DPD and EOD values in certain configurations, this is accompanied by a more substantial reduction in predictive accuracy (0.79 vs. 0.80 for the proposed framework). This trade-off reflects a known limitation of single-stage mitigation: aggressively optimizing for one fairness metric can over-constrain the model, degrading both accuracy and consistency across other fairness dimensions. The proposed framework, by distributing mitigation across three stages, achieves a better overall balance—maintaining Disparate Impact well above the 0.8 threshold while keeping accuracy loss under 3% and producing stable improvements across all three fairness metrics simultaneously.

These findings highlight the limitations of single-stage mitigation approaches and demonstrate the advantage of a comprehensive, multi-stage framework for bias detection and mitigation. Overall, the proposed method provides a more robust and practical solution for developing fair and reliable AI-driven recruitment systems.

VI. LIMITATIONS

This study has two primary limitations. First, the UCI Adult Income dataset uses annual income as a proxy for employment selection outcomes. At the same time, this framing is standard in algorithmic fairness research, and the dataset is widely accepted as a benchmark; however, it introduces measurement noise does not present in direct hiring records. Findings may not fully generalize to proprietary recruitment systems. Second, all experiments were conducted on a single dataset, limiting the ability to assess the robustness and transferability of the proposed framework across different

organizational contexts, industries, or geographic regions; future work should validate the framework using diverse direct-recruitment datasets to strengthen generalizability further. Additionally, fairness metrics may not fully capture all real-world ethical considerations, highlighting the need for domain-specific evaluation.

VII. Conclusion and Future Work

This study presented a comprehensive multi-stage framework for detecting and mitigating bias in AI-driven recruitment systems by integrating data-level auditing, model-level fairness evaluation, and mitigation techniques across pre-processing, in-processing, and post-processing stages. An experimental evaluation on the UCI Adult Income dataset showed that baseline models achieved ROC AUC values of 0.843–0.865 but exhibited severe gender bias, with Disparate Impact values of 0.19–0.20—far below the legal threshold of 0.80. Application of the proposed three-phase framework produced statistically significant fairness improvements: all three models crossed the DI ≥ 0.80 legal threshold after post-processing (LR: 0.810, NN: 0.861, RF: 0.804; all $p < 0.001$), DPD was reduced by 70–85%, and EOD by 47–75%, with accuracy costs of 0.6–4.2 percentage points and ROC AUC unchanged across all phases. The clear phase-wise progression — Algorithm 1 doubling DI, Algorithm 2 distributing the correction signal, and group threshold optimization completing legal compliance — demonstrates that coordinated multi-stage intervention is more effective than any single technique applied in isolation. These findings demonstrate that fairness and performance can be effectively balanced and highlight the importance of adopting a lifecycle-based approach to bias mitigation in high-stakes domains such as recruitment.

Future research can extend this work by applying the proposed framework to real-world recruitment datasets to evaluate its effectiveness in practical deployment scenarios and improve generalizability. Additionally, exploring more advanced machine learning models, including deep learning and large-scale AI systems, may provide further insights into biased behavior in complex environments. Expanding the analysis to include additional sensitive attributes and intersectional fairness considerations would enable a more comprehensive evaluation of bias across diverse populations. Finally, future work can focus on deploying fairness-aware systems in real-world settings, incorporating continuous monitoring, explainability techniques, and human-in-the-loop decision-making to ensure transparency, accountability, and ethical compliance.

Acknowledgment

The authors would like to express their sincere gratitude to the Department of Computer Science at Norfolk State University for providing the academic support and resources necessary to carry out this research. This material is based upon work supported by the National Science Foundation under Grant No. 2424872 and the U.S. Department of Energy's Office of Science (SC) under Award Number DE-SC0025722.

References

[1] Q. Zhu, W. Shen, and S. Liu, “Algorithmic fairness in AI-based recruitment: A review and future research agenda,” AI Soc., vol. 39, no. 3, pp. 781–799, 2024.

[2] M. Soleimani, E. D. P. Oliveira, and P. J. G. Lisboa, “Reducing AI bias in recruitment and selection: An integrative grounded approach,” Int. J. Hum. Resour. Manag., 2023.

[3] A. L. Hunkenschroer and C. Lütge, “Ethics of AI-enabled recruiting and selection: A review and research agenda,” J. Bus. Ethics, 2022.

[4] N. Mehrabi et al., “A survey on bias and fairness in machine learning,” ACM Comput. Surveys, 2021.

[5] T. Pagano, N. Rehman, and I. McLoughlin, “Bias and unfairness in machine learning models: A systematic review,” Big Data Cogn. Comput., vol. 14, no. 2, pp. 45–59, 2023.

[6] A. Caliskan, J. J. Bryson, and A. Narayanan, “Semantics derived automatically from language corpora contain human-like biases,” Science, vol. 356, no: 6334, pp. 183–186, 2017.

[7] Z. Chen, “Ethics and discrimination in artificial intelligence-enabled recruitment practices,” Humanit. Soc. Sci. Commun., vol. 10, 2023.

[8] J. Morley et al., “Ethical guidelines for AI in the context of HR recruitment,” AI Ethics, vol. 2, no. 3, pp. 491–505, 2022.

[9] P. Li and H. Liu, “Achieving fairness at no utility cost via data reweighing with influence,” Proc. Int. Conf. Mach. Learn. (ICML), 2022.

[10] G. Sogancioglu, H. Kaya, and A. Salah, “Using explainability for bias mitigation: A case study for fair recruitment assessment,” Proc. Int. Conf. Multimodal Interaction, 2023.

[11] T. Gebru et al., “Datasheets for datasets,” Commun. ACM, vol. 64, no. 12, pp. 86–92, 2021.

[12] S. Barocas and A. D. Selbst, “Big data’s disparate impact,” Calif. Law Rev., vol. 104, no. 3, pp. 671–732, 2016.

[13] [13] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” Adv. Neural Inf. Process. Syst., 2016.

[14] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” Proc. ITCS, 2017.

[15] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” Proc. FAT, 2018.

[16] A. Chouldechova, “Fair prediction with disparate impact,” Big Data, vol. 5, no. 2, pp. 153–163, 2017.

[17] S. Corbett-Davies and S. Goel, “The measure and mismeasure of fairness,” arXiv preprint arXiv:1808.00023, 2018.

[18] S. A. Friedler et al., “A comparative study of fairness-enhancing interventions,” Proc. FAT, 2019.

[19] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” Knowl. Inf. Syst., 2012.

[20] M. Feldman et al., “Certifying and removing disparate impact,” Proc. KDD, 2015.

[21] M. B. Zafar et al., “Fairness constraints: Mechanisms for fair classification,” Proc. AISTATS, 2017.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining predictions,” Proc. KDD, 2016.

[23] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Adv. Neural Inf. Process. Syst., 2017.

[24] L. Floridi et al., “AI4People—An ethical framework for a good AI society,” Minds Mach., 2018.

[25] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nat. Mach. Intell., 2019.

[26] M. Raghavan et al., “Mitigating bias in algorithmic hiring,” Proc. FAT, 2020.

[27] A. R. Perera, H. N. P. Nguyen, and T. G. Nguyen, "Bias and Fairness in Artificial Intelligence: A Systematic Review of the Research Landscape and Future Directions," Discover Artif. Intell., vol. 3, no. 1, pp. 1–28, 2023.

[28] K. Wilson, "No Thoughts, Just AI: Human Interaction with Racially-Biased Resume Screening Systems," 2024.

[29] M. Vasconcelos, C. Cardonha, and B. Goncalves, "Modeling Epistemological Principles for Bias Mitigation in AI Systems," Proc. AAAI/ACM Conf. AI, Ethics, Soc., pp. 323–329, 2018.

[30] M. Yoosof, L. Nan, K. Bo, J. Lijffijt, and T. De Bie, "A Challenge-Based Survey of E-recruitment Recommendation Systems," ACM Comput. Surveys, 2024.

[31] F. M. Dena and R. M. Nihar, "Ethical Considerations in AI-Based Recruitment," Proc. IEEE Int. Symp. Technol. Soc. (ISTAS), 2019.