

FairHR-X: An Explainable and Bias-Aware Artificial Intelligence Framework for Transparent Employee Recruitment and Selection

Ahmad Nader Aloqaily^{1,*}

¹ Department of Human Resource Management, Business School, Jadara University, Irbid, Jordan.

* Corresponding author: a.aloqaily@jadara.edu.jo

Received: 21 May 2026 Accepted: 2 September 2026 Published: 15 October 2026

DOI: 10.15849/ijasca.184 License: Open access under CC BY 4.0. Authors retain copyright.

Publisher: Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan.

Cite this article: A. N. Aloqaily, "FairHR-X: an explainable and bias-aware artificial intelligence framework for transparent employee recruitment and selection," *Int. J. Advance Soft Compu. Appl.*, vol. 18, no. 3, pp. 140–156, 2026, doi: <https://doi.org/10.15849/ijasca.184>.

ABSTRACT

Artificial intelligence can improve recruitment efficiency, but models trained on historical hiring data may reproduce direct and proxy discrimination. This paper presents FairHR-X, an explainable and bias-aware lifecycle framework for human-supervised employee recruitment and selection that integrates predictive modelling, fairness auditing, mitigation, calibration, explainability, human oversight, and continuous monitoring. The framework was evaluated on 10,000 recruiting records regenerated from the CDEI data-generating process, using 6,000 training, 2,000 validation, and 2,000 held-out test observations. Logistic regression achieved the best predictive performance, with a test ROC-AUC of 0.941, PR-AUC of 0.911, and accuracy of 87.35%. The selected fairness-aware configuration, based on intersectional reweighing and Platt calibration, retained 86.45% accuracy and reduced race demographic-parity difference from 0.263 to 0.213, race equalised-odds difference from 0.090 to 0.042, sex demographic-parity difference from 0.163 to 0.119, and the intersectional demographic-parity gap from 0.436 to 0.340. Bootstrap analysis confirmed stable fairness improvements, while SHAP rankings showed high robustness across 1,000 resamples. For external validation, a recruiter-grouped protocol is defined for the 2026 PATHS2INCLUDE dataset comprising 15,036 candidate evaluations from 2,506 recruiters across four European countries. FairHR-X is intended as a transparent decision-support and governance framework rather than an autonomous hiring authority.

Keywords: Algorithmic hiring; explainable artificial intelligence; fairness-aware machine learning; recruitment analytics; bias mitigation; SHAP; human-in-the-loop; MLOps.

1 Introduction

Artificial intelligence (AI) is increasingly used to support candidate sourcing, ranking, screening, interviewing, and employee selection [1]. Its potential benefits include consistent processing, reduced administrative workload, and scalable analysis of large applicant pools [2]. Recruitment decisions, however, affect access to income, career development, and professional opportunity. Historical hiring records may encode human preferences, structural inequalities, or organisational practices that are not legitimate indicators of future job performance [3]. Consequently, a model optimised exclusively for predictive accuracy may reproduce existing discriminatory patterns [4].

Bias cannot be addressed merely by deleting protected variables. Features such as education, career gaps, referral channels, institutional prestige, or socioeconomic information may remain statistically associated with sex, race, nationality, age, disability, or family status. Such proxy pathways can generate unequal selection rates even when protected characteristics are absent from the deployed feature vector [5, 6].

Fairness is also multidimensional. Demographic parity evaluates differences in positive decision rates, equal opportunity compares true-positive rates, and equalised odds additionally considers false-positive rates [5]. These criteria represent different normative objectives and cannot always be simultaneously satisfied when group base rates differ [7, 8].

Explainability introduces a complementary requirement. Recruiters, candidates, auditors, and model-risk reviewers require intelligible information about model behaviour, but an explanation alone does not

establish fairness. A transparent model can remain discriminatory, while aggregate fairness metrics may conceal harm to intersectional subgroups. A deployable recruitment system must therefore consider predictive quality, fairness, calibration, explainability, human oversight, and monitoring jointly.

This paper introduces FairHR-X, an explainable and bias-aware lifecycle framework for human-supervised recruitment decision support. The main contributions are:

1. A lifecycle architecture integrating data governance, predictive modelling, fairness auditing, mitigation, calibration, explainability, human review, and post-deployment monitoring.
2. A multi-objective selection strategy that separates model-training optimisation from fairness–utility Pareto analysis.
3. An intersectional fairness protocol covering single and combined protected groups with explicit minimum-support requirements.
4. An independent reproducible evaluation based on 10,000 records regenerated from the published CDEI recruiting data-generating process.
5. A comparison of five predictive models and multiple fairness configurations under a fixed train/validation/test protocol.
6. Quantitative SHAP stability and proxy-feature auditing rather than qualitative explanation claims.
7. A leakage-resistant external-validation protocol for the PATHS2INCLUDE dataset using recruiter-grouped splits and country-aware evaluation.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 introduces the FairHR-X architecture. Section 4 describes the datasets and governance assumptions. Section 5 defines the methodology. Section 6 presents the experimental design. Section 7 reports the empirical results. Section 8 examines explainability and intersectional behaviour. Section 9 defines the monitoring lifecycle. Section 10 discusses limitations, and Section 11 concludes the paper.

2 Related Work

2.1 Algorithmic recruitment and discrimination

Algorithmic hiring systems can standardise parts of recruitment, but their validity depends on target definition, data quality, subgroup representation, organisational context, and how model recommendations are used. Raghavan et al. [1] showed that commercial claims concerning bias mitigation often lack sufficient transparency regarding validation and fairness assumptions. Drage and Mackereth [3] similarly argued that technical debiasing cannot substitute for analysis of institutional discrimination.

Recent multidisciplinary work describes algorithmic hiring as a socio-technical system in which model design, recruiter behaviour, organisational practices, and legal requirements interact [4]. Empirical labour-market studies also report persistent differences in hiring outcomes across demographic groups [9, 10].

The PATHS2INCLUDE project extends this evidence base through a cross-national factorial survey experiment. Buttler et al. [11] released 15,036 candidate evaluations produced by 2,506 recruiters across Germany, Norway, Poland, and Romania, with controlled variation in nationality, gender, and care obligations.

2.2 Fairness-aware machine learning

Fairness interventions are commonly classified as pre-processing, in-processing, and post-processing methods. Reweighting modifies the influence of training examples to reduce dependence between protected membership and historical outcomes [12]. Constrained reductions transform fair classification

into a sequence of cost-sensitive optimisation problems [13]. Adversarial debiasing trains a predictor jointly with an adversary that attempts to recover protected information from model outputs [14].

Software frameworks such as AI Fairness 360 [15] and Fairlearn [16] provide reusable fairness metrics and mitigation procedures.

2.3 Explainable and accountable AI

SHAP provides additive feature attributions derived from cooperative game theory [17]. Generalised additive models with interactions provide an intrinsically interpretable alternative to post-hoc explanations [18], while InterpretML operationalises several glass-box and post-hoc methods [19]. Model cards provide a complementary governance mechanism by documenting intended use, training data, subgroup performance, limitations, and deployment assumptions [20].

Table 1 positions FairHR-X relative to representative approaches.

Table 1. Positioning of FairHR-X relative to representative approaches.

Approach	Prediction	Fairness	XAI	Monitoring	Context	Limitation addressed by FairHR-X
Accuracy-only screening	✓	–	–	–	Limited	Ignores subgroup harm, explanation quality, and governance.
Reweighting [12]	✓	✓	–	–	Limited	Provides a single pre-processing intervention without lifecycle controls.
Adversarial debiasing [14]	✓	✓	Partial	–	Limited	Addresses training-time fairness but not complete operational governance.
Model cards [20]	–	Partial	Partial	Partial	General	Provides documentation without an end-to-end recruitment decision pipeline.
CDEI benchmark [21]	✓	✓	Partial	–	Recruiting	Provides mitigation comparisons but not an integrated XAI–MLOps lifecycle.
FairHR-X	✓	✓	✓	✓	Recruiting	Integrates prediction, fairness, calibration, explanation, human oversight, and monitoring.

3 The FairHR-X Framework

FairHR-X is designed for human-supervised decision support. It does not issue an autonomous final employment decision. Instead, the framework produces a calibrated suitability score, fairness diagnostics, explanation information, uncertainty indicators, and a review status. Human reviewers retain authority over employment decisions, and overrides are recorded for audit.

Figure 1 presents the lifecycle architecture.

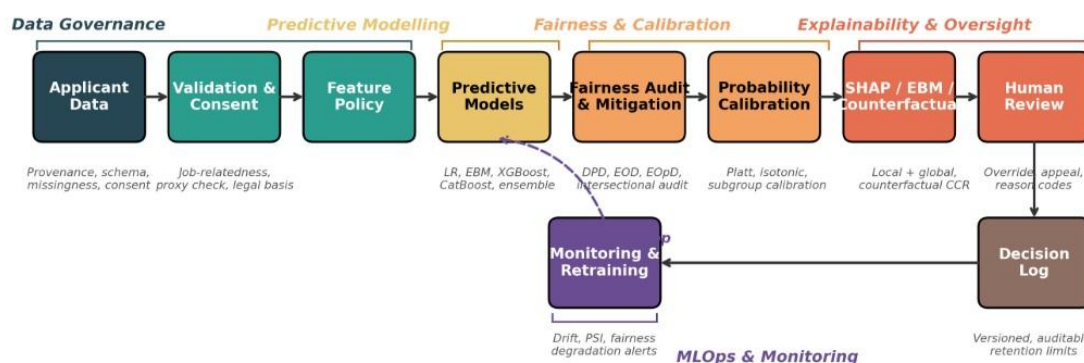


Fig. 1. Conceptual architecture of the FairHR-X lifecycle framework. Source: Created by the author.

3.1 Lifecycle layers

The framework consists of seven connected layers:

1. Data governance: provenance, schema validation, missingness control, access policy, subgroup representation, and prohibited-feature checks.
2. Predictive modelling: transparent and high-performance model candidates are evaluated using the same locked partitions.
3. Fairness auditing: demographic parity, disparate impact, equal opportunity, equalised odds, calibration by group, and intersectional gaps are measured.
4. Bias mitigation: reweighing, constrained learning, adversarial learning, and threshold-based post-processing can be incorporated.
5. Explainability: SHAP, additive-model explanations, proxy analysis, and counterfactual consistency are available for model review.
6. Human oversight: borderline or fairness-sensitive recommendations can be escalated to an authorised reviewer.
7. MLOps and monitoring: models, feature policies, fairness metrics, calibration, drift, and incidents are versioned and monitored.

3.2 Why fairness through unawareness is insufficient

Proposition 3.1. Let $A \in \{0,1\}$ be a protected attribute, let X be an apparently non-protected feature, and let $\hat{Y} = g(X)$ be a classifier that does not receive A . If $P(X | A = 0) \neq P(X | A = 1)$, and g is non-constant on a region where the two conditional distributions differ, demographic parity is not guaranteed.

Proof. By the law of total probability,

$$P(\hat{Y} = 1 | A = a) = \int \mathbb{I}\{g(x) = 1\} dP(x | A = a). \quad (1)$$

When the conditional distributions of X differ across values of A , the two integrals can differ although A is absent from the model. Therefore, protected-attribute removal does not imply equal selection rates.

This result formalises the proxy-feature problem. In recruitment, referral channels, career gaps, educational measures, and socioeconomic indicators may carry information correlated with protected-group membership.

4 Datasets and Ethical Protocol

4.1 CDEI recruiting benchmark

The primary controlled benchmark is based on the synthetic recruiting data-generating process released by the United Kingdom Centre for Data Ethics and Innovation (CDEI) [21, 22]. Public notebooks and generation resources are available through the CDEI repository [23].

The process contains job-related and contextual variables together with protected attributes for sex and race. Predictor groups include years of experience, referral status, GCSE and A-level indicators, Russell Group attendance, honours, volunteering, income, IT skills, career gaps, and CV quality. The binary target `employed_yes` encodes the generated historical employment outcome.

For the present study, 10,000 records were regenerated with a fixed reproducibility seed of 20260831. The overall positive-class proportion was 37.52%. The data were divided into 6,000 training, 2,000 validation, and 2,000 held-out test observations.

4.2 PATHS2INCLUDE external-validation dataset

External validation is structured around the open PATHS2INCLUDE “Hiring Discrimination in the Organisational Context” dataset [11, 24]. The dataset was collected in 2024–2025 and contains 15,036 evaluations of job-candidate profiles made by 2,506 recruiters in Germany, Norway, Poland, and Romania. Each recruiter evaluated six candidate profiles for one of five medium-skilled occupations: ICT technician, bookkeeping clerk, office clerk, secretary, or sales worker. Candidate factors include

nationality, gender, care obligations, language skills, qualifications, and referral information. The dataset also contains recruiter- and organisation-level contextual variables.

Because six observations originate from each recruiter, row-wise random splitting can leak recruiter-specific preferences across training and test sets. FairHR-X therefore defines recruiter ID as the grouping variable for external validation.

4.3 Feature governance

Protected characteristics are excluded from the ordinary predictor set but retained in a controlled audit store. A feature is approved only when it has a documented relationship to the recruitment task, acceptable measurement quality, and an appropriate governance basis.

The recruitment recommendation remains advisory, and model predictions should be accompanied by human-review and correction mechanisms.

Table 2. Datasets used in the FairHR-X study design.

Dataset	Scale	Outcome	Protected/context variables	Role
CDEI recruiting benchmark [21]	10,000 independently regenerated observations	Binary employment outcome	Sex, race, qualifications, experience, referral, skills, career gaps	Primary controlled experiment and fairness evaluation
PATHS2INCLUDE [11]	15,036 candidate evaluations; 2,506 recruiters; four countries	Interview and hiring likelihood	Gender, nationality, care obligations, recruiter, organisation, occupation	Recruiter-grouped external-validation dataset

5 Methodology

Let

$$D = \{(x_i, a_i, y_i)\}_{i=1}^n, \quad (2)$$

where $x_i \in \mathbb{R}^p$ contains approved applicant features, a_i denotes one or more protected-group memberships, and $y_i \in \{0,1\}$ is the employment outcome. A model f_θ produces

$$s_i = f_\theta(x_i) \in [0,1], \quad (3)$$

and the classification decision is

$$\hat{y}_i = \mathbb{I}\{s_i \geq \tau\}. \quad (4)$$

5.1 Predictive metrics

Accuracy is

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \quad (5)$$

Balanced accuracy is

$$BAcc = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right). \quad (6)$$

Precision, recall, F1, ROC–AUC, and PR–AUC are also reported. Calibration is measured using the Brier score

$$BS = \frac{1}{n} \sum_{i=1}^n (s_i - y_i)^2, \quad (7)$$

and expected calibration error

$$ECE = \sum_{b=1}^B \frac{|I_b|}{n} |acc(I_b) - conf(I_b)|. \quad (8)$$

5.2 Fairness metrics

Demographic-parity difference is

$$DPD = |P(\hat{Y} = 1 | A = 1) - P(\hat{Y} = 1 | A = 0)|. \quad (9)$$

The disparate-impact ratio is

$$DIR = \frac{P(\hat{Y} = 1 | A = d)}{P(\hat{Y} = 1 | A = r)}, \quad (10)$$

where d denotes the monitored group and r the pre-specified reference group.

Equal-opportunity difference is

$$EOpD = |TPR_{A=1} - TPR_{A=0}|. \quad (11)$$

Equalised-odds difference is

$$EOD = \max\{|TPR_{A=1} - TPR_{A=0}|, |FPR_{A=1} - FPR_{A=0}|\}. \quad (12)$$

For multiple protected dimensions,

$$G = A_1 \times A_2 \times \dots \times A_m, \quad (13)$$

and the intersectional selection-rate gap is

$$\Delta_{intersection} = \max_{g,h \in G} |P(\hat{Y} = 1 | G = g) - P(\hat{Y} = 1 | G = h)|. \quad (14)$$

The audit requires at least

$$n_g \geq \max(30, 0.05 n_{eval}) \quad (15)$$

for quantitative interpretation of an intersectional subgroup. Smaller groups are retained in the audit report but are marked as underpowered.

5.3 Fairness-aware model selection

FairHR-X separates model optimisation from governance selection. A general fairness-aware training objective can be expressed as

$$\mathcal{L}_{FairHR}(\theta) = \mathcal{L}_{pred}(\theta) + \lambda_{DP} DPD(\theta) + \lambda_{EO} EOD(\theta) + \lambda_{cal} BS(\theta). \quad (16)$$

Candidate models are subsequently compared on a non-dominated Pareto set

$$P = \{f \in F: \nexists f' \text{ such that } U(f') \geq U(f), D(f') \leq D(f)\}, \quad (17)$$

where U represents predictive utility and D the fairness disparity. Thus, Equation (16) describes a training objective, whereas Equation (17) describes the subsequent governance comparison across trained candidates.

Deployment constraints are expressed generically as

$$AUC \geq \alpha, \quad DPD \leq \epsilon_{DP}, \quad EOD \leq \epsilon_{EO}, \quad ECE \leq \epsilon_{cal}. \quad (18)$$

The values of α , ϵ_{DP} , ϵ_{EO} , and ϵ_{cal} are application-dependent and must be fixed before final test evaluation.

5.4 Bias mitigation

The empirical study evaluates: an unconstrained model containing protected attributes; a fairness-through-unawareness model excluding protected attributes; race-based reweighing; intersectional sex–

race reweighing; validation-optimised demographic-parity thresholding; and validation-optimised equalised-odds thresholding.

Reweighting follows the principle introduced by Kamiran and Calders [12]. FairHR-X can additionally accommodate constrained reductions [13] and adversarial debiasing [14]. For intersectional reweighting, the training weight for observation i is

$$w_i = \frac{P(G = g_i) P(Y = y_i)}{P(G = g_i, Y = y_i)}. \quad (19)$$

After mitigation, Platt scaling is fitted on the validation set and applied to the locked test probabilities.

5.5 Explainability and proxy-feature inspection

SHAP represents a prediction as

$$f(x) = \phi_0 + \sum_{j=1}^p \phi_j, \quad (20)$$

where ϕ_j is the feature contribution [17]. For the selected standardised logistic model, the linear SHAP contribution is

$$\phi_{ij} = \beta_j z_{ij}, \quad (21)$$

where z_{ij} is the standardised feature value and β_j the fitted coefficient.

A secondary proxy audit measures the absolute Spearman association between each ordinary predictor and both protected attributes. Features with

$$\max(|\rho(X_j, A_{race})|, |\rho(X_j, A_{sex})|) \geq 0.20 \quad (22)$$

are escalated for governance review. This threshold is used as a screening criterion rather than as evidence that a feature is intrinsically prohibited.

5.6 SHAP stability

Let $r^{(0)}$ denote the global SHAP importance ranking on the held-out sample and $r^{(b)}$ the ranking from bootstrap replicate b . Rank stability is

$$S_\rho = \frac{1}{B} \sum_{b=1}^B \rho_S(r^{(0)}, r^{(b)}), \quad (23)$$

where ρ_S is Spearman's rank correlation. Top- k stability is measured by

$$S_J = \frac{1}{B} \sum_{b=1}^B \frac{|T_k^{(0)} \cap T_k^{(b)}|}{|T_k^{(0)} \cup T_k^{(b)}|}. \quad (24)$$

6 Experimental Design

6.1 Data generation and partitioning

The CDEI recruiting process was independently regenerated with random seed 20260831, yielding 10,000 observations. A fixed partition was used:

$$N_{train} = 6000, \quad N_{validation} = 2000, \quad N_{test} = 2000. \quad (25)$$

The corresponding positive-class proportions were 37.45%, 37.85%, and 37.40%, respectively. Model selection and threshold optimisation were performed exclusively on training and validation data. The test set remained untouched until final evaluation.

Algorithm 1. FairHR-X training, auditing, and model selection**Require:** Dataset D , protected attributes A , candidate models M **Ensure:** Selected model f^* , threshold policy τ^* , audit report R

8. Validate schema, missingness, ranges, labels, and subgroup support
9. Separate protected audit attributes from ordinary predictors
10. Create training, validation, and locked test partitions
11. Fit all preprocessing transformations using training data only
12. **for all** $m \in M$ **do**
13. Train predictive model f_m using training data
14. Evaluate predictive utility and calibration on validation data
15. **end for**
16. Select the predictive backbone using validation ROC–AUC
17. Generate fairness-through-unawareness and reweighted candidates
18. Optimise post-processing thresholds using validation data only
19. Calibrate mitigated probabilities using validation data only
20. Evaluate utility, calibration, and fairness of all candidates on validation data
21. Construct validation fairness–utility Pareto frontiers
22. Apply the pre-specified governance selection rule
23. Select and lock f^* and τ^* before accessing the test set
24. Evaluate the locked f^* once on the held-out test set
25. Compute bootstrap confidence intervals from the locked test predictions
26. Compute SHAP explanations and proxy-feature associations for f^*
27. Generate the final audit report, model card, and monitoring baseline
28. **return** f^*, τ^*, R

6.2 Predictive models

Five predictive model families were evaluated.

Table 3. Predictive model configurations used in the empirical evaluation.

Model	Configuration
Logistic regression	StandardScaler; L2 regularisation; maximum 2,000 iterations
Random Forest [25]	500 trees; maximum depth 8; minimum leaf size 5
Histogram Gradient Boosting [26]	250 iterations; learning rate 0.05; maximum 15 leaves; L2 regularisation 1.0
XGBoost [27]	350 estimators; depth 3; learning rate 0.04; subsample 0.85; column subsample 0.85; $\lambda = 2.0$; $\alpha = 0.1$
CatBoost [28]	350 iterations; depth 5; learning rate 0.04; L2 regularisation 4

The predictive model with the highest validation ROC–AUC was used as the base classifier for the fairness experiments.

Fairness-aware model selection was also performed exclusively on the validation set. The selection rule was specified before final test evaluation. First, candidates with an accuracy decrease greater than one percentage point relative to the fairness-through-unawareness validation baseline were excluded from

the balanced operating-point comparison. Among the remaining fairness-aware candidates, preference was given to Pareto-efficient configurations that minimised intersectional equalised-odds difference, with intersectional demographic-parity difference and the race- and sex-specific equalised-odds differences used as secondary criteria.

On the validation set, fairness through unawareness achieved an accuracy of 0.8570, race DPD of 0.2648, race EOD of 0.0866, sex DPD of 0.1976, and intersectional EOD of 0.1860. Race reweighing retained an accuracy of 0.8505 and reduced intersectional EOD to 0.1099. Intersectional reweighing achieved an accuracy of 0.8530 while further reducing race DPD to 0.2211, race EOD to 0.0438, sex DPD to 0.1434, sex EOD to 0.0335, intersectional DPD to 0.3501, and intersectional EOD to 0.0532.

The demographic-parity and equalised-odds thresholding candidates reached validation accuracy of 0.8370 and therefore exceeded the one-percentage-point utility-loss limit used for selection of the balanced operating point. Accordingly, intersectional reweighing was selected and locked before the held-out test set was evaluated.

6.3 Probability calibration

The final intersectionally reweighted model was calibrated by fitting a logistic mapping between validation-set logits and observed outcomes, following the Platt-scaling principle [29]. Calibration parameters were estimated using validation data only.

6.4 Uncertainty analysis

Final-model confidence intervals were estimated through 1,000 non-parametric bootstrap resamples of the held-out test set. A paired bootstrap was also used to compare the selected fairness-aware model with fairness through unawareness. Both models were evaluated on identical bootstrap samples, preserving dependence between predictions.

6.5 PATHS2INCLUDE grouped evaluation

For PATHS2INCLUDE, all six candidate evaluations produced by a single recruiter are assigned to the same partition. The external-validation protocol comprises: recruiter-grouped 60/20/20 train/validation/test splitting; five-fold GroupKFold sensitivity analysis; leave-one-country-out evaluation; recruiter-level cluster bootstrap confidence intervals; and gender–nationality–care-obligation intersectional auditing.

7 Results and Discussion

7.1 Predictive model comparison

Table 4 reports the predictive results.

Table 4. Predictive performance on the regenerated CDEI recruiting benchmark.

Model	Val. AUC	Test AUC	PR–AUC	Accuracy	Brier	ECE
Logistic regression	0.9339	0.9405	0.9107	0.8735	0.0926	0.0135
XGBoost	0.9312	0.9369	0.9030	0.8700	0.0957	0.0173
CatBoost	0.9305	0.9368	0.9036	0.8710	0.0958	0.0261
Histogram Gradient Boosting	0.9280	0.9366	0.9026	0.8655	0.0958	0.0197
Random Forest	0.9250	0.9305	0.8940	0.8525	0.1052	0.0631

Logistic regression achieved the highest validation ROC–AUC (0.9339) and retained the highest test ROC–AUC (0.9405). Its test accuracy was 87.35%, with PR–AUC 0.9107 and ECE 0.0135.

The strong performance of the linear model is consistent with the structure of the CDEI data-generating mechanism, in which the employment probability is derived from an additive logistic score. Logistic regression was therefore retained as the predictive backbone instead of selecting a more complex model solely on model capacity.

7.2 Fairness–utility comparison

After the intersectionally reweighted configuration had been selected and locked using the validation procedure described in Section 6, the held-out test set was opened for final evaluation. Table 5 reports the test performance of the evaluated configurations for comparison; these test results were not used to revise the previously selected FairHR-X model.

Table 5. Utility and fairness results on the locked CDEI test set.

Configuration	Acc.	AUC	Race DPD	Race EOD	Sex DPD	Sex EOD	Int. DPD	Int. EOD
Protected-attribute baseline	0.8730	0.9433	0.3016	0.1149	0.1824	0.0938	0.4959	0.2400
Unawareness	0.8735	0.9405	0.2626	0.0901	0.1634	0.0593	0.4365	0.1469
Race reweighing	0.8690	0.9386	0.2257	0.0654	0.1583	0.0749	0.3934	0.1035
Intersectional reweighing	0.8645	0.9373	0.2126	0.0422	0.1194	0.0214	0.3402	0.0971
DP thresholding	0.8500	0.9405	0.1178	0.1348	0.1742	0.1039	0.2992	0.2194
EO thresholding	0.8580	0.9405	0.1624	0.0716	0.1636	0.0950	0.3340	0.1518

Fairness through unawareness reduced race DPD relative to the protected-attribute baseline, from 0.3016 to 0.2626, but substantial disparity remained. This empirically supports Proposition 3.1.

Race reweighing further reduced race DPD to 0.2257 and race EOD to 0.0654. Consistent with the validation-based selection rule, intersectional reweighing provided the selected balanced operating point across both protected dimensions on the held-out test set. Compared with unawareness, race DPD fell from 0.2626 to 0.2126, corresponding to a 19.0% relative reduction. Race EOD fell from 0.0901 to 0.0422, a 53.2% relative reduction.

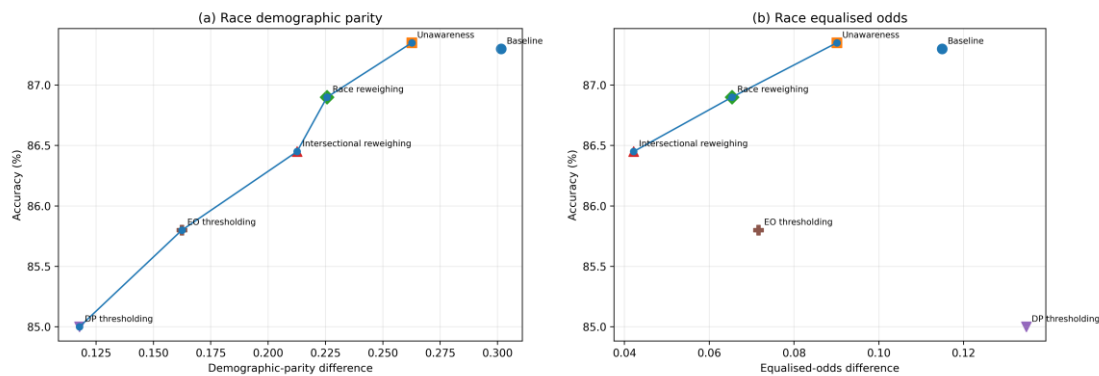


Fig. 2. Empirical fairness–utility Pareto analysis for the regenerated CDEI recruiting benchmark. Panel (a) evaluates race demographic-parity difference; Panel (b) evaluates race equalised-odds difference. Source: Author's calculations.

Sex DPD decreased from 0.1634 to 0.1194, corresponding to a 26.9% reduction, and sex EOD decreased from 0.0593 to 0.0214, corresponding to a 63.9% reduction. The intersectional DPD decreased from 0.4365 to 0.3402, a 22.1% reduction, while intersectional EOD decreased from 0.1469 to 0.0971, a 33.9% reduction. These reductions were obtained with an accuracy change from 87.35% to 86.45%, or 0.90 percentage points.

Post-processing optimised specifically for demographic parity produced a lower race DPD of 0.1178, but this improvement was accompanied by lower accuracy and a larger race equalised-odds gap. This result demonstrates why selection based on one fairness metric alone can produce undesirable trade-offs.

7.3 Bootstrap uncertainty

Table 6 reports 95% bootstrap confidence intervals for the selected FairHR-X configuration.

Table 6. Selected FairHR-X model: test estimates and 95% bootstrap confidence intervals.

Metric	Estimate	95% CI lower	95% CI upper
Accuracy	0.8645	0.8495	0.8785
Balanced accuracy	0.8533	0.8374	0.8689
F1	0.8170	0.7956	0.8378
ROC-AUC	0.9373	0.9270	0.9469
PR-AUC	0.9059	0.8886	0.9222
Brier score	0.0956	0.0879	0.1035
ECE	0.0136	0.0133	0.0334
Race DPD	0.2126	0.1745	0.2548
Race EOD	0.0422	0.0169	0.0817
Sex DPD	0.1194	0.0786	0.1629
Sex EOD	0.0214	0.0047	0.0806
Intersectional DPD	0.3402	0.2884	0.3944
Intersectional EOD	0.0971	0.0521	0.2153

The paired bootstrap comparison between intersectional reweighing and unawareness produced an accuracy difference of -0.0091 with a 95% interval of $[-0.0175, 0.0000]$. For race DPD, the paired difference was -0.0504 (95% CI $[-0.0688, -0.0330]$), and for race EOD it was -0.0435 (95% CI $[-0.0697, -0.0033]$). Sex DPD improved by -0.0440 (95% CI $[-0.0620, -0.0266]$), while the intersectional DPD difference was -0.0968 (95% CI $[-0.1254, -0.0698]$).

The confidence intervals therefore support systematic reductions in race demographic parity, race equalised odds, sex demographic parity, and intersectional demographic parity while the accuracy interval reaches zero at its upper boundary.

7.4 Comparison with published CDEI reference values

The independently regenerated experiment is complemented by the previously reported CDEI reference results in Table 7. These values are retained as external reference points rather than mixed with the results of the present run.

Table 7. Published CDEI recruiting reference values [21, 22].

Configuration	Criterion	Accuracy (%)	Reported gap
Unconstrained baseline	DPD	86.2	0.327
Fairness through unawareness	DPD	85.4	0.267
Reweighting	DPD	84.3	0.190
Reject-option post-processing	DPD	79.4	0.055
Reductions approach	DPD	80.0	0.041
Adversarial-specific baseline	DPD	86.2	0.173
Adversarial debiasing	DPD	84.5	0.007
Reject-option post-processing	EOD	84.1	0.032
Reject-option post-processing	EOPD	84.1	0.010

The ordering of the trade-offs in the independent experiment is compatible with the published CDEI evidence: stronger parity-oriented post-processing can reduce demographic disparity substantially, whereas multi-objective mitigation is needed when accuracy, error-rate fairness, and multiple protected characteristics are considered simultaneously.

8 Explainability and Intersectional Audit

8.1 Global SHAP analysis

Figure 3 reports global SHAP contributions for the selected intersectionally reweighted logistic model together with protected-feature association screening.

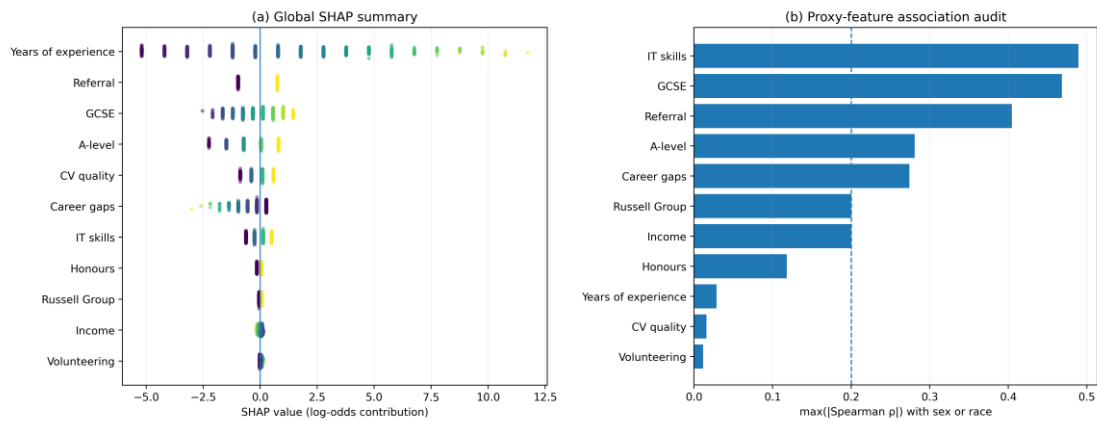


Fig. 3. Global SHAP explanation and proxy-feature association audit for the selected FairHR-X model on the held-out CDEI test set. Source: Author's calculations.

The largest mean absolute SHAP contribution was associated with years of experience (2.304), followed by referral (0.851), GCSE score (0.583), A-level score (0.569), CV quality (0.323), career gaps (0.313), and IT skills (0.307). Years of experience was highly predictive but showed little association with protected attributes, with maximum absolute Spearman correlation of 0.029.

By contrast, several predictors were identified for secondary proxy review. The maximum protected-attribute correlations were 0.489 for IT skills, 0.468 for GCSE score, 0.405 for referral, 0.281 for A-level score, and 0.274 for career gaps. Russell Group attendance and income were both close to 0.20.

These findings illustrate why feature importance and proxy analysis should be reported separately. A feature may be important without strongly encoding a protected characteristic, while another variable may have moderate predictive importance but a substantial protected-group association.

8.2 Explanation stability

Global SHAP rankings were evaluated across 1,000 bootstrap resamples of the held-out explanation set. The mean Spearman correlation between the original and bootstrap feature-importance rankings was

$$S_\rho = 0.996, \quad (26)$$

with a 95% bootstrap interval of

$$[0.982, 1.000]. \quad (27)$$

For the three highest-ranked features, the mean Jaccard overlap was

$$S_J = 0.940. \quad (28)$$

The explanation hierarchy was therefore highly stable under resampling.

8.3 Intersectional subgroup analysis

Figure 4 reports the sex–race intersectional audit of the selected FairHR-X configuration.

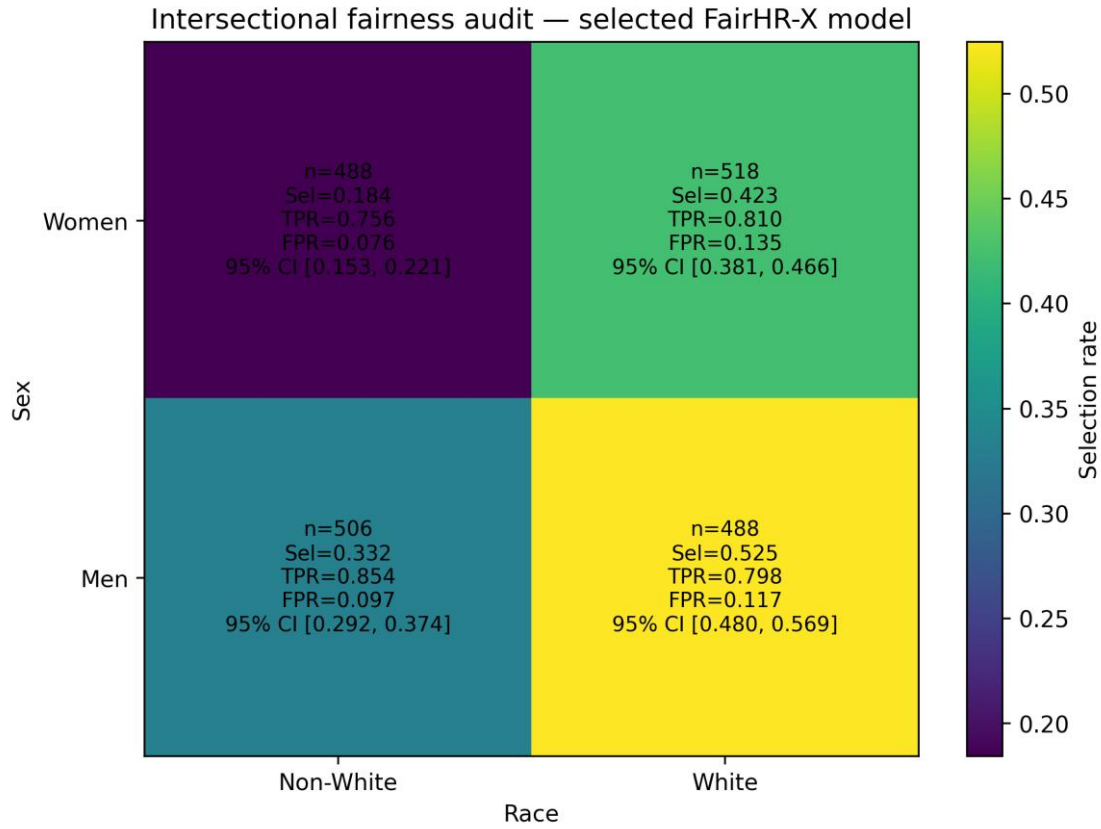


Fig. 4. Intersectional subgroup audit for the selected FairHR-X model. Each cell reports subgroup size, selection rate, true-positive rate, false-positive rate, and the 95% Wilson confidence interval for the selection rate. Source: Author's calculations.

The held-out test sample contained 488 non-White women, 518 White women, 506 non-White men, and 488 White men. Selection rates were 0.184 for non-White women, 0.423 for White women, 0.332 for non-White men, and 0.525 for White men. The corresponding true-positive rates were 0.756, 0.810, 0.854, and 0.798, whereas false-positive rates were 0.076, 0.135, 0.097, and 0.117, respectively.

The resulting intersectional DPD was 0.340 and intersectional EOD was 0.097. Although mitigation reduced both values relative to unawareness, the residual selection-rate differences confirm that fairness improvement should not be interpreted as complete parity.

9 Continuous Monitoring and HRMLOps

A model that satisfies validation criteria can degrade after deployment when candidate populations, recruitment channels, job requirements, or data pipelines change. Production machine-learning systems are also exposed to versioning, hidden dependencies, delayed labels, and feedback loops [30].

FairHR-X therefore stores a monitoring baseline for features, predictions, calibration, explanations, and subgroup metrics. For a binned feature, Population Stability Index is

$$PSI = \sum_{b=1}^B (q_b - p_b) \ln \left(\frac{q_b}{p_b} \right), \quad (29)$$

where p_b and q_b are the reference and current proportions. PSI is treated as a descriptive drift indicator rather than a significance test. It is complemented by distributional diagnostics such as the Kolmogorov–Smirnov statistic for continuous variables.

Fairness drift for metric M is

$$D_M(t) = |M_t - M_{ref}|. \quad (30)$$

To reduce one-window false alarms, the governance policy uses persistence: a non-critical drift or fairness breach must occur in two consecutive monitoring windows before model-risk review is automatically opened. Critical schema, security, privacy, or prohibited-feature violations trigger immediate review.

Every production prediction record contains the model version, feature-policy version, timestamp, calibrated probability, threshold policy, and explanation version. Retraining requires renewed label validation, fairness evaluation, calibration testing, and model approval rather than being triggered automatically by drift.

10 Deployment Governance

Table 8 defines the FairHR-X deployment gate.

Table 8. FairHR-X deployment-gate requirements.

Control dimension	Requirement	Evidence
Predictive validity	Performance must satisfy the pre-registered minimum and remain statistically stable	Locked test evaluation, bootstrap intervals, subgroup error analysis
Fairness	Approved group and intersectional fairness limits must be satisfied	DPD, DIR, EOpD, EOD, confidence intervals, intersectional audit
Calibration	Probability estimates must remain calibrated overall and by monitored group	Brier score, ECE, reliability analysis
Explainability	Decision reasons must remain interpretable and sufficiently stable	SHAP diagnostics, bootstrap stability, proxy-feature audit
Operational quality	Pipeline must be reproducible, versioned, and rollback-ready	Data hash, model version, feature-policy version, monitoring logs
Human oversight	Affected recommendations must support authorised review	Review record, override reason, escalation and correction mechanism

FairHR-X does not define a universal numerical fairness threshold. Deployment criteria must be fixed before final evaluation and should reflect the decision context, error costs, sample size, applicable regulation, and organisational governance policy.

Algorithm 2. FairHR-X continuous monitoring and response

Require: Production window B_t , reference profile P_0 , deployed model f^*

Ensure: Decision $z_t \in \{\text{continue}, \text{review}, \text{rollback}\}$

29. Validate schema, missingness, ranges, and feature-policy compliance
30. Compute prediction volume, score distribution, and service health
31. Compute PSI, KS, and explanation-distribution drift
32. When labels are available, compute predictive performance and calibration
33. Compute DPD, DIR, EOpD, EOD, and supported intersectional metrics
34. **if** critical schema, privacy, security, or prohibited-feature violation **then**
35. **return** rollback
36. **else if** a monitored non-critical threshold fails in two consecutive windows **then**
37. Open human and model-risk review
38. **return** review
39. **else**
40. Store the monitoring report and continue supervised operation
41. **return** continue

42. **end if**

11 Limitations and Threats to Validity

Several limitations remain. First, the primary empirical dataset is synthetic. The CDEI benchmark is valuable because its protected attributes and biased outcome-generation process allow controlled fairness evaluation, but it does not capture the full complexity of observed recruitment.

Second, the generated target represents a synthetic historical hiring outcome rather than independently validated future employee performance. Fairness relative to this label therefore cannot establish fairness of real employment decisions.

Third, PATHS2INCLUDE provides a stronger organisational context but measures recruiter responses to experimentally constructed candidate profiles. Its external-validation protocol should preserve recruiter grouping and country structure.

Fourth, fairness metrics depend on the validity of the target. Equal opportunity, for example, is difficult to interpret when the observed positive label itself contains historical discrimination.

Fifth, intersectional fairness estimates become unstable for small subgroups. FairHR-X therefore uses minimum-support rules and confidence intervals, but larger real-world datasets may require hierarchical modelling or partial pooling.

Sixth, post-processing rules that depend on protected-group information require careful legal and governance review. The technical ability to reduce a fairness metric does not itself establish the legitimacy of a decision policy.

Seventh, statistical fairness criteria can conflict [7, 8]. A model with a low demographic-parity gap may have less favourable error-rate parity, as observed in the thresholding experiment.

Finally, SHAP measures predictive attribution rather than causality. Explanation quality must therefore be combined with domain validation, feature governance, fairness analysis, and human oversight.

12 Conclusion

This paper presented FairHR-X, an explainable and bias-aware artificial intelligence lifecycle framework for transparent and human-supervised employee recruitment and selection.

An independent experiment on 10,000 records regenerated from the published CDEI recruiting process demonstrated that predictive accuracy alone is insufficient for model selection. Logistic regression achieved the strongest predictive performance, with ROC–AUC 0.941 and accuracy 87.35%, but fairness-through-unawareness retained substantial race, sex, and intersectional disparities.

Intersectional reweighing combined with validation-only probability calibration was selected as the balanced operating point using the pre-specified validation-set selection rule and subsequently evaluated on the locked test set. It retained 86.45% accuracy and ROC–AUC 0.937 while reducing race DPD by 19.0%, race EOD by 53.2%, sex DPD by 26.9%, sex EOD by 63.9%, intersectional DPD by 22.1%, and intersectional EOD by 33.9% relative to fairness through unawareness.

Bootstrap analysis supported the principal reductions in race demographic parity, race equalised odds, sex demographic parity, and intersectional demographic parity. The explanation layer was also stable: global SHAP feature rankings achieved mean bootstrap Spearman correlation of 0.996 and top-three Jaccard overlap of 0.940.

The results support the central design principle of FairHR-X: recruitment AI should not be selected by predictive performance or one fairness metric in isolation. Predictive quality, multi-attribute fairness, calibration, explainability, uncertainty, human review, and post-deployment monitoring must be assessed jointly.

The PATHS2INCLUDE dataset provides a suitable next stage for cross-national external validation because it contains recruiter-level candidate evaluations, organisational context, and multiple protected characteristics across four European countries.

FairHR-X is intended as a decision-support and governance framework. Final employment decisions remain subject to accountable human review.

Data Availability

The CDEI recruiting data-generation resources, processed artifacts, and bias-mitigation notebooks are publicly available through the CDEIUK/bias-mitigation repository [23]. The PATHS2INCLUDE “Hiring Discrimination in the Organisational Context” dataset, questionnaire, and codebook are available through the Open Science Framework under DOI 10.17605/OSF.IO/74Q69 [24].

Ethics Statement

FairHR-X is intended for research, auditing, and human-supervised recruitment decision support. The framework should not autonomously reject candidates. Protected attributes used for fairness auditing require appropriate access control and purpose limitation. Statistical fairness and explainability metrics do not by themselves establish legal compliance.

Acknowledgements. The authors are grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

REFERENCES

- [1] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. 2020 Conf. Fairness, Accountability, and Transparency*, 2020, pp. 469–481.
- [2] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanities and Social Sciences Communications*, vol. 10, Art. no. 567, 2023.
- [3] E. Drage and K. Mackereth, "Does AI debias recruitment? Race, gender, and AI's elimination of difference," *Philosophy & Technology*, vol. 35, Art. no. 89, 2022.
- [4] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius, and A. J. Biega, "Fairness and bias in algorithmic hiring: A multidisciplinary survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, Art. no. 16, pp. 1–54, 2025, doi: 10.1145/3696457.
- [5] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3315–3323.
- [6] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, 2019.
- [7] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [8] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5680–5689.
- [9] L. Lippens, S. Vermeiren, and S. Baert, "The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments," *European Economic Review*, vol. 151, Art. no. 104315, 2023.
- [10] D. R. Galos and A. Coppock, "Gender composition predicts gender bias: A meta-reanalysis of hiring discrimination audit experiments," *Science Advances*, vol. 9, no. 18, Art. no. eade7979, 2023.
- [11] D. Buttler, C. Imdorf, V. Bjørnshagen, A. Cosciug, K. Lipowska, I. Magda, M. Palczyńska, M. Smoter, S. Ayllón, J. Bell, M. Krzakała, I. Precupetu, R. M. Radogna, J. Rogstad, R. Samuel, L. Tufa, O. Valls, and N. Wessel, "Hiring discrimination in the organisational context: A harmonized factorial survey dataset from four European countries (Germany, Norway, Poland, Romania)," *Scientific Data*, vol. 13, Art. no. 896, 2026, doi: 10.1038/s41597-026-07187-2.
- [12] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.
- [13] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proc. 35th Int. Conf. Machine Learning*, vol. 80, 2018, pp. 60–69.
- [14] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, 2018, pp. 335–340.
- [15] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

- [16] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, 2020.
- [17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
- [18] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2013, pp. 623–631.
- [19] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A unified framework for machine learning interpretability," arXiv:1909.09223, 2019.
- [20] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, and Transparency*, 2019, pp. 220–229.
- [21] Centre for Data Ethics and Innovation, "Machine learning bias mitigation: Recruiting data and interventions," UK Government, 2020. [Online]. Available: <https://cdeiuk.github.io/bias-mitigation/>. Accessed: Aug. 31, 2026.
- [22] Centre for Data Ethics and Innovation, *Review into Bias in Algorithmic Decision-Making*, UK Government, London, 2020.
- [23] Centre for Data Ethics and Innovation, "CDEIUK/bias-mitigation: Machine learning bias mitigation," GitHub repository, 2020. [Online]. Available: <https://github.com/CDEIUK/bias-mitigation>. Accessed: Aug. 31, 2026.
- [24] D. Buttler et al., "Data for the article 'Hiring Discrimination in the Organisational Context: A Harmonised Factorial Survey Dataset from Four European Countries'," *Open Science Framework*, 2026, doi: 10.17605/OSF.IO/74Q69.
- [25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 6638–6648.
- [29] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds., MIT Press, 1999, pp. 61–74.
- [30] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 2503–2511.