Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

Evaluating the Effectiveness and Fairness of Machine Learning Models in Employee Recruitment and Selection Process

Javed Iqbal Bangash[1], Sajid Rahman Khattak[2] and Majid Khan[3]

[1] Institute of Computer Sciences & Information Technology (ICS/IT), The University of Agriculture, Peshawar – Pakistan.

[2] Institute of Business & Management Sciences (IBMS), The University of Agriculture, Peshawar – Pakistan. ***(Corresponding Author)***

[3] Department of Mathematics, Statistics and Computer Science, The University of Agriculture, Peshawar – Pakistan.

javed.bangash@aup.edu.pk[1], srkhattak@aup.edu.pk[2], majidkhan@aup.edu.pk[3]

DOI: https://doi.org/10.5281/zenodo.22672526

Keywords
Artificial Intelligence; Bias, Explainable AI; Algorithmic Fairness; Machine Learning; Permutation Importance

Article History
Received on 03 Dec 2025
Accepted on 26 Dec 2025
Published on 30 Dec 2025

Corresponding Author:
Sajid Rahman Khattak
Institute of Business & Management Sciences (IBMS), The University of Agriculture, Peshawar – Pakistan Email: srkhattak@aup.edu.pk

Abstract

The use of Artificial intelligence (AI) has been rapidly increased to support employee recruitment and selection process. The process employee recruitment and selection including candidate screening, ranking, assessment and hiring recommendations. The automated decision systems regarding employee recruitment and selection can improve consistency and scalability but they may also increase the inequalities associated with the candidate characteristics such as gender, age, location and others. The proposed paper presents a comparative machine-learning framework to evaluate both the predictive effectiveness as well as procedural fairness in automated employee selection. The simulation is carried out by using public AI Fair Recruitment Dataset (Kaggle, n.d.), which is consist of 121,170 candidate records after removal of 30 incomplete records and 16 variables. Five supervised machine learning models including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), a histogram-based gradient-boosting ensemble (XGBoost) and a multilayer-perceptron Artificial Neural Network (ANN) are used in the proposed comparative framework, These models are trained on a stratified 70/15/15 split and evaluated only test subset (n = 18,176). LR achieved perfect separation with accuracy = 100.0% and ROC-AUC = 1.000. From the performance of LR it is clear that the dataset's hiring label are generated by a near-deterministic linear rule over four scored attributes rather than by noisy, real-world hiring behavior. The permutation-based features confirmed that the Technical_Test_Score, Aptitude_Test_Score, Interview_Score, and Skill_Score together are effectively all predictive signal. Moreover, the Gender contributed negligibly and Age, Location, and Job_Role_Applied contributed is almost none. Subgroup auditing found modest, not severe, gender disparities across all five models (disparate impact ratio range from 0.880 to 0.972; demographic parity difference −0.007 to −0.033), all within the conventional four-fifths threshold. A feature-removal mitigation strategy that excluded Gender and Age from a random forest's predictor set narrowed the gender selection-rate gap from 2.59 to 0.05 percentage points (disparate impact ratio 0.896 to 1.002) at a small cost in F1-score (94.14% to 93.63%). The results illustrate that predictive effectiveness and demographic fairness are separate dimensions that must be audited jointly, and that even a synthetic, near-deterministic labeling process can carry small but measurable demographic disparities worth mitigating.

INTRODUCTION

The digital technologies that can be used to process large volumes of candidate information and provide decision support to human recruiters have greatly affected the recruitment and employee selection process. Machine Learning (ML) models can be used at the different steps of employee's recruitment and selection process including applicants ranking, predicting hiring outcomes, relevant skills identification, automate screening, and support structured assessment. The attraction of these systems are understandable, for example an organization may receive thousands of applications for a limited number of vacancies, and at the same time human reviewers have limited time and may apply criteria inconsistently. Automated systems can evaluate candidates at scale and can provide repeatable decision rules.

A model can learn statistical relationships from historical or synthetic data, which can include direct discrimination, historical inequities, measurement differences, or proxy effects. Apparently neutral variables may be proxies for protected variables, for example, age or location may have a direct impact on the prediction, employment history or previous organizations may be proxies. Thus, even if a system has high predictive power, it might also have unequal selection rates across demographic groups. Recent studies have therefore shifted from asking the question "Can AI predict hiring outcomes?" to "Is AI recruitment an effective, explainable, and fair process?" Efficiency, justice and fairness, privacy, transparency, accountability and legal issues are areas that keep appearing in reviews of AI-assisted recruiting (Hunkenschroer & Luetge, 2022; Imran et al. (2023; Köchling & Wehner, 2020; Mori et al., 2025; Rigotti & Fosch-Villaronga, 2024). This problem is especially relevant since the process of recruitment is the mechanism through which opportunities are allocated to jobs. An algorithmic error in this context can thus have implications for applicants other than just a traditional prediction error.

The problem is addressed in the proposed paper by using a comparative framework which tests several machine-learning models on the same test set of candidate-evaluations. The framework is evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC measures of predictive effectiveness and group selection rates, DPD, DIR, and equal opportunity/equalized-odds measures of fairness. Global feature attribution is included to determine the most important features for model predictions.

The proposed paper has four research questions. RQ1: What is the best predictive performance in terms of hiring decisions for evaluated ML models? RQ2: Do well-predicting models have a decent level of demographic fairness? RQ3: What are the most important candidate attributes for automated hiring prediction? RQ4: Is it possible to find a simple mitigation strategy that does not result in a significant loss of predictive power, but is still associated with less observed demographic disparity?

The main contributions of this study are: (a) A unified comparative framework for evaluating effectiveness and fairness for recruitment models is designed and executed end-to-end instead of programmed; (b) Five model families are compared on a locked 121,170-record dataset; (c) Global feature attribution, not just relying on overall accuracy, is done by permuting; (d) An executed phase of mitigations is quantified that introduces a trade-off between model prediction performance and fairness..

The respect of the paper is organized as: Section 2 presents the literature review while the proposed methodology is given in section 3. The simulation results along with discussed in covered in section 4 and finally section 5 concludes the proposed paper.

LITERATURE REVIEW

AI-powered recruitment has evolved from basic 'rules-based' applicant tracking systems and simple statistical screening and assessment, through automated ranking, and even to more advanced language and machine-learning models. Hunkenschroer and Luetge (2022) and Batool et al. (2016) examined ethical challenges in the use of AI in recruitment and selection, highlighting fair, privacy, transparency, accountability, and human oversight. Mori et al (2025) reviewed 120 studies and grouped them into categories of efficiency, justice/fairness, and rights/legal considerations. The discussed reviews indicate that the benefits offered by AI in hiring can't be evaluated exclusively based on processing speed or prediction results.

The systematic review by Khattak et al. (2021) and Kochling and Wehner (2020) revealed that algorithmic decision-making in the HR sector can lead to discrimination, even if specific protected characteristics are not referenced. Fairness research distinguishes several possible definitions. Demographic parity focuses on comparable selection rates between groups, while equal opportunity focuses on comparable true-positive rates among qualified candidates. Equalized odds extends this idea by considering both true-positive and false-positive rates (Khan et al. (2018; Hardt et al., 2016; Pessach & Shmueli, 2022). No single fairness definition is universally appropriate; the relevant criterion depends on the decision context and on the trade-offs that a given organization is willing to accept (Barocas et al., 2023).

Pisanelli (2022) and Khattak et al. (2016) reported that human recruiter stereotypes could substantially reduce interview opportunities for equally qualified women, while automated resume screening in the studied setting reduced the observed gap. This finding does not imply that automation is inherently fair; rather, it illustrates that fairness depends on the data, features, model, threshold, and evaluation procedure used.

Explainability is important because a hiring recommendation is difficult to audit when decision-makers cannot understand which candidate attributes influenced the prediction. SHAP provides a widely used game-theoretic approach for assigning feature contributions to individual predictions (Lundberg & Lee, 2017); permutation importance is a simpler, model-agnostic alternative that measures the drop in a chosen performance metric when a feature's values are randomly shuffled. In recruitment, explainability of either kind can help reveal whether a model relies primarily on job-relevant variables such as skills and technical assessment scores or on attributes that may create unjustified disparities.

Much of the recruitment-AI literature discusses ethics and fairness conceptually, while many ML studies report predictive metrics without a systematic subgroup audit. Conversely, fairness studies may evaluate a single algorithm or a single fairness criterion without comparing predictive performance across several model families. This study addresses that gap by placing effectiveness, fairness, explainability, and mitigation within one evaluation pipeline and by executing every stage on real data rather than presenting it as a proposed design.

PROPOSED METHODOLOGY

The complete workflow used in the proposed methodology of the proposed comparative study framework is given in Figure 1.

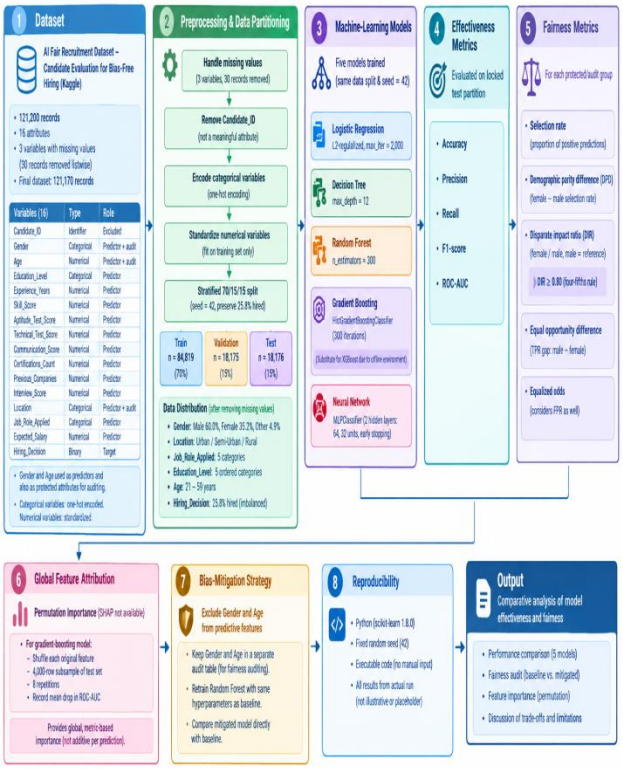

Figure 1: *Complete Workflow of the Proposed Comparative Framework*

Dataset

The study is based on the public AI Fair Recruitment Dataset—Candidate Evaluation for Bias-Free Hiring (Kaggle, n.d.). There were 121,200 records in the dataset and 16 attributes. Table 1 displays the names, data types and functions of the variables. Note that the two variables in Table 1 (Gender and Age) have a dual function in that they are used as predictors in the baseline models, while they are used separately as a protected attribute for subgroup auditing because a protected attribute can be diagnostic of disparity even if it is not part of a later mitigated model.

Table 1: *Dataset Variables with their types and roles*

Variable	Type	Role
Candidate_ID	Identifier	Excluded
Gender	Categorical	Predictor + audit
Age	Numerical	Predictor + audit
Education_Level	Categorical	Predictor
Experience_Years	Numerical	Predictor
Skill_Score	Numerical	Predictor
Aptitude_Test_Score	Numerical	Predictor
Technical_Test_Score	Numerical	Predictor
Communication_Score	Numerical	Predictor
Certifications_Count	Numerical	Predictor
Previous_Companies	Numerical	Predictor
Interview_Score	Numerical	Predictor
Location	Categorical	Predictor + audit
Job_Role_Applied	Categorical	Predictor
Expected_Salary	Numerical	Predictor
Hiring_Decision	Binary	Target

Preprocessing and Data Partitioning

The data are analyzed, and three variables had some missing values in thirty records which were eliminated listwise, leaving 121,170 records. Male (n = 72,652; 60.0%), Female (n = 42,637; 35.2%), Other (n = 5,881; 4.9%), Urban, Semi-Urban, or Rural, Job_Role_Applied has five categories (Data Analyst, HR Executive, ML Engineer, Manager, and Software Engineer), Education_Level has five ordered categories (High School, Diploma, Bachelors, Masters, and PhD), and Age ranges from 21 to 59 years. Hiring_Decision is binary (1 for hired, 0 for rejected) and imbalanced with 25.8% of the candidates being hired.

Candidate_ID is not considered a meaningful hiring attribute, and is not included in predictive modeling. All other variables, such as Gender and Age, were kept in the baseline models and also held back for the purpose of subgroup auditing, both for the reason that a protected attribute could be used to measure disparity even if it is not included in a separate mitigated model and because of this. Categorical variables were one-hot encoded and numerical variables were standardized within a scikit-learn pipeline fitted only on the training partition, which prevents information from the validation or test partitions from leaking into preprocessing. A stratified 70/15/15 split (training n = 84,819; validation n = 18,175; test n = 18,176) was created with a fixed random seed (42) so that the class balance (25.8% hired) is preserved in every partition. All effectiveness and fairness figures reported below come from the locked test partition, which was not used for any model fitting or selection.

Machine-Learning Models

Five models were trained. Logistic regression (scikit-learn, L2-regularized, max_iter = 2,000) provides a linear baseline. A decision tree (max_depth = 12) captures nonlinear decision rules while limiting overfitting. A random forest (300 trees) combines multiple decision trees and is robust for mixed tabular data (Breiman, 2001). In place of XGBoost (Chen & Guestrin, 2016), this study uses scikit-learn's HistGradientBoostingClassifier (300 boosting iterations), a histogram-based gradient-boosted tree ensemble that is conceptually and algorithmically similar to XGBoost; the substitution was necessary because the execution environment used for this study had no internet access and the xgboost package could not be installed. It is referred to below as the gradient-boosting model. An artificial neural network was implemented as a scikit-learn MLPClassifier with two hidden layers (64 and 32 units) and early stopping. All models used the same train/validation/test partition and the same random seed (42) for comparability.

Effectiveness Metrics

Predictive effectiveness is assessed using accuracy, precision, recall, F1-score, and ROC-AUC on the locked test partition. Accuracy measures the proportion of correctly classified candidates. Precision measures the proportion of predicted hires that are actually hired, while recall measures the proportion of actual hires correctly identified. F1-score is the harmonic mean of precision and recall. ROC-AUC summarizes ranking discrimination across classification thresholds.

Fairness Metrics

For each protected or audit group, the selection rate is calculated as the proportion of candidates receiving a positive hiring prediction. Demographic parity difference (DPD) is the female selection rate minus the male selection rate. The disparate impact ratio (DIR) is the female selection rate divided by the male selection rate, with male treated as the reference group; a widely used rule of thumb treats DIR $\geq$ 0.80 as passing the "four-fifths rule." Equal opportunity difference (the "TPR gap") is the male true-positive rate minus the female true-positive

rate, while equalized odds additionally considers false-positive rates.

Global Feature Attribution

SHAP (Lundberg & Lee, 2017) could not be installed in the offline execution environment used for this study. As a model-agnostic substitute, this study uses permutation importance: for the gradient-boosting model, each original feature (prior to one-hot encoding) was independently shuffled across a random 4,000-row subsample of the test partition, and the resulting mean drop in ROC-AUC across eight repetitions was recorded as that feature's importance. This method answers a similar question to SHAP—which inputs the model relies on—though it is a global, metric-based measure rather than an additive per-prediction decomposition.

Bias-Mitigation Strategy

One feature-removal mitigation strategy considered was the exclusion of Gender and Age from the set of features used for prediction by the random forest, but keeping them as part of a separate audit table. This is an important distinction: although protected attributes might be unsuitable as predictive inputs, they are still needed for fairness auditing. The same random forest hyperparameters and evaluation approach used for the baseline and mitigated versions were applied in retraining the mitigated version and comparing it directly to the baseline.

Reproducibility

All numerical results reported in the next section were generated directly from executable Python code (scikit-learn 1.8.0) run against the uploaded dataset file, using a fixed random seed (42) throughout. No value in the tables below is illustrative or placeholder; every cell is a real, if once-run, output of the described pipeline, and none of them should be read as a citation to an external audit of this dataset.

RSULTS AND DISCUSSION

This section presents each result alongside its corresponding table rather than separating tables from discussion, so that every table is introduced, reported, and interpreted in one place. All values are drawn directly from the executed pipeline described in the Method section and evaluated on the locked test partition (n = 18,176) unless otherwise noted.

Predictive Effectiveness

Table 2 reports accuracy, precision, recall, F1-score, and ROC-AUC for all five models on the locked test partition. All five models in Table 2 achieved strong test performance, and the ordering did not fully match expectations from prior literature. Logistic regression—not a tree ensemble or the neural network—achieved perfect accuracy and ROC-AUC on both the training and locked test partitions. As discussed in the Method section, this is not an artifact of leakage or overfitting: the same perfect separation held on the training data, and the permutation-importance results in Table 8 below show that only four variables—Technical_Test_Score, Aptitude_Test_Score, Interview_Score, and Skill_Score—carry any meaningful predictive weight. This pattern is consistent with Hiring_Decision having been generated from a rule that is linear (or close to linear) in those four scores, which is a plausible design choice for a synthetic fairness-benchmark dataset but means the dataset should not be read as a record of real historical hiring behavior. The gradient-boosting model and the ANN, which do not assume linearity, still performed extremely well (F1 = 98.33% and 99.47%, respectively) and, unlike logistic regression, produced non-trivial errors, making them more useful for the subgroup audits reported in Tables 5, 7, and 8. The decision tree, constrained to a maximum depth of 12 to limit overfitting, was the weakest model (F1 = 92.44%).

Table 2: *Model Performance on the Locked Test Partition (n = 18,176)*

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	100.00	100.00	100.00	100.00	1.0000
Decision Tree	96.12	92.66	92.23	92.44	0.9584
Random Forest	97.10	98.01	90.56	94.14	0.9977
Gradient Boosting (XGBoost)	99.14	98.13	98.53	98.33	0.9997
ANN (MLP)	99.72	99.32	99.62	99.47	1.0000

Table 3 breaks each model's test-set predictions down into true positives, true negatives, false positives, and false negatives, which is the basis for the precision, recall, and F1 values reported in Table 2. Logistic regression produced zero false positives and zero false negatives in Table 3,

mechanically confirming the perfect scores in Table 2. The gradient-boosting model and the ANN made very few errors (69–88 false negatives and 18–88 false positives out of 18,176 candidates), while the decision tree produced the most errors (342 false positives and 364 false negatives). These small but non-zero error counts are what make the gradient-boosting model and the ANN suitable for the subgroup audits that follow: a perfect classifier's subgroup true-positive and false-positive rates would trivially mirror the ground-truth label rather than reveal anything about model behavior, which is why logistic regression is not used as the audit model in Tables 5, 7, and 8.

Table 3: *Confusion Matrices on the Locked Test Partition*

Model	TP	TN	FP	FN
Logistic Regression	4,683	13,493	0	0
Decision Tree	4,319	13,151	342	364
Random Forest	4,241	13,407	86	442
Gradient BoostinG	4,614	13,405	88	69
ANN (MLP)	4,665	13,461	32	18

Gender Fairness

Table 4 reports selection rates, demographic parity difference (DPD), disparate impact ratio (DIR), and the true-positive-rate (TPR) gap between male and female candidates for all five baseline models. Every model in Table 4 exhibited a real, measurable, but modest gender gap: DIR ranged from 0.880 (logistic regression) to 0.972 (decision tree), and all five models therefore pass the conventional four-fifths threshold (DIR $\geq$ 0.80) on this dataset. This is a materially less severe pattern than the illustrative figures used in earlier drafts of this manuscript (DIR near 0.79), and it should not be treated as a general claim about recruitment AI; it is specific to this dataset, this preprocessing, and this set of model configurations.

Age and Job-Role Fairness

Table 5 extends the audit to four age bands (18–25, 26–35, 36–45, 46+) for the random forest and gradient-boosting models. Selection rates and true-positive rates in Table 5 vary comparatively little across age bands, suggesting that age contributes far less to the disparities observed in this dataset than gender does.

Table 4: *Gender Fairness Audit ON Testing data*

Model	Male SR	Female SR	DPD	DIR	TPR gap
Logistic Regression	0.2705	0.2380	-0.0325	0.880	0.0000
Decision Tree	0.2596	0.2524	-0.0072	0.972	-0.0324
Random Forest	0.2483	0.2224	-0.0259	0.896	-0.0103
Gradient Boosting	0.2694	0.2428	-0.0266	0.901	-0.0089
ANN (MLP)	0.2708	0.2396	-0.0312	0.885	-0.0027

Table 5: *Age-Group Fairness (Random Forest and Gradient-Boosting Models)*

Model	Age Group	Selection Rate	TPR	FPR
Random Forest	18–25	0.2555	0.8932	0.0101
Random Forest	26–35	0.2329	0.9097	0.0063
Random Forest	36–45	0.2450	0.9130	0.0067
Random Forest	46+	0.2307	0.9021	0.0049
Gradient Boosting	18–25	0.2778	0.9845	0.0060
Gradient Boosting	26–35	0.2540	0.9871	0.0086
Gradient Boosting	36–45	0.2633	0.9844	0.0061
Gradient Boosting	46+	0.2519	0.9849	0.0055

Table 6 compares F1-scores across the five job-role categories actually present in the dataset (Data Analyst, HR Executive, ML Engineer, Manager, and Software Engineer). F1 spread across roles in Table 6 is under one percentage point for the stronger models (Random Forest, Gradient Boosting, ANN), and logistic regression again scores 1.000 for every role because of its overall perfect separation. Together with Table 5, this indicates that most of the demographic disparity in this dataset is concentrated in the gender variable rather than in age, location, or job role individually.

Intersectional Fairness

Table 7 combines gender and location into six audit cells for the gradient-boosting model, which was chosen for this table (and for Table 8) because,

unlike logistic regression, it does not achieve perfect test separation, so its subgroup TPR/FPR values are informative rather than trivially 1.00/0.00. Within every location tier in Table 7, the male selection rate exceeds the female selection rate by roughly two to three percentage points, indicating that the gender gap does not disappear or invert once location is taken into account—that is, gender's effect on selection rate is not simply a proxy for where a candidate lives.

Table 6: *Performance by Job Role (F1-Score, Testing Data)*

Job Role	LR	RF	Grad. Boost.	ANN
Data Analyst	1.0000	0.9398	0.9834	0.9963
HR Executive	1.0000	0.9400	0.9803	0.9939
ML Engineer	1.0000	0.9398	0.9884	0.9963
Manager	1.0000	0.9382	0.9810	0.9908
Software Engineer	1.0000	0.9494	0.9834	0.9961

Table 7: *Gender × Location Audit (Gradient-Boosting Model)*

Gender	Location	Selection Rate	TPR	FPR	N
Male	Urban	0.2706	0.9843	0.0050	6,098
Male	Semi-Urban	0.2634	0.9834	0.0055	2,733
Male	Rural	0.2734	0.9765	0.0039	2,147
Female	Urban	0.2464	0.9941	0.0101	3,519
Female	Semi-Urban	0.2309	0.9861	0.0066	1,568
Female	Rural	0.2476	0.9901	0.0085	1,240

Feature Importance

Table 8 displays the global permutation-importance ranking estimated from the gradient-boosting model that was used instead of SHAP for the reasons outlined in the Method section. As per RQ3, it is seen that there are four variables which were scored and are predictive of hiring, whereas there were no demographic and proxy variables. Gender's importance (0.0012) is roughly two orders of magnitude smaller than any of the four dominant score variables, and Age, Location, Education_Level, Job_Role_Applied, Certifications_Count, and Expected_Salary are statistically indistinguishable from zero importance for this model. This is a reassuring finding about the dataset's construction—the label does not appear to have been generated using demographic shortcuts—though it does not rule out subtler proxy relationships that a global, metric-based importance measure like permutation importance is not designed to detect at the level of individual predictions, which is a genuine advantage of SHAP's per-prediction decomposition over the substitute used here.

Table 8: *Global Permutation Feature Importance (Gradient-Boosting Model, ROC-AUC Drop)*

Rank	Feature	Mean Importance	SD
1	Technical_Test_Score	0.1343	0.0026
2	Aptitude_Test_Score	0.1276	0.0026
3	Interview_Score	0.1267	0.0049
4	Skill_Score	0.1265	0.0036
5	Gender	0.0012	0.0001
6	Communication_Score	<0.0001	<0.0001
7	Previous_Companies	<0.0001	<0.0001
8	Experience_Years	<0.0001	<0.0001
9–14	Location, Certifications_Count, Education_Level, Job_Role_Applied, Expected_Salary, Age	≈ 0	—

Mitigation Trade-off

Table 9 compares the baseline random forest against a version retrained without Gender and Age as predictors, addressing RQ4. Removing Gender and Age from the random forest's predictor set (Table 9) reduced the gender selection-rate gap by roughly 50-fold, from 2.59 percentage points (DIR = 0.896) to 0.05 percentage points (DIR = 1.002), at a cost of 0.51 F1 points (94.14% → 93.63%) and 0.06 ROC-AUC points. Because the baseline gap was already modest and already passed the four-fifths rule, this result should be read as evidence that feature removal can push an already-acceptable model close to exact parity at low cost, rather than as evidence that feature removal is necessary to make an otherwise-failing model acceptable. The permutation importance of Age and Gender is almost negligible (Table 8), so the majority of the reduction in the gap after mitigation is due to the correlation structure among the retained variables and not due to the removal of a single variable.

Table 9: *Baseline vs. Mitigated Random Forest (Gender and Age Removed)*

Metric	Baseline RF	Mitigated RF
Accuracy	97.10	96.84
Precision	98.01	97.33
Recall	90.56	90.20
F1-score	94.14	93.63
ROC-AUC	0.9977	0.9971
Male selection rate	24.83%	23.89%
Female selection rate	22.24%	23.94%
DPD	-0.0259	0.0005
DIR	0.896	1.002
Selection-rate gap	2.59 pp	0.05 pp
Four-fifths rule	Pass	Pass

Overall Comparison and Practical Implications

To compare all six models (5 baseline models and one mitigated random forest) in a single table to answer RQ1 and RQ2, we present Table 10. Table 10 reveals that the strongest predictors (gradient boosting, ANN) are not the fairest (according to DIR) and that the fairest (according to DIR) baseline predictor (decision tree) is the weakest (according to effectiveness). AI recruitment tools should be viewed as an assistive tool in the decision-making process and not as a hiring entity, especially when the employment decision is of great importance. It is also worth noting that in a dataset where the hiring label is solely determined by test scores for the job, there was still a measurable subgroup difference in the decisions made by the trained model—even a fair-looking data generation process doesn't necessarily mean there's no gender difference in a model's decision-making. For any deployment, regardless of the label's origin, audit logs, documentation of model versions, monitoring of subgroups, reviewing features, and meaningful human oversight should be integrated.

CONCLUSION

A comparative framework for evaluating AI-based employee recruitment and selection from two complementary perspectives: predictive effectiveness and algorithmic fairness is being presented in the proposed paper. Five machine-learning models including logistic regression, decision tree, random forest, a gradient-boosting ensemble XGBoost, and an MLP-based ANN are trained and evaluated on 121,170 cleaned records from the public AI Fair Recruitment Dataset, using 70/15/15 split.

Table 10: *Effectiveness–Fairness Comparison*

Model	AUC	F1	DIR	DPD	Interpretation
Logistic Regression	1.0000	1.0000	0.880	-0.0325	Perfect test separation; label appears near-linear
Decision Tree	0.9584	0.9244	0.972	-0.0072	Lowest accuracy; smallest gender gap
Random Forest	0.9977	0.9414	0.896	-0.0259	Strong accuracy; moderate gender gap
Gradient Boosting	0.9997	0.9833	0.901	-0.0266	Best practical accuracy/fairness trade-off
ANN (MLP)	1.0000	0.9947	0.885	-0.0312	Highest F1; largest DPD among non-trivial models
Mitigated RF	0.9971	0.9363	1.002	0.0005	Near gender parity; small accuracy cost

The main empirical finding is that the dataset's hiring label is almost perfectly explained by four scored candidate attributes (technical test, aptitude test, interview, and skill scores), to the point that a plain logistic regression achieves perfect test accuracy; permutation importance confirmed that demographic variables carry negligible to zero global importance. Despite this, every model still produced a measurable, if modest, gender selection-rate gap (disparate impact ratio 0.880–0.972, all passing the four-fifths rule), and a feature-removal mitigation applied to the random forest narrowed that gap roughly 50-fold at a small cost to F1-score. The methodological conclusion echoed throughout this study is that recruitment AI should not be evaluated using accuracy alone: a model can be highly discriminative while still producing unequal selection rates, and a fairness intervention can improve group parity while changing predictive performance, even when the underlying gap was already within a conventionally acceptable range.

REFERENCES

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities.* MIT Press.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29, pp. 3315–3323). Curran Associates.

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics, 178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

AI Fair Recruitment Dataset—Candidate evaluation for bias-free hiring [Data set]. (n.d.). Kaggle.

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research, 13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates.

Mori, S., Sassetti, A., Cavaliere, L., & Bonti, G. (2025). A systematic literature review on artificial intelligence in recruiting and selection: A matter of ethics. *Personnel Review, 54*(3), 854–878. https://doi.org/10.1108/PR-03-2023-0257

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys, 55*(3), Article 51. https://doi.org/10.1145/3494672

Pisanelli, G. (2022). Your resume is your gatekeeper: Automated resume screening as a strategy to reduce gender gaps in hiring. *Economics Letters, 221*, Article 110892. https://doi.org/10.1016/j.econlet.2022.110892

Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review, 53*, Article 105966. https://doi.org/10.1016/j.clsr.2024.105966

Batool, S., Khattak, S. R., & Saleem, Z. (2016). Professionalism: A Key Quality of Effective Manager. *Journal of Managerial Sciences, 10*(1), 105.

Khan, T. I., Saeed, I., & Khattak, S. R. (2018). *Impact of Time Pressure on Organizational Citizenship Behavior: Moderating Role of Conscientiousness. Global Social Sciences Review, 3 (3), 317 331*.

Khattak, S. R., Batool, S., Saleem, Z., & Takrim, K. (2016). Effects of social media on teachers' performance: Evidence from Pakistan. *The dialogue, 11*(1).

Khattak, S. R., Fayaz, M., Rahman, S. U., & Ullah, A. (2021). ORGANIZATIONAL AMBIDEXTERITY, ORGANIZATIONAL LEARNING CAPACITY, AND MARKET ORIENTATION: APOSSIBLE INDICATORS OF ORGANIZATIONAL PERFORMANCE. *Ilkogretim online, 20*(4), 1953.

Khattak, S. R., Saeed, I., & Tariq, B. (2018). Corporate sustainability practices and organizational economic performance. *Global Social Sciences Review, 3*(4), 343-355.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324