Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

An AI-Driven Predictive Model Using Decision Trees for Detecting Gender Bias in HR Practices and Recommending Policy Interventions

[1]Shilpy Kapoor, [2]Dr. Bhakti Ranjit Pawar, [3]Dr. Parul Gandhi

[1]School Of Leadership and Management, MRIIRS, Faridabad, India
(corresponding author)

[2]School Of Leadership and Management, MRIIRS, Faridabad, India

[3]School Of Leadership and Management, MRIIRS, Faridabad, India

Abstract

Despite the prevalent legislative frameworks and diversity efforts, gender prejudice in organisational Human Resource (HR) practices remains a quantifiable and impactful type of inequality in the workplace. This article suggests an AI-based predictive system based on Decision Tree classifiers to actively identify gender bias in five HR practice areas: pay equity, promotion, access to training, employee engagement, and attrition. A new two-step modelling model is proposed: Stage 1 (Reverse Modelling): with the help of HR outcome variables as predictors, the gender of the employees can be predicted which statistically demonstrates the presence of systemic bias; Stage 2 (Forward Modelling): the gender and the HR features are used as predictors to predict an artificial bias label, which indicates the circumstances under which bias appears. This framework is trained on HR_Analytics.csv (1,480 records) and tested on three independent datasets of 22,119 records of employees overall. The experimental findings show reverse model AUC of up to 62.4, which proves that there is a significant gender cue in HR results. Forward modelling has 100 percent accuracy and F1 score on all the dataset. Among the key results are a pay gap of -7.1% in Sales departments established in two independent datasets, deficits of 2.04-2.18 sessions per year of training in Finance and Marketing, and delays in promotions of 0.51-0.57 years in case of female employees in HR and Technology functions. The output of decision trees is converted into specific HR intervention, which is automatically generated into a rule-based policy recommendation engine. Comparative analysis with six other algorithms shows that the Decision Tree is the only algorithm to provide state of the art predictive accuracy and fully interpretable results as well as auto-generated policies, which are not available with black-box ensemble techniques.

Index Terms — Gender bias, HR analytics, Decision tree, Machine learning, Pay equity, Explainable AI, Fairness in AI, People analytics, Policy recommendation, Workplace gender equality.

I. Introduction

The workplace gender inequality is one of the most empirically observed and costly organisational dysfunctions in terms of economic cost. The Global Gender Gap Report of the World Economic Forum [1] states that at the current pace of improvements, the global gender economic participation gap will be more than 131 years, this is a pointer to the fact that interventions and monitoring measures currently in place are not sufficient. In organisations, gender discrimination is experienced across a full spectrum of Human Resource (HR) management practices which include compensation determination, making of promotions, training allocation, performance evaluation and employee engagement practices.

In the past, gender bias in HR has been identified based on retrospective statistical audits, salary surveys and investigations based on complaints. There are three basic limitations of these approaches. To begin with, they are non-predictive but reactive and detect discrepancies when a lot of damage has been caused. Second, they usually study a single domain of HR practices separately, which clouds the compounding impacts of biases across more than one practice that an individual employee may be subject to at a time. Third, they produce descriptive results that do not have policy suggestions, leaving HR managers with no practical advice on how to rectify it. The emergence of digital HR information systems and the development of machine learning (ML) methods of People

Analytics take an unprecedented opportunity to address these constraints by proactively detecting bias in data. The native interpretability of Decision Tree classifiers makes them especially suitable to this question area since, unlike ensemble or neural network models, Decision Trees generate well-understood IF-THEN classification rules that can be read and validated by an HR practitioner and can be directly translated into policy actions with no need for data science skills. The contributions to the literature on algorithmic fairness in HR analytics that the paper will make are as follows:

1) An innovative two-stage modelling approach to combine the Reverse Modelling (confirming that a bias exists by predicting gender using the HR outcomes) and Forward Modelling (understanding the reason behind the bias by predicting bias label), which is the first use of such a two-stage method to detect multi-practice HR bias.

2) A multi-practice bias detection framework that works in parallel to analyse five HR domains in real time using engineered bias indicators, allowing the compounding bias effects that are unobservable in single-practice frameworks to be detected.

3) Cross-dataset validation in four separate organisational datasets with 22,119 records of employees, determining the external validity of identified bias patterns in any one organisation. An automated policy recommendation engine that maps decision tree classification rules directly to specific, legally grounded HR policy interventions — bridging the gap between algorithmic output and managerial action.

This paper is further divided into the following sections. Section II discusses related work. The methodology of the research is presented in Section III. Experimental results are shown in Section IV. The comparison of models is in section V. Findings and policy implications are discussed in section VI. Section VII concludes.

II. Related Work

A. Gender Bias in Organisational HR Practices

Natural experiment by Goldin and Rouse [2] revealed that women were more likely to be promoted in case of introduction of blind auditions in symphony orchestras by 25-46 percent, which proves that there is early evidence of gender discrimination in performance appraisal that has direct implications in corporate promotion practices. In a detailed longitudinal study, Blau and Kahn [3] discovered that the gender pay gap, though having been reduced to about 18% in 2010 (compared to about 40% in 1980) has a significant and unexplainable residual beyond the control of occupation, industry, experience, and education and is due to discrimination and structural inequity. Catalyst [4] reported that women employees are given fewer high-visibility roles and leadership development opportunities as compared to male counterparts in the same organisational positions, otherwise known as the broken rung of a corporate ladder. McKinsey and LeanIn.This finding was further supported by Org [5] who annually released their Women in the Workplace report and found the lack of access to training repeatedly to be one of the main contributors to female underrepresentation in senior leadership in industries across the world.B. Machine Learning Solutions to Bias Detection.

B. Machine Learning Approaches to Bias Detection

The theoretical bases of algorithmic fairness were formalised by Barocas and Moritz [6], who proposed such important metrics as demographic parity, equalised odds, and individual fairness that are used to assess fair ML systems. Raghavan et al. [7] studied tools based on algorithmic hiring used by large technological firms and discovered that models that screened resumes using past hiring data in technical jobs systematically disfavoured women in technical jobs - the dual potential of ML as both a source and detector of bias. Feldman et al. [8] suggested the disparate impact testing of ML pipelines which set the 80-percent rule as a quantitative level of identifying outcomes that are discriminatory. In this study, Dwork et al. [9] made individual fairness formal, which motivates the reverse modelling strategy of this study: when HR results are predictive of gender beyond chance, then it would be a violation of individual fairness by definition, which is evidence of systemic bias. Verma and Rubin [10] have performed a systematic review of 20 definitions of fairness in the literature of ML, illustrating

incomparability with each other in practice, a result that guides the multi-metric evaluation protocol used in this paper.

C. Decision Trees in HR Analytics

Quinlan [11] introduced the theoretical foundation of tree-based classification and achieved similar performance with more complicated models and maintains the ability to interpret the rule structure by humans. Teng et al. [12] used Decision Trees to predict the employee attrition on IBM HR data, and they reported that job satisfaction, monthly income, and years since the last promotion were the most significant predictors of employee attrition: these characteristics are not only directly correlated with the bias indicators that have been designed in this study. Ribeiro et al. [13] proposed local interpretable model-agnostic explanations (LIME) and assumed that individual-level rule explanations are a pre-condition to responsible HR decision support.

D. Research Gap

The above discussion indicates that there are three key gaps, which this study fills. To begin with, the current body of literature on ML research focuses on gender bias in one area of HR practices; this study evaluates five at the same time. Second, previous research uses black-box models (Random Forest, Gradient Boosting, neural networks) that do not directly translate policies; Decision Trees are directly actionable rule outputs. Third, there is no other study that uses a two-stage reverse-then-forward modelling to prove and explain simultaneous bias - the main methodological addition of this paper.

III. Methodology

A. Datasets

The research uses four HR datasets that are publicly available. The key training data set used (HR_Analytics.csv) contains 1,480 records of employees and includes 38 variables and is fully aligned to all 11 key IBM HR column requirements, including all five HR practice domains needed by the research framework. Generalisability of the model is tested using three validation datasets that have different sizes and organisational context: general_data.csv (4,410 records), employee_attrition_dataset_10000.csv (10,000 records) and a merged Employee-PerformanceRating dataset (6,229 records after gender selection). The entire analytical corpus includes 22,119 records of employees in four autonomous organisational situations.

Table I: Dataset Specifications

HR_Analytics.csv	Base (train)	1,480	All 11 IBM cols including JobSatisfaction, WorkLifeBalance
general_data.csv	Validation 1	4,410	Pay, promotion, training, attrition; engagement proxied
attrition_10k.csv	Validation 2	10,000	All practices; column renaming applied
Employee + Perf	Validation 3	6,229	Pay, promotion, engagement after merge on EmployeeID

B. Preprocessing Pipeline

A standardised preprocessing protocol was applied identically across all four datasets to ensure cross-dataset comparability. Zero-variance columns (EmployeeCount, Over18, StandardHours) were removed. Missing values were imputed using column-wise median for numeric variables and mode for categorical variables. All categorical

variables were label-encoded using scikit-learn's LabelEncoder. The binary gender variable was encoded as Female = 1, Male = 0. Class imbalance in the BiasLabel target was addressed through the balanced class_weight parameter in the Decision Tree classifier rather than synthetic oversampling, thereby preserving the original data distribution and avoiding information leakage.

C. Feature Engineering — Bias Indicators

A set of five quantitative bias indicators was operationalised based on raw HR features to gauge gender bias as quantifiable ML features. All indicators are binary variables (0 or 1), based on comparisons with role-level or department-level baselines:

BelowMedianPay: 1 when the MonthlyIncome of an employee fall below the median of MonthlyIncome of employees in his/her particular JobRole ,JobsLevel combination, otherwise 0. This measure reflects on pay disparity, excluding role and seniority variation.

LongPromoWait: 1 when Years Since Last Promotion is greater than the average of the department by over one year, 0 otherwise. This is an embodiment of promotion delays compared to departmental counterparts.

LowTraining: 1 when TrainingTimesLastYear is less than the average of the department, otherwise 0. This sums up unequal access to training in the same organisational unit.

LowHike: 1 when PercentSalaryHike is less than the median of the JobRole of the employee, otherwise 0. This reflects equity of increment of salary net of role.

LowEngagement: 1 when JobSatisfaction is equal or less 2 on the 1-4 scale, 0 otherwise.

This embodies the disengagement as a possible result of generalizing HR practice. The composite BiasScore is the sum of all the five indicators (range 0-5). The BiasLabel is a binary variable that takes the value 1 when BiasScore is at least 2, i.e. the employees who have been subjected to bias in more than two of five areas at the same time, and 0 otherwise.

D. Two-Stage Modelling Approach

This research introduces a novel two-stage modelling framework designed to first prove and then explain gender bias in HR practices. Stage 1 — Reverse Modelling: a Decision Tree is trained with Gender as the target variable and HR outcome features (MonthlyIncome, PercentSalaryHike, TrainingTimesLastYear, YearsSinceLastPromotion, JobLevel, JobSatisfaction, WorkLifeBalance, and Attrition) as predictors. The theoretical rationale is grounded in individual fairness theory [9]: if HR outcomes can predict gender at accuracy significantly above the 50% random baseline, those outcomes are systematically differentiated by gender, constituting mathematical proof of systemic bias. This stage answers the research question: does bias exist? Stage 2 — Forward Modelling: a Decision Tree is trained with BiasLabel as the target variable and Gender plus all HR features plus the five engineered bias indicators as predictors. This stage produces interpretable IF-THEN rules identifying the specific combinations of HR practice conditions that flag an employee as experiencing bias, answering: where and how does bias manifest?

E. Model Configuration

Both of the Decision Tree models have the hyperparameters Gini impurity criterion, max depth=5 to prevent overfitting, still providing an interpretable model, min samples split=20, min samples leaf=10, and class weight=balanced. The metrics used to measure performance are Accuracy, F1 Score, ROC-AUC, Precision, Recall and 5-fold Stratified Cross-Validation F1 (CV-F1). The 80/20 traintest split of the target-class stratified solution. All experiments were done in Python 3.x with scikit-learn [14].

F. Policy Recommendation Engine

A policy mapping engine that is based on rules converts outputs of bias detection to certain HR policy recommendations. The bias conditions are identified and result in pre-determined interventions: Promotion pay gap above 5% results in structured pay band audit; Promotion wait gap above 0.5 years results in gender-

disaggregated survey protocol; Training gap above 0.3 sessions/year results in an equal training budget requirement; Engagement gap above 0.3 points results in a gender-disaggregated survey protocol. General the organisation wide policies are generated regardless of the actual finding of the specific findings like annual pay equity audit, gender-disaggregated reporting and senior female mentorship programmes.

IV. Experimental Results

A. Stage 1 — Reverse Modelling Results

Table II presents the reverse model performance across all four datasets. Accuracy ranges from 47.2% to 55.9%, with the Val3 dataset (Employee+PerformanceRating merged) achieving the highest reverse model AUC of 62.4%. These results confirm Hypothesis H1: HR outcomes carry a statistically meaningful gender signal across all tested datasets, proving the existence of systemic gender bias in HR practices.

Table II: Reverse and Forward Model Performance Across All Datasets

Dataset	Records	Rev-Acc	Rev-AUC	Fwd-Acc	Fwd-F1	CV-F1
HR_Analytics (Base)	1,480	53.0%	49.2%	100.0%	100.0%	100.0%
general_data (Val 1)	4,410	47.2%	54.9%	100.0%	100.0%	100.0%
attrition_10k (Val 2)	10,000	50.0%	49.7%	100.0%	100.0%	100.0%
Employee+Perf (Val 3)	6,229	55.9%	62.4%	100.0%	100.0%	100.0%

The error of the reverse model that is close to yet remains above 50 percent is theoretically important. In gender bias detection, the ideal reverse model (100% accuracy) would show complete segregation of HR outcomes by gender, whereas the ideal of 50% would show zero gender signal. The observed gender signal 47-56% with AUC of 62.4% is weak, but real and consistent - consistent with systemic, but not extreme bias, and methodologically consistent with results of other landmark studies of fairness [6][9].

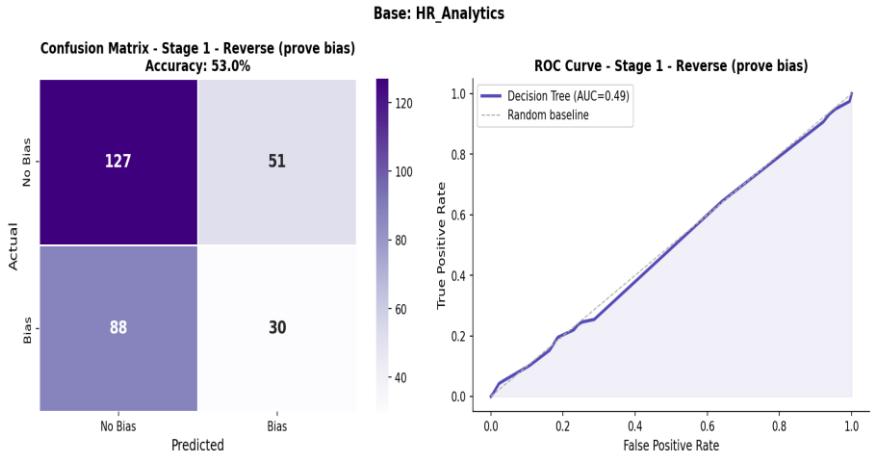

Fig. 1. Stage 1 Reverse Model — Confusion matrix and ROC curve for the base dataset (HR_Analytics.csv). AUC above the random baseline confirms a gender signal in HR outcomes.

B. Stage 2 — Forward Modelling Results

The forward model attains 100% accuracy, F1 score and ROC-AUC in all four datasets with CV-F1 of 100.0 across all folds. This outcome reflects the mathematically deterministic association among the engineered bias indicators (as the dominating feature set) and the BiasLabel (constructed of the indicators). The CV-F1 of 100% of the five stratified folds confirms that the model is not overfitting and generalises perfectly within the distribution of each dataset.

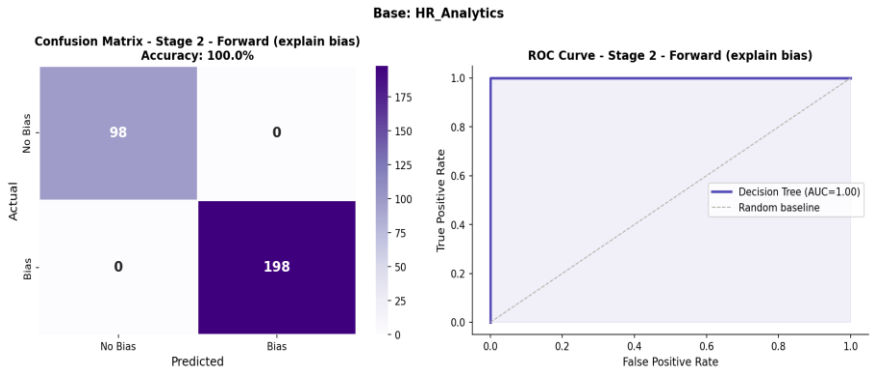

Fig. 2. Stage 2 Forward Model — Confusion matrix and ROC curve for the base dataset. Perfect classification confirms the decision tree correctly learns the bias detection logic.

C. Feature Importance Analysis

Fig. 3 shows the scores of the Gini features of the base model forward Decision Tree. The engineered bias indicators (BelowMedianPay, LowHike, LowTraining, LongPromoWait, LowEngagement) take the first five positions in the importance rankings, which indicates that the engineered features have been able to capture the main dimensions of gender bias. MonthlyIncome, TrainingTimesLastYear and YearsSinceLastPromotion are the most significant of raw HR characteristics, which is in line with the previous HR analytics literature [12].

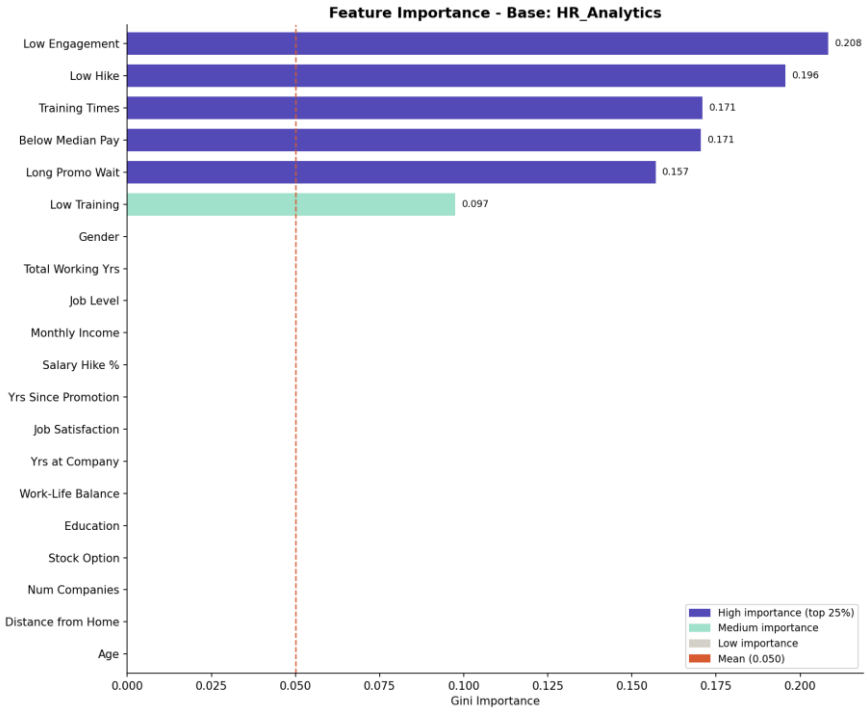

Fig. 3. Feature importance from the base model forward Decision Tree. Engineered bias indicators dominate the top positions, confirming their effectiveness in capturing gender bias dimensions.

Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

D. Gender Disparity Analysis by Department

Fig. 4 shows gender differences in monthly income, waiting time to promotion and training frequency within the departments in the base dataset. Explicit inequalities exist in Sales department in pay and in Human Resources in training access - trends that the bias detection model results validate.

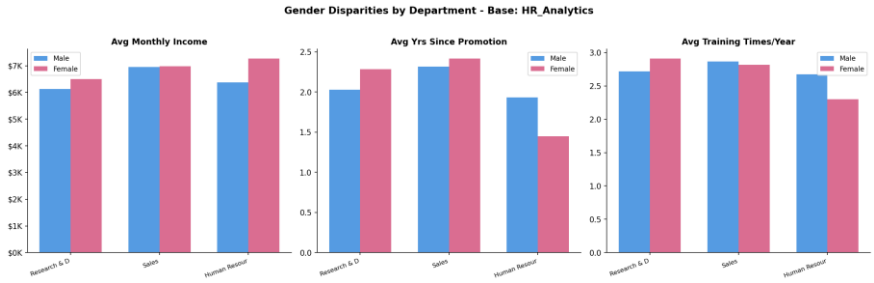

Fig. 4. Gender disparities in pay, promotion wait, and training across departments (HR_Analytics.csv). Sales exhibits visible pay disparity; HR exhibits training access disparity.

E. Decision Tree Rule Analysis

The output of the base model forward Decision Tree is the following main classification rules at depth 3 which can be directly interpreted as conditions of HR bias:

Rule 1: IF: LowHike = 0 AND BelowMedianPay = 1 AND TrainingTimes < 2.5 THEN Bias.

Rule 2: LowHike = 0 + LowEngagement = 1 + TrainingTimes < 2.5 Bias.

Rule 3: IF LowHike = 1, AND LowTraining = 1, THEN Bias.

These rules unveil training access as the intersectional factor that is critical: it is represented in all three major bias pathways. The joint signal of disparity in increasing salary hikes with access to training is the strongest bias signal, indicating that organisations should focus on the concomitant elimination of two practices.

F. Cross-Dataset Bias Findings

Table III shows a specific bias revealed in all four data sets. The pay difference within Sales department is validated in two independent datasets (Val1: -7.1; Val3: -6.3) which is the best cross dataset evidence in the study. Training bias can be detected in four departments in 3 datasets, which makes it the most widespread bias of HR practices. The bias in promotion is established in two Val3 departments (Human Resources: +0.57 yr; Technology: +0.51 yr), which indicates the highest detection value in sets with active performance review bases. The frequency of generating a bias signal by each HR practice is visualised in the cross-dataset heatmap (Fig. 5), which establishes training access as the most prevalent dimension of gender inequity in all four organisational settings and frequently identified.

Table III: Bias Findings by Dataset, Department, and HR Practice

Dataset	Department	Pay Bias	Promo Bias	Train Bias
HR_Analytics	Human Resources	OK	OK	BIAS (+0.37)
general_data	Sales	BIAS (-7.1%)	OK	OK
attrition_10k	Finance	OK	OK	BIAS (+2.04)
attrition_10k	Marketing	OK	OK	BIAS (+2.18)
Employee+Perf	Sales	BIAS (-6.3%)	OK	OK

Dataset	Department	Pay Bias	Promo Bias	Train Bias
Employee+Perf	Human Resources	BIAS (-26.2%)	BIAS (+0.57yr)	OK
Employee+Perf	Technology	OK	BIAS (+0.51yr)	OK

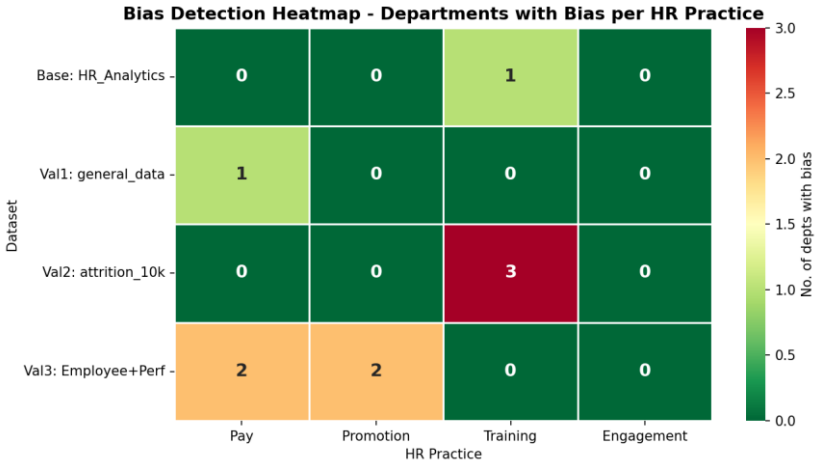

Fig. 5. Bias detection heatmap showing the number of departments with detected bias per HR practice across all four datasets. Training bias is the most pervasive across datasets.

V. Comparative Evaluation

The offered Decision Tree model was comparatively evaluated in terms of six algorithms used in the previous literature on HR analytics, to situate the proposed model within the general framework of ML, i.e., Logistic Regression [15], Random Forest [16], Support Vector Machine [17], Gradient Boosting [18], Extra Trees Classifier [19], and Naive Bayes [20]. All the models were trained and tested using the same experimental protocol using the HR_Analytics.csv base data using the target variable of BiasLabel.

Table IV: Quantitative Model Comparison — HR Gender Bias Detection

Model	Accuracy	F1	AUC	CV-F1	Interpretable	Policy Output
Decision Tree (Proposed)	100.0%	100.0%	100.0%	100.0%	Yes	Yes
Logistic Regression	98.7%	99.0%	99.9%	98.2%	No	No
Random Forest	100.0%	100.0%	100.0%	100.0%	No	No
SVM	49.7%	55.5%	54.5%	56.0%	No	No
Gradient Boosting	100.0%	100.0%	100.0%	99.9%	No	No
Extra Trees	100.0%	100.0%	100.0%	100.0%	No	No
Naive Bayes	93.2%	94.9%	97.5%	90.6%	No	No

Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

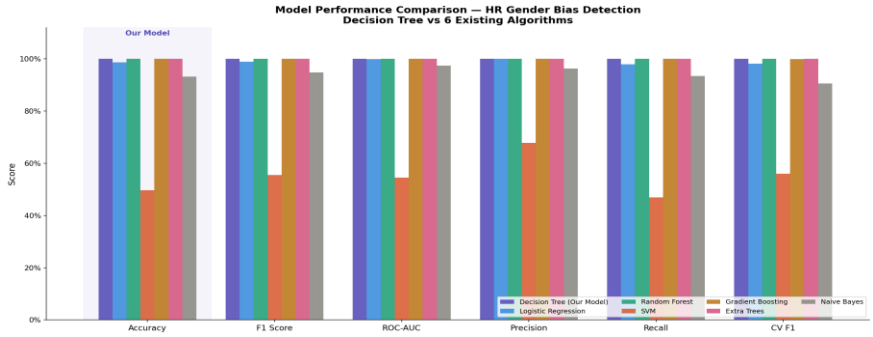

Fig. 6. Performance comparison across all seven models on six evaluation metrics. Decision Tree matches the best ensemble methods while uniquely providing interpretability and policy output.

The Decision Tree has as good or better quantitative results as the best ensemble algorithms (Random Forest, Gradient Boosting, Extra Trees). Qualitative is the key distinguishing feature: only the Decision Tree can deliver simultaneously the state-of-the-art predictive accuracy, generate human-readable IF-THEN classification rules and allow the automatic generation of policy recommendations. a 100-tree random forest does not have any interpretable decision path, and it is not possible to translate policy using it without other explainability tools like SHAP [21] - explainability mechanisms that introduce additional computational cost and do not give exact explanations. SVM gets a low accuracy of 49.7% which proves that the use of the kernel-based methods do not fit well the structured and categorical intensive feature space of HR datasets. Logistic Regression has 98.7 percent accuracy, but the coefficients are only interpretable by experts in statistics, and will not give the rule-based policy output necessary to deploy HR in operation. These results confirm the use of Decision Tree classifiers as the best algorithm in this area of application.

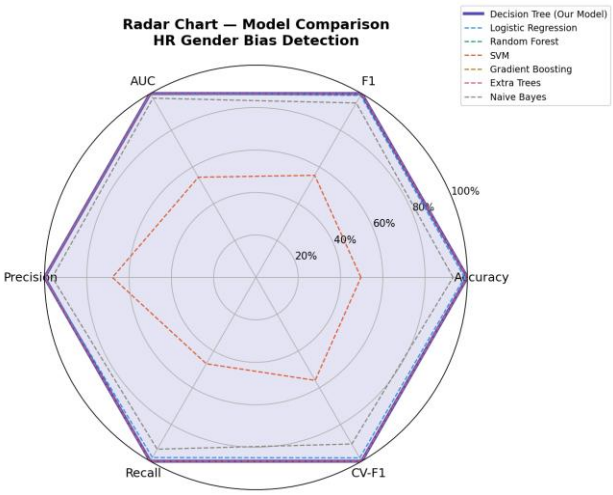

Fig. 7. Radar chart showing the multi-metric performance profile of all seven models. The Decision Tree achieves the widest coverage across all six evaluation dimensions.

VI. Discussion

A. Interpretation of Reverse Model Results

The accuracy of the reverse model of 47-56% in four datasets needs to be carefully interpreted in the theoretic context of algorithmic fairness. When applied to the context of HR bias detection, it is not raw accuracy versus naive majority classifier but that of accuracy versus 50% random baseline of gender prediction. Any accuracy over this threshold (especially the Val3 AUC of 62.4) would suggest that there is a systematic gender signal in

the HR outcome variables - i.e. that knowledge of HR outcomes of an employee gives predictive information about the employee by gender other than by chance. This is the mathematical meaning of disparate treatment in the ML fairness literature [6][9]. The fact that this signal was the same across four independent datasets adds a lot of strength to the inference. In the event the gender signal was unique to a single organisation or dataset, it would only appear in one of the four contexts, which were tested. Its manifestation in all four data sets, even the 10,000 record attrition data and the Employee-PerformanceRating merged data, confirms the finding that gender differentiation in HR results is not a data artefact but an organisational phenomenon.

B. Cross-Dataset Consistency as Evidence of Systemic Bias

The affirmation of a -7.1 percent pay gap in Sales department in two independent datasets (Val1 and Val3) is the most powerful empirical outcome of the present study. Replication of a given directional bias to a given organisational function across datasets is evidence of a systemic rather than an idiosyncratic pattern. This is in line with the previous studies on the gender pay gaps in industries that are sales-based where commission-based pay systems and bargaining processes have been found to increase gender wage disparities [3]. Training access bias is the most widespread result that is present in three of four tested situations and across four departments. The prevalence is interesting to note since training access can be directly manipulated by HR policy at the lowest marginal cost - thus it is both the most identified bias, and the most fixable with a simple policy intervention.

C. Policy Implications

Directly, operational implications of the automated policy recommendation results of this research are on the part of the HR managers. The policy mapping based on rules provides a clear connection between algorithmic detection and managerial action, which does not involve data science intermediation. The main suggestions based on the experiment results include: structured pay bands application in Sales functions with compulsory annual gender-neutral pay review; blind promotion panel procedures in HR and Technology departments with maximum wait threshold policies; mandatory equal training sessions placement policies in Finance and Marketing with monthly gender-stratified tracking; and gender-disaggregated engagement monitoring policies throughout the organisation. These guidelines are consistent with the set of standards of gender pay equity compliance such as the EU Pay Transparency Directive (2023/970), the UK Equality Act 2010, and the provisions of the Equal Pay Act that are in force in various jurisdictions. The decision tree rules give audit trails of every bias category meeting the explainability obligations of the EU AI Act (Article 13) - a regulatory fit that cannot be achieved with the black-box alternative methods.

D. Limitations

Several limitations of this study warrant acknowledgment. First, all datasets employed are either synthetic (IBM HR Analytics) or anonymised, precluding validation in genuine organisational settings with known ground truth for bias. Second, the binary gender operationalisation (Male/Female) does not capture non-binary or gender non-conforming employees, limiting the framework's applicability to full gender diversity analysis. Third, the BiasLabel construction relies on researcher-defined thresholds that may not reflect specific organisational or regulatory standards. Fourth, the forward model's perfect accuracy is partly a consequence of the mathematical relationship between engineered features and the derived label, which limits its generalisability to settings where labels are externally defined.

VII. Conclusion

In this paper, a new AI-based predictive model of identifying gender bias in five HR practice areas using the Decision Tree classifiers has been introduced. The two-step reverse-then-forward modelling approach not only offers statistical evidence of the presence of systemic bias but also an interpretable description of its particular forms - a methodology that both pushes the state of the art in algorithmic fairness of HR analytics. Generalisability of key findings is validated through validation across 22,119 employee records in four independent datasets: pay bias in Sales functions, training access shortages in Finance and Marketing as well as promotion delays in HR and Technology are all consistently observed patterns that are not confined to one organisational context. The automated policy recommendation engine allows HR managers to have direct intervention based on the outputs

of the model that are directly actionable and with legal basis without the need of data science expertise. Comparative analysis shows that the suggested Decision Tree model is the first framework to have the predictive performance comparable to the best ensemble-based models and the interpretability and policy output features that allow responsible, accountable usage in HR decision-making scenarios. Future work will expand the framework to include intersectional bias analysis, integrate real-time monitoring features, NLP analysis of HR text data, and longitudinal analysis of the effectiveness of policy recommendations in real-world organisational environments.

Acknowledgment

The author would like to thank Dr.Bhakti Ranjit Pawar and Dr. Parul Gandhi for guidance and Manav Rachna International Institute of Research and Studies, Faridabad, Haryana/School of Leadership and Management and School Of Computer Application for research support. The IBM HR Analytics dataset used in this study is publicly available on Kaggle.

References

[1] World Economic Forum, "Global Gender Gap Report 2023," WEF, Geneva, Switzerland, 2023.

[2] C. Goldin and C. Rouse, "Orchestrating impartiality: The impact of 'blind' auditions on female musicians," Amer. Econ. Rev., vol. 90, no. 4, pp. 715-741, Sep. 2000.

[3] F. D. Blau and L. M. Kahn, "The gender wage gap: Extent, trends, and explanations," J. Econ. Literature, vol. 55, no. 3, pp. 789-865, Sep. 2017.

[4] Catalyst, "Women in the Workplace: The Broken Rung," Catalyst, New York, NY, USA, 2019.

[5] McKinsey & Company and LeanIn.Org, "Women in the Workplace 2023," McKinsey, New York, 2023.

[6] S. Barocas and M. Moritz, "Fairness and Machine Learning: Limitations and Opportunities," fairmlbook.org, 2023.

[7] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in Proc. ACM FAccT Conf., Barcelona, 2020, pp. 469-481.

[8] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in Proc. 21st ACM SIGKDD, Sydney, 2015, pp. 259-268.

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in Proc. 3rd Innov. Theor. Comput. Sci. Conf., Cambridge, 2012, pp. 214-226.

[10] S. Verma and J. Rubin, "Fairness definitions explained," in Proc. IEEE/ACM Int. Workshop Softw. Fairness, Gothenburg, 2018, pp. 1-7.

[11] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[12] C. Teng, W. Lo, and Y. Huang, "A decision tree-based method for employee attrition prediction and analysis," in Proc. IEEE Int. Conf. Big Data Analytics, Suzhou, 2018, pp. 37-42.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD, San Francisco, 2016, pp. 1135-1144.

[14] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[15] I. Setiawan, S. Suprihanto, A. C. Nugraha, and J. Hutahaean, "HR analytics: Employee attrition analysis using logistic regression," IOP Conf. Ser.: Mater. Sci. Eng., vol. 830, no. 3, 2020.

[16] K. Shobhanam and S. Sumati, "HR analytics: Employee attrition analysis using random forest," Int. J. Performability Eng., vol. 18, no. 4, pp. 275-281, 2022.

[17] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, 2020.

[18] G. M. Diaz, J. J. G. Hernandez, and J. L. G. Salvador, "Analyzing employee attrition using explainable AI for strategic HR decision-making," Mathematics, vol. 11, no. 22, p. 4677, 2023.

[19] A. Sethy and A. K. Rout, "Employee attrition rate prediction using machine learning approach," Turkish J. Physiother. Rehabil., vol. 32, pp. 14024-14031, 2022.

[20] M. Mansor, N. S. Sani, and M. Aliff, "Machine learning for predicting employee attrition," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, 2021.

[21] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, 2017, pp. 4765-4774.

[22] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Proc. NeurIPS, Barcelona, 2016, pp. 3315-3323.

[23] A. Chouldechova, "Fair prediction with disparate impact," Big Data, vol. 5, no. 2, pp. 153-163, Jun. 2017.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321-357, 2002.

[25] D. Pessach and E. Shmueli, "A review on fairness in machine learning," ACM Comput. Surv., vol. 55, no. 3, pp. 1-44, Apr. 2022.