Innovation in Computer and Data Sciences
2(1), 11–23
DOI: https://doi.org/10.64389/icds.2026.02162
https://sphinxsp.org/journal/index.php/icds/

Research article

Bias and Fairness in AI-Based Employee Attrition Prediction Using Random Forest

Idowu Adesoji Oladipupo[1], Serifat Adedamola Folorunso[1,*], Sadiq Olusegun Balogun[2], Fatima Iganya Suleiman[1], Olufunke Catherine Olayemi[3], Joseph Maugbe Jacob[1]

[1,*] Department of Applied Data Science, Teesside University, Middlesbrough TS1 3BA, United Kingdom

[2] Leeds Institute for Data Analytics, Leeds LS2 9JT, United Kingdom

[3] Department of Computer Science, Teesside University, Middlesbrough TS1 3BX, United Kingdom

* **Correspondence:** serifatf005@gmail.com

ARTICLE INFO

Keywords:
Artificial intelligence (AI)
Employee attrition
Algorithmic bias
Fairness in machine learning
Workforce analytics
Human resource technology
Random Forest
Gender disparity

Mathematics Subject Classification:
68T07, 90B30, 62H30, 90C90, 93C95

Important Dates:
Received: 27 December 2025
Revised: 22 January 2026
Accepted: 25 January 2026
Online: 26 January 2026

ABSTRACT

Artificial intelligence is increasingly employed to predict employee attrition, enabling organisations to improve talent retention and workforce planning. However, without explicit consideration of fairness, these models risk embedding and amplifying societal biases. This study examines bias in AI-based attrition prediction using a Random Forest classifier applied to the IBM HR Analytics employee attrition dataset. Although the model demonstrates high predictive performance (92.3 percent) and an area under the curve of 0.97, subgroup analysis reveals disparities in prediction performance across gender. Fairness assessments based on equal accuracy, demographic parity, and equality of opportunity show that predictions for female employees achieve higher precision and recall than those for male employees, suggesting differential predictive performance across gender groups. The findings highlight organisational risks associated with such disparities, including the risk of unjust decision-making, reduced employee trust, and hindered diversity and inclusion efforts. To mitigate these challenges, the study recommends fairness-aware strategies such as balanced sampling, established fairness metrics, post-processing approaches (e.g., equalised odds), and continuous model auditing. This research underscores the ethical importance of aligning AI systems in human resource management with principles of equity, transparency, and accountability.

1. Introduction

Employee attrition prediction via AI algorithms has garnered attention as organisations seek to refine their workforce management strategies. Yet, this reliance raises concerns regarding bias and its repercussions. Bias in attrition prediction denotes systematic inaccuracies or unfairness in AI models, potentially resulting in discriminatory outcomes or reinforcing workplace disparities [4].

AI is also increasingly employed in performance management and employee development. Tools powered by artificial intelligence analyse data from performance reviews, peer feedback, and other employee-related inputs to assess individual strengths, identify areas for improvement, and forecast growth potential [2]. While these applications aim to enhance objectivity and support development, they too risk embedding bias, especially when trained on historical data that may reflect organisational inequities or subjective evaluations.

According to [7]. Biased employee attrition predictions can profoundly impact organisations and individuals. Organisational repercussions include suboptimal decision-making, inefficient resource allocation, reduced productivity, and impaired morale. Additionally, biased predictions may thwart diversity, equity, and inclusion efforts, perpetuating workforce inequalities. Biased attrition predictions can negatively impact employees by influencing decisions on hiring, promotion, and termination unrelated to job performance. This can foster feelings of unfairness, erode trust in the organisation, and ultimately lead to disengagement or attrition [3]. Biased AI systems may also reinforce societal biases, worsening social inequalities and systemic discrimination [8].

To address these concerns, it is imperative to critically examine the existence and implications of bias in employee attrition prediction models. By identifying and mitigating sources of bias, organisations can enhance the fairness, transparency, and effectiveness of their workforce management practices. This study aims to explore the presence of bias in AI-driven employee attrition prediction and its impact on organisational outcomes and individual experiences.

Numerous studies have highlighted the prevalence of bias in AI systems across different domains, including hiring, lending, and criminal justice. For instance, research by Obermeyer et al. [5] demonstrated racial bias in a healthcare algorithm that led to systematic underestimation of the healthcare needs of Black patients. Similarly, investigations by Angwin et al. [1] uncovered gender bias in hiring algorithms used by major companies, resulting in discrimination against female job applicants.

The impact of biased AI models extends beyond organisational contexts, affecting individuals' opportunities, livelihoods, and well-being [4]. Biased predictions can perpetuate systemic inequalities, reinforce stereotypes, and undermine trust in AI systems. Therefore, addressing bias in employee attrition prediction models is not only essential for enhancing organisational decision-making but also for promoting fairness, equity, and inclusivity in the workplace.

2. METHODOLOGY

This section outlines the methodological framework adopted in this study, detailing the procedures followed for data preparation, model development, and evaluation. The subsequent subsections describe each stage of the analytical pipeline employed to investigate bias and fairness in employee attrition prediction.

The Random Forest (RF) classifier was selected as the primary predictive model due to its robustness, ability to capture complex non-linear relationships, and effectiveness in handling high-dimensional data.

These characteristics make the algorithm particularly suitable for employee attrition prediction, where behavioural patterns are often subtle and influenced by multiple interacting factors [6].

Rather than relying on a single decision tree, the Random Forest algorithm aggregates the outputs of multiple trees trained on random subsets of data and features, thereby reducing variance and mitigating overfitting. This ensemble-based approach enhances predictive stability and generalisation performance, which is essential for reliable evaluation of model outcomes and fairness across demographic groups.

The methodological process comprises data sourcing, cleaning and preprocessing, exploratory data analysis, feature engineering, imbalance handling, feature selection, and model training and evaluation. Each of these stages is described in detail in the subsections that follow.

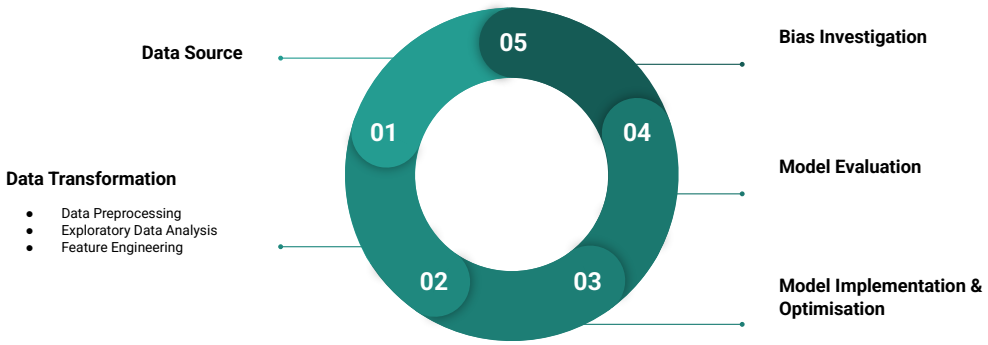


Figure 1. Workflow of the proposed system

2.1. Data Scource

The present study utilizes the IBM HR Analytics Employee Attrition dataset, publicly accessible through Kaggle. This dataset comprises 1,470 employee records, each described by 35 attributes capturing demographic characteristics, job-related factors, and indicators of workplace engagement. Widely employed as a benchmark for employee attrition prediction, the dataset reflects a simulated organisational context and exhibits the class imbalance commonly observed in attrition scenarios. To ensure robust and fair evaluation, preprocessing and resampling strategies were applied during model development to address this imbalance and prevent potential biases in predictive performance. Its use enables reproducibility while providing a realistic framework for evaluating predictive models and fairness analyses.

2.2. Data Cleaning and Preprocessing

To enhance data integrity, we eliminated duplicate entries, assigned suitable data types to columns, and addressed missing values. This preprocessing ensures the dataset's reliability and facilitates streamlined analysis for accurate results.

2.3. Exploratory Data Analysis

Employee attrition analysis reveals several key factors influencing turnover. Firstly, the distribution of income is bimodal, with most employees earning between 2000-6000 dollars, and higher earners generally staying. However, exceptions exist, suggesting potential biases in retention strategies. Secondly, younger employees exhibit higher attrition, reflecting possible age-related biases in career stability. Proximity to the workplace affects attrition, hinting at location-based disparities. Prior work experiences and promotions also impact attrition, indicating potential biases in career advancement. Job satisfaction, marital status, and work-life balance influence attrition as well, raising questions about biases in workplace culture. Recognizing and addressing these biases is essential for fostering fair organisational decision-making and ensuring equitable individual experiences.

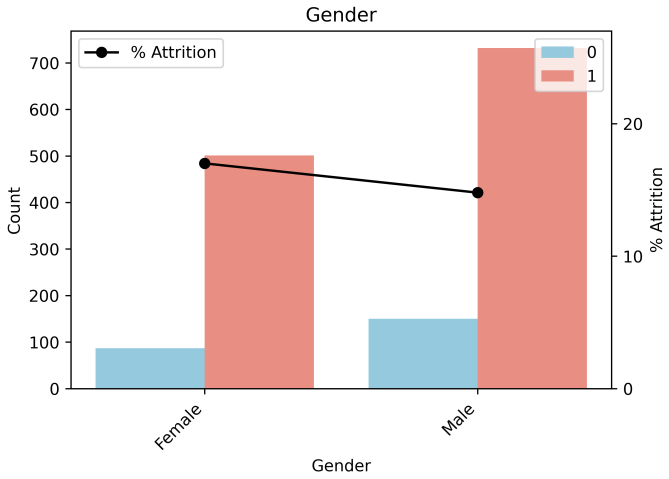


Figure 2. Gender and Attrition bar chart

2.4. Feature Engineering

In this research study, certain techniques are necessary to prepare our data for machine learning models. One crucial step involves removing unwanted columns, such as the case ID, which does not contribute to the predictive task at hand.

- Label Encoding: Accurately encoding categorical variables is crucial in data analysis, especially for machine learning tasks. Since most ML algorithms operate on numerical data, the proper encoding of Categorical variables are essential for effective analysis.

- One Hot Encoding: One-hot encoding is indispensable in machine learning for managing categorical variables with multiple categories. This technique transforms categorical data into binary format, enabling algorithms to interpret and utilize information effectively in predictive modeling, thereby boosting model accuracy and performance.

- Outliers: EDA provides insight into customer churn, revealing no outliers or abnormalities in numerical columns.

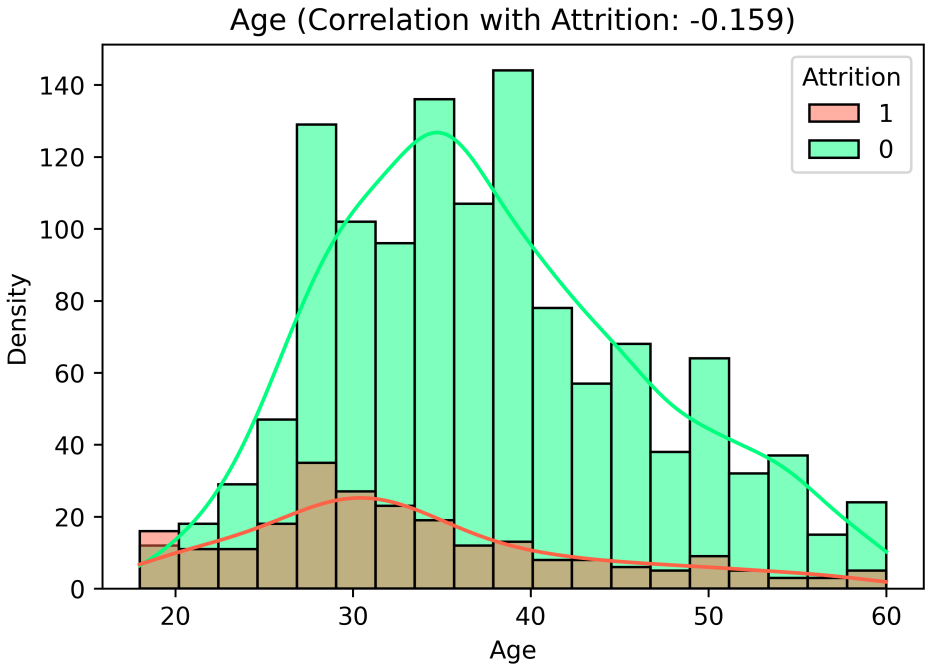


Figure 3. Age relationship with Attrition

2.5. Handling Imbalance Data

Addressing class imbalance is critical in attrition prediction, as models trained on skewed data may become biased toward the majority class. To mitigate this issue, Synthetic Minority Over-sampling Technique combined with Tomek Links (SMOTE-Tomek) was applied to the training data. This approach simultaneously oversamples minority class instances and removes ambiguous samples near class boundaries, leading to improved class separability. By applying resampling only during training, the study ensures that performance metrics reflect genuine generalisation rather than artifacts introduced by imbalance correction. This strategy also supports fairer evaluation across demographic subgroups, particularly when assessing gender-based disparities.

2.6. Feature Selection

Feature selection is a crucial step in enhancing model accuracy by eliminating unnecessary data that may lead to overfitting and reduced performance. Its advantages include mitigating overfitting, improving accuracy by removing misleading data, and reducing training time by simplifying algorithm complexity.

2.7. Fitting Machine Learning Model

To ensure model generalizability, we split our dataset into training and testing subsets, with a test size of 0.2 (80% training, 20% testing). Setting the random state to 42 ensures consistent random splits across runs.

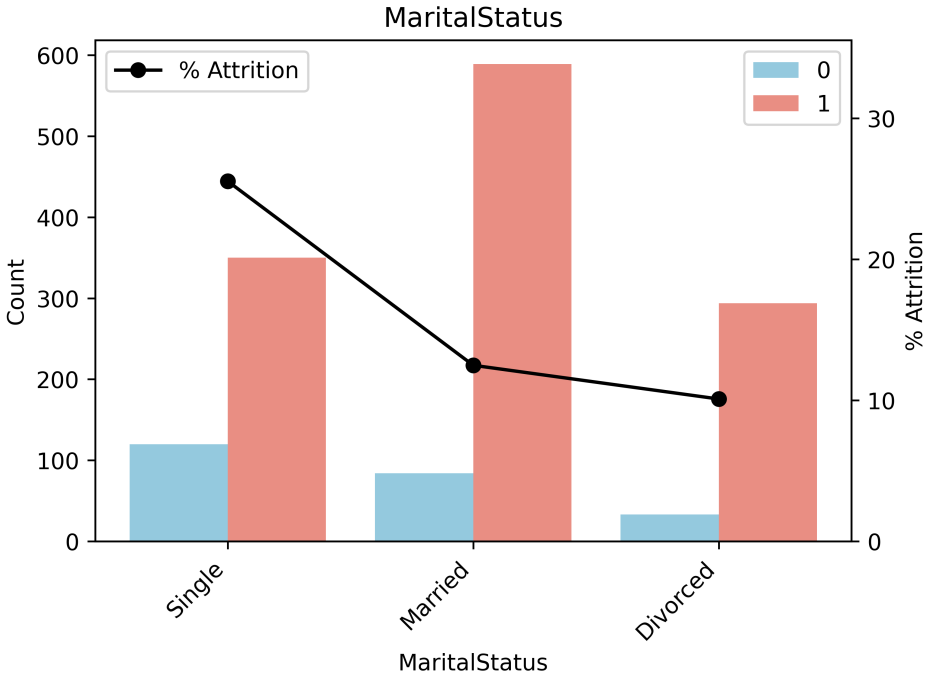


Figure 4. Marital Status with Attrition

2.7.1. Scaling

Standardizing dataset features is crucial due to their varying scales. Using StandardScaler ensures uniformity in feature scales, thereby enhancing machine learning algorithm performance and accuracy by centering features around a zero mean with unit variance.

2.7.2. Model Implementation

In this study, the Random Forest (RF) classifier was employed as the primary predictive model for employee attrition due to its robustness, ability to model complex non-linear relationships, and resistance to overfitting. Random Forest is an ensemble learning technique that constructs multiple decision trees during training and combines their outputs to produce a final prediction. By aggregating predictions from diverse trees trained on bootstrapped samples and random subsets of features, the model achieves improved generalisation performance compared to single decision trees.

Formally, the prediction of a Random Forest classifier is obtained through majority voting across individual decision trees and can be expressed as:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_k(x)\}$$

where $h_i(x)$ represents the prediction of the i-th decision tree for input feature vector x, k denotes the total number of trees in the ensemble, and $\hat{y}$ is the final predicted class.

Prior to model training, the dataset was divided into training and testing subsets using an 80:20 split to ensure robust evaluation of model performance. Feature scaling was applied using standardisation to normalise feature distributions and improve model stability. The Random Forest classifier was trained on the processed training data and evaluated using standard performance metrics, including accuracy, precision,

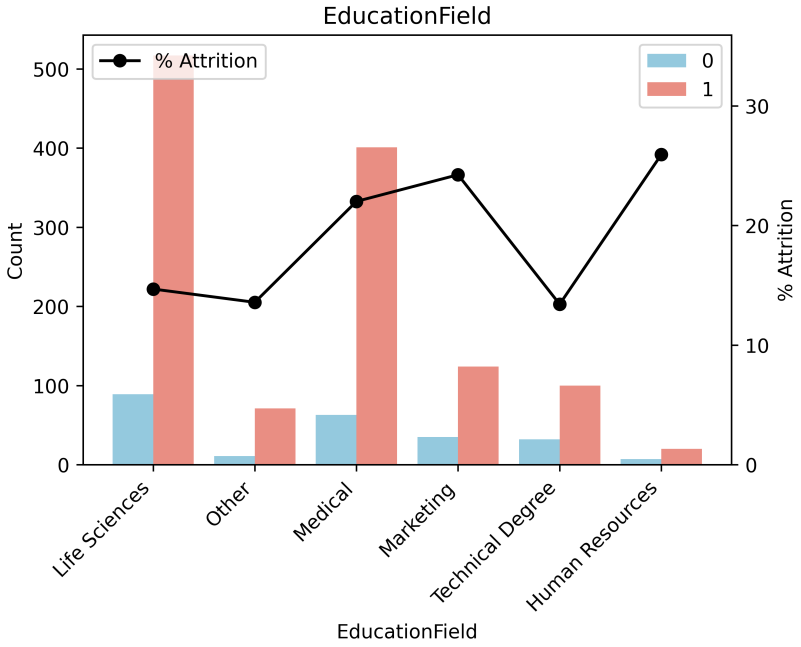


Figure 5. Educational Field with Attrition

recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC).

This implementation framework enables the Random Forest model to effectively capture underlying patterns in employee attrition while maintaining interpretability and suitability for fairness evaluation across demographic groups.

3. RESULT

Random Forest Classifier exhibits robust performance, achieving high accuracy (0.912) and balanced precision, recall, and F1-scores across both employee classes. With a remarkable AUC of 0.97, the classifier demonstrates excellent discriminatory power. Overall, it serves as an effective tool for predicting employee attrition, offering valuable insights for organisational decision-making.

3.1. Receiver Operating Characteristic (ROC)

achieving a Receiver Operating Characteristic (ROC) curve with an area under the curve (AUC) of 97 signifies strong predictive performance. This high AUC indicates that the model effectively balances the trade-off between true positive and false positive rates across different discrimination thresholds. In practical terms, it means that the model can accurately distinguish between employees who are likely to leave the company (positives) and those who are likely to stay (negatives), with minimal misclassification. This robust predictive capability is essential for organisational decision-making, as it enables proactive measures to retain valuable employees and mitigate attrition risks, ultimately enhancing organisational stability and performance.

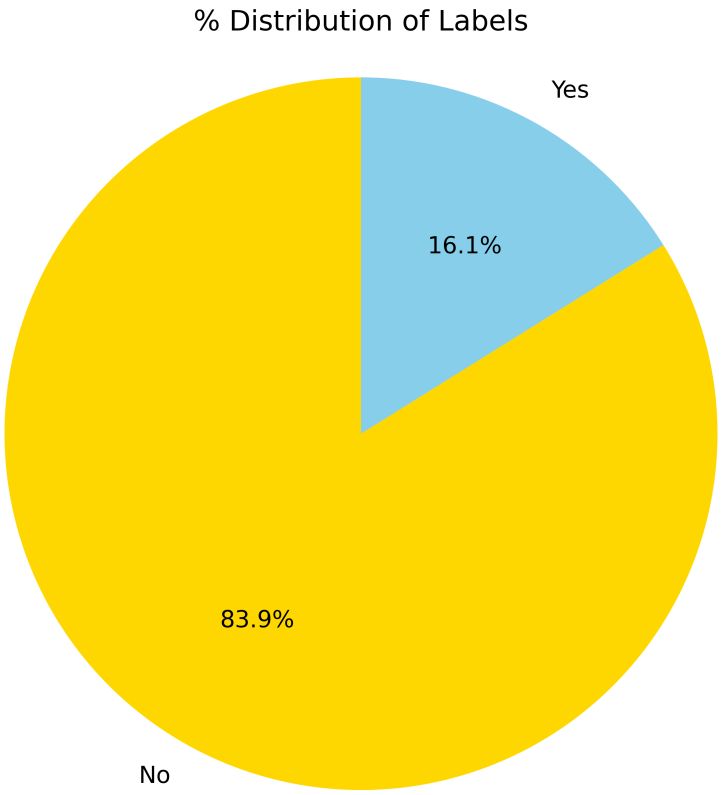


Figure 6. Distribution of label (Attrition)

3.2. Model Optimization

3.2.1. Hyperparameter Tuning

The integration of hyperparameter tuning through grid search, complemented by 5-fold cross-validation, distinctly amplified the Random Forest classifier's predictive prowess. This meticulous optimization procedure culminated in notable enhancements across pivotal performance metrics: accuracy, precision, recall, and F1-score. Initially, the model exhibited commendable performance with an accuracy of 0.912, precision of 0.910, and recall of 0.910. However, post-tuning, the model demonstrated substantial improvements, yielding an accuracy of 0.923, precision of 0.920, and recall of 0.920. Furthermore, the refined model showcased heightened robustness, as evidenced by its elevated ROC curve and AUC score. This comprehensive refinement underscores the profound impact of hyperparameter tuning in augmenting predictive efficacy, affirming its indispensable role in bolstering model performance and discernment accuracy across positive and negative instances.

3.3. Fairness Evaluation

3.3.1. Equal Accuracy

Equal accuracy ensures that the prediction model performs equally well across different demographic groups. In the context of employee attrition prediction, this means that the model's accuracy in predicting attrition should be consistent for all employees, regardless of their demographic characteristics. Comparing

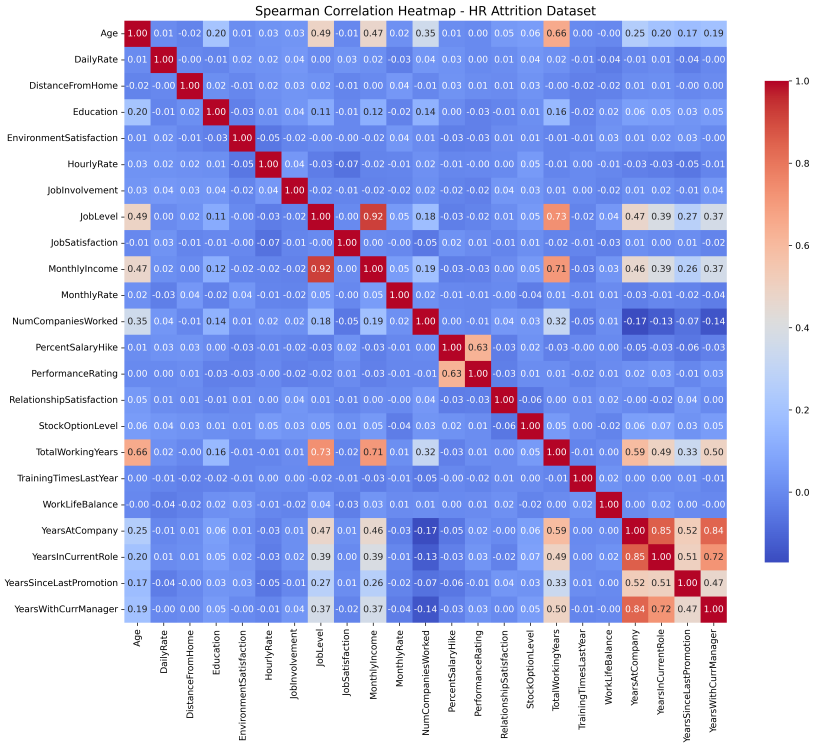


Figure 7. Correlation Plot

the accuracy metrics for the male and female groups, we find that the accuracy for both groups is high, with 0.8701 for males and 0.9530 for females. This initially suggests that the model performs similarly well in predicting attrition for both male and female employees. However, it's important to acknowledge that while the overall accuracy appears high for both groups, there is a discrepancy in accuracy between males and females. Despite the high accuracy for females, it's crucial to recognize and address any disparities in model performance to ensure fairness and equity in predicting attrition for all employees.

3.3.2. Group Fairness (Demographic/Statistical Parity)

Group fairness, also known as demographic or statistical parity, focuses on ensuring that the outcomes of the prediction model are balanced across different demographic groups. In the context of employee attrition prediction using random forest classifier, this means that the attrition rates should be similar for all demographic groups. Comparing the positive rate (precision) metrics for the male and female groups, we find that they are both high, with 0.8901 for males and 0.9615 for females. Initially, this suggests that the attrition rates are balanced across gender groups, indicating fairness in the model's outcomes. However, upon closer inspection, it's important to note that the positive rate for females is higher than that for males, revealing a potential imbalance in the model's predictions between the two gender groups.

3.3.3. Equality of Opportunity

Equality opportunity ensures that individuals from different demographic groups have an equal chance of receiving accurate predictions about their likelihood of attrition. In the context of employee attrition prediction, this means that the false positive and false negative rates should be balanced across demographic

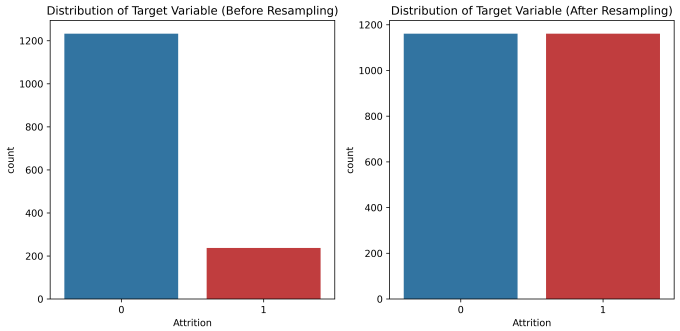


Figure 8. Imbalance and balanced data before and after resampling

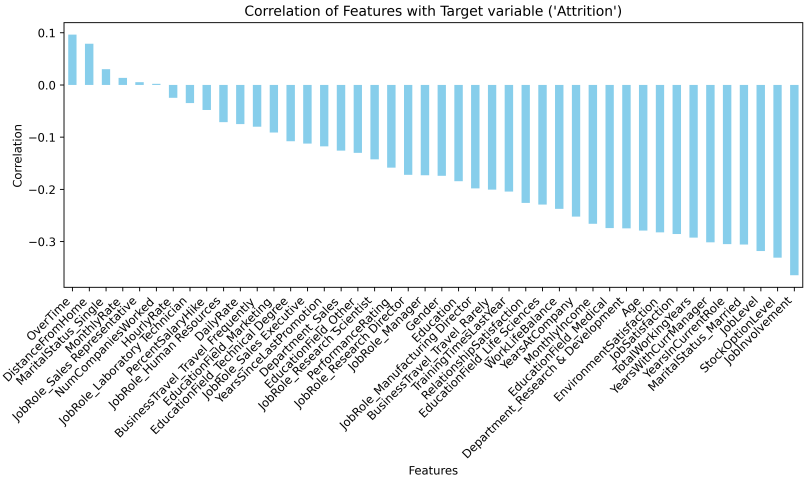


Figure 9. Correlation between features and target variable

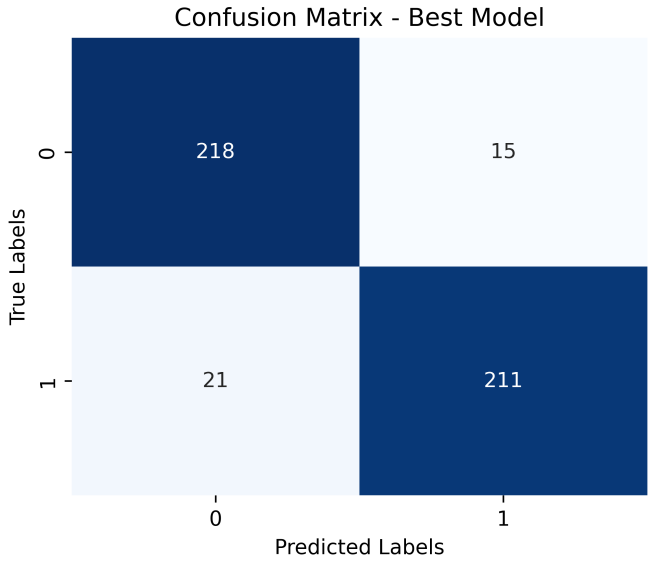
groups. Comparing the recall (sensitivity) metrics for the male and female groups, we find that they are also similar, with 0.8020 for males and 0.9542 for females. This initially suggests that individuals from both gender groups have an equal chance of receiving accurate predictions about their likelihood of attrition. However, it's crucial to recognize that there is a significant difference in recall between males and females, highlighting a potential disparity in the model's predictions and indicating the need for further investigation into gender-based biases.

4. DISCUSSION

While the Random Forest model demonstrates high predictive performance, fairness evaluation highlights notable gender-based disparities. Female employees consistently achieved higher accuracy, precision, and recall, suggesting that the model is more attuned to patterns associated with female attrition. This outcome may arise from imbalanced class distributions, differential feature correlations, or historical biases present in organisational data.

These findings align with existing literature indicating that AI models, even when statistically robust, can inadvertently reinforce or amplify inequities [5] [1]. In HR contexts, such disparities carry organisational risks, including unequal treatment in retention efforts, reduced trust, and potential setbacks for diversity and inclusion initiatives. Addressing these disparities requires the integration of fairness-aware techniques,

	precision	recall	f1-score	support
0	0.91	0.94	0.92	233
1	0.93	0.91	0.92	232
accuracy			0.92	465
macro avg	0.92	0.92	0.92	465
weighted avg	0.92	0.92	0.92	465



(a) **(b)**

Figure 10. Classification report and confusion matrix of best model

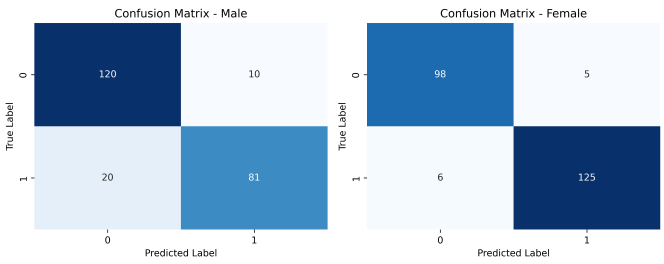
such as balanced sampling, algorithmic adjustments, or post-hoc bias correction. Moreover, incorporating explainable AI methods (e.g., SHAP) could clarify which features drive gender-specific predictions, enhancing transparency and supporting equitable HR decision-making.

5. CONCLUSION

In conclusion, predictive models for employee attrition offer substantial value for workforce planning and organisational decision-making. However, this study demonstrates that strong predictive performance does not guarantee fairness, as gender-based disparities were observed across multiple evaluation metrics. Integrating fairness considerations alongside performance evaluation is therefore essential to ensure equitable outcomes, maintain employee trust, and support responsible human resource management. By combining robust predictive modelling with fairness-aware evaluation, organisations can move toward more transparent, inclusive, and ethically grounded applications of AI in the workplace.

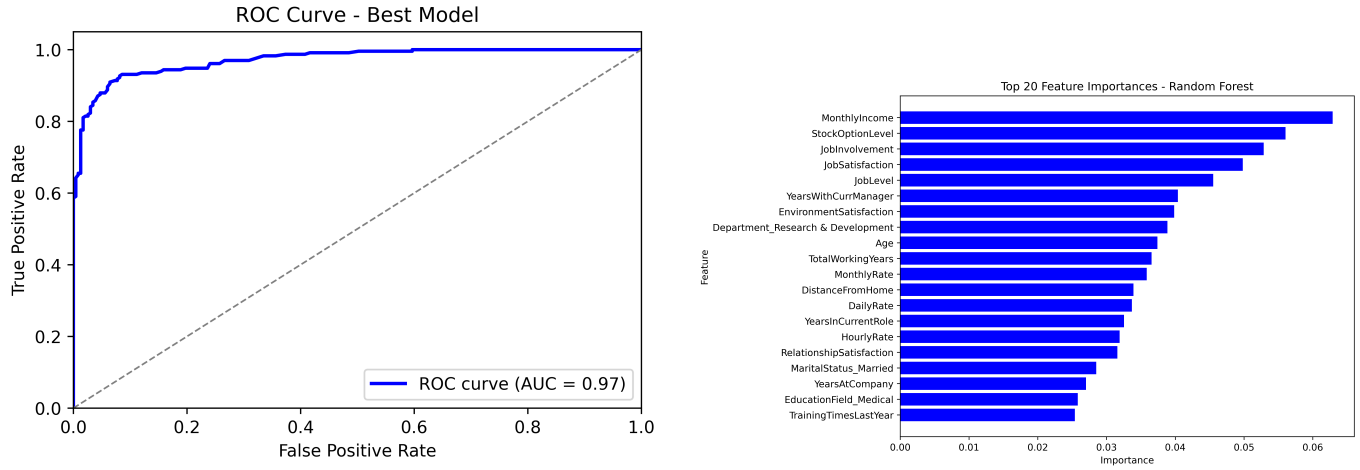
Figure showing the metric performance of Male and Female

Group	Accuracy	Precision	Recall	F1-score
Male	0.87	0.89	0.8	0.84
Female	0.95	0.96	0.95	0.96



(a) **(b)**

Figure 13. Gender-based metrics and confusion matrix



(a) **(b)**

Figure 11. (a) ROC Curve of best model (b) Feature importance from Random Forest

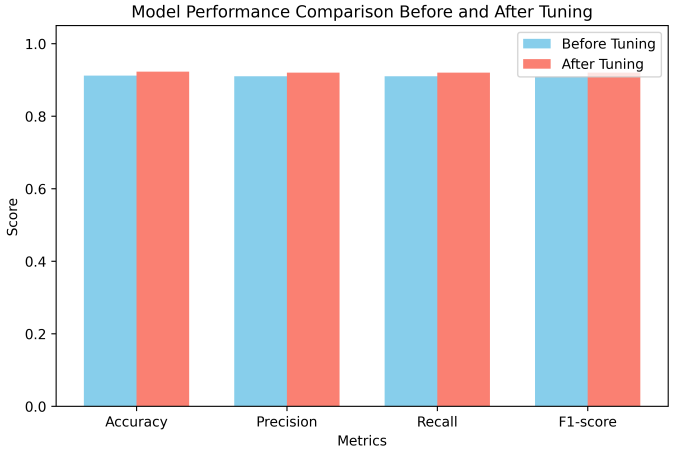


Figure 12. Model metrics before and after fine-tuning

6. LIMITATIONS AND FUTURE WORK

Despite the contributions of this study, several limitations highlight important directions for future research. First, although fairness was evaluated using multiple subgroup-level metrics, the analysis was restricted to gender as the protected attribute. Bias in organisational settings is often intersectional, and future studies should extend fairness assessments to additional attributes such as age, ethnicity, marital status, and job role to better capture compound and overlapping sources of disparity.

Second, while the study identifies performance disparities across demographic groups, it does not examine feature-level drivers of these differences. Incorporating explainable AI techniques, such as SHapley Additive exPlanations (SHAP), would enable deeper insight into how specific features influence attrition predictions across subgroups, thereby improving transparency and supporting responsible human resource decision-making.

Third, the dataset used represents a static snapshot of employee information within a simulated organisational context. Future work should explore longitudinal fairness monitoring to assess whether model

performance and fairness remain stable over time as workforce composition and organisational policies evolve.

Finally, future research should investigate human-in-the-loop approaches, examining how HR practitioners interpret, trust, and act upon fairness-aware predictive outputs. Understanding the interaction between algorithmic recommendations and human judgement is critical to ensuring that AI systems are deployed ethically and equitably in real-world organisational settings.

Conflict of Interest

The authors declare there is no existing conflict of interest.

References

1.Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications.

2.Huang, M.-H. and Rust, R. T. (2021). Artificial intelligence in service. *Journal of the Academy of Marketing Science*, 49(1):7–21.

3.Kim, B., Kim, M., and Lee, J. (2024). The impact of an unstable job on mental health: the critical role of self-efficacy in artificial intelligence use. *Current Psychology*, pages 1–18.

4.Köchling, A. and Wehner, M. (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.

5.Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

6.Olasehinde, O., Johnson, O., and Fakoya, J. (2018). Computational efficiency analysis of customer churn prediction using spark and caret random forest classifier. *IKMt*, 8(2):8–16.

7.Rughoobur-Seetah, S. (2023). An assessment of the impact of emotional labour and burnout on the employees' work performance. *International Journal of Organizational Analysis*.

8.West, S., Whittaker, M., and Crawford, K. (2019). Discriminating systems. Technical report, AI Now.