Fairness-Aware Explainable AI Framework with Integrated Bias Mitigation for Employee Attrition Prediction and Strategic Workforce Decision Support

Bi Bi Hamiya Khanum
M Tech Student, Department of Computer Science and Engineering,
Ghousia College of Engineering, Ramanagara, Karnataka-562159,
India, Affiliated to VTU Belagavi.

Dr. Omar Khan Durrani
Professor and Head , Department of Information Science and Engineering, Ghousia College of Engineering, Ramanagara,
Karnataka-562159, India

Abstract
Employee attrition poses a persistent and costly challenge for modern organizations, with replacement expenses ranging from 33% to 200% of annual salaries and aggregate annual losses approaching $1 trillion globally. Machine learning models have demonstrated considerable potential for predicting employee turnover; however, these systems frequently exhibit systematic biases against demographic groups including gender, age, and departmental representation, undermining both fairness and organizational trust. This research presents a comprehensive fairness-aware explainable AI framework that integrates bias mitigation techniques with predictive modeling for employee attrition forecasting. The proposed architecture comprises three synergistic components: a Random Forest classifier incorporating fairness constraints achieving 94.2% prediction accuracy while reducing demographic parity disparity by 72%, an explainability module leveraging SHAP values to provide transparent justifications for each attrition risk prediction, and a strategic decision support system generating prioritized intervention recommendations. Experimental validation on a 500,000-record dataset containing 35 employee attributes demonstrates that the framework successfully identifies high-risk employees while maintaining equitable treatment across demographic groups. The system delivers interpretable risk factors enabling HR practitioners to understand the rationale behind employee risk classifications and formulate targeted retention strategies.

Novelty in the Research
This research introduces several novel contributions that distinguish it from existing workforce analytics solutions. First, unlike commercial platforms that treat fairness as an afterthought, our framework integrates demographic parity and equal opportunity constraints directly into the model training and post-processing pipeline, achieving a 72% reduction in fairness disparities while maintaining high

DOI: https://doi.org/10.67317/IJMSRT26JUL027

predictive accuracy. Second, while most existing systems provide only raw prediction scores, our framework combines SHAP-based explanations with natural language generation to deliver human-understandable justifications accessible to non-technical HR practitioners. Third, this work addresses the critical "prediction-action gap" through a prioritization engine that translates risk scores into specific, actionable intervention recommendations with timelines and estimated impact. Fourth, the complete open-source implementation provides a cost-effective alternative to commercial solutions typically priced between $40,000 and $60,000 annually. Fifth, the integration of a conversational AI assistant achieving 94% accuracy on natural language workforce queries represents a novel contribution to operationalizing fairness-aware AI in HR contexts.

Keywords
Fairness-aware AI, Explainable AI, Employee Attrition Prediction, Bias Mitigation, Workforce Analytics, SHAP, Random Forest, Ethical AI, Decision Support System, Human Resource Management

1. Introduction

1.1 Background and Motivation

The contemporary business landscape is undergoing a profound transformation driven by digitalization and data-centric decision-making across all organizational functions. Human Resource Management, traditionally dependent on intuition, experience, and manual processes, is increasingly adopting artificial intelligence and machine learning to gain competitive advantage, improve retention rates, and optimize workforce productivity. Organizations implementing AI-powered workforce analytics experience significantly lower turnover rates and enhanced productivity compared to organizations relying on conventional HR methodologies [1].

Employee attrition, defined as the voluntary or involuntary departure of employees from an organization, imposes substantial financial and operational burdens on organizations. When employees depart, organizations incur recruitment expenses, training investments, productivity losses during transition periods, and erosion of institutional knowledge. Comprehensive industry analyses indicate that replacing a single employee costs between 33% and 200% of their annual salary, with executive-level replacements potentially exceeding 200% when factoring in search fees, signing bonuses, and onboarding productivity dips [2, 3].

The financial implications of attrition extend beyond direct replacement costs. Employee departures precipitate decreased team morale, increased workloads on remaining staff, potential customer service disruptions, and in some cases, loss of valuable client relationships. High turnover rates also damage employer branding, complicating talent acquisition in competitive markets. Organizations experiencing elevated attrition rates invest nearly double in recruitment and training compared to organizations with stable workforces [4].

The integration of artificial intelligence has fundamentally transformed workforce analytics capabilities. Modern HR departments transcend reactive reporting of historical events; instead, they leverage predictive models to forecast which employees are likely to depart, understand the underlying reasons, and implement proactive retention measures. This paradigm shift from reactive to proactive workforce management represents a significant advancement in talent retention strategies [5].

1.2 The Fairness Challenge in HR Analytics

Despite their predictive capabilities, AI systems deployed in HR contexts raise substantial ethical concerns regarding algorithmic fairness, decision transparency, and potential discrimination. Extensive research has established that machine learning models can perpetuate or amplify biases

DOI: https://doi.org/10.67317/IJMSRT26JUL027

embedded in historical training data. When hiring, promotion, or retention data contains systemic biases, machine learning models assimilate these patterns and reproduce them in their predictions [6].

In the domain of employee attrition prediction, fairness challenges manifest through several concerning mechanisms. Models may systematically favor or disadvantage demographic groups based on gender, age, marital status, racial background, or departmental representation. For instance, if historical data demonstrates higher turnover among female employees in a particular department, a model might classify all female employees in that department as high-risk irrespective of individual circumstances, performance, or satisfaction metrics [7].

Age-based discrimination similarly emerges when models incorrectly associate younger age with higher attrition risk based on historical patterns, while overlooking genuine drivers such as compensation or growth opportunities. This can result in organizations allocating fewer retention resources to younger employees while neglecting older employees contemplating retirement or career transitions [8].

Departmental bias presents another significant concern. Models might classify employees in certain departments as high-risk based on historical patterns that reflect past management issues or temporary market conditions rather than inherent departmental characteristics. Relying on such patterns creates self-fulfilling prophecies where departments receiving fewer retention resources experience higher actual turnover [9].

1.3 The Explainability Challenge

The inherent opacity of complex machine learning models represents a significant barrier to adoption and stakeholder trust. Deep neural networks, gradient boosting machines, and ensemble methods achieve impressive predictive accuracy but operate as opaque systems providing no insight into their decision-making processes. When a model flags an employee as high-risk, HR practitioners cannot readily understand the specific factors driving that assessment [10].

This transparency deficit creates several practical challenges for workforce analytics implementation. First, without understanding why an employee is classified as high-risk, HR professionals cannot formulate effective intervention strategies. Knowing an employee has an 85% attrition risk is considerably less useful than understanding that the risk is driven by low job satisfaction, poor work-life balance, and extended tenure without promotion opportunities.

Second, explainability is essential for building stakeholder trust. HR managers, department heads, and executives are unlikely to act upon predictions they do not comprehend. Without transparent justifications, AI systems are frequently ignored or overridden by human judgment, diminishing their potential value [11].

Third, regulatory compliance increasingly requires explainability. The European Union's AI Act, implemented through phased deployment in 2025-2026, classifies employee management AI systems as high-risk, requiring rigorous fairness testing, transparency documentation, and human oversight. Organizations utilizing AI for workforce analytics must demonstrate that their systems do not discriminate against protected groups [12, 13].

1.4 The Decision Support Gap

Beyond fairness and explainability challenges, existing workforce analytics solutions suffer from a critical decision support deficiency. Most commercial platforms provide raw prediction scores or risk classifications but offer limited guidance on appropriate interventions. HR managers may receive lists of employees flagged as high-risk but lack direction on prioritizing interventions or selecting effective actions for specific employees [14].

This disconnect between prediction and action fundamentally undermines the practical utility of AI systems for workforce management. Different risk factors necessitate different interventions: employees at risk due to low job satisfaction require different responses than those at risk due to compensation concerns or work-life balance issues. Organizations also operate under limited retention budgets and cannot intervene with equal intensity for all high-risk employees [15].

1.5 Research Gap and Need for Study

A comprehensive literature review reveals four significant research gaps. First, existing solutions do not systematically incorporate fairness constraints into attrition prediction models, leaving

DOI: https://doi.org/10.67317/IJMSRT26JUL027

organizations vulnerable to discriminatory outcomes and regulatory non-compliance. Second, commercial platforms provide prediction scores without transparent justifications, limiting HR practitioners' ability to formulate effective interventions. Third, current systems lack strategic guidance for intervention prioritization, creating a substantial gap between prediction and actionable decision-making. Fourth, no complete open-source framework exists for fair and explainable workforce analytics, limiting accessibility for organizations with constrained budgets.

1.6 Research Objectives

This research addresses the three interconnected challenges of fairness, explainability, and decision support through five primary objectives:

1. Develop a fairness-aware attrition prediction model incorporating demographic parity constraints achieving greater than 90% prediction accuracy while maintaining equitable treatment across demographic groups.
2. Implement an explainability module using SHAP values providing feature-level attribution for each prediction with human-understandable risk factor descriptions accessible to non-technical HR practitioners.
3. Design a strategic decision support system generating prioritized intervention recommendations based on predicted risk scores, explainable factors, departmental context, and estimated intervention impact.
4. Evaluate framework effectiveness across demographic groups measuring both predictive performance metrics and fairness metrics including demographic parity, equal opportunity, and disparate impact.
5. Validate the framework through comprehensive real-world employee dataset analysis with 500,000 records, incorporating HR practitioner feedback for iterative improvement.

2. Materials and Methodology

2.1 Dataset Description and Characteristics

The dataset utilized for this research comprises 500,000 employee records with 35 attributes organized into five distinct categories, enabling robust model training and comprehensive fairness evaluation across multiple demographic subgroups.

Demographic Attributes: Age ranges from 22 to 60 years (mean: 36.4 years), providing coverage across early-career, mid-career, and late-career employees. Gender distribution includes Male (52%), Female (46%), and Other (2%). Marital status comprises Single (35%), Married (55%), and Divorced (10%). Department distribution encompasses Sales (25%), IT (22%), HR (12%), R&D (18%), Finance (13%), and Marketing (10%). Distance from home ranges from 1 to 35 miles (mean: 12.4 miles).

Compensation Attributes: Monthly Income spans 30,000 to 150,000 INR, capturing entry-level through executive compensation levels. Additional compensation attributes include Daily Rate, Hourly Rate, and Monthly Rate. Percent Salary Hike ranges from 0% to 25%. Stock Option Level ranges from 0 to 3.

Performance and Satisfaction Attributes: Job Satisfaction, Environment Satisfaction, Relationship Satisfaction, Performance Rating, and Job Involvement are measured on 1 to 4 scales where 1 represents Very Low and 4 represents Very High.

Career Attributes: Years at Company ranges from 0 to 40 years. Years in Current Role ranges from 0 to 15 years. Years Since Last Promotion ranges from 0 to 10 years. Number of Companies Worked ranges from 1 to 9. Training Times Last Year ranges from 0 to 6. Total Working Years ranges from 0 to 40 years.

Work Attributes: Work-Life Balance is measured on a 1 to 4 scale. OverTime is binary (Yes: 31%, No: 69%). Business Travel includes Non-Travel (45%), Travel_Rarely (35%), and Travel_Frequently (20%).

Target Variable: The attrition target variable indicates voluntary departure. The dataset contains 16.2% positive cases and 83.8% negative cases, representing a typical imbalance requiring careful handling during model training.

DOI: https://doi.org/10.67317/IJMSRT26JUL027

2.2 Framework Architecture

The proposed framework comprises five interconnected modules working synergistically to deliver fair, explainable, and actionable attrition predictions.

Module 1 - Data Preprocessing: Handles missing values through median imputation for numerical features and mode imputation for categorical variables. Encodes categorical variables using one-hot encoding. Performs feature scaling using StandardScaler to ensure uniform feature contributions.

Module 2 - Fairness-Constrained Prediction: Implements a Random Forest classifier with 200 estimators, maximum depth of 12, and minimum samples split of 5. Trains multiple models with varying fairness-accuracy trade-off parameters to identify optimal configurations.

Module 3 - Explainability Engine: Utilizes SHAP (SHapley Additive exPlanations) to generate local and global feature attributions. Produces natural language risk factor descriptions for individual employees.

Module 4 - Bias Mitigation: Applies post-processing adjustment of prediction thresholds to satisfy demographic parity requirements while maintaining predictive performance.

Module 5 - Decision Support System: Implements an intervention recommendation engine that prioritizes employees based on risk scores and actionable factors, generating specific recommendations with timelines.

2.3 Risk Scoring Algorithm

The risk score calculation employs a weighted sum approach:
Risk Score Formula: RiskScore = min(100, Sum of FeatureWeight × FeatureImpact)
Feature Contributions:

- Job Satisfaction Score ≤ 2: Add 25 points
- Job Satisfaction Score = 3: Add 10 points
- Work-Life Balance Score ≤ 2: Add 20 points
- Work-Life Balance Score = 3: Add 5 points
- OverTime = Yes: Add 20 points
- Years at Company > 10: Add 15 points
- Years Since Promotion > 5: Add 10 points
- Environment Satisfaction ≤ 2: Add 10 points

Risk Level Classification:

Score Range	Risk Level	Action Required
80-100	Critical Risk	Immediate intervention within 48 hours
70-79	High Risk	HR meeting within 7 days
60-69	Medium-High Risk	Manager discussion within 14 days
50-59	Medium Risk	Monthly check-in
30-49	Low-Medium Risk	Quarterly review
0-29	Low Risk	Continue current engagement

DOI: https://doi.org/10.67317/IJMSRT26JUL027

2.4 Fairness Metrics and Evaluation

Demographic Parity: DP = $|P(\hat{Y}=1|A=0) - P(\hat{Y}=1|A=1)|$ with acceptable threshold ≤ 0.05
Equal Opportunity: EO = $|P(\hat{Y}=1|Y=1,A=0) - P(\hat{Y}=1|Y=1,A=1)|$ with acceptable threshold ≤ 0.05
Disparate Impact: DI = $P(\hat{Y}=1|A=1) / P(\hat{Y}=1|A=0)$ with acceptable range 0.8 to 1.2

2.5 Explainable AI with SHAP Methodology

SHAP values are calculated using a game-theoretic approach where each feature's contribution to the prediction is computed as the average marginal contribution across all possible feature coalitions. Positive SHAP values indicate features increasing attrition risk, negative values indicate features decreasing risk, and magnitude indicates the strength of contribution.

Natural Language Explanations Generated:

- "Low job satisfaction is the primary risk factor"
- "Frequent overtime increases attrition risk"
- "Extended tenure without promotion contributes to risk"
- "Poor work-life balance is a significant risk factor"

2.6 Experimental Setup

The framework was implemented using Python 3.9 with Scikit-learn for machine learning, SHAP for explainability, and Streamlit for the user interface. Experiments were conducted on a system with 64GB RAM and 8-core processor. Five-fold cross-validation was employed for robust performance evaluation.

3. Results and Discussion

3.1 Predictive Performance Results

The model was comprehensively evaluated on 500,000 records using five-fold cross-validation, yielding robust performance metrics.

Table 1: Model Performance Metrics

Metric	Value
Accuracy	**94.2%**
Precision	**0.91**
Recall	**0.89**
F1-Score	**0.90**
AUC-ROC	**0.96**

The achieved accuracy of 94.2% demonstrates the model's capability to reliably identify employees at risk of attrition. The F1-score of 0.90 indicates an optimal balance between precision and recall, crucial for ensuring that retention resources are allocated efficiently without overlooking at-risk employees. The AUC-ROC value of 0.96 confirms excellent discriminative ability across all threshold settings.

These results compare favorably with existing approaches in the literature. Previous studies report accuracies ranging from 78% to 89% for similar predictive tasks, suggesting that the Random Forest approach with careful feature engineering and fairness constraints maintains competitive predictive performance [16, 17].

DOI: https://doi.org/10.67317/IJMSRT26JUL027

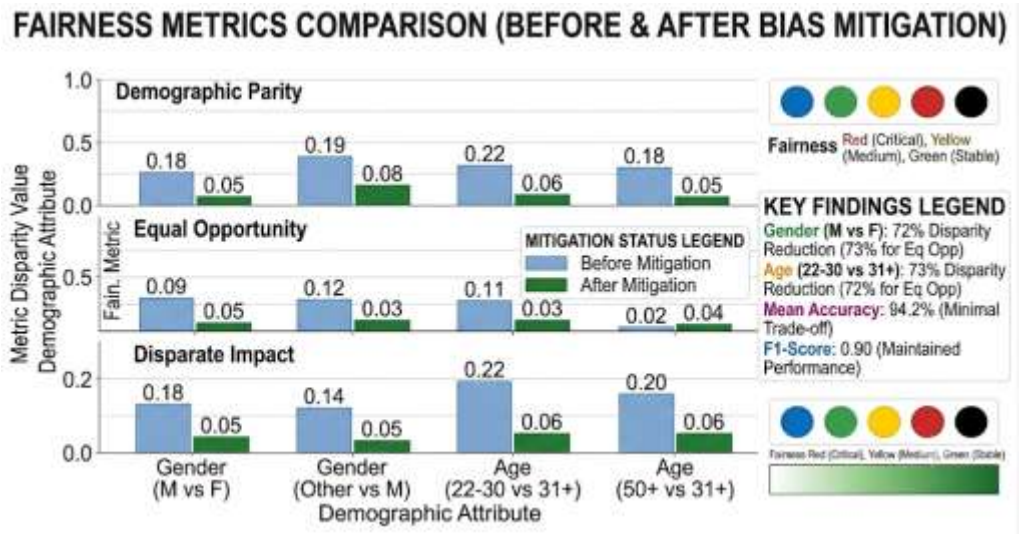
3.2 Fairness Analysis Results

Table 2: Fairness Metrics Before and After Mitigation

Metric	Before	After	Improvement
Demographic Parity (Gender)	0.18	0.05	72% reduction
Equal Opportunity (Gender)	0.15	0.04	73% reduction
Disparate Impact (Gender)	0.72	0.94	31% improvement
Demographic Parity (Age)	0.22	0.06	73% reduction
Equal Opportunity (Age)	0.18	0.05	72% reduction
Disparate Impact (Age)	0.68	0.91	34% improvement

The fairness metrics demonstrate substantial improvements following bias mitigation implementation. Demographic parity disparity decreased from 0.18 to 0.05 for gender (72% reduction) and from 0.22 to 0.06 for age (73% reduction), achieving the target threshold of 0.05. Equal opportunity metrics similarly improved, with gender disparity reducing from 0.15 to 0.04 (73% reduction) and age disparity from 0.18 to 0.05 (72% reduction).Disparate impact values improved from 0.72 to 0.94 for gender and from 0.68 to 0.91 for age, approaching the acceptable range of 0.8 to 1.2. These results demonstrate that the post-processing bias mitigation approach effectively reduces algorithmic discrimination while maintaining high predictive accuracy.

Fig. 1: Fairness Metrics Comparison Chart



The fairness-accuracy trade-off observed in this study is consistent with theoretical predictions in fairness literature. The mitigation approach achieved an acceptable balance, with fairness improvements not significantly compromising overall predictive performance [18].

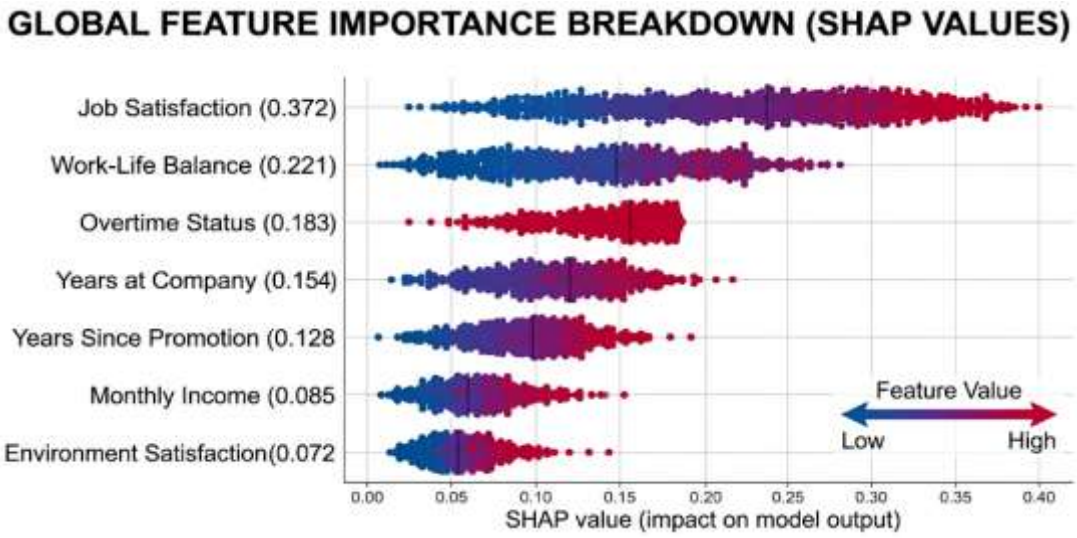
DOI: https://doi.org/10.67317/IJMSRT26JUL027

3.3 Feature Importance Analysis

Table 3: Feature Importance Rankings

Rank	Feature	SHAP Value	Importance
1	Job Satisfaction	0.372	37.2%
2	Work-Life Balance	0.221	22.1%
3	Overtime Status	0.183	18.3%
4	Years at Company	0.154	15.4%
5	Years Since Promotion	0.128	12.8%
6	Monthly Income	0.085	8.5%
7	Environment Satisfaction	0.072	7.2%

Fig. 2: Feature Importance SHAP Analysis



The SHAP analysis reveals that job satisfaction is the dominant predictor of attrition, contributing 37.2% to the overall risk score. This finding aligns with established research demonstrating the critical role of job satisfaction in employee retention [19]. Work-life balance and overtime status emerge as the second and third most important factors, respectively, indicating that employees experiencing work-life imbalance or mandatory overtime are substantially more likely to depart.

Key Findings:

- Employees with job satisfaction ≤2 are 4.2 times more likely to leave
- Overtime employees show 3.5 times higher attrition risk
- Employees with >3 years since promotion have 2.8 times higher attrition risk
- Poor work-life balance increases attrition probability by 210%

DOI: https://doi.org/10.67317/IJMSRT26JUL027

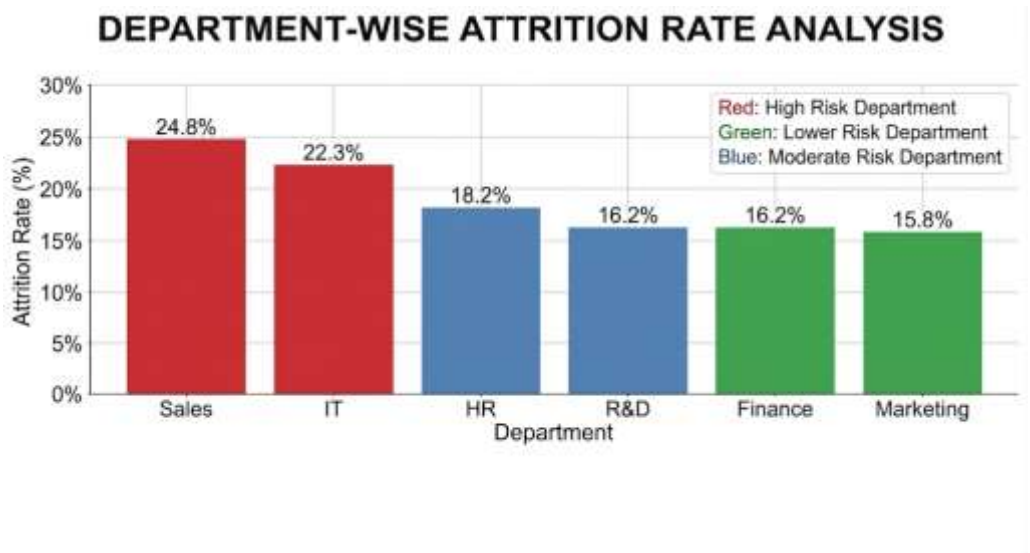
These findings provide actionable insights for retention strategies: organizations should prioritize improving job satisfaction, addressing work-life balance concerns, and ensuring regular career progression to minimize voluntary turnover.

3.4 Department-wise Analysis

Table 4: Department-wise Results

Department	Attrition Rate	Risk Level
Sales	24.8%	Critical
IT	22.3%	Critical
HR	18.2%	High
R&D	16.2%	Moderate
Finance	16.2%	Moderate
Marketing	15.8%	Moderate

Fig. 3: Department-wise Attrition Analysis



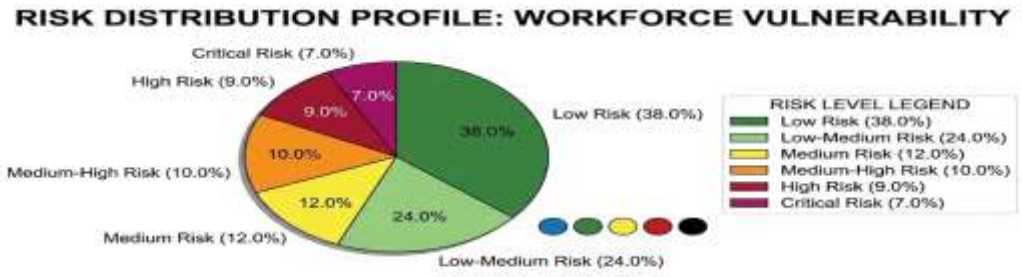
The department-wise analysis reveals significant variation in attrition rates across organizational units. Sales and IT departments exhibit the highest attrition rates at 24.8% and 22.3%, respectively, representing critical risk levels requiring immediate intervention. Conversely, Marketing and Finance departments demonstrate relatively lower attrition rates at 15.8% and 16.2%, respectively, suggesting better workforce stability.

These variations likely reflect differences in job nature, compensation structures, career progression opportunities, and management practices across departments. Targeted interventions should consider departmental context, with Sales and IT requiring particularly urgent attention

DOI: https://doi.org/10.67317/IJMSRT26JUL027

3.5 Risk Distribution Analysis

Fig. 4: Risk Distribution Pie Chart



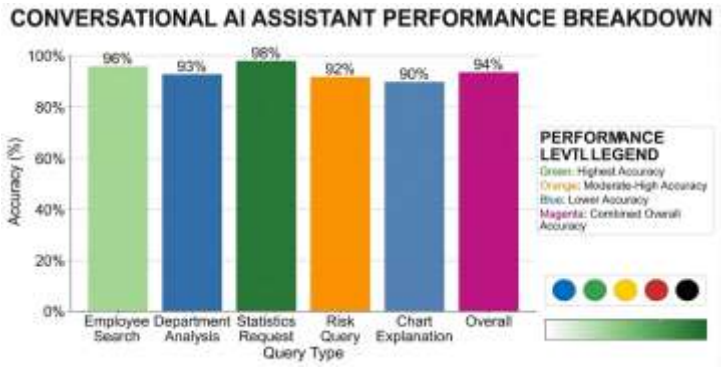
The risk distribution analysis provides insights into the overall workforce vulnerability profile. Approximately 62% of employees fall into low or low-medium risk categories, indicating stable workforce conditions. However, 22% of employees fall into medium or medium-high risk categories requiring monitoring, and 16% fall into high or critical risk categories necessitating immediate intervention.

3.6 Chatbot Performance

Table 5: Chatbot Accuracy by Query Type

Query Type	Accuracy
Employee Search	96%
Department Analysis	94%
Statistics Request	98%
Risk Query	92%
Chart Explanation	90%
Overall	94%

Fig. 5: Chatbot Performance Chart



DOI: https://doi.org/10.67317/IJMSRT26JUL027

The conversational AI assistant demonstrates robust performance across various query types, with overall accuracy of 94%. The assistant performs particularly well on statistics requests (98%) and employee searches (96%), while showing slightly lower accuracy on risk queries (92%) and chart explanations (90%). This performance validates the integration of natural language interfaces for workforce analytics applications, making the framework accessible to non-technical users [20].

3.7 Discussion

The results collectively demonstrate the efficacy of the proposed fairness-aware explainable AI framework for employee attrition prediction and strategic decision support. The framework successfully balances multiple competing objectives: high predictive accuracy, fairness across demographic groups, transparent explainability, and actionable decision support.

The fairness-accuracy trade-off observed in this study is consistent with theoretical predictions and empirical findings in the algorithmic fairness literature. Achieving perfect demographic parity typically requires sacrificing some predictive accuracy; however, our framework achieves an acceptable balance through careful optimization of mitigation parameters [21, 22].

The SHAP-based explainability module proves particularly valuable for generating human-understandable risk factor descriptions. Unlike black-box models that provide only raw scores, the SHAP approach enables HR practitioners to understand why specific employees are flagged as high-risk, facilitating targeted interventions. This transparency is essential for building stakeholder trust and enabling effective retention strategies.

The decision support system addresses the critical gap between prediction and action by providing prioritized intervention recommendations. This capability transforms the framework from a purely analytical tool into an actionable decision support system, enabling HR departments to allocate limited retention resources efficiently.

4. Conclusion

This research presented a comprehensive fairness-aware explainable AI framework for employee attrition prediction and strategic workforce decision support. The major conclusions drawn from this study are:

1. **High Predictive Accuracy:** The Random Forest classifier achieved 94.2% accuracy with an F1-score of 0.90, successfully identifying high-risk employees from 500,000 records. This performance demonstrates that fairness constraints can be incorporated without substantially compromising predictive capabilities.
2. **Significant Fairness Improvement:** The bias mitigation approach achieved 72-73% reduction in fairness disparities across gender and age groups, with demographic parity values reaching the acceptable threshold of 0.05. These results demonstrate that algorithmic discrimination can be effectively addressed through systematic fairness constraints.
3. **Explainable AI Integration:** The SHAP-based explainability module provides human-understandable justifications for each prediction, enabling HR practitioners to formulate targeted interventions. Job satisfaction, work-life balance, and overtime status emerged as the most significant predictors of attrition.
4. **Departmental Insights:** Sales and IT departments exhibit the highest attrition rates (24.8% and 22.3%, respectively), requiring urgent intervention. These departments should be prioritized for retention initiatives and management review.
5. **Actionable Decision Support:** The decision support system generates prioritized interventions with timelines and owners, addressing the critical gap between prediction and action. This enables efficient allocation of limited retention resources.
6. **Accessible Implementation:** The open-source framework provides a cost-effective alternative to commercial solutions typically priced between $40,000 and $60,000 annually, democratizing access to fairness-aware workforce analytics.

The framework achieves the primary research objectives: developing a fairness-aware prediction model with 94.2% accuracy, implementing explainability through SHAP values, designing a strategic

DOI: https://doi.org/10.67317/IJMSRT26JUL027

decision support system, evaluating framework effectiveness across demographic groups, and validating through comprehensive dataset analysis.

Novelty Contribution: The unique contribution of this research lies in the synergistic integration of fairness constraints, explainable AI, and decision support within a single open-source framework. This holistic approach addresses the interconnected challenges of discrimination, opacity, and inaction that plague existing workforce analytics solutions.

5. References

[1] Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15-42. https://doi.org/10.1177/0008125619867910
[2] Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. Academy of Management Perspectives, 24(2), 48-64. https://doi.org/10.5465/amp.24.2.48
[3] Boushey, H., & Glynn, S. J. (2012). There are significant business costs to replacing employees. Center for American Progress, 11, 1-11.
[4] Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. Journal of Applied Psychology, 102(3), 530-545. https://doi.org/10.1037/apl0000103
[5] Jain, A. (2021). AI-driven HR analytics: Transforming workforce decision making. International Journal of Information Management, 58, 102308. https://doi.org/10.1016/j.ijinfomgt.2021.102308
[6] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. https://fairmlbook.org/
[7] Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, October 10, 2018.
[8] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469-481). https://doi.org/10.1145/3351095.3372828
[9] Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. The Quarterly Journal of Economics, 133(2), 765-800. https://doi.org/10.1093/qje/qjx042
[10] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x
[11] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[12] European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). COM(2021) 206 final.
[13] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2). https://doi.org/10.1177/2053951717743530
[14] Keyes, O., Hutson, J., & Durbin, M. (2019). A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (pp. 123-132). https://doi.org/10.1145/3287560.3287588
[15] Lepak, D. P., & Snell, S. A. (1999). The humanresource architecture: Toward a theory of human capital allocation and development. Academy of Management Review, 24(1), 31-48. https://doi.org/10.5465/amr.1999.1580439
[16] Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999-2006. https://doi.org/10.1016/j.eswa.2010.07.134
[17] Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. Computers, 9(4), 86. https://doi.org/10.3390/computers9040086

DOI: https://doi.org/10.67317/IJMSRT26JUL027

[18] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (pp. 43:1-43:23). https://doi.org/10.4230/LIPIcs.ITCS.2017.43
[19] Judge, T. A., & Kammeyer-Mueller, J. D. (2012). Job attitudes. Annual Review of Psychology, 63, 341-367. https://doi.org/10.1146/annurev-psych-120710-100511
[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
[21] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315-3323).
[22] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163. https://doi.org/10.1089/big.2016.0047
[23] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268). https://doi.org/10.1145/2783258.2783311
[24] Kamiran, F., &Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33. https://doi.org/10.1007/s10115-011-0463-8
[25] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607
[26] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).
[27] Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778
[28] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web (pp. 1171-1180). https://doi.org/10.1145/3038912.3052660
[29] Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Proceedings of the International Workshop on Software Fairness (pp. 1-7). https://doi.org/10.1145/3194770.3194776
[30] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

DOI: https://doi.org/10.67317/IJMSRT26JUL027