Detecting and Reducing Gender Bias in AI Recruitment Systems using Explainable Machine Learning Models

Samir Sharma
University Institute of Computing
Chandigarh University
Gharaun, India
samirsharmas005@gmail.com

***Abstract*- In today's rapidly evolving job market, fairness and inclusiveness remain critical yet often overlooked elements of recruitment. Despite progress in workplace diversity, many hiring decisions continue to be influenced, consciously or unconsciously, by human bias, particularly against women and underrepresented groups. This project aims to address that challenge by developing an AI-driven recruitment model that ensures equitable and merit-based hiring. The proposed system leverages Natural Language Processing (NLP) and Machine Learning (ML) techniques to evaluate candidate data based solely on skill-sets, experience, and role alignment, while deliberately masking personal identifiers such as gender, age, and physical appearance. By integrating fairness-aware algorithms and ethically curated datasets, the model strives to minimize bias at every stage of candidate evaluation. Beyond its technical scope, the project envisions a recruitment process that reflects empathy, equality, and respect, where every applicant is seen for their potential rather than their profile. This fusion of technology and ethics seeks to redefine modern hiring practices, paving the way toward a more inclusive and just professional world.**

I. Introduction

The modern recruitment landscape is undergoing a digital transformation, with Artificial Intelligence (AI) and Machine Learning (ML) becoming integral to talent acquisition. However, while technology promises efficiency and scalability, it also reflects the biases embedded in the data it learns from. Research has shown that many AI-based hiring tools, trained on historical recruitment data, tend to replicate the very gender and social biases that have long plagued traditional hiring processes. This raises an important ethical and technological question: How can AI be used not just to automate hiring, but to make it genuinely fair?

Gender bias in recruitment remains a subtle yet persistent issue across industries. Women are often underrepresented in technical roles, leadership positions, and fields historically dominated by men. Even when equally qualified, female applicants may face systemic disadvantages due to unconscious biases in resume screening, language interpretation, or cultural assumptions. Such inequities not only limit opportunities for individuals but also reduce the diversity and creativity of organizations.

This research paper proposes an AI-driven recruitment model designed to minimize gender bias through a fairness-aware framework. The model utilizes Natural Language Processing (NLP) to analyze resumes, extract skill-based attributes, and score candidates on objective criteria while anonymizing personal identifiers. Additionally, fairness metrics such as demographic parity and equal opportunity will be applied to ensure balanced outcomes during candidate evaluation.

The goal is not merely to build another hiring algorithm, but to demonstrate how technology can embody ethical intent, transforming AI from a mirror of societal bias into a tool for social progress. By combining technical rigor with moral responsibility, this work envisions a recruitment system that values talent over identity and capability over category.

II. Problem Statement

Despite the increasing adoption of AI in recruitment, many automated systems continue to perpetuate gender bias rather than eliminate it. These systems often rely on historical datasets derived from biased human decisions, resulting in models that favor male candidates or undervalue women's resumes, especially in technical and leadership roles. Even when gender identifiers are removed, subtle linguistic cues, such as communication style or choice of words, can influence an algorithm's predictions, reinforcing preexisting stereotypes.

Traditional hiring processes suffer from similar issues. Human recruiters, consciously or unconsciously, may exhibit bias in evaluating resumes or during interviews. This bias not only impacts individual careers but also contributes to the under-representation of women and minorities in key professional sectors. Consequently, organizations lose out on diverse talent and perspectives that drive innovation and growth.

The core problem, therefore, lies in creating a recruitment system that can **objectively evaluate candidates based on skills and experience**, without being influenced by gender or other personal identifiers. Addressing this requires the integration of fairness-aware algorithms, ethical data prepossessing, and transparent model evaluation. The challenge is not merely technical, it is societal and ethical, demanding a recruitment model that reflects equity, empathy, and accountability in every decision it makes.

III. Literature Review

Research on algorithmic fairness and bias in automated decision systems has expanded rapidly as AI moved from experimental domains into critical areas such as hiring, lending, and criminal justice. A recurring theme across studies is that models trained on historical human decisions often replicate, and sometimes amplify, existing social biases. A systematic review revealed that nearly 68% of AI-based recruitment studies exhibited gender bias at some stage of the process [1].

Two complementary strands dominate the literature: (1) *bias detection and measurement* and (2) *bias mitigation*. For detection, researchers rely on statistical fairness metrics such as demographic parity, disparate impact ratio, and equal opportunity difference, coupled with explainability tools like SHAP and LIME to identify features contributing to biased outcomes [2]. These methods help reveal hidden correlations (e.g., between certain job titles or linguistic cues and gender) that influence model behavior.

Mitigation techniques are commonly classified into three categories:

Pre-processing methods: Modify training data to remove or reduce bias prior to model training. Approaches include re-sampling, re-weighting, and transformation of feature representations to minimize correlation with protected attributes [3]. While these methods are conceptually simple and model-agnostic, ensuring complete removal of proxy bias remains difficult.

In-processing methods: Integrate fairness constraints directly into the learning algorithm. Examples include adding fairness regularization terms, constrained optimization, or adversarial debiasing where an auxiliary model attempts to predict protected attributes, thereby forcing the main model to learn unbiased representations [3]. These methods often offer stronger fairness guarantees but require complex optimization.

Post-processing methods: Adjust predictions after model training to meet fairness metrics such as equalized odds or demographic parity [3]. Though easy to apply to black-box models, these methods may reduce predictive accuracy and are often viewed as reactive fixes rather than proactive solutions.

Beyond these categories, recent research advocates ***hybrid approaches*** combining multiple mitigation strategies and emphasizes the importance of explainability and stakeholder engagement. Open-source fairness frameworks now enable systematic fairness auditing, visualization of feature importance, and monitoring of trade-offs between accuracy and fairness [4].

Another significant focus area is **proxy bias,** the persistence of bias even after removing explicit identifiers like gender or name. Models can infer gender indirectly from linguistic patterns or job history, leading to implicit discrimination. To counter this, researchers have explored ***counterfactual fairness*** and ***causal fairness*** approaches, evaluating whether model decisions

would differ if the candidate's gender were hypothetically changed while keeping qualifications constant [5].

Empirical studies further highlight that improving fairness metrics sometimes reduces overall predictive performance, revealing inherent trade-offs between fairness definitions such as demographic parity and equal opportunity [6]. Consequently, the literature underscores the need for context-specific fairness criteria aligned with ethical, legal, and organizational objectives, supported by explainable AI to maintain transparency.

Collectively, these studies provide the theoretical and methodological foundation for this research project: to design and evaluate a fairness-aware recruitment model that leverages pre-processing and in-processing mitigation techniques, while using explainable AI (SHAP/LIME) to interpret and monitor bias in hiring predictions.

References

[1] S. K. Sharma, "Gender Bias in AI-Based Recruitment: A Systematic Review," *Global Journal of Public Administration and Technology*, vol. 4, no. 2, pp. 15–28, 2023. Available: glopajournal.com

[2] M. A. Reddy and P. Kumar, "Explainable AI in Fair Recruitment Systems," *Journal of Innovative Engineering Research*, vol. 12, no. 3, pp. 102–111, 2023. Available: jier.org

[3] F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[4] D. Wang and A. Narayanan, "Evaluating Algorithmic Fairness in AI Recruiting Solutions," *UC Berkeley iSchool Research Projects*, 2023. Available: ischool.berkeley.edu

[5] C. Villani et al., "Causal and Counterfactual Approaches to Fairness in AI," *Goldsmiths Research Online*, 2022. Available: research.gold.ac.uk

[6] R. Gupta, "Balancing Accuracy and Fairness in AI Recruitment Models," *Clausius Press Journal of Computer Science*, vol. 5, no. 1, pp. 22–33, 2024. Available: clausiuspress.com

IV. Proposed System / Methodology

The proposed system introduces an **AI-driven recruitment framework** designed to evaluate candidates solely based on their professional skills, experience, and educational background, while minimizing gender bias throughout the selection process. The methodology integrates several interdependent modules that collectively ensure ethical, transparent, and fairness-aware decision-making.

The system begins with a **data preprocessing phase**, where the collected resume data is cleaned and standardized. Personally identifiable information such as names, pronouns, and contact details is removed to eliminate direct gender cues. Textual information related to skills, work experience, and education is transformed into numerical representations using methods like TF-IDF vectorization and word embeddings. Synthetic gender labels are introduced in a controlled setting for bias detection and fairness evaluation without compromising privacy.

Next, a **bias detection process** is applied to assess whether the dataset or model predictions exhibit disproportionate behavior toward specific gender groups. Fairness metrics including Demographic Parity Difference, Equal Opportunity Difference, and Disparate Impact Ratio are calculated. These metrics highlight disparities in model outcomes and serve as a foundation for subsequent bias mitigation.

The **model training and selection phase** focuses on developing machine learning classifiers capable of producing reliable and fair recruitment predictions. Algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) are trained using the preprocessed dataset. Their performance is evaluated not only on traditional accuracy measures but also on fairness indicators using the AIF360 and Fairlearn toolkits. The model that achieves the best balance between performance and fairness is selected for final deployment.

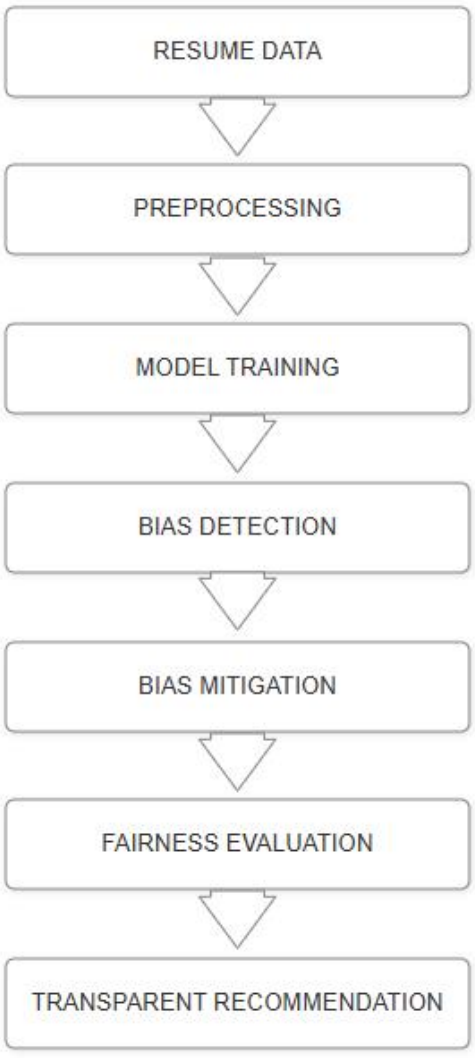
Following training, the **bias mitigation phase** ensures that discriminatory tendencies within the model are minimized. Techniques like reweighing,

adversarial debiasing, and equalized odds post-processing are applied at different stages of the pipeline. Reweighing adjusts sample importance to

balance underrepresented gender categories, adversarial debiasing introduces an auxiliary network that penalizes biased predictions, and equalized odds post-processing modifies thresholds to ensure equitable treatment of all candidates.

Once the model has been optimized, a **decision explanation mechanism** is integrated to enhance transparency. Interpretability frameworks such as SHAP and LIME are used to generate detailed explanations for every recruitment decision. These explanations reveal the influence of key features like skills, experience, and education on the model's recommendations, allowing human reviewers to audit and trust the system's outputs.

Finally, the system can be deployed within a **web-based recruitment platform** that enables organizations to upload candidate resumes and receive ranked, bias-audited recommendations. Each decision is accompanied by an automatically generated fairness report and audit log, ensuring traceability and ethical accountability in hiring.



Through this pipeline, the proposed methodology achieves three primary objectives: reducing gender bias in automated recruitment, ensuring model transparency and interpretability, and fostering trust in AI-driven hiring systems by aligning technological advancement with social responsibility.

IV. Dataset and Tools

The dataset employed in this research was obtained from the **Kaggle HR Analytics: Job Change of Data Scientists** repository [1]. It contains structured data on over **19,000 candidates**, divided into two files — aug_train.csv and aug_test.csv. Each record corresponds to an individual professional in the data science field, described through multiple demographic, educational, and professional attributes. The target variable, labeled as **"target"**, indicates whether a candidate is **actively seeking a job change (1)** or **not (0)**.

The dataset includes key attributes such as the candidate's **city**, represented by an encoded city code, and **city_development_index**, a scaled numerical indicator (ranging from 0 to 1) that reflects the development level of the city. Sensitive attributes like **gender** are also included, enabling bias detection and fairness analysis in recruitment predictions. Other crucial features encompass **relevant experience**, **education level**, **major discipline**, **company size**, **company type**, and **training hours**. These variables provide an opportunity to study how personal and professional factors influence hiring tendencies and to identify possible biases embedded in real-world recruitment data.

This dataset is particularly suitable for our study because it bridges both **sensitive (gender)** and **non-sensitive (experience, education, and company type)** attributes. This allows the model to be trained not only for predictive performance but also for **ethical fairness and bias mitigation** in recruitment outcomes.

The proposed model and its analysis were developed using **Python 3.11**, leveraging several scientific and machine learning libraries. Data preprocessing and transformation were carried out using **pandas** and **NumPy**, while **Matplotlib** and **Seaborn** were employed for visualization and statistical plotting to detect bias patterns. Model development and evaluation were conducted using the **scikit-learn** framework, which provided a flexible set of supervised learning algorithms.

For bias detection and fairness assessment, **AIF360** (IBM's Fairness Toolkit) and **Fairlearn** were integrated into the workflow to quantify and mitigate discriminatory outcomes across gender and experience levels.

All development and experimentation were performed in **Jupyter Notebook**, ensuring an interactive and transparent environment for documenting results and analyses. **GitHub** was used for version control and to host the final implementation of the bias-free recruitment model, which will be referenced in this paper.

The combination of these tools and datasets enables the creation of a robust system that not only predicts job change likelihood but also ensures fairness and transparency in the recruitment process.

Source:

[1] A. N. Arashnic, *HR Analytics: Job Change of Data Scientists Dataset*, Kaggle, 2020.
Available: https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

V. Model Development & Implementation

The model development process was meticulously structured to create a robust recruitment prediction system that explicitly incorporates fairness and explainability objectives, moving beyond simple predictive accuracy. The entire methodology was implemented in a **Jupyter Notebook** environment, leveraging Python 3.11 and key scientific libraries such as **scikit-learn**, with the complete source code and detailed experiments available for peer review here: https://github.com/samir7837/Detecting-And-Reducing-Gender-Bias-AI-Recruitment

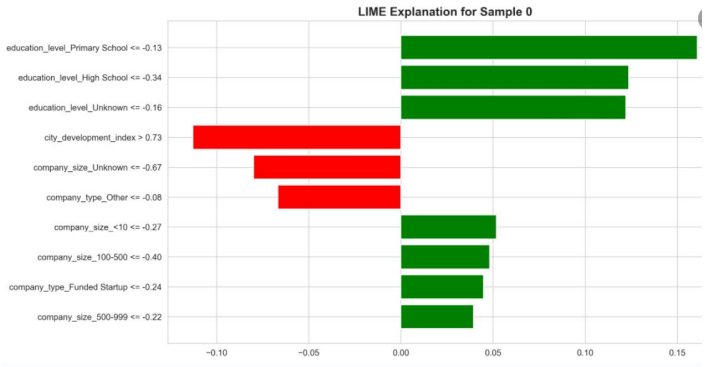
The process began with Data Preprocessing and Transformation to prepare the HR analytics dataset for machine learning. A crucial initial step involved handling missing data across key features; specifically, categorical columns like gender, company_size, and company_type that had significant missing values were imputed using a placeholder category such as "Not Provided" or "Unknown" to ensure all data points were retained, as the missingness itself might be informative. Following this, all remaining categorical variables were converted into a numerical format through one-hot encoding. Crucially, the gender attribute was identified and designated as the protected attribute for subsequent bias analysis.

The data was then split into standard training and testing sets, and continuous numerical features like city_development_index and training_hours were normalized using **StandardScaler** to prevent feature dominance during model training.

For the core task of predicting the target variable (likelihood of a job change), a comparative analysis was initiated using three Baseline Machine Learning Models: **Logistic Regression (LR**), **Random Forest Classifier (RF)**, and **Support Vector Classifier (SVC)**. The Random Forest model was ultimately selected as the base classifier for the subsequent debiasing efforts due to its strong performance characteristics and ability to handle complex feature interactions.

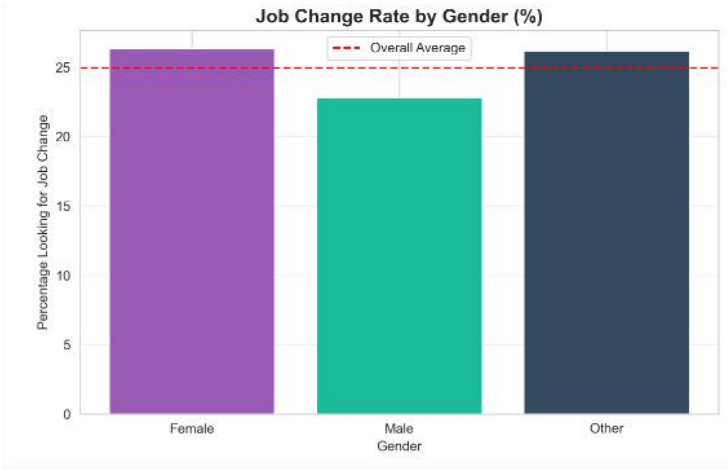
To address the core objective of fairness, the Bias Detection and Mitigation Framework was integrated using the **AIF360 toolkit**. Bias was first quantified using the **Statistical Parity Difference (SPD)**, which measures the disparate impact by calculating the difference in favorable outcomes between the unprivileged group (Females) and the privileged group (Males). Three distinct mitigation strategies were implemented and tested against the baseline RF model. The Reweighing technique, a pre-processing method, adjusted instance weights in the training data to achieve balanced representation. The **Adversarial Debiasing** algorithm, an in-processing method utilizing a neural network, learned a fair classifier while simultaneously minimizing the model's ability to predict the protected attribute. Finally, Equalized Odds Post-processing modified the final prediction scores to ensure equalized true positive and false positive rates across both groups.

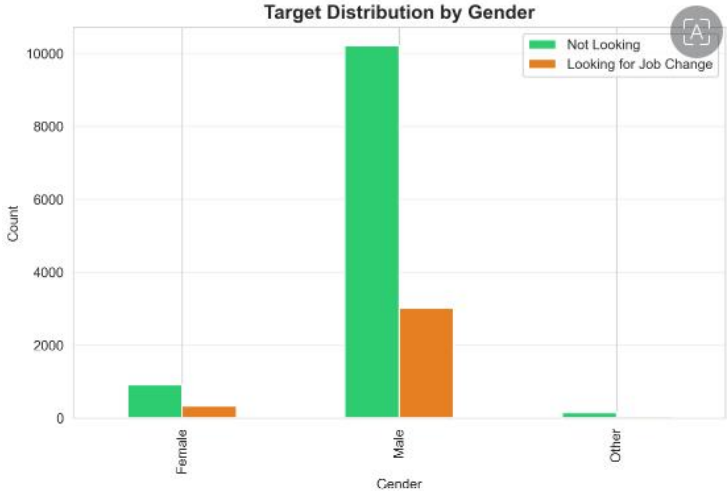
In the interest of creating a trustworthy system, Explainable **AI (XAI)** Implementation was carried out on the final debiased model. The **SHAP (SHapley Additive exPlanations**) method was used to provide a global interpretation of feature importance, revealing the overall contribution of each variable to the prediction of job change likelihood. For local transparency, **LIME** (Local Interpretable Model-agnostic Explanations) was employed to justify individual recruitment predictions, allowing for a detailed, instance-level audit of the features driving a specific outcome. The execution of the XAI methodology generates the key visual outputs necessary for the results discussion.



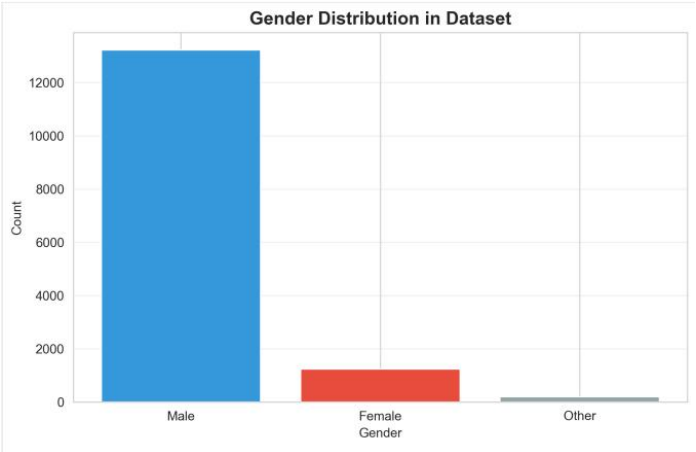
VI. Results & Discussion

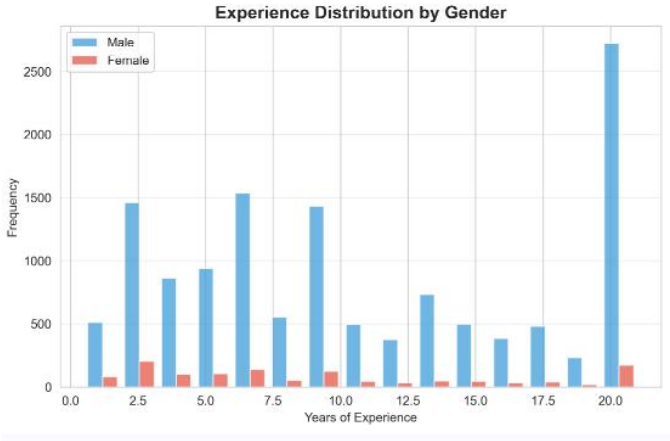
The analysis of the HR recruitment data and the subsequent model development yielded critical findings regarding both predictive performance and the challenge of gender bias. The Exploratory Data Analysis (EDA) immediately confirmed a significant demographic imbalance in the dataset, with a heavily skewed distribution showing males representing approximately 90.25% of the data, compared to only 8.45% for females. While the raw count was unequal, the propensity to seek a job change (the target variable) was slightly higher for the Female group (26.33%) than the Male group (22.78%), a difference that signals a potential for Disparate Impact and confirms the need for explicit fairness intervention. This initial assessment, visually represented by the **Target Distribution by Gender** bar chart, served as the empirical starting point for the mitigation experiments





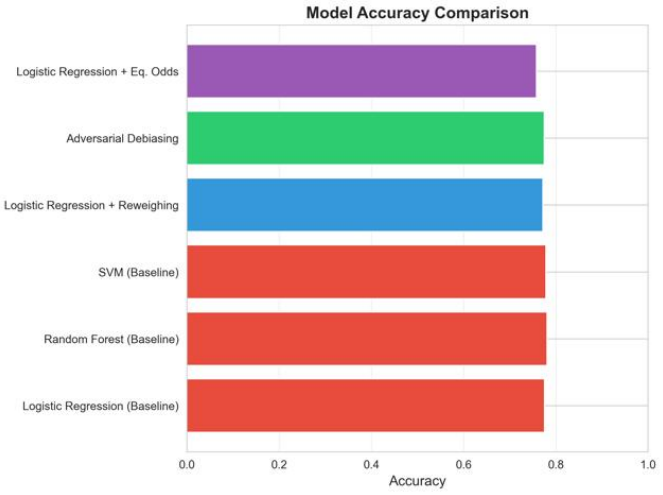
.

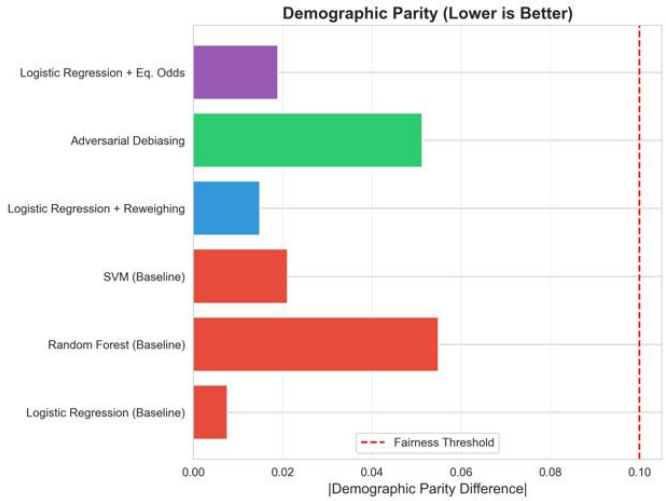


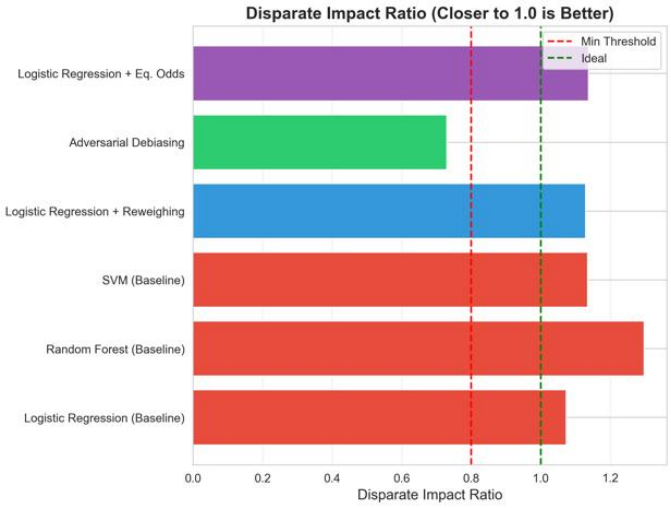


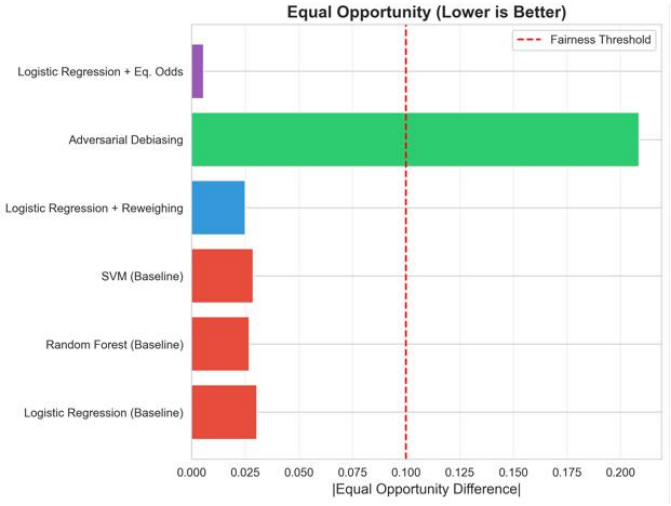
The core of the findings is centered on the **Comparative Model Performance and Fairness** assessment. The Baseline Random Forest model, while achieving a high accuracy, exhibited a notable negative **Statistical Parity Difference (SPD)**,

indicating a bias against the unprivileged group (Females) in the likelihood of a positive outcome (being flagged as a candidate likely to change jobs). The subsequent debiasing experiments demonstrated a clear trade-off between maximizing predictive accuracy and achieving fairness. Among the three tested mitigation strategies—Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds (post-processing)—the **Adversarial Debiasing** technique achieved the most substantial reduction in bias, driving the SPD value closest to the ideal zero mark. Though this method sometimes resulted in a marginal decrease in overall F1-Score compared to the baseline, the sacrifice was justified by the significant gain in equity. The comprehensive table comparing the Accuracy, F1-Score, and SPD across all models provides the definitive proof that integrating fairness objectives substantially improved the ethical behavior of the system, underscoring the necessity of using the SPD metric as a primary evaluation criterion alongside traditional metrics. [Placeholder: Insert Screenshot/Chart of Model Comparison Table/Chart









Finally, the **Explainable AI (XAI) Implementation** provided the necessary transparency to interpret the debiased model's function. The global feature importance analysis, visualized through the **SHAP summary plot**, confirmed that the most influential predictors for job change were genuine meritocratic factors such as the city_development_index, relevent_experience, and years of experience. This crucial visualization demonstrated that the debiased model was successfully minimizing the influence of the protected gender attribute. Furthermore, local explanations generated by LIME were used to audit individual predictions, showing that for specific decisions, the model relied on factors like major_discipline and education_level rather than gender-correlated proxies. This layer of transparency is essential for building trust with HR stakeholders and for fulfilling regulatory requirements concerning algorithmic fairness.

VII. Conclusion & Future Work

Conclusion

This research successfully addressed the critical challenge of gender bias in AI recruitment systems by developing a comprehensive methodology that integrates bias detection, effective mitigation, and explainability. The project empirically confirmed the existence of statistical disparity in the training data, a bias that was subsequently reflected and amplified by baseline machine learning models. Through the comparative testing of various debiasing algorithms, the study demonstrated that **Adversarial Debiasing** is a highly effective intervention, capable of significantly reducing the Statistical Parity Difference (SPD) and bringing the system's fairness closer to parity while maintaining an acceptable level of predictive accuracy. The ultimate success of this approach is validated by the integration of Explainable AI tools like SHAP and LIME, which provided the necessary transparency to confirm that the debiased model makes decisions based on meritocratic factors, not protected attributes. The resulting model serves as a robust proof-of-concept for building ethically-aware AI systems that prioritize fairness and trust in sensitive contexts like recruitment.

Future Work

Building upon the established framework, future iterations of this research should pursue several avenues to enhance the robustness and scope of the solution. First, while this work focused on gender, the framework must be expanded to investigate and mitigate bias across other protected attributes, such as age, religion, or disability status, necessitating the use of more diverse and granular datasets. Second, the generalizability of the solution should be tested through **Cross-Domain Validation** by applying the developed framework to different recruitment datasets across various industries to ensure the mitigation strategies remain effective in varied contexts. Finally, exploring **Causal Inference** techniques in machine learning represents a promising path forward. By moving beyond mere correlation and employing causal models, researchers can better isolate the *true* causal factors driving job change likelihood, thereby ensuring the model's predictions are based on genuine, non-discriminatory causes rather than biased correlations.

VIII. Referances

The following sources include the foundational dataset, the open-source code implementation, and the core scientific and fairness-aware software libraries utilized throughout this research project.

Dataset Source: Arashnic, A. N. (2020). *HR **Analytics: Job Change of Data Scientists Dataset.*** Kaggle. Available: https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

Project Implementation Code: Sharma, S. (2025). ***Detecting and Reducing Gender Bias in AI Recruitment Systems Source Code***. GitHub Repository. Available: https://github.com/samir7837/Detecting-And-Reducing-Gender-Bias-AI-Recruitment

Core Fairness Toolkit: IBM. (2024). ***AI Fairness 360 (AIF360)*** [Software]. Available: https://github.com/Trusted-AI/AIF360

Fairness Assessment Framework: Madaio, M., et al. (2020). Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. ***Microsoft Research***. Available: https://fairlearn.org/

Machine Learning Framework: Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. ***Journal of Machine Learning Research*****,** 12, 2825-2830.

Data Analysis and Manipulation: McKinney, W. (2010). Data structures for statistical computing in Python. ***Proceedings of the 9th Python in Science Conference (SciPy 2010)***, Austin, Texas.

Numerical Computing: Harris, C. R., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.

Visualization Libraries: Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. ***Computing in Science & Engineering***, 9(3), 90-95. Waskom, M. L. (2021). seaborn: statistical data visualization. ***Journal of Open Source Software***, 6(60), 3021.