Resume Bias Detection and Mitigation using AI/ML

Disha Barapatre[1], Anushka Adak[2], Aarya Nalawade[3], Gargi Pawar[4], Prof. Sayali Joshi[5]
[1,2,3,4,5]*Marathwada Mitra Mandal's College of Engineering,Pune*

***Abstract*—Automated resume screening software, although highly efficient, often perpetuates historical biases associated with gender, geography, and educational backgrounds. The current work develops a stage fairness pipeline framework involving detection, mitigation through several techniques, and audit based interpretability. Using four empirical studies' results, this paper shows that adversarial debiasing outperforms any other form of debiased pre-processing as it yields DPD/EOD equal to 0 without any loss of predictive accuracy. We introduce two new concepts Checkpoint and Automated Bias Audit Report as part of the paper. Intersectional bias is recognized as a major overlooked issue in algorithmic discrimination in recruitment practices.**

***Index Terms*— AI Bias, Recruitment Algorithms, Disparate Impact, Algorithmic Fairness, Machine Learning Ethics, HR Technology**

I. INTRODUCTION

If an eligible engineer graduated from a regional institution and applies to fifty positions, her resume does not even pass the first round of automated filtering - not because of insufficient professional skills but due to algorithmic prejudice built in the program. This particular story is far from unique; such instances are becoming increasingly common in various sectors of economy and highlighting the core paradox of using AI in recruitment process. Today, automatic resume filtering programs are applied in more than 75% of large corporations. Machine learning algorithms analyze patterns found in the data collected during previous hiring and identify preferred candidates based on their technical skillset, elite education, or geographical location.. When bias is algorithmic, the accountability gap is severe because neither the recruiter nor the candidate is usually aware of it.

The three coordinated stages of bias quantification using standardized fairness metrics, multi-technique debiasing with comparative evaluation, and SHAP-based interpretability with automated audit reporting are all integrated in this paper's structured detection and mitigation framework. An architectural requirement that was missing from earlier work is the inclusion of a mandatory checkpoint between mitigation and deployment.

1.1 Research Objectives

The study pursues four primary objectives:

• Using DIR, DPD, and EOD metrics calculated for each job category and proxy group, create a methodical bias detection protocol.
• To determine the most reliable strategy, assess and contrast preprocessing, in processing, and post processing mitigation techniques.

• Create a pipeline that incorporates interpretability, mitigation, and detection into a continuous loop.

• To create a repeatable and methodical framework for bias assessment that businesses can use and implement regularly, as well as to offer workable tactics and recommendations to lessen bias in AI-powered hiring systems.

1.2 Scope and Objectives

The identification and reduction of algorithmic and human introduced bias in resume evaluation procedures across both conventional and artificial intelligence-powered hiring methods. The study looks at discriminatory trends that arise from non job related characteristics like gender, age, race, ethnicity, and educational background and that have a disproportionate impact on hiring decisions during the shortlisting process. Establishing a merit based framework for candidate evaluation that adheres to the values of equity, diversity, and inclusion throughout the hiring process is the main goal. Beyond reducing bias, the study aims to improve hiring practices' dependability, openness, and moral integrity by making sure they comply with current equal

opportunity and anti-discrimination laws. Another contribution is the creation of explainable, white box recruitment models that assess candidates solely on the basis of their verified credentials, relevant work experience, and demonstrated competencies completely unrelated to personal or demographic traits that have no bearing on job performance.

II. LITERATURE REVIEW

The stages of the pipeline are covered by mitigation techniques. In empirical evaluations, pre-processing techniques like Disparate Impact Remover and Reweighing consistently perform poorly, leaving bias metrics essentially unchanged. By introducing an adversarial network to prevent sensitive attribute recovery from model representations, Adversarial Debiasing (Zhang et al., 2018) is an in-processing technique that maintains 88.15% accuracy while achieving DPD = 0 and EOD = 0. According to SHAP analysis, post-debiasing models move their emphasis from demographic proxies to reliable predictors like professional experience and educational attainment. There is ample documentation of the transition from manual screening to ML driven systems Roy et al. (2020) utilized Linear SVM for presenting the phase classification and recommendations system with 78.53% precision rate; Roy et al., (2025) further improved on the technique, incorporating BERT embeddings to enhance semantics of job-resume matching. These innovations help to enhance productivity, yet at the same time raise concerns of bias. There is more than one source of biases present. Mehta (2025), analyzing actual resume-to-job matching, found that graduates from elite universities had advantage in scores ranging between 15-20% higher. Furthermore, non-technical skills profiles demonstrated DIs below 0.62 level when set at 0.7 selection threshold, meaning that the selection was under legal requirement for disparate impact. Having worked with 70,000 applicants in total, Mishra (2025) identified country-of-origin bias in classifier leading to 1.0 DPD and EOD rates.

Approach	Method	Application	Limitation
Survey Based Analysis	Collaborative & Content Based Filtering	Job Recommendation Systems	Cold start problem, data sparsity
Historical & Analytical Study	Feature Based Machine Learning	Employee Recruitment Screening	Embedded bias, privacy concerns
Ensemble Learning	Biodata Driven Selection	High Precision Candidate Filtering	High computational overhead
ML for HR Classification	NLP and ML Integration	Automated Candidate Shortlisting	Dependency on social/auxiliary data
Model Stability Testing	Cross Format Validation	Stability and Robustness Analysis	Susceptibility to popularity bias

III. METHODOLOGY

The first element of the proposed bias detection and mitigation approach to AI/ML-based resume evaluation involves gathering of resumes dataset with job-relevant and population-relevant features. To enable precision of any subsequent fairness interventions, particular attention should be paid to sensitive features like gender, nationality, age, and others.

3.1 Data Collection and Preprocessing

A structured NLP pipeline is deployed to process resume data gathered from different document formats, like PDF and DOCX. After extraction, text undergoes several cleansing operations, including normalization, tokenization, and stop word elimination. Relevant text semantics are retained while extraneous information is removed systematically.

3.2 Feature Extraction

Named Entity Recognition is one of the sophisticated natural language processing (NLP) methods utilized to mine structured technical skills, educational qualifications, work experience, and certification information from unstructured resumes. Semantic-ware vector representations are generated for each resume by transformer-based pre-trained language models. The extent of compatibility between candidate resumes and job requirements is quantified by comparing resume vectors with job description vectors.

3.3 Machine Learning Model for Classification and Ranking

The machine ingests the preprocessed and enriched features for learning tasks related to relevance ranking and binary classification, both performed sequentially.

A candidate's application is assessed for suitability in relation to a specific job role using Support Vector Machines (SVM). For relevance ranking, K-Nearest Neighbors (KNN) and ensemble stacking models are applied, producing a ranked list for recruiters to consider by calculating the relevance scores of each candidate based on extracted competencies and alignment with the job specification criteria.

3.4 Evaluation and Testing

A labelled resume dataset is employed in evaluating the performance of the model in terms of a comprehensive range of metrics, among which are accuracy, precision, recall, and F1-score. To make sure that the model demonstrates stability and predictability of its classification and ranking capabilities irrespective of resume format and the category of jobs considered, cross-validation techniques have been applied.

3.5 Deployment and User Interface

Deployment takes place via an interface of a web application running on server side. Recruiters and other specialists involved in personnel hiring can use the interface to upload resumes, compose

job descriptions, and see the results immediately. They will be able to review candidate rankings provided to them through the convenient, informative interface. Job fit scores and profiles of candidate competencies are designed to facilitate evidence based decision making without compromising accessibility for non-specialists.

IV. ANALYSIS AND DISCUSSION

4.1 Impact of NLP Techniques on Resume Parsing

The professional expertise, qualifications, and capabilities were successfully extracted from the unstructured content of resumes using Named Entity Recognition combined with vector embedding techniques generated by language models built on transformer-based neural architectures. They showed excellent results across multiple resume templates regardless of terminology differences, formatting preferences, and linguistic nuances without significantly affecting the precision of extraction. Keyword-based matching algorithms often struggle to identify similar capabilities conveyed using unconventional language and specific industry terminology, while transformer-based models' superior semantic capabilities are particularly relevant here.

4.2 Candidate Ranking and Model Performance

Ranking of candidates was efficiently handled by machine learning algorithms including Support Vector Machine (SVM), XGBoost, K-Nearest Neighbours (KNN), and hybrid ensemble architectures. Capability to rank candidates was proven through performance on evaluation datasets with labeled outcomes. The overall accuracy rate was above 90% on average after k-fold cross-validation and hyperparameter optimization. This indicates the ability to generalize well and accurately predict the ranking order of applicants of various profiles. Among all methods tested, the ensemble-based stacking approach was found to be the most effective. The best-performing models yielded stable ranking scores due to their complementary strengths as individual classifiers. These results indicate that the ML-based approach to ranking job candidates is practical and outperforms manual selection.

4.3 Comparative Evaluation Against Traditional Screening Methods

The automated pipeline offers significant operational benefits in a number of areas when compared to traditional manual resume review. Large candidate pools can be evaluated in timeframes that would be operationally impossible for human reviewers thanks to a significant increase in processing throughput. More importantly, by assessing applicants based on consistently applied, competency anchored criteria, the algorithmic approach adds a structural restraint on the expression of subjective bias. It also lessens the impact of irrelevant demographic traits that usually add variation to human-led shortlisting decisions. The outcome is a screening procedure that is concurrently quicker, more scalable, and more equitable, meeting the efficiency and fairness requirements that drive the use of AI in hiring.

4.4 Scope for Future Enhancement

However, despite the positive findings, several opportunities for substantial advancements remain. Semantic comprehension and precision in matching in different professional fields might potentially be enhanced using more advanced transformer models or their adaptations for recruitment-related data sets. Adding multilingual NLP support will increase the applicability of the system to foreign employment cases, in which case applicant pools usually comprise individuals from multiple linguistic backgrounds with different educational credentials frameworks. Furthermore, enhancing long-term relevance and minimizing the probability of model drift will be achieved by incorporating adaptive feedback mechanisms that will enable continuous learning by the model and the possibility of refining the matching criteria based on recruiters' recommendations and employment market dynamics. As a result, all of these enhancements will make the system a globally scalable, continuously growing platform of recruitment intelligence rather than a single-purpose evaluation tool.

V. EXPERIMENTAL RESULT SUMMARY

5.1 Baseline Model Performance and Bias Detection

The initial KNN algorithm demonstrated satisfactory predictive accuracy, with an accuracy rate of 88.15%. Equalized Odds Difference (EOD) and Demographic Parity Difference (DPD) were both assessed at 1.0, which suggests a massive imbalance in decision-making processes favoring candidates from certain nations and institutions in terms of tier rank. Misclassification disparities were also confirmed by confusion matrix analysis and affected underrepresented geographic populations. This establishes the core concept that fair decision-making cannot be guaranteed by predictive accuracy alone.

5.2 Comparative Mitigation Analysis

The biased base line was used to examine three methods of mitigations. Preprocessing methods Reweighing and Disparate Impact Remover made slight distribution changes in the data set, yet had no significant impact on reducing DPD and EOD. The problem here is that predictive correlations introduced by bias into the model cannot be altered by rebalancing the structure and distribution of the data set. As an in-processing technique, adversarial debiasing completely eliminated biases, resulting in DPD and EOD being equal to 0.0 while preserving the accuracy rate at 88.15%. This proves that properly implemented processing techniques may achieve both equity and utility goals without any compromise between them. Thus, it refutes the prevalent belief in the inevitable trade-off between fairness and performance.

5.3 Threshold Sensitivity and Legal Implications

It was concluded that the amount of bias is threshold dependent. Disparities between groups were within compliance requirements at the selection threshold of 0.5; however, at the threshold of 0.7, the Disparate Impact Ratio exceeded 0.8 (four fifths rule of EEOC) and thus crossed the point of legally actionable discrimination. This demonstrates how unreliable single bias threshold evaluation may be. To identify compliance risks obscured by standard evaluation, multiple thresholds should be examined.

5.4 SHAP Based Interpretability Analysis

The SHAP analysis was conducted both pre-debiasing and post-debiasing to verify that there was an improvement in the quality of decision-making rather than merely changes in the metrics. Prior to adversarial debiasing, proxy variables related to geography and institutions, such as country of origin and organization levels, had a dominant role in the predictive modeling process, while legitimate variables like skills and education were accorded lesser priority. Post-adversarial debiasing, on the other hand, proxy variables were accorded lesser priority.

5.5 Confusion Matrix Equalization

After debiasing, the confusion matrix revealed that there was a more uniform distribution of false positives and false negatives for different demographic groups. There were no longer high error rates for underrepresented geographic regions, indicating that adversarial debiasing had improved the fairness of the decision boundary rather than just improving overall model accuracy.

5.6 The Intersectionality Gap

Predictive accuracy and fairness are unique properties that need to be assessed separately. The pre-processing mitigation technique, by itself, is inherently flawed as a method for dealing with biases that may appear in the learned predictive correlations. Adversarial debiasing can be effective in eliminating detectable biases without affecting model performance. Fairness assessment depends on the choice of thresholds and must be performed at multiple operating points. SHAP interpretability provides evidence for causal enhancement in feature weighting, as opposed to superficial performance metrics. Intersectional biases are a critical blind spot that needs more attention through the development of multi-attribute fairness auditing techniques.

5.7 Key Insights

The key limitation that was identified in the analysis above is the intersectionality gap. This category of candidates who were disadvantaged on more than one sensitive dimension geography, institutional tier, and non technical skills were found to have been subjected to a higher level of discrimination in the baseline model compared to what the fairness metrics were able to capture individually. Currently, auditing mechanisms are able to evaluate each sensitive dimension individually but are unable to account for the intersectionality gap; this is an interesting gap that should be explored in future research on fairness auditing mechanisms.

VI. LIMITATIONS

6.1 Methodological Limitations

Our suggested framework does a great job of minimizing bias. We have a few things to consider. We employed a variety of techniques to determine whether bias existed, such as examining the type of institution a person attended and the skills they possessed. This approach can let us know if there are any issues. It is not as effective as using people's direct information. To see if our approach could lessen bias, we also conducted some tests. The real world is more complex, and these tests were conducted in a controlled setting. Additionally, our framework must be able to examine the model, which is not feasible with certain commercial recruitment platforms that prevent us from making modifications. We did not focus as much on bias related to gender, ethnicity, and age and how these factors intersect; instead, we focused primarily on bias related to a person's place of origin and the institution from which they came.

6.2 Technical and Deployment Limitations

Real-world resumes frequently come in more complex formats like PDF and DOCX, so we need to figure out how to handle these formats better. Our system was trained on data that was in a simple format. For our tests, our dataset was adequate. It would be better if it included a wider variety of industries and people. Our system's information summarization component may omit some important details, so we must figure out how to improve it. Additionally, we must test our system to see how well it functions in the real world, ensure that it can support a large number of users, and integrate with current HR platforms.

6.3 Future Research Directions

We must create metrics for measuring fairness that consider the intersections of various attributes, such as bias and fairness and the Framework. The Framework and black box auditing are two methods we must use to audit recruitment platforms that do not provide us with access to their models. We must create methods for continuously assessing fairness in real-world settings, such as multi threshold fairness monitoring and the Framework. Our system, such as the Framework and multilingual and cross-cultural recruitment, needs to be expanded to accommodate individuals who speak languages and come from diverse cultural backgrounds. Our system must be integrated with tools that assist businesses in adhering to AI-related regulations.

VII. CONCLUSION

Fairness in AI hiring is a long-term issue. It's a process that requires management. The suggested pipeline offers a means of identifying, resolving, and demonstrating the resolution of problems. It aids in addressing and bringing discrimination to light. This

is crucial for employers, job seekers, and regulators. Adversarial debiasing assists in eliminating bias without compromising the accuracy of the system. We can see how it is getting better thanks to SHAP analysis. The hiring process is not only quicker but also more equitable thanks to these tools. With these elements, the hiring process is genuinely more effective. Fair hiring practices for AI demand attention.

REFERENCES

[1] Sahu, B. et al. (2025). AI-Powered Resume Screening: Enhancing Efficiency and Fairness in Recruitment. IJNRD, 10(6), a592–a595.

[2] Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. Procedia Computer Science, 167, 2318–2327.

[3] Mishra, A. (2025). Exploring Bias in AI-Driven Resume Screening: A Fairness Analysis and Mitigation Approach. SSRN Working Paper. ssrn-5160444.

[4] Mehta, P. (2025). Algorithmic Bias Detection in AI-Powered Recruitment Tools. IRJET, 12(8), 398–401.

[5] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning . AAAI/ACM Conference on AI, Ethics, and Society.

[6] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. MIT Press. https://fairmlbook.org/

[7] Mehrabi, N. et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), Article 96.

[8] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS, 30, 4765–4774.

[9] NYC Dept. of Consumer & Worker Protection. (2023). Automated Employment Decision Tools. Local Law 144.