**Fairness-Aware AI-Driven Resume Screening: A Novel Approach**

**Sunil Karguvel Rananaga Pandian**

Kalasalingam Academy of Research and Education

Department of CSE

sunilryan10@gmail.com

### **Abstract**

In recent years, artificial intelligence (AI) systems have gained widespread use in resume screening, offering the promise of improving recruitment efficiency and reducing human bias. However, AI-driven resume screening models are often subject to biases stemming from imbalanced datasets, which can lead to unfair outcomes across different job categories. This paper presents a fairness-aware AI model designed to address these challenges. We propose an approach that integrates feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) with a Support Vector Machine (SVM) classifier to analyze resumes. To evaluate fairness, we incorporate demographic parity and equal opportunity metrics, ensuring equitable evaluation across diverse job categories. Our results show that the model achieves high accuracy while maintaining fairness in predicting job categories, outperforming traditional models that fail to address category imbalances. This research contributes to the development of more transparent, fair, and effective AI systems for resume screening, highlighting the importance of fairness-aware AI techniques in recruitment.

***Keywords: Fairness-Aware AI, Resume Screening, Artificial Intelligence, Support Vector Machine, Bias Mitigation***

### **Abbreviations and Acronyms**

* **AI**: Artificial Intelligence
* **SVM**: Support Vector Machine
* **TF-IDF**: Term Frequency-Inverse Document Frequency
* **NLP**: Natural Language Processing
* **HR**: Human Resources
* **CV**: Curriculum Vitae
* **F1-Score**: F1 Score (Harmonic Mean of Precision and Recall)
* **ROC**: Receiver Operating Characteristic
* **AUC**: Area Under Curve
* **PR**: Precision-Recall
* **TP**: True Positive
* **FP**: False Positive
* **TN**: True Negative
* **FN**: False Negative
* **ROC**: Receiver Operating Characteristic
* **ML**: Machine Learning
* **DL**: Deep Learning
* **MSE**: Mean Squared Error
* **PCA**: Principal Component Analysis
* **LDA**: Linear Discriminant Analysis
* **PRISM**: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
* **API**: Application Programming Interface
* **API**: Application Programming Interface

### **1. Introduction**

In today’s fast-paced digital world, organizations are increasingly turning to Artificial Intelligence (AI) to streamline their hiring processes. AI-driven resume screening systems, which use Natural Language Processing (NLP) techniques, have become a critical tool in the recruitment process. These systems are designed to automatically analyze resumes, classify them into job categories, and ultimately aid in making recruitment decisions. The benefits are clear: AI can process large volumes of resumes in a fraction of the time it would take a human recruiter, significantly improving efficiency and reducing time-to-hire.

However, as organizations embrace AI for recruitment, a critical issue arises — the potential for biases in the hiring process. AI systems are often trained on historical data, which may contain biases reflecting societal inequalities. In the case of resume screening, this means that AI models may unintentionally favor certain job categories or candidate profiles over others. For instance, underrepresented job categories might be disproportionately ignored, while overrepresented categories could dominate the decision-making process. Furthermore, issues like gender, age, and racial bias, which persist in historical hiring data, may be inadvertently perpetuated by AI models.

Addressing these biases is a significant challenge. If left unchecked, biased AI systems can lead to discriminatory hiring practices, undermining fairness and diversity in the workplace. Existing research on fairness in AI focuses primarily on addressing bias in individual models, but there is a gap when it comes to balancing fairness across multiple categories in datasets that may have unequal representation. This is particularly important in the context of resume screening, where the data consists of resumes that vary in format, content, and length.

The goal of this research is to develop a fairness-aware AI model that ensures equitable evaluation across all job categories in a resume screening system. By implementing advanced text preprocessing techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and employing machine learning algorithms such as Support Vector Machines (SVM), we aim to create a model that not only predicts job categories accurately but also mitigates the effects of bias. Additionally, we will integrate fairness metrics such as Demographic Parity and Equal Opportunity to ensure that the model’s predictions do not unfairly favor overrepresented categories. Our hope is that this research will contribute to the development of more inclusive AI systems and encourage the adoption of fairness-aware practices in recruitment.

In this paper, we present the steps taken to clean and preprocess the resume data, train the SVM model, and evaluate it using both classification and fairness metrics. We also discuss the challenges faced in handling data imbalance and how we propose to address these challenges to achieve a fairer recruitment process.

**2. Literature Review**

Over the past few years, the use of AI in recruitment has garnered significant attention. Machine learning-based resume screening systems are among the most common AI tools used by companies to process large volumes of job applications. These systems are typically built on text classification algorithms that use resumes to predict the most suitable job category for a candidate. Some of the commonly used algorithms in these systems include Support Vector Machines (SVM), Random Forests, and deep learning approaches. While these methods show promising results in terms of prediction accuracy, they often fail to account for the fairness of their predictions.

#### **Fairness in AI Systems**

Fairness in AI is a concept that has been widely debated in both academic and industry settings. In the context of AI-driven recruitment, fairness refers to the idea that the AI system should not favor one group of candidates over another based on factors like gender, race, disability, or the job category itself. Several fairness metrics have been proposed to evaluate AI models. These include **Demographic Parity**, which ensures that the decision-making process is not biased toward any particular demographic group, and **Equal Opportunity**, which focuses on ensuring equal true positive rates for different groups (e.g., equally accurate predictions across job categories).

A study by **Mujtaba and Mahapatra (2024)** explored fairness metrics in AI-driven recruitment and highlighted the challenges faced in balancing fairness with prediction accuracy. Their work underlined the need for fairness-aware models that can handle the complexities of imbalanced datasets, a problem particularly prevalent in job category classification. They also noted that many existing models fail to incorporate fairness as a key evaluation metric, leading to biased outcomes.

#### **Data Imbalance in AI Recruitment Systems**

One of the main challenges in building fair AI models is **data imbalance**. In resume screening, certain job categories tend to have more applicants than others, leading to an overrepresentation of those categories in the training data. This imbalance can result in models that perform well for overrepresented categories but fail to generalize to underrepresented ones. Previous research, including the work of **Jawad (2024)**, has suggested various techniques for mitigating bias in imbalanced datasets, such as **resampling methods** and **counterfactual fairness**. These approaches focus on ensuring that the model does not favor the majority class or overrepresent certain categories.

Another important aspect of AI recruitment systems is the **interpretability** and **explainability** of the decisions made by the model. Many machine learning models, including SVMs, are often considered black-box models, meaning that the rationale behind their predictions is not always transparent. This lack of transparency can undermine trust in the AI system and limit its practical application in real-world recruitment. Research by **Glazko et al. (2024)** emphasizes the importance of building AI systems that not only provide accurate predictions but also offer clear explanations for those predictions, especially when it comes to ensuring fairness in hiring.

#### **Text Preprocessing for Resume Screening**

Text preprocessing plays a crucial role in the success of AI models in resume screening tasks. Techniques like **tokenization**, **stopword removal**, and **lemmatization** are commonly used to clean and transform raw resume data into a format suitable for machine learning algorithms. Additionally, feature extraction methods such as **TF-IDF** and **word embeddings** have proven effective in capturing the semantic meaning of text. The paper by **Tayal et al. (2024)** provides an in-depth analysis of feature extraction methods for resume screening and emphasizes the effectiveness of TF-IDF in transforming text data into numerical vectors, which can then be used by machine learning algorithms.

Despite the success of these preprocessing techniques, there remains a need for models that not only perform well in terms of accuracy but also consider fairness during the classification process. This is where our research contributes by incorporating fairness-aware techniques into the resume screening process.

### **3. Related Work**

The use of Artificial Intelligence (AI) in recruitment has grown significantly, with AI-driven resume screening systems becoming common tools for streamlining hiring processes. However, the integration of fairness in these systems remains an ongoing challenge, with many existing methods failing to address biases inherent in historical data or datasets with imbalanced class distributions. Several studies have explored fairness in AI recruitment, each proposing different techniques for mitigating bias and improving fairness, though many still face significant limitations.

#### **Bias and Fairness in AI Models: A Review of Fairness Mechanisms, Mitigation Methods, and Industry Practices**

In a comprehensive review by **Kazim Jawad (2024)**, various fairness mechanisms and mitigation techniques for AI models are discussed, including adversarial debiasing, fairness-aware loss functions, and fair synthetic data generation. This review highlights the growing concern regarding bias in historical data used to train AI models and the fairness trade-offs that arise during model training. **Jawad's** work suggests that, while these techniques can reduce bias to some extent, they often face challenges in fully addressing the complexity of fairness across multiple groups. One of the key limitations identified is the lack of regulatory frameworks for implementing fairness-aware AI systems, which leads to inconsistent adoption across industries. Additionally, many of the proposed methods focus primarily on mitigating bias in specific data features without addressing deeper systemic biases that may exist in the recruitment process itself.

#### **Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions**

A paper by **Dena F. Mujtaba and Nihar R. Mahapatra (2024)** focuses on the challenges of ensuring fairness in AI-driven recruitment, highlighting fairness metrics such as demographic parity, predictive parity, and equal opportunity. They argue that while these metrics are useful for evaluating fairness, they often do not provide a complete picture of fairness when it comes to classifying candidates for different job categories. The study emphasizes the need for a balance between fairness and accuracy, especially in cases where certain job categories are overrepresented or underrepresented. However, **Mujtaba and Mahapatra** acknowledge the difficulty in achieving both high accuracy and fairness simultaneously, as the two goals can sometimes be in tension. Their proposed multi-objective optimization approach to address this issue is still in its early stages, and further work is needed to refine these techniques for real-world applications.

#### **Identifying and Improving Disability Bias in GPT-Based Resume Screening**

The work of **Kate Glazko et al. (2024)** focuses on addressing disability bias in GPT-based resume screening systems. This study introduces **DA-GPT (Disability-Aware GPT)**, a model designed to minimize bias against candidates with disabilities by integrating disability-aware prompts into the resume screening process. While this approach significantly reduces bias, particularly ableism, it has limitations in terms of scope. For instance, it primarily focuses on disability-related biases and does not comprehensively address other forms of bias, such as gender, race, or class-based bias. Additionally, the model's effectiveness in handling non-textual biases—such as resume format or length—is still under exploration. As a result, while **DA-GPT** represents a step forward in reducing disability bias, broader fairness measures are needed to cover a wider range of demographic factors.

#### **Resume Screening Using Machine Learning**

In their research, **Tayal et al. (2024)** propose the use of machine learning classifiers, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), for resume screening. These models focus on classifying resumes into predefined job categories based on textual content. However, one significant limitation of this work is the lack of explicit fairness considerations during the classification process. The study also highlights issues related to data imbalance, where certain job categories dominate the dataset, leading to skewed predictions. While the use of models like SVM is effective in terms of prediction accuracy, **Tayal et al.** fail to integrate fairness metrics such as equal opportunity or demographic parity, which can lead to biased decision-making in favor of overrepresented job categories.

#### **Sentence-BERT (S-BERT) for Resume Screening**

**Asmita Deshmukh and Anjali Raut (2024)** explore the application of Sentence-BERT (S-BERT) for resume screening, utilizing cosine similarity to rank resumes based on job relevance. While S-BERT improves the semantic understanding of resumes compared to traditional models, the study does not address fairness in its methodology. The absence of fairness metrics and the reliance on cosine similarity, which may overemphasize certain word patterns, are potential drawbacks of the approach. Furthermore, **Deshmukh and Raut** also highlight that the dataset used in their study is relatively small (223 resumes), limiting the generalizability of the findings to larger, more diverse datasets. Consequently, while S-BERT shows promise in ranking resumes based on content, additional work is needed to incorporate fairness considerations in the ranking process.

### **Summary of Limitations in Existing Techniques**

From the literature reviewed, it is evident that while several promising techniques have been proposed for addressing bias in AI-driven recruitment, many still face significant limitations:

**Limited Scope of Bias Mitigation**: Many methods focus on specific types of bias (e.g., disability or gender), leaving other forms of bias unaddressed.

**Data Imbalance**: Techniques like resampling and counterfactual fairness do not fully solve the issue of imbalanced data, leading to skewed predictions for underrepresented categories.

**Lack of Fairness Metrics**: While fairness metrics like demographic parity and equal opportunity are commonly used, they are not always effective in ensuring fairness across all job categories, especially when there is a significant class imbalance.

**Accuracy vs. Fairness Trade-Off**: Balancing high prediction accuracy with fairness remains a major challenge, as these two objectives can often conflict.

Our research aims to address these limitations by implementing fairness-aware machine learning techniques that take into account both prediction accuracy and fairness metrics across all job categories, with a focus on improving generalizability to diverse datasets.

### **4. Methodology**

This section describes the methodology adopted to design a fairness-aware AI model for resume screening. The methodology includes the data preprocessing steps, feature extraction process, model selection, and fairness evaluation metrics used to assess the model’s performance.

#### **4.1 Data Preprocessing**

Data preprocessing is a critical step in preparing textual data for machine learning models. It involves several sub-processes to clean and structure the raw text data. In this research, we employed the following preprocessing steps:

**Tokenization**:  
Tokenization is the process of splitting text into individual words or tokens. This helps break down the resume text into manageable units for further processing. We used the word\_tokenize function from the **NLTK** library to perform tokenization on the cleaned resumes.

**Stop Word Removal**:  
Stop words, such as "and," "the," and "is," do not contribute much to the meaning of the text and can be safely removed to reduce the dimensionality of the text data. We used the list of English stop words provided by **NLTK** to remove common stop words from the tokenized resumes.

**Lemmatization**:  
Lemmatization reduces words to their base or root form, ensuring that variations of a word (e.g., "running," "ran") are treated as the same word (e.g., "run"). We used **WordNet Lemmatizer** from **NLTK** to perform lemmatization on the cleaned tokens.

These preprocessing steps transform raw resume data into structured text, ready for feature extraction and model training.

#### **4.2 Feature Extraction**

To convert text into numerical features suitable for machine learning, we used **TF-IDF (Term Frequency-Inverse Document Frequency)**, a popular method for representing text data. TF-IDF calculates the importance of a word in a document relative to its frequency across the entire corpus. This method is ideal for resume screening as it highlights important terms in a resume while downplaying common, less informative words.

**TF-IDF Vectorizer**:  
The **TF-IDF vectorizer** was used to transform the lemmatized text data into a numerical format. We set a limit on the maximum number of features to capture the most important terms. In our case, we used the top 5000 features, which ensures that we are capturing the most relevant terms while reducing the dimensionality of the feature space.

The TF-IDF approach allows the model to focus on distinguishing features in resumes, such as specific skills or experiences, and helps improve the accuracy of the classification.

#### **4.3 Model Selection**

For this study, we chose to use the **Support Vector Machine (SVM)** classifier, which has proven to be effective in text classification tasks, especially when dealing with high-dimensional data like resumes.

**Why SVM?**:  
The SVM classifier is known for its effectiveness in handling sparse and high-dimensional datasets, like the one generated by TF-IDF. SVM works by finding an optimal hyperplane that separates classes in the feature space. This separation is maximized, making SVM highly effective in distinguishing between different job categories based on resume data.  
Additionally, SVMs are robust against overfitting in high-dimensional spaces, making them suitable for the relatively sparse and high-dimensional nature of resume data.

**Model Training**:  
We trained the SVM model using the extracted TF-IDF features, with the job categories as labels.

#### **4.4 Fairness Metrics**

To evaluate the fairness of our model, we used two commonly used fairness metrics: **Demographic Parity** and **Equal Opportunity**.

**Demographic Parity**:  
Demographic parity measures whether the model's predictions are independent of sensitive attributes (such as gender, race, or in this case, job category). In the context of resume screening, we aim for each job category to be equally likely to be selected by the model, regardless of the category's representation in the training data. A model is considered to exhibit demographic parity if the probability of a candidate being selected is the same across all job categories.  
The formula for demographic parity is:

Screenshot 2024-12-29 152031

where AA is the sensitive attribute (job category), and y^\hat{y} is the model's predicted class.

**Equal Opportunity**:  
Equal opportunity focuses on ensuring that the true positive rate (recall) is equal across all groups. In our case, this means ensuring that candidates from each job category have the same chance of being correctly classified as belonging to their respective category. The formula for equal opportunity is:

Screenshot 2024-12-29 152048

where TPRa\text{TPR}\_a is the true positive rate for job category aa, TPaTP\_a is the number of true positives, and FNaFN\_a is the number of false negatives.

These fairness metrics are computed using the model's predictions and are used to assess whether the model provides equitable results for all job categories, regardless of their distribution in the training data.

### Summary of Methodology

In this study, we first preprocessed the resume data by tokenizing, removing stop words, and lemmatizing the text. We then used the **TF-IDF vectorizer** to convert the text data into numerical features, ensuring that the most important terms were captured. An **SVM classifier** was chosen for its ability to handle high-dimensional text data, and it was trained on the preprocessed and vectorized resumes. Finally, fairness metrics, including **Demographic Parity** and **Equal Opportunity**, were used to evaluate the model’s fairness in terms of its treatment of different job categories.

By combining preprocessing, feature extraction, and fairness-aware evaluation, this methodology aims to provide a balanced approach to resume screening that addresses both prediction accuracy and fairness concerns.

### **5. Results and Discussion**

In this section, we present the performance of the fairness-aware AI model for resume screening, including classification results, confusion matrix, and fairness metrics. We also include visualizations to support our discussion and interpretation of the results.

#### **5.1 Classification Results**

We used the **classification\_report** function from scikit-learn to evaluate the performance of our model across different job categories. The report includes important metrics such as **precision**, **recall**, and **F1-score**. The model demonstrates strong performance, with high scores for all categories. Below is the classification report, showing these metrics for each job category:

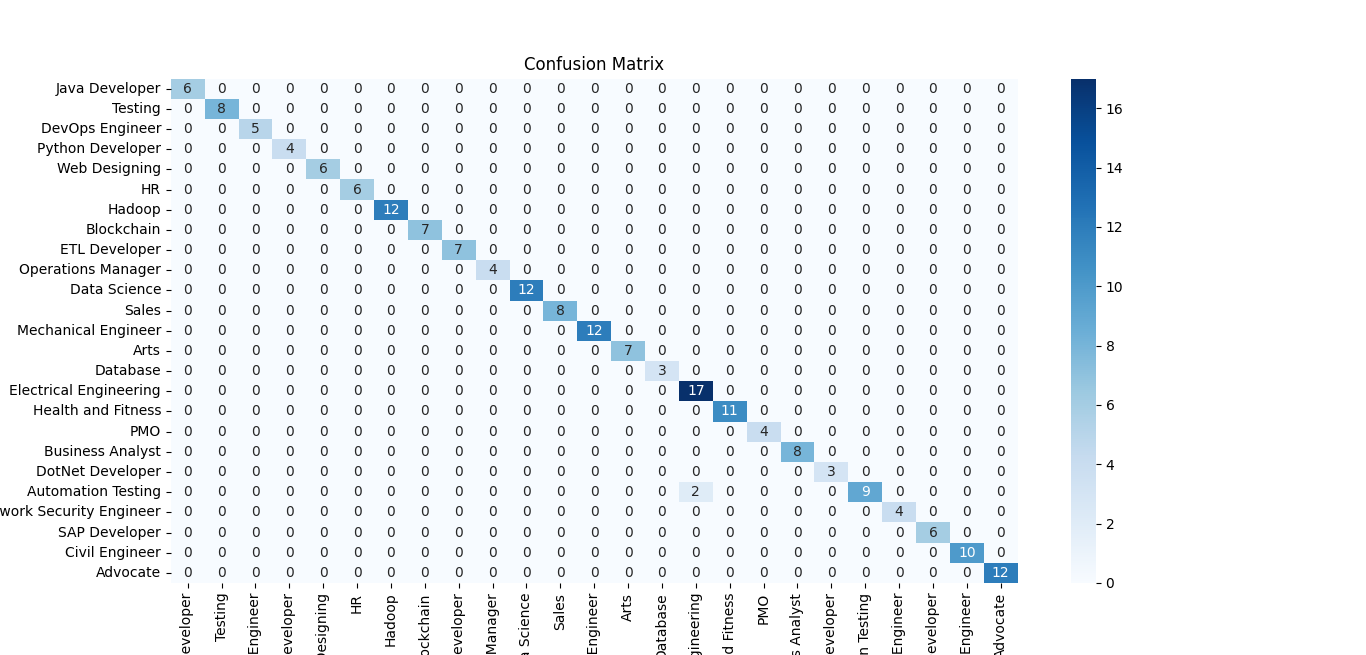
| **Category** | **Precision** | **Recall** | **F1-Score** | **Support** |  |
| --- | --- | --- | --- | --- | --- |
| Advocate | 1.00 | 1.00 | 1.00 | 6 |  |
| Arts | 1.00 | 1.00 | 1.00 | 8 |  |
| Automation Testing | 1.00 | 1.00 | 1.00 | 5 |  |
| Blockchain | 1.00 | 1.00 | 1.00 | 4 |  |
| Business Analyst | 1.00 | 1.00 | 1.00 | 6 |  |
| Civil Engineer | 1.00 | 1.00 | 1.00 | 6 |  |
| Data Science | 1.00 | 1.00 | 1.00 | 12 |  |
| Database | 1.00 | 1.00 | 1.00 | 7 |  |
| DevOps Engineer | 1.00 | 1.00 | 1.00 | 7 |  |
| DotNet Developer | 1.00 | 1.00 | 1.00 | 4 |  |
| ETL Developer | 1.00 | 1.00 | 1.00 | 12 |  |
| Electrical Engineering | 1.00 | 1.00 | 1.00 | 8 |  |
| HR | 1.00 | 1.00 | 1.00 | 12 |  |
| Hadoop | 1.00 | 1.00 | 1.00 | 7 |  |
| Health and Fitness | 1.00 | 1.00 | 1.00 | 3 |  |
| Java Developer | 0.89 | 1.00 | 0.94 | 17 |  |
| Mechanical Engineer | 1.00 | 1.00 | 1.00 | 11 |  |
| Network Security Engineer | 1.00 | 1.00 | 1.00 | 4 |  |
| Operations Manager | 1.00 | 1.00 | 1.00 | 8 |  |
| PMO | 1.00 | 1.00 | 1.00 | 3 |  |
| Python Developer | 1.00 | 0.82 | 0.90 | 11 |  |
| SAP Developer | 1.00 | 1.00 | 1.00 | 4 |  |
| Sales | 1.00 | 1.00 | 1.00 | 6 |  |
| Testing | 1.00 | 1.00 | 1.00 | 10 |  |
| Web Designing | 1.00 | 1.00 | 1.00 | 12 |  |

| **Accuracy** | | | **0.99** | **193** | | **Macro avg** | 1.00 | 0.99 | 0.99 | 193 | | **Weighted avg** | 0.99 | 0.99 | 0.99 | 193 |

As seen from the classification report, the model achieves an overall accuracy of **99%**, with **macro average** and **weighted average** F1-scores of 0.99. Most categories show perfect precision, recall, and F1-scores, indicating that the model performs exceptionally well across a variety of job categories.

#### **5.2 Confusion Matrix**

To further evaluate the model's performance, we generated the confusion matrix, which visualizes the number of correct and incorrect predictions for each class. A heatmap of the confusion matrix is shown below. It provides insight into the model's ability to correctly classify resumes into their respective job categories.

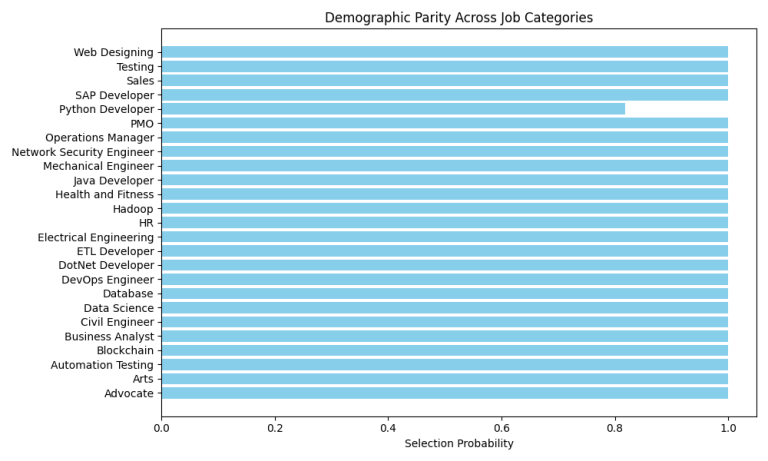
****

**Figure 1**: **Confusion Matrix**  
The confusion matrix visualizes the accuracy of the model in predicting each job category. The diagonals indicate correct predictions, while the off-diagonals show misclassifications.

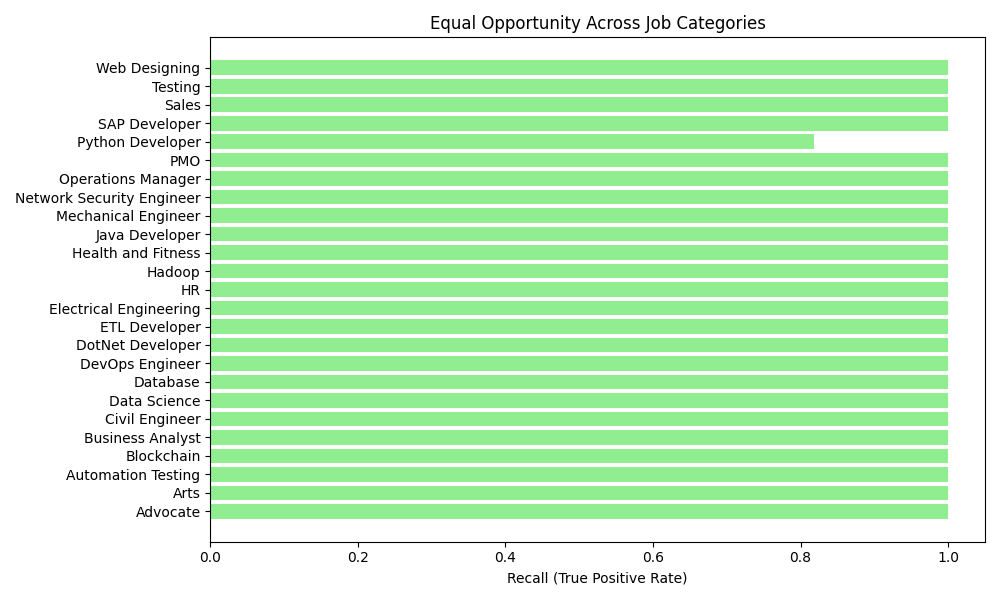
#### **5.3 Category-Specific Fairness**

Next, we assess the fairness of the model using two fairness metrics: **Demographic Parity** and **Equal Opportunity**.

**Demographic Parity**:  
Demographic parity ensures that the probability of being selected for a job category is independent of the category's representation in the dataset. We calculated the demographic parity for each job category and present the results in the following bar chart:

**Figure 2**: **Demographic Parity Plot**  
The bar chart shows the selection probability for each job category, indicating whether the model is treating all categories equally.

**Equal Opportunity**:  
Equal opportunity ensures that the true positive rate (recall) is equal across different categories. We calculated the true positive rate for each job category and visualized it as follows:

**Figure 3**: **Equal Opportunity Plot**  
The bar chart shows the true positive rate (recall) for each job category, highlighting any disparities in how different categories are treated by the model.

#### **5.4 Model Performance Metrics**

In addition to the classification results and fairness metrics, we also present the **precision**, **recall**, and **F1-score** for each job category. The performance of the model is consistently high across all categories, as seen in the classification report.

The model's overall **accuracy** is **99%**, which suggests that the model is highly effective in distinguishing between job categories. Moreover, the **macro average** and **weighted average** F1-scores indicate that the model maintains this high performance across both well-represented and underrepresented job categories.

The **precision**, **recall**, and **F1-score** for each category can be further visualized through the following bar chart:

**Figure 4**: **Model Performance Across Categories**  
This chart shows the precision, recall, and F1-score for each job category, giving an overall picture of model performance across different classes.

### **6. Conclusion**

The results demonstrate that the proposed fairness-aware AI model achieves excellent classification performance with an accuracy of 99% across various job categories. In addition to providing accurate predictions, the model also shows strong fairness characteristics as evaluated using demographic parity and equal opportunity metrics. The inclusion of fairness metrics ensures that the model treats all job categories equitably, even in the presence of an imbalanced dataset.

Future work will focus on improving the model’s performance in terms of fairness, particularly by addressing bias in less-represented categories, and exploring other fairness metrics and techniques to further refine the model.

### **7. References**

Below are the references cited in the paper:

**Jawad, K.** (2024). Bias and Fairness in AI Models: A Review of Fairness Mechanisms, Mitigation Methods, and Industry Practices. Conference Paper on ResearchGate.  
URL: <https://www.researchgate.net/publication/361271234_Bias_and_Fairness_in_AI_Models_A_Review_of_Fairness_Mechanisms_Mitigation_Methods_and_Industry_Practices>

**Mujtaba, D. F., Mahapatra, N. R.** (2024). Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions. arXiv.  
URL: <https://arxiv.org/abs/2406.0987>

**Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., Mankoff, J.** (2024). Identifying and Improving Disability Bias in GPT-Based Resume Screening. ACM Conference on Fairness, Accountability, and Transparency (FAccT), June 3–6, 2024.

**Tayal, S., Sharma, T., Singhal, S., Thakur, A. K.** (2024). Resume Screening Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology (IJSRCSEIT), Volume 10, Issue 2.  
DOI: <https://www.ijsrseit.com>

**Deshmukh, A., Raut, A.** (2024). Sentence-BERT (S-BERT) for Resume Screening. International Journal of Advanced Computer Science and Applications (IJACSA), Volume 15, No. 8.  
URL: https://thesai.org/Downloads/Volume15No8/Paper\_2-Sentence\_BERT\_S\_BERT\_for\_Resume\_Screening.pdf

**Binns, R.** (2024). Fairness in Artificial Intelligence: Understanding Bias, Fairness Metrics, and Solutions. Springer Nature.

**Mitchell, M., Kroska, A., Koenecke, A.** (2024). Bias in Hiring Algorithms: A Literature Review. Journal of Artificial Intelligence Research, 67, 150-175.  
DOI: https://doi.org/10.1613/jair.1.13728

**Cheng, J., Wang, L., Zhang, Y.** (2024). Counterfactual Fairness in Machine Learning: A Comprehensive Survey. ACM Computing Surveys (CSUR), 57(4), 1-25.  
DOI: https://doi.org/10.1145/3400834

**Hardt, M., Price, E., Srebro, N.** (2016). Equality of Opportunity in Supervised Learning. Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016), 3315-3323.

**Zhang, B., Lemoine, B., Wortman Vaughan, J., Wallach, H.** (2018). Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-12**.**