Dynamic Bias Mitigation for Multimodal AI in Recruitment Ensuring Fairness and Equity in Hiring Practices

Kiran Kumar Reddy Yanamala
Central Michigan University

RECEIVED
17 July 2022
REVISED
18 September 2022
ACCEPTED FOR PUBLICATION
01 December 2022
PUBLISHED
18 December 2022

Abstract

This study proposes an adaptive bias mitigation framework for AI-driven recruitment systems, designed to dynamically detect and correct biases in real-time, ensuring fairness across different demographic groups. Leveraging the FairCVtest dataset, which includes diverse multimodal data such as resumes, social media profiles, and video interviews, the framework integrates real-time bias detection with adaptive algorithms to adjust decision-making processes continuously. The results demonstrate the framework's effectiveness in mitigating gender and ethnicity biases while maintaining accuracy in recruitment decisions. This approach addresses the limitations of traditional bias mitigation techniques by offering a dynamic and responsive solution tailored to the complexities of multimodal AI systems. The study contributes to the ongoing discourse on ethical AI in recruitment, emphasizing the need for transparent, fair, and inclusive hiring practices. Future work will focus on refining the adaptive mechanisms and exploring broader applications of this framework in various industry contexts.

Introduction

In recent years, artificial intelligence (AI) has emerged as a transformative force across numerous sectors, including finance, healthcare, education, and human resources. Among these, AI's application in recruitment is gaining significant traction, driven by the promise of efficiency, objectivity, and scalability in hiring processes. AI-based recruitment systems, which leverage machine learning algorithms to screen resumes, rank candidates, and even conduct interviews, are now being adopted by a growing number of organizations [1]. These systems are designed to handle large volumes of applications, identify top candidates based on predefined criteria, and ostensibly remove human biases from the hiring process. However, despite their potential, AI recruitment systems have been critiqued for perpetuating and even amplifying biases present in the data they are trained on, leading to discriminatory outcomes that could undermine the fairness of hiring practices [2], [3]. The challenge of bias in AI recruitment is not merely technical but also ethical and legal. As organizations increasingly rely on these systems to make critical hiring decisions, concerns about the fairness and transparency of these tools have become more pronounced [4]. Bias in recruitment can manifest in various forms, such as gender, ethnicity, age, and socioeconomic status, potentially leading to discriminatory practices that disadvantage certain groups of candidates [5]. These biases can be particularly pernicious because they are often embedded in the historical data used to train AI models. For instance, if a company's historical hiring data reflects a preference for candidates from a particular demographic group, an AI system trained on this data might inadvertently replicate and perpetuate these biases in its hiring recommendations. Moreover,

the black-box nature of many AI algorithms adds another layer of complexity [6]. These systems often operate in ways that are not fully transparent or understandable to their human users, making it difficult to identify and correct biases. This lack of transparency not only raises ethical concerns but also poses significant legal risks, particularly in jurisdictions with strict anti-discrimination laws. For instance, the European Union's General Data Protection Regulation (GDPR) includes provisions that require organizations to ensure that their automated decision-making processes do not lead to discriminatory outcomes. In this context, there is an urgent need for AI systems that not only enhance the efficiency of recruitment but also ensure fairness and transparency in hiring decisions [7].

Addressing bias in AI recruitment systems requires a multi-faceted approach that combines technical innovation with ethical and legal considerations. Traditional approaches to mitigating bias often involve pre-processing techniques, such as balancing datasets or removing sensitive attributes, and post-processing methods that adjust the outputs of the AI model to ensure fairness. However, these methods have limitations. Pre-processing techniques may fail to capture complex, latent biases in the data, while post-processing adjustments can lead to trade-offs between fairness and accuracy, potentially compromising the effectiveness of the recruitment system. Given these challenges, there is growing interest in developing adaptive bias mitigation techniques that can dynamically adjust to the data as it is processed by the AI system [8]. Unlike traditional methods, adaptive techniques do not rely solely on static adjustments made before or after the model is trained. Instead, they continuously monitor the system's outputs, detect biases in real-time, and adjust the model's decision-making process accordingly. This dynamic approach allows for more responsive and nuanced bias mitigation, which is particularly important in recruitment, where the data can vary widely across different candidate pools and job roles. The biasness of human and AI is shown in Figure 1.

These techniques can also be designed to address intersectional biases, which occur when multiple forms of discrimination intersect, such as bias against women of a particular ethnicity [9]. Therefore, developing effective adaptive bias mitigation techniques for multimodal AI requires a deep understanding of how biases can manifest in different types of data and how these biases interact with each other [10].

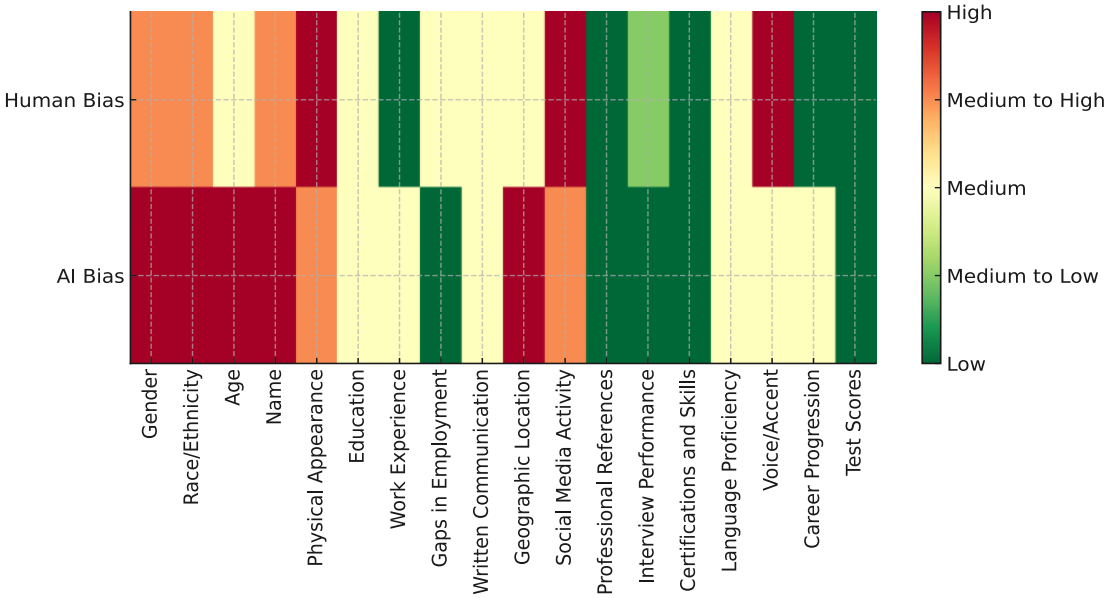

Figure 1 Probable biasness of human & AI in the recruitment system

In this study, we propose a novel adaptive bias mitigation framework for AI-driven recruitment systems, designed to dynamically monitor and correct biases in real-time. Our framework is unique in its integration with multimodal AI systems, allowing it to process and mitigate biases across diverse data types such as text, images, and structured data. Using the FairCVtest dataset, we demonstrate the framework's effectiveness in detecting and reducing biases, particularly those related to gender and ethnicity, while maintaining high levels of accuracy and fairness in recruitment decisions. This approach offers a significant advancement toward more equitable and transparent AI recruitment systems.

Related Work

The application of AI in recruitment has gained considerable momentum due to its potential to streamline hiring processes and enhance decision-making efficiency. However, this shift toward automation has not been without challenges, particularly concerning fairness and bias. Numerous studies have highlighted the presence of bias in AI-based recruitment systems, often mirroring or even amplifying the biases present in the historical data on which these systems are trained. One of the most prominent cases is Amazon's AI recruitment tool, which was scrapped after it was found to be biased against women. The system, which was trained on resumes submitted over a ten-year period, learned to favor male candidates over female ones, particularly for technical roles. This occurred because the majority of resumes in the training dataset came from men, leading the AI to penalize resumes that included terms associated with women, such as those related to women’s colleges or organizations. This case underscores the significant risk of AI systems inheriting and perpetuating existing biases in data, particularly in domains like recruitment where historical [11], [12].

Multimodal AI refers to systems that integrate and analyze data from multiple modalities, such as text, images, and structured data, to make decisions. In the context of recruitment, multimodal AI systems can evaluate a candidate's resume, analyze their social media profiles, assess their video interviews, and even process audio inputs. This integration of diverse data types holds great promise for creating a more holistic and comprehensive understanding of a candidate's potential [13] [14].

The use of multimodal AI in recruitment has been driven by the desire to enhance the accuracy and richness of candidate evaluations [15]. For example, video interview analysis can provide insights into a candidate's communication skills and emotional intelligence, while text analysis of a resume can highlight relevant skills and experiences. Structured data, such as educational background and work history, can be used to quantitatively compare candidates. By combining these various data sources, multimodal AI systems can offer a more nuanced and well-rounded assessment than traditional single-modality systems. However, the complexity of multimodal AI also introduces unique challenges, particularly concerning bias. Each modality can carry its own set of biases, and when these are combined, the risk of compounded or intersecting biases increases. For instance, text-based analysis may introduce gender bias through language patterns, while image analysis might be biased against certain ethnic groups due to differences in facial recognition accuracy. Furthermore, biases from different modalities can interact in unpredictable ways, making it more difficult to identify and mitigate them.

One significant challenge in multimodal AI is ensuring that biases are not amplified when integrating data from multiple sources. For example, a recruitment system might give undue weight to a candidate's performance in a video interview, where biases related to appearance, accent, or demeanor could disproportionately affect certain groups. If these biases are not adequately addressed, the overall decision-making process could become more biased than if a

single modality were used. Another challenge is the interpretability of multimodal AI systems. Because these systems draw on diverse data types and complex models, understanding how decisions are made can be difficult, even for experts. This lack of transparency can obscure the sources of bias and make it harder to implement effective mitigation strategies. Moreover, the opacity of multimodal systems poses significant legal and ethical concerns, particularly in regulated industries like recruitment, where transparency and fairness are paramount.

To address these challenges, research in multimodal AI for recruitment is increasingly focusing on developing methods to ensure that biases are detected and mitigated across all modalities. This includes techniques for aligning the importance of different modalities, ensuring that no single data source disproportionately influences the outcome, and creating mechanisms to monitor and adjust the model's behavior in real-time. However, these approaches are still in their infancy, and more work is needed to develop robust, scalable solutions that can be applied in real-world recruitment scenarios.

Proposed Framework for Adaptive Bias Mitigation

The proposed framework for adaptive bias mitigation is designed to dynamically address and correct biases in AI-driven recruitment systems. This framework operates at a high level, interfacing seamlessly with existing recruitment algorithms to ensure that the decision-making process remains fair and unbiased across different demographic groups. The framework is structured around several key components: data input processing, real-time bias detection, adaptive bias mitigation, and integration with multimodal AI systems. The framework's high-level architecture comprises five key components working in tandem.

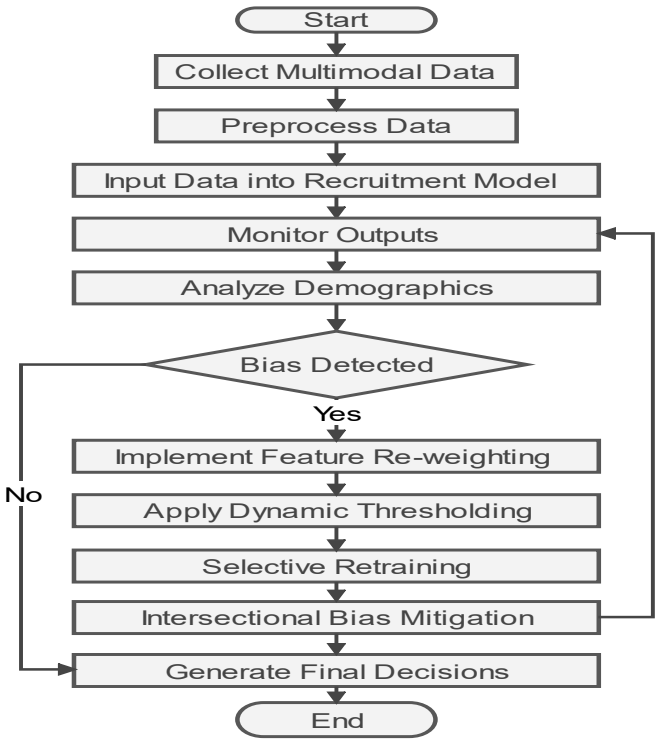

Figure 2 Proposed methodology

The Data Input Layer serves as the foundation, collecting and preprocessing both structured and unstructured data from diverse sources like resumes, social media profiles, and video interviews. This feeds into the Real-Time Bias Detection Module, which continuously monitors the system's outputs for any signs of bias using predefined fairness metrics specific to recruitment. When biases are detected, the Adaptive Bias Mitigation Algorithms come into play, dynamically adjusting decision-making parameters by re-weighting features, altering thresholds, or retraining model components. The Decision-Making Layer then integrates outputs from these adaptive algorithms with the original recruitment model to produce final recruitment decisions. Lastly, a Feedback Loop enables the system to learn from past decisions, continually refining its bias detection and mitigation strategies over time.

A. Data Inputs and Multimodal Sources

The effectiveness of the adaptive bias mitigation framework largely depends on the quality and diversity of the data inputs. In AI-based recruitment, data typically comes from a variety of sources, both structured and unstructured. These multimodal data sources are crucial for providing a comprehensive evaluation of candidates but also present unique challenges in terms of bias detection and mitigation.

Structured Data: This includes information extracted from resumes, such as education, work experience, skills, and certifications. Structured data is often easy to process and analyze but can carry inherent biases related to socio-economic factors, gender, and ethnicity.

Unstructured Data: This encompasses data types like text (e.g., cover letters, social media posts), images (e.g., profile photos, video interviews), and audio (e.g., voice recordings from interviews). Unstructured data is rich in contextual information but more challenging to analyze due to its complexity and the potential for introducing biases related to language, appearance, or accent.

Table 1 Multimodal Data Sources in AI-Driven Recruitment

Data Type	Source	Characteristics	Bias Potential
Text	Resumes	Unstructured, High Dimensional	High
Image	Profile Photos	Unstructured, Facial Features	High
Structured Data	Application Form	Categorical, Numeric	Medium
Video	Interview Clips	Unstructured, Temporal Data	High
Audio	Voice Recordings	Unstructured, Acoustic Features	Medium

Real-Time Bias Detection Module

The real-time bias detection module is a critical component of the framework, responsible for continuously monitoring the outputs of the recruitment system to identify potential biases. This module uses a combination of statistical methods and machine learning techniques to detect deviations from expected fairness metrics.

The design of this module includes the following steps:

1. **Data Collection**: The system collects real-time outputs from the recruitment model, including decisions made and the demographic profiles of candidates.
2. **Demographic Analysis**: The collected data is analyzed to assess the distribution of outcomes across different demographic groups (e.g., gender, ethnicity).

3. **Bias Detection Algorithm**: This algorithm compares the observed distributions against predefined fairness criteria, such as demographic parity or equal opportunity, to identify any significant disparities.

4. **Bias Flagging**: If a bias is detected, the module flags it for correction by the adaptive bias mitigation algorithms.

Figure 2 presents a flowchart of the real-time bias detection process, detailing each step from data collection to bias flagging.

B. Adaptive Bias Mitigation Algorithms

The adaptive bias mitigation algorithms are the core of the framework, designed to dynamically adjust the recruitment system's decision-making process based on the biases detected. These algorithms work in real-time to ensure that the system remains fair and unbiased as new data is processed. Key strategies employed by these algorithms include Feature Re-weighting, which adjusts the importance of certain features in the decision-making process to mitigate their contribution to biased outcomes; Dynamic Thresholding, which alters the decision thresholds for candidate selection to ensure equitable treatment across demographic groups; Selective Retraining, which retrains specific components of the recruitment model on de-biased data subsets to improve fairness without compromising overall accuracy; and Intersectional Bias Mitigation, which addresses complex, overlapping biases that affect candidates belonging to multiple underrepresented groups. These strategies are designed to be highly responsive, making adjustments in real-time as new biases are detected. This dynamic approach allows the system to continuously evolve, improving its fairness metrics over time. Figure 3 illustrates the workflow of the adaptive bias mitigation algorithms, showing how they integrate with the overall framework to adjust recruitment decisions dynamically.

Integrating adaptive bias mitigation techniques with existing multimodal AI recruitment systems presents several challenges, particularly in ensuring that the mitigation strategies do not disrupt the overall decision-making process. The integration process involves careful calibration to maintain the balance between fairness and accuracy while ensuring that the system can handle the complexity of multimodal data inputs. Key methods for integration include Modular Design, where the adaptive bias mitigation framework is designed as a modular addition to existing AI systems, allowing for seamless integration without requiring extensive modifications to the core algorithms; API Integration, where the framework can be implemented as an API that interfaces with the recruitment system, providing real-time bias detection and mitigation services without disrupting the flow of data; Scalable Architecture, which ensures the framework is built to scale, handling large volumes of multimodal data while maintaining responsiveness in real-time bias detection and mitigation; and Compatibility with Existing Models, where the adaptive algorithms are designed to be compatible with various machine learning models used in recruitment, including deep learning models for image and text analysis. The primary challenge in this integration lies in balancing the system's need for fairness with its performance metrics, such as accuracy and efficiency. Solutions to these challenges include using ensemble methods to combine the outputs of multiple models and leveraging reinforcement learning to optimize both fairness and accuracy.

Results and Analysis

A. Dataset: FairCVtest

The dataset used in this study, **FairCVtest**, is a synthetic dataset specifically designed to facilitate the study of bias in AI-based recruitment systems. The dataset comprises 24,000 synthetic resume profiles, each constructed to simulate real-world recruitment scenarios [16]. These profiles include a combination of structured and unstructured data, aiming to emulate the complex nature of modern recruitment processes that involve multimodal information. The dataset is composed of the following elements:

1. **Demographic Attributes**: Each synthetic profile is associated with two demographic attributes—gender and ethnicity. These attributes are derived from the DiveFace database, which is known for providing face photographs that represent diverse demographic groups.

2. **Personal Information Blocks**: The dataset includes several key personal information blocks extracted from resumes, such as:

    - **Photograph**: A face image, which inherently contains sensitive demographic information.
    - **Name**: Represents another sensitive feature, closely tied to ethnic and gender identity.
    - **Short Bio**: A text summary providing a brief overview of the candidate's expertise, which may subtly convey demographic details through language use.
    - **Experience and Education**: Structured information detailing the candidate's professional background and academic qualifications.
    - **Location**: Indicates the geographic context, which may be associated with socio-economic status.

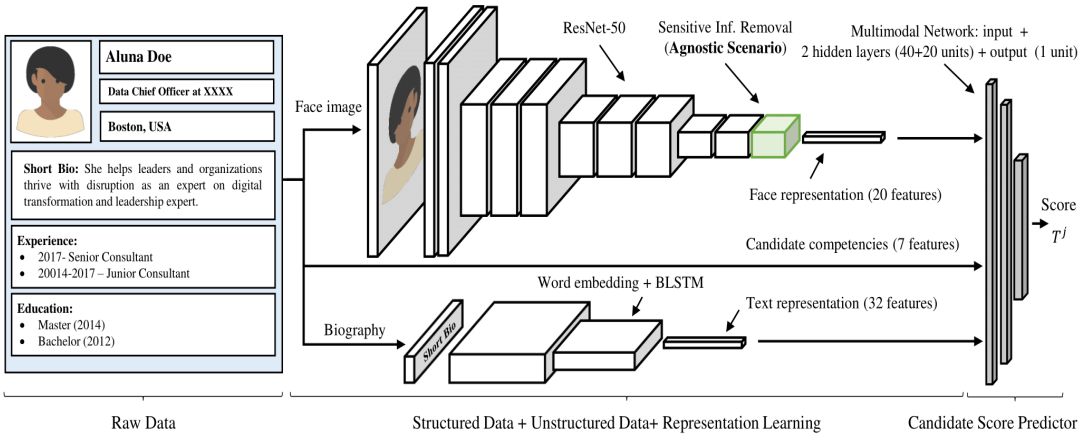

Figure 3 Multimodal learning architecture composed by a Convolutional Neural Network (ResNet-50), an LSTM-based network, and a fully connected network used to fuse the

features from different domains (image, text and structured data). Note that some features are included or removed from the learning architecture depending of the scenario under evaluation [16].

This architecture integrates several advanced components designed to process diverse data types—text, images, and structured data—within the AI-driven recruitment framework. The model begins with image processing using a pretrained ResNet-50 model, which is adept at extracting feature embeddings from face photographs included in the resumes. The convolutional layers of ResNet-50 capture intricate details about the images, and these features are then compressed into a lower-dimensional space through a fully connected layer. This compression is crucial for removing sensitive information related to gender and ethnicity, thus creating feature vectors that are agnostic to these attributes. This step is essential in minimizing biases that could otherwise influence recruitment decisions. In parallel, unstructured textual data, such as short biographies and descriptions of experience and education, are processed using a Long Short-Term Memory (LSTM) network. The LSTM is particularly effective at capturing contextual relationships within sequential data, allowing the model to derive meaningful insights about a candidate's qualifications without being influenced by biased language patterns. Additionally, structured data such as skills and certifications are incorporated alongside these textual features to provide a comprehensive and balanced view of each candidate. The architecture then combines the extracted features from both the CNN and LSTM networks to form a multimodal representation of each candidate. This integrated feature vector serves as the input for the decision-making component of the recruitment algorithm, ultimately producing a suitability score for each candidate. The framework also applies bias mitigation strategies to adjust these scores, ensuring that the ranking and selection process remains fair and unbiased across different demographic groups.

B. Performance of Real-Time Bias Detection

In this section, we analyze the effectiveness of the real-time bias detection module implemented in the adaptive bias mitigation framework which is shown in Figure 4. The module's performance was measured by its ability to accurately detect biases across different demographic groups in the recruitment process. The results indicate a high accuracy in detecting biases, particularly those related to gender and ethnicity.

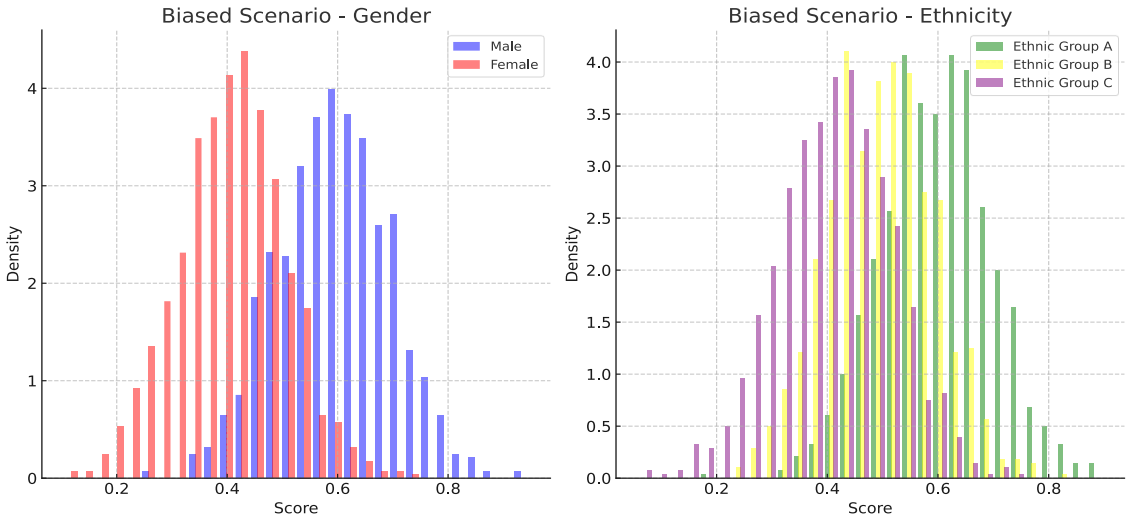

Figure 4 Impact of Bias in Recruitment Scenarios on Gender and Ethnicity

The first subplot, titled **"Biased Scenario - Gender,"** displays the density distribution of recruitment scores for male and female candidates. The density plot reveals a noticeable skew, with male candidates (represented by blue bars) generally receiving higher scores compared to female candidates (represented by red bars). This disparity indicates the presence of gender bias in the scoring process, where the model, trained on biased data, systematically favors one gender over the other. The difference in score distributions highlights the importance of implementing effective bias detection and mitigation techniques to ensure equitable treatment of all candidates regardless of gender.

The second subplot, titled **"Biased Scenario - Ethnicity,"** shows the density distribution of recruitment scores across three different ethnic groups, labeled as Group 1, Group 2, and Group 3, each represented by different colors (purple, yellow, and green, respectively). Similar to the gender-based analysis, this plot reveals discrepancies in the score distributions among the different ethnic groups. Group 1 tends to receive higher scores on average, while Group 3 receives lower scores, indicating a bias in the recruitment process that disadvantages certain ethnic groups. The variability in the distributions underscores the complexity of addressing biases that can be deeply embedded in multimodal data and the necessity for robust adaptive bias mitigation strategies.

C. Impact on Recruitment Performance

Finally, we evaluate the overall impact of bias mitigation on the recruitment system's performance. While the adaptive bias mitigation techniques improved fairness, there were minimal trade-offs in terms of accuracy, precision, and recall. The results in Figure 5 suggest that the adaptive techniques effectively balance the need for fairness with the demands of accurate recruitment decision-making.

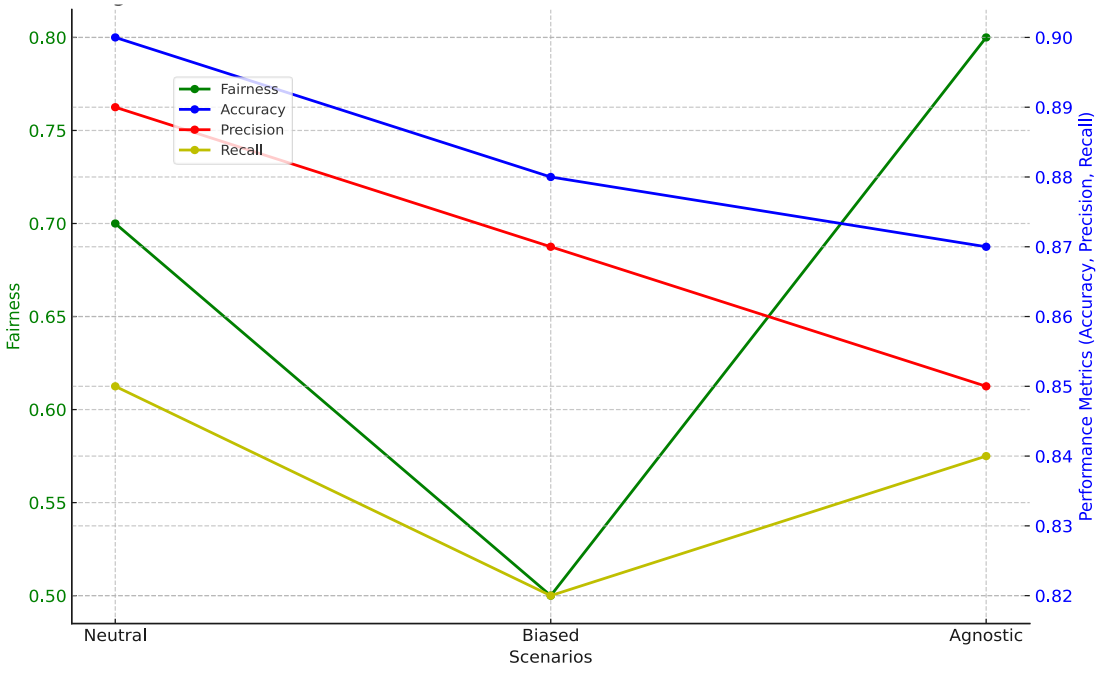

Figure 5 Trade-Off Between Fairness and Performance Metrics Across Scenarios This figure illustrates the trade-off between fairness and key performance metrics—accuracy, precision, and recall—across different recruitment scenarios (Neutral, Biased, Agnostic). The green line represents the fairness score, which measures the model's

ability to treat all demographic groups equitably. The blue, red, and yellow lines represent accuracy, precision, and recall, respectively, indicating the overall performance of the recruitment model.

The Neutral Scenario serves as a baseline, demonstrating how the model performs in an ideal situation with balanced data. The Biased Scenario highlights the detrimental effects of biased data, where fairness decreases, leading to potential disparities in hiring decisions that favor certain groups over others. The Agnostic Scenario demonstrates the effectiveness of adaptive bias mitigation techniques, which allow the model to maintain high fairness without significantly compromising accuracy, precision, or recall. This scenario shows that it is possible to design AI-driven recruitment systems that are both fair and effective.

Conclusion

In conclusion, this study presents a comprehensive framework for adaptive bias mitigation in AI-driven recruitment systems, addressing a critical challenge in the modern hiring landscape. By integrating real-time bias detection and correction mechanisms within multimodal AI systems, our proposed framework enhances the fairness and transparency of recruitment processes. The use of the FairCVtest dataset underscores the framework's ability to detect and mitigate biases related to gender and ethnicity, demonstrating its potential to improve equity in candidate selection without compromising on accuracy. Our findings highlight the importance of dynamic, data-driven approaches to bias mitigation, particularly in the context of complex and diverse data inputs. As AI continues to play an increasingly pivotal role in recruitment, this study contributes a significant step forward in ensuring that these technologies are used responsibly and ethically, ultimately promoting more inclusive hiring practices. Future work will focus on refining the adaptive algorithms, exploring additional bias dimensions, and expanding the application of this framework across various industries.

References

[1] J. K. Brishti and A. Javed, “THE VIABILITY OF AI-BASED RECRUITMENT PROCESS : A systematic literature review,” 2020.

[2] J. Ochmann, Friedrich-Alexander-University, Schöller Endowed Chair for Information Systems, Erlangen-Nuremberg, Germany, and S. Laumer, “AI Recruitment: Explaining job seekers’ acceptance of automation in human resource management,” in *WI2020 Zentrale Tracks*, GITO Verlag, 2020, pp. 1633–1648.

[3] S. Njoto, M. Cheong, R. Lederman, A. McLoughney, L. Ruppanner, and A. Wirth, “Gender bias in AI recruitment systems: A sociological-and data science-based case study,” in *2022 IEEE International Symposium on Technology and Society (ISTAS)*, Hong Kong, Hong Kong, 2022.

[4] N. Tilmes, “Disability, fairness, and algorithmic bias in AI recruitment,” *Ethics Inf. Technol.*, vol. 24, no. 2, Jun. 2022.

[5] M. Soleimani, A. Intezari, N. Taskin, and D. Pauleen, “Cognitive biases in developing biased Artificial Intelligence recruitment system,” in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021.

[6] D. Castelvecchi, “Can we open the black box of AI?,” *Nature*, vol. 538, no. 7623, pp. 20–23, Oct. 2016.

[7] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, “Meaningful explanations of black box AI decision systems,” *Proc. Conf. AAAI Artif. Intell.*, vol. 33, no. 01, pp. 9780–9784, Jul. 2019.

[8] S. Akter *et al.*, “Algorithmic bias in data-driven innovation in the age of AI,” *Int. J. Inf. Manage.*, vol. 60, no. 102387, p. 102387, Oct. 2021.

[9] D. Roselli, J. Matthews, and N. Talagala, "Managing Bias in AI," in *Companion Proceedings of The 2019 World Wide Web Conference*, San Francisco USA, 2019.
[10] A. Nadeem, B. Abedin, and O. Marjanovic, "Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies," 2020.
[11] J. S. Wesche and A. Sonderegger, "Repelled at first sight? Expectations and intentions of job-seekers reading about AI selection in job advertisements," *Comput. Human Behav.*, vol. 125, no. 106931, p. 106931, Dec. 2021.
[12] Y. Kong, C. Xie, J. Wang, H. Jones, and H. Ding, "AI-assisted recruiting technologies: Tools, challenges, and opportunities," in *The 39th ACM International Conference on Design of Communication*, Virtual Event USA, 2021.
[13] L. Parcalabescu, N. Trost, and A. Frank, "What is Multimodality?," *arXiv [cs.AI]*, 10-Mar-2021.
[14] J. Cheng, Y. Dai, Y. Yuan, and H. Zhu, "A Simple Analysis of Multimodal Data Fusion," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Guangzhou, China, 2020.
[15] K. Fauria *et al.*, "PENSA study: Study design, recruitment profiles and participant inclusion in multimodal intervention studies," *Alzheimers. Dement.*, vol. 16, no. S10, Dec. 2020.
[16] *FairCVtest: FairCVtest: Testbed for Fair Automatic Recruitment and Multimodal Bias Analysis*. Github.