Fair Learning for Bias Mitigation and Quality Optimization in Paper Recommendation

Uttamasha Anjally Oyshi, Susan Gauch
Department of Electrical Engineering & Computer Science
University of Arkansas
Fayetteville, USA
e-mails: {uoyshi, sgauch}@uark.edu

***Abstract*—Despite frequent double-blind review, demographic biases of authors still disadvantage the underrepresented groups. We present Fair-PaperRec, a MultiLayer Perceptron (MLP) based model that addresses demographic disparities in post-review paper acceptance decisions while maintaining high-quality requirements. Our methodology penalizes demographic disparities while preserving quality through intersectional criteria (e.g., race, country) and a customized fairness loss, in contrast to heuristic approaches. Evaluations using conference data from ACM Special Interest Group on Computer-Human Interaction (SIGCHI), Designing Interactive Systems (DIS), and Intelligent User Interfaces (IUI) indicate a 42.03% increase in underrepresented group participation and a 3.16% improvement in overall utility, indicating that diversity promotion does not compromise academic rigor and supports equity-focused peer review solutions.**

Keywords-Fairness-aware recommendation; Paper selection; Demographic bias mitigation

I. INTRODUCTION

Double-blind review often does not eradicate systemic biases linked to authors' demographics, reputations, or institutional affiliations, despite attempts to ensure impartiality [1]–[4]. Recent data indicates that even the most stringent anonymization techniques can be undermined by analyzing writing style or cross-referencing previous articles [5], [6]. This tendency can sustain biases against particular groups, including women, racial minorities, and researchers from underrepresented areas [3], [7]–[9]. Simultaneously, there is a growing dependence on recommendation algorithms to optimize processes such as paper selection, grant distribution, and significant publication identification [10]–[12]. While these systems can accelerate decision-making, they also pose a danger of perpetuating biases present in the training data, particularly if they focus only on predictive accuracy [13]–[15]. Therefore, it is imperative to devise novel methodologies that explicitly include demographic justice, preventing the perpetuation of historical inequalities.

In this paper, we introduce Fair-PaperRec, a fairness-aware recommendation framework specifically designed to mitigate post-review bias. Unlike previous heuristic-based approaches that often handle single-attribute fairness constraints or overlook intersectionality, in our approach:

- We surpass single-attribute approaches by incorporating multiple demographic attributes (e.g., race, country) and constructing multi-dimensional profiles that capture underlying biases.
- After a double-blind review, a specialized fairness penalty is implemented to address demographic disparities, thereby correcting latent biases without the need to replace existing processes.
- Our method ensures that the quality of the paper is maintained throughout by ensuring demographic parity, thereby obtaining equitable representation without compromising academic rigor.

Our results demonstrate improved representation in the participation of underrepresented groups, as well as an enhancement in overall paper quality, as indicated by the h-index. Notably, these findings reveal that enhanced inclusivity need not diminish academic rigor; a fairness-driven approach can yield greater demographic parity while simultaneously preserving, and at times even *enhancing*, the quality of accepted papers.

By mitigating biases in paper selection, our strategy promotes a richer academic discourse and amplifies the representation of marginalized communities, thereby paving the way toward more equitable, high-quality conferences. The paper includes the folling sections, where in Section 2, we review related work. Section 3 presents the proposed methodology. Section 4 explains our experimental setup and metrics. Section 5 provides results and analysis. Finally, the Section 6 concludes the paper.

II. RELATED WORK

We begin by examining double-blind review and bias in academic paper selection, then explore fairness in recommender systems, and finally discuss recent advancements in neural approaches for fair selection.

A. Double-Blind Review and Bias in Academic Paper Selection

Although double-blind review conceals identities [1]–[3], it often fails to eliminate biases in gender, race,

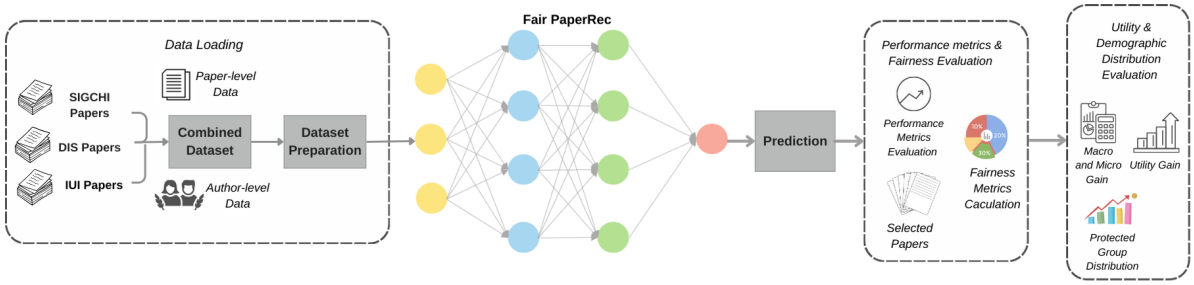


Figure 1. Overview of the Fair-PaperRec Architecture.

or geography [9], [16]. While authorship-attribution can rectify advanced anonymization [5], high-prestige institutions continue to receive favorable reviews [17]. As a result, underrepresented groups, including women and racial minorities, continue to be marginalized [18], and substantial acceptance rate disparities persist [7], [8].

B. Fairness in Recommendation Systems

When optimizing solely for accuracy, recommenders frequently exacerbate biases [11], [19]. Although some fairness issues are addressed by multi-objective [13], adversarial [14], and re-ranking methods [15], the majority of these methods concentrate on single attributes or user-item data, leaving intersectional biases in paper acceptance unaccounted for. In academic settings, *provider fairness* is equivalent to *author fairness*, which protects minority researchers [20]. There are very few algorithms that resolve post-review bias, not to mention, multi-attribute fairness [12], [21].

C. Post-Review Bias Mitigation and Neural Approaches

Some heuristic methods attempt to rebalance accepted papers after reviews [20], but they risk local optima and often fail to consider multi-attribute fairness. Neural-based solutions such as *DeepFair* [11] or *Neural Fair Collaborative Filtering* [22] demonstrate that fairness can align with accuracy, yet they typically target commercial recommendations rather than the nuances of academic peer review. Meanwhile, multi-stakeholder optimization [23], [24] highlights the need for more contextual fairness definitions within scholarly publishing. Although certain approaches (e.g., Bulut et al. [25]) employ text-based features like Term Frequency–Inverse Document Frequency (TF-IDF) to improve relevance, they often disregard the imperative of equity for authors from historically marginalized groups.

III. Methodology

Our approach tackles demographic biases in conference data by employing a simple Multilayer Perceptron (MLP) to enforce fairness post-review. We highlight two fundamental principles: (1) revealing and alleviating biases instead of eliminating them, and (2) implementing a straightforward, yet efficient neural architecture that harmonizes equality and utility.

A. Data Collection and Pre-processing

Real-world datasets—particularly those drawn from academic conference submissions—often contain latent biases that mirror systemic imbalances in the scholarly community (e.g., underrepresentation of certain demographics). We utilize datasets from SIGCHI 2017, DIS 2017, and IUI 2017 [20], which naturally reflect systemic disparities (e.g., skewed demographics). Instead of eliminating such biases, our objective is to *recognize and rectify* them.

We describe the process of collecting and preparing the data used in our experiments. The dataset consists of academic papers submitted to conferences, and we employ a variety of pre-processing steps to ensure the data are suitable for training our model.

TABLE I. DEMOGRAPHIC PARTICIPATION FROM PROTECTED GROUPS IN THREE CONFERENCES.

Conference	Gender (%)	Race (%)	Country (%)
SIGCHI	41.88	6.84	21.94
DIS	65.79	35.09	24.56
IUI	43.75	51.56	39.06
Average	50.47	31.16	28.52

1) Data Description: We gathered detailed information at the paper and author levels, resulting in a robust combined dataset. Every paper record has a title, authors, and a conference designation (1 = IUI, 2 = DIS, 3 = SIGCHI). Author records encompass demographic information (gender, race, nationality, career stage), for detailed analysis. We classify SIGCHI 2017 articles as a standard for high-impact research, whereby *Overall* includes all submissions and *Selected* refers to those identified by our algorithms.

2) Data Pre-processing: Several preprocessing steps were undertaken to prepare the dataset for training:

- *Categorical Encoding:* Gender, Country, and Race are subjected to one-hot encoding. Gender is binary (0 = male, 1 = female), Country is categorized as *developed* or *underdeveloped*, and Race comprises

{White, Asian, Hispanic, Black}, with Hispanic and Black designated as protected groups (Table I).

- *Normalization:* Numerical attributes (e.g., h-index) employ min-max scaling for consistent magnitude.
- *Training and Validation Division:* An 80%/20% stratified division guarantees equitable distribution of labels and protected attributes in both subsets.

B. Problem Definition

This study develops a *fairness-aware paper recommendation system* that ensures demographic parity with respect to authors' race and country, while preserving high academic standards. We frame acceptance decisions as a *recommendation* task, where *conference organizers (users)* seek to select from *530 papers (items)* spanning SIGCHI, DIS, and IUI. Each paper (item) includes an *h-index* for quality, demographic data (race, country), and a conference rating (SIGCHI: 1, DIS: 2, IUI: 3).

Our approach enforces fairness constraints on race and country independently, excluding *gender* due to its relatively balanced distribution (see Table I). By leveraging historical acceptance patterns and explicit diversity goals, the system balances the *need for high-quality research* with the *requirement to address demographic biases* in the final recommendation of papers.

Let D represent the dataset of submitted papers, where each paper $p \in D$ is associated with a set of features X_p (e.g., race, country, h-index) and a target variable y_p indicating acceptance (1) or rejection (0). The *race* attribute R_p and *country* attribute C_p are the protected attributes.

We aim to optimize a predictive model $f : X_p \to \hat{y}_p$ that minimizes the following objective function:

$$\min_f \left(\mathcal{L}(f(X_p), y_p) + \lambda \cdot \mathcal{L}_{\text{fairness}}(f, D)\right) \tag{1}$$

Here, $\mathcal{L}(f(X_p), y_p)$ is the *prediction loss* (e.g., Binary Cross-Entropy Loss), $\mathcal{L}_{\text{fairness}}(f, D)$ is the *fairness loss*, penalizing deviations from demographic parity across race and country and λ is a hyperparameter that balances the trade-off between prediction accuracy and fairness.

C. Demographic Parity

We aim to ensure that the probability of a paper being accepted is independent of the protected attributes:

$$P(\hat{y}_p = 1 \mid R_p = r) = P(\hat{y}_p = 1), \quad \forall r \in \text{Race}$$

$$P(\hat{y}_p = 1 \mid C_p = c) = P(\hat{y}_p = 1), \quad \forall c \in \text{Country}$$

Utilizing these equations ensures that the papers authored by individuals from different races and countries have an equal probability of acceptance.

Algorithm 1. FAIR-PAPERREC LOSS FUNCTION.

Input: Model M, Epochs E, Batch size B, Data D, Protected attributes A, Hyperparameter λ
Output: Trained Model M
Initialize Model M
for each $e \in E$ **do**
  Shuffle Data D
  for each batch $\{(X, Y)\} \in D$ with size B **do**
    Predict $\hat{Y} \leftarrow M(X)$
    Calculate Loss:
      $L_{\text{prediction}} \leftarrow \text{PredictionLoss}(Y, \hat{Y})$
      $L_{\text{fairness}} \leftarrow \text{FairnessLoss}(A, \hat{Y})$
    Calculate Total Loss:
      $L_{\text{total}} \leftarrow \lambda \cdot L_{\text{fairness}} + L_{\text{prediction}}$
    Compute gradients $\nabla L_{\text{total}} \leftarrow \frac{\partial L_{\text{total}}}{\partial M}$
    Update Model parameters: $M \leftarrow M - \alpha \nabla L_{\text{total}}$
  end for
end for

D. Fairness Loss

The fairness loss from the objective function in Equation 1 is constructed to minimize statistical parity differences between the protected and non-protected group:

$$\mathcal{L}_{\text{fairness}} = \left(P(\hat{y}_p = 1 \mid G_{\text{p}}) - P(\hat{y}_p = 1 \mid G_{\text{np}})\right)^2 \tag{2}$$

Here, $P(\hat{y}_p = 1 \mid G_{\text{p}})$ denotes the acceptance probability for the protected group and $P(\hat{y}_p = 1 \mid G_{\text{np}})$ is the acceptance probability for the non-protected group.

E. Combined Fairness Loss

Furthermore, we define a combined fairness loss to minimize statistical parity differences across race and country attributes between the protected and unprotected groups, as shown in Equation 3.

$$\begin{aligned}\mathcal{L}_{\text{fairness}} = W_r \left(\frac{1}{N_r} \sum_{p \in G_r} \hat{y}_p - \frac{1}{N} \sum_{p=1}^{N} \hat{y}_p\right)^2 \\ + W_c \left(\frac{1}{N_c} \sum_{p \in G_c} \hat{y}_p - \frac{1}{N} \sum_{p=1}^{N} \hat{y}_p\right)^2\end{aligned} \tag{3}$$

G_{r} and G_{c} denote the race and country groups, respectively. N_r and N_c are the number of papers in each group and weights W_r and W_c reflect group distributions.

F. Total Loss

The total loss is the combination of prediction and fairness losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{prediction}} + \lambda \cdot \mathcal{L}_{\text{fairness}}$$

G. Constraints and Considerations

We assess fairness by training our model separately on *race* and *country*, as well as jointly on both attributes to evaluate selection fairness across multiple dimensions.

TABLE II. GAIN CALCULATIONS FOR COUNTRY AND RACE FEATURES WITH UTILITY GAIN (UG_{i}).

λ	Country Feature			Race Feature		
	Macro Gain (%)	Micro Gain (%)	UG_{i} (%)	Macro Gain (%)	Micro Gain (%)	UG_{i} (%)
1	7.71	8.67	3.16	24.81	31.11	0.35
2	10.77	13.23	1.05	33.54	46.30	1.75
2.5	**12.67**	**22.96**	**1.75**	39.25	54.81	1.40
3	13.60	16.96	0.35	**42.03**	**56.48**	**3.16**
5	14.80	19.97	-0.35	43.04	56.11	-0.70
10	13.86	18.73	2.46	52.91	64.81	-0.70

a) Exclusion of Protected Attributes: Race R_p and country C_p are excluded from the input feature set X_p to mitigate direct bias amplification. To achieve joint fairness, both attributes are omitted during training, preventing the model from learning acceptance outcomes influenced by race or country.

b) Indirect Bias Mitigation: A fairness loss promotes demographic parity, addressing indirect biases associated with features related to race or country. The model maintains neutrality by penalizing selection disparities, even in the absence of protected attributes.

c) Scalability: Our method supports datasets of varying scales and complexities, demonstrating strong performance across various academic fields. This scalability ensures fairness across various use cases.

IV. Model Overview

To achieve demographic parity while preserving quality in paper selection, we present a MLP-based neural network (See Figure 1), explicitly engineered to balance the trade-off between fairness and accuracy. It illustrates the correlations between input features, like author demographic attributes and paper quality, while alleviating biases during selection.

A unique fairness loss function was employed to ensure equity, imposing penalties on the model for substantial differences in selection rates between protected and non-protected groups. This loss function is integrated with the conventional prediction loss to attain a balance between diversity and accuracy; the algorithm is shown in Algorithm 1.

The acceptance probabilities for submitted papers are generated by the MLP, which are subsequently ranked to guarantee that the final selection meets both quality and fairness objectives. By selecting top papers according to these probabilities, we ensure equal representation of authors from both protected and non-protected groups while upholding the requisite standard of academic excellence.

A. Selection Mechanism

The model calculates acceptance probabilities for all submitted papers after training. After calculating acceptance odds, the algorithm ranks candidate papers.

Algorithm 2. FAIRNESS-AWARE PAPER SELECTION MECHANISM.

Input: Dataset D, Model M, Number of Accepted Papers N_a, Total Papers N_t
Output: Selected Papers P_{selected}
Initialize: $P_{\text{selected}} \leftarrow \emptyset$
Step 1: Apply trained model M to the entire dataset D
for each paper $p \in D$ **do**
    Compute acceptance probability: $\hat{y}_p \leftarrow M(p)$
end for
Step 2: Rank all papers p by acceptance probability $\hat{y}_p$
Sort D in descending order of $\hat{y}_p$
Step 3: Select top N_a papers:
    $P_{\text{selected}} \leftarrow \{p \mid \hat{y}_p \geq \hat{y}_{(N_a)}\}$
Step 4: Ensure Fairness Constraints
Return P_{selected}

This rating phase ensures underrepresented groups are represented in final admission decisions. Representing this as a suggestion list preserves the peer-review process and corrects residual biases. Algorithm 2 selects the best papers based on probability, ensuring fairness and preserving the desired number of accepted papers.

- *Prediction Aggregation:* The trained MLP model is applied to the entire dataset to obtain predicted acceptance probabilities $\hat{y}_p$ for each paper.
- *Ranking:* Papers are ranked in descending order based on their predicted probabilities.
- *Selection:* The papers with the highest predicted probabilities are selected for acceptance, ensuring the total number of selected papers matches the required acceptance quota.

Mathematically, the selection process is represented as:

$$\text{Selected Papers} = \left\{p \in D \mid \hat{y}_p \geq \hat{y}_{(N_a)}\right\}$$

Here, $\hat{y}_{(N_a)}$ is the N_a-th highest predicted probability in the set $\{\hat{y}_p \mid p \in D\}$ while N_a is the total number of accepted papers and N_t is the total number of submitted papers, where $N_a \leq N_t$.

This approach ensures that the selection process is both informed by the model's predictions and constrained to uphold demographic parity, fostering an equitable and meritocratic paper selection environment.

V. Evaluation and Experiments

This section presents the experimental evaluation of our proposed Fair-PaperRec model on the chosen datasets. To guide the exploration of fairness and quality in our proposed paper recommendation system, we pose the following research questions:

- *RQ1:* How do fairness constraints affect the overall quality (utility) of recommended papers, as measured by metrics, such as the h-index?
- *RQ2:* Does handling race and country as separate protected attributes differ from treating them jointly in terms of fairness outcomes and selection decisions?

- *RQ3:* How do varying weight assignments to multiple protected attributes (race and country) influence the trade-off between fairness and utility?

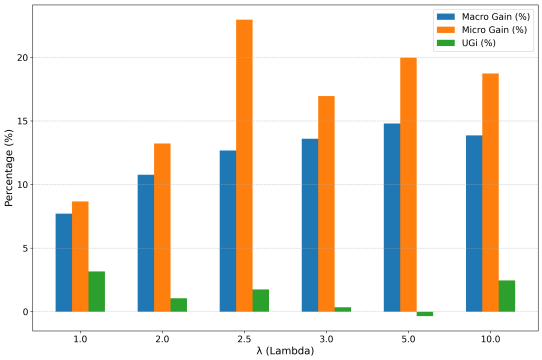


Figure 2. Comparison of Macro and Micro Gains for Country Across Different Fairness Configurations.

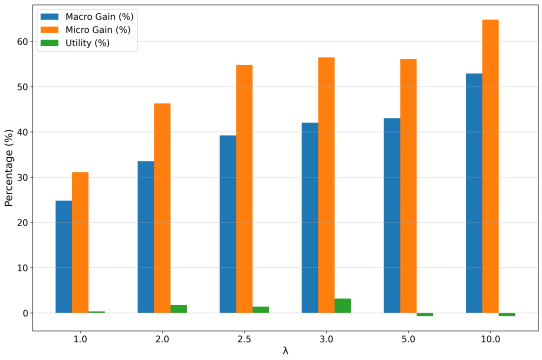


Figure 3. Comparison of Macro and Micro Gains for Race Across Different Fairness Configurations.

A. Experimental Setting

We evaluate Fair-PaperRec using datasets from prominent academic conferences, contrasting it with baseline approaches and examining the trade-off between fairness and selection quality. Each experiment is conducted 5 times individually, with *standard deviations* provided for *consistency*.

TABLE III. DISTRIBUTION OF RECOMMENDED PAPERS FROM EACH CONFERENCE.

Label	Country	Race	Multi-Fair
SIGCHI	92.02%	92.00%	92.02%
DIS	4.84%	7.69%	7.40%
IUI	3.14%	0.31%	0.56%
# Papers	351	351	351

1) Implementation Details: All experiments use `PyTorch` on a high-performance machine with two NVIDIA Quadro RTX 4000 Graphics Processing Units (GPUs). Our model is a two-hidden-layer MLP (Rectified Linear Unit (ReLU) activations, Batch Normalization), ending in a sigmoid output for acceptance probabilities. We train for 50 epochs using Adam (learning rate = 0.001), applying early stopping if no improvement occurs over 10 epochs. The fairness regularization parameter λ is tuned to balance utility and demographic parity. Each dataset is split 80/20 (training/validation) via stratified sampling, and each run is repeated five times with different random seeds to average performance metrics and capture variance.

2) Baseline: We compare our model against a baseline Demographic-Blind Model which is a conventional (MLP) model that prioritizes quality and ignores fairness constraints. This model selects the original list of papers chosen by the SIGCHI 2017 program committee.

3) Parameters: A hyperparameter λ is used for controlling the trade-off between prediction accuracy and fairness. Higher values emphasize fairness more strongly.

The weights W_c, W_r respectively denote the weighting factors assigned to the country and race attributes in the fairness loss function, as shown in Equation 3.

B. Evaluation Metrics

Diversity is assessed at both the *paper level* and the *author level*. In particular:

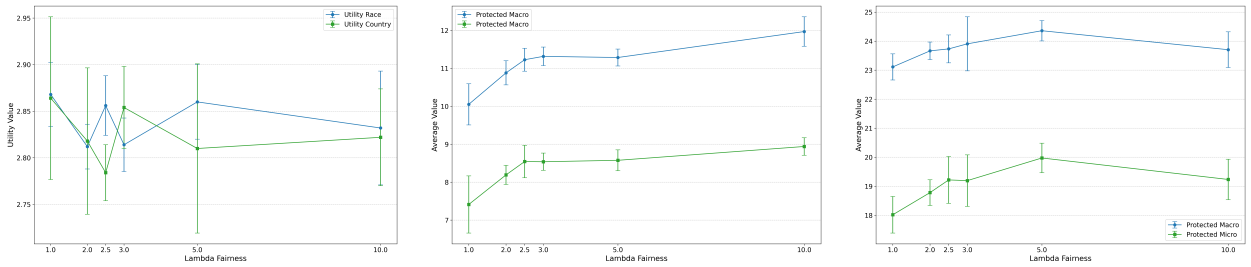
- *Macro Gain* represents the percentage increase in the diversity of each feature within the selected papers compared with the baseline, assessing the overall representation of protected groups.
- *Micro Gain* is the percentage increase in the diversity of each feature among authors of the selected papers, providing more detailed perspective on inclusivity.

A *Diversity Gain* [20] further normalizes these macro-level changes (Equation 4), capping each feature at 100 to avoid any single attribute skewing the total. The *F - measure* [20] (Equation 5) then combines this diversity improvement with the resulting utility, offering a harmonic balance between fairness gains and paper quality.

To ensure that enhancements in diversity do not compromise the quality of papers, we assess *Utility Gain* (UG_i). The utility is represented by the weighted h-index corresponding to an author's career stage—Professor, Associate Professor, Lecturer, Post-Doctoral Researcher, or Graduate Student—indicating their distribution within the dataset. Analyzing the values of the h-index in relation to a baseline determines whether equity initiatives compromise academic quality.

$$D_G = \frac{\sum_{i=1}^{n} \min(100, \text{Macro Gain}_{G_i})}{n} \tag{4}$$

$$F = 2 \times \frac{D_G \times (100 - UG_i)}{D_G + (100 - UG_i)} \tag{5}$$



(a) Utility Gain. (b) Macro/Micro for Race. (c) Macro/Micro for Country.

Figure 4. Comparison of gains across different fairness configurations.

C. Interpretation of the Results

The fairness regularization parameter (λ) was evaluated using values from 1 to 10 to examine its impact on fairness, utility, and diversity (see Table II). *RQ1*, which investigates how fairness constraints affect the utility of paper recommendations, was addressed through Figures 2 and 3. For the protected attribute "race," a λ value of 3 achieved an effective balance between diversity (both micro and macro) and utility. For "country," the optimal λ value was 2.5, which performed best across metrics. As λ increased, both micro and macro diversity gain improved, but utility decreased, indicating a reduction in the quality of recommended papers. This observation highlights the trade-off between increasing fairness and maintaining high utility, providing a clear answer to *RQ1*.

The varying optimal λ values for race and country reflect the different disparity ratios between these protected groups. This directly addresses *RQ2*, which examines how independent consideration of race and country affects fairness outcomes. The higher disparity ratio for race, which results from the smaller fraction of protected racial groups in the initial pool, requires a higher λ to achieve a balance between fairness and utility compared to country. Adjusting λ based on the specific levels of disparity in each protected group is essential to achieving optimal results. Overall, fairness interventions led to positive diversity outcomes in both micro and macro measures compared to the baseline, indicating the benefit of targeted fairness constraints.

Figure 4 presents three comparisons: (a) Utility Gain, (b) Race Fairness, and (c) Country Fairness, providing insights into utility values and diversity indicators across various λ values. The first graph shows that utility remained relatively stable for race but fluctuated significantly for country, especially at higher λ values, with larger error bars indicating greater uncertainty. Utility tended to decrease for both attributes as λ increased, further emphasizing the trade-off between fairness and utility discussed in *RQ1*.

The second and third graphs, which illustrate the protected macro and micro diversity measures for race and country, reveal that increasing λ consistently improved macro diversity for both attributes, with race showing more steady growth. In contrast, micro diversity measures, particularly for country, displayed more variability and less predictable improvement. These results suggest that macro diversity benefits are easier to achieve under higher fairness constraints, while micro-level improvements, especially for country, may require more targeted interventions. This finding is relevant to *RQ2*, as it highlights the differential effects of fairness interventions across protected attributes and the need for careful calibration of fairness constraints.

In summary, the results indicate a clear trade-off between fairness (as measured by micro and macro diversity gains) and utility, with the optimal λ values differing between race and country. This suggests that fairness policies should be tailored to the specific characteristics of each protected group to balance equity and quality effectively.

Table II presents the percentage of recommended papers from SIGCHI, DIS, and IUI across various fairness constraints. Regardless of the application of country-only, race-only, or multi-attribute fairness, SIGCHI papers maintain a dominant acceptance rate of approximately 92%, indicative of their elevated baseline acceptance rates. DIS and IUI contribute a modest but significant share of recommendations, suggesting that while SIGCHI retains prominence, the fairness constraints facilitate the inclusion of papers from smaller conferences without substantially affecting the overall distribution.

D. Ablation Study: Multi-Demographic Fairness

The objective of our ablation study was to evaluate the model's performance when optimizing fairness across multiple demographic attributes simultaneously, specifically with respect to both *country* and *race*. This ablation was conducted to address *RQ3*, which explores the impact of varying fairness weights for each attribute when multiple fairness attributes are considered together.

To ensure fairness, we removed these attributes from the input space, preventing the model from learning

TABLE IV. GAIN CALCULATIONS FOR COUNTRY AND RACE FEATURES WITH UTILITY GAIN.

λ	Weights	Country Feature		Race Feature		$UG_i(\%)$	Avg. $D_G(\%)$	Avg. F (%)
		Macro Gain (%)	Micro Gain (%)	Macro Gain (%)	Micro Gain (%)			
1	$W_r = 0.32, W_c = 0.68$	6.17	6.34	30.51	46.30	3.16	44.66	53.71
	$W_r = 1, W_c = 2$	6.73	9.15	-0.25	0.37	2.81	6.48	13.77
	$W_r = 2, W_c = 1$	7.43	11.43	12.91	16.11	3.16	25.63	40.36
2	$W_r = 0.32, W_c = 0.68$	**13.60**	**24.43**	30.51	42.22	**4.21**	55.38	68.47
	$W_r = 1, W_c = 2$	5.24	6.88	15.45	17.96	0.70	20.69	21.58
	$W_r = 2, W_c = 1$	8.36	12.86	39.49	54.26	1.75	26.31	21.58
2.5	$W_r = 0.32, W_c = 0.68$	8.63	17.33	36.58	50.37	2.46	56.46	66.31
	$W_r = 1, W_c = 2$	9.89	14.00	30.63	46.30	2.81	40.52	62.09
	$W_r = 2, W_c = 1$	9.60	17.11	42.53	56.48	1.40	59.25	69.98
3	$W_r = 0.32, W_c = 0.68$	7.15	11.42	39.49	53.89	1.40	55.98	63.45
	$W_r = 1, W_c = 2$	10.16	21.17	33.29	43.89	0.70	43.45	47.63
	$W_r = 2, W_c = 1$	9.60	18.35	42.53	55.37	2.81	61.90	47.63
5	$W_r = 0.32, W_c = 0.68$	10.80	19.38	**45.82**	**58.52**	0.70	**65.09**	**72.92**
	$W_r = 1, W_c = 2$	4.69	3.88	33.92	40.19	0.35	38.61	15.73
	$W_r = 2, W_c = 1$	7.43	11.90	39.49	52.96	5.26	52.26	15.73
10	$W_r = 0.32, W_c = 0.68$	9.60	18.34	42.53	55.37	1.40	62.92	70.89
	$W_r = 1, W_c = 2$	7.43	13.91	24.94	25.19	4.91	32.37	34.88
	$W_r = 2, W_c = 1$	7.43	11.72	35.44	47.41	-4.21	40.53	34.88

direct associations between them and the paper acceptance decisions. Instead, demographic parity loss was computed for each attribute during training, capturing deviations from fairness. The parity losses for both country and race were combined by assigning weights: W_c for country and W_r for race, with the initial weights set to $W_c = 0.68$ and $W_r = 0.32$, reflecting the distribution of protected groups.

To further explore the model's behavior and answer *RQ3*, we varied these weights, first increasing W_c while keeping W_r constant, and then increasing W_r while keeping W_c fixed. Additionally, we experimented with different values of the fairness regularization parameter λ, which controls the trade-off between fairness and utility. These experiments allowed us to observe how different weight configurations and fairness constraints influenced the model's ability to achieve demographic fairness while maintaining utility and the quality of selected papers.

The results of the ablation study, shown in Table IV, reveal that at $\lambda = 1$, assigning equal weights to both race and country ($W_r = 0.32, W_c = 0.68$) produced significant gains for race, with a Macro Gain of 30.51% and a Micro Gain of 46.3%, while country showed relatively smaller improvements (6.17% and 6.34%, respectively). However, when the weight for country was increased ($W_c = 2 \times 0.68$), diversity gains for race dropped sharply, with a negative Macro Gain (-0.25%), while country experienced slight improvements. Conversely, increasing the weight for race ($W_r = 2 \times 0.32$) resulted in improved diversity for both race and country, indicating that assigning more weight to race enhances diversity for both attributes to some degree.

At $\lambda = 2.5$, the model achieved the best balance between diversity and utility. Equal weights for race and country yielded Macro and Micro Gains of 36.58% and 50.37% for race, and 8.63% and 17.33% for country, with a low utility loss of 2.46%. This suggests that $\lambda = 2.5$ is optimal for balancing fairness and utility. As λ increases further, race diversity continues to improve (reaching 45.82% Macro Gain at $\lambda = 5$), but at the cost of decreasing utility. The different optimal λ values for race and country suggest that disparity ratios impact how fairness constraints should be weighted, with race requiring a higher λ due to its higher disparity ratio. This leads to greater race diversity gains at higher λ values, whereas country achieves optimal results at moderate λ values, such as 2.5.

These findings directly address *RQ3*, demonstrating that fairness weights must be carefully calibrated for each protected attribute. Assigning greater weight to race tends to improve diversity for both race and country, whereas increasing the weight for country may result in reduced fairness for race. The optimal balance between fairness and utility is achieved when fairness weights and λ values are adjusted based on the unique disparity ratios of each attribute.

VI. Conclusion and Future Work

This study introduces a fairness-oriented paper recommendation methodology that enhances demographic parity for race and country while maintaining academic quality. Our findings indicate that adjusting fairness requirements, including the regularization parameter λ and demographic weights, improves diversity while maintaining selection criteria.

Ablation experiments indicate that variations in race and country necessitate more stringent fairness requirements for optimal inclusion. Although beneficial, our technique lacks explicit causal modeling, which could enhance bias reduction. Investigating sophisticated designs such as Variational AutoEncoders (VAE) or graph-based models could enhance fairness and precision.

Incorporating institutional connections and combining causal fairness may improve bias mitigation. Confronting these obstacles will enhance fairness-oriented proposals, promoting a more inclusive peer review process.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under Award number OIA-1946391, Data Analytics that are Robust and Trusted (DART).

REFERENCES

[1] A. Tomkins, M. Zhang, and W. Heavlin, "Reviewer bias in single- versus double-blind peer review," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, pp. 12 708–12 713, 2017. DOI: 10.1073/pnas.1707323114.

[2] E. Shmidt and B. Jacobson, "Double-blind reviews: A step toward eliminating unconscious bias," *Clinical and Translational Gastroenterology*, vol. 13, no. 1, e00443, 2022. DOI: 10.14309/ctg.0000000000000443.

[3] V. P. Giannakakos, T. S. Karanfilian, A. D. Dimopoulos, and A. Barmettler, "Impact of author characteristics on outcomes of single- versus double-blind peer review: A systematic review of comparative studies in scientific abstracts and publications," *Scientometrics*, vol. 130, pp. 399–421, 2025. DOI: 10.1007/s11192-024-05213-x.

[4] C. Mebane, "Double-blind peer review is detrimental to scientific integrity," *Environmental Toxicology and Chemistry*, vol. 44, pp. 318–323, 2025. DOI: 10.1093/etojnl/vgae046.

[5] L. Bauersfeld, A. Romero, M. Muglikar, and D. Scaramuzza, "Cracking double-blind review: Authorship attribution with deep learning," *PLoS ONE*, vol. 18, no. 6, e0287611, Jun. 2023. DOI: 10.1371/journal.pone.0287611.

[6] N. B. Shah, "The role of author identities in peer review," *PLOS ONE*, vol. 18, no. 6, e0286206, Jun. 2023. DOI: 10.1371/journal.pone.0286206.

[7] J. Huber *et al.*, "Nobel and novice: Author prominence affects peer review," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 41, e2205779119, 2022. DOI: 10.1073/pnas.2205779119.

[8] E. Frachtenberg and K. McConville, "Metrics and methods in the evaluation of prestige bias in peer review: A case study in computer systems conferences," *PLoS ONE*, vol. 17, e0264131, 2022. DOI: 10.1371/journal.pone.0264131.

[9] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin, "Bias in peer review," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 1, pp. 2–17, 2013. DOI: 10.1002/asi.22784.

[10] C. L. Goues *et al.*, "Effectiveness of anonymization in double-blind review," *Communications of the ACM*, vol. 61, pp. 30–33, 2017. DOI: 10.1145/3208157.

[11] J. Bobadilla, R. Lara-Cabrera, Á. González-Prieto, and F. Ortega, "DeepFair: Deep learning for improving fairness in recommender systems," *Information Processing & Management*, vol. 58, no. 3, p. 102 547, May 2021. DOI: 10.1016/j.ipm.2021.102547.

[12] Y. Peng, X. Qian, and W. Song, "A re-ranking approach for two-sided fairness on recommendation systems," in *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, May 2023, pp. 312–316. DOI: 10.1145/3603781.3603836.

[13] K. Morik *et al.*, "Controlling fairness and bias in dynamic learning-to-rank," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, 2020, pp. 267–276. DOI: 10.1145/3340531.3412875.

[14] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, *Data decisions and theoretical implications when adversarially learning fair representations*, FAT/ML 2017 Workshop, 2017.

[15] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," in *Advances in Neural Information Processing Systems (NeurIPS)*, Focuses on equality of opportunity and calibration in CF., vol. 30, 2017.

[16] J. A. Bol, A. Sheffel, N. Zia, and A. Meghani, "How to address the geographical bias in academic publishing," *BMJ Glob Health*, vol. 8, no. 12, e013111, Dec. 2023. DOI: 10.1136/bmjgh-2023-013111.

[17] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin, "Bias in peer review," *J. Assoc. Inf. Sci. Technol.*, vol. 64, pp. 2–17, 2013. DOI: 10.1002/ASI.22784.

[18] W. M. Williams and S. J. Ceci, "National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track," eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 17, pp. 5360–5365, Apr. 2015, ISSN: 1091-6490. DOI: 10.1073/pnas.1418878112.

[19] R. Burke, "Multisided fairness for recommendation," in *Proceedings of the ACM RecSys '17 Workshop on Responsible Recommendation*, Como, Italy: ACM, Aug. 2017, pp. 1–4. DOI: 10.1145/3109859.3109962.

[20] R. Alsaffar and S. Gauch, "Multidimensional demographic profiles for fair paper recommendation," in *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021)*, Online: SCITEPRESS - Science and Technology Publications, Oct. 2021, pp. 199–208, ISBN: 978-989-758-533-3. DOI: 10.5220/0010655800003064.

[21] Z. Fu *et al.*, "Fairness-aware explainable recommendation over knowledge graphs," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp. 69–78. DOI: 10.1145/3397271.3401051.

[22] R. Islam, K. N. Keya, Z. Zeng, S. Pan, and J. Foulds, "Neural Fair Collaborative Filtering," in *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21)*, Proposes a multi-task learning approach for fairness in neural collaborative filtering., Amsterdam, Netherlands: ACM, Sep. 2021, pp. 148–159. DOI: 10.1145/3460231.3474603.

[23] H. Wu, C. Ma, B. Mitra, F. Diaz, and X. Liu, "A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation," *ACM Transactions on Information Systems (TOIS)*, vol. 41, no. 2, 47:1–47:29, 2022. DOI: 10.1145/3564285.

[24] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 41, no. 3, 52:1–52:43, 2023. DOI: 10.1145/3547333.

[25] B. Bulut, B. Kaya, R. Alhajj, and M. Kaya, "A paper recommendation system based on user's research interests," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 911–915. DOI: 10.1109/ASONAM.2018.8508313.