Social Network Analysis and Mining (2026) 16:92
https://doi.org/10.1007/s13278-026-01635-y

ORIGINAL ARTICLE

Fairness at no cost: how data debiasing improves equity without compromising accuracy in machine learning models

Junhua Liu[1] · Roy Ka-Wei Lee[2] · Kwan Hui Lim[3]

Received: 30 November 2025 / Accepted: 19 July 2026
© The Author(s) 2026

Abstract

Fairness in both machine learning (ML) and human decision-making is crucial, yet both are inherently prone to distinct forms of bias: algorithmic or data-driven in ML models, and subjective or inconsistency-related in humans. This study investigates fairness within the context of university admissions using a real-world dataset of 1,121 applicant profiles, of which 870 correspond to students with local qualifications. We evaluate three ML models, namely the Extreme Gradient Boosting (XGB), Bidirectional Long Short-Term Memory (Bi-LSTM), and k-Nearest Neighbors (KNN), enhanced with BERT embeddings to capture textual features from application materials. To assess *individual fairness*, we employ a consistency-based metric that measures the agreement between predictions made by ML models and evaluations from human experts of diverse backgrounds. Results demonstrate that ML models exhibit superior consistency compared to human evaluators, with improvements of more than 14%. For *group fairness*, we propose a gender-debiasing pipeline that systematically reduces explicit gender-indicative linguistic cues while preserving predictive performance. Empirically, debiasing does not degrade classification accuracy for any of the ML models, with point-estimate accuracy improving in some cases, suggesting that fairness and predictive accuracy need not be competing objectives. Overall, our findings demonstrate the potential of hybrid human–ML frameworks to promote more consistent and transparent decision-making processes in high-stakes domains such as university admissions.

1 Introduction

Fairness has become a central concern in both the predictive outcomes of machine learning (ML) models and the judgments made by human experts (Wang et al. 2023; Li et al. 2023b; Caton and Haas 2024; Mehrabi et al. 2021). In ML systems, unfairness often arises from biased data distributions, unbalanced feature representations, or algorithmic design limitations that can unintentionally reinforce societal inequities. Conversely, human decision-making is inherently shaped by subjective interpretations, contextual influences, and evaluator heterogeneity, making it susceptible to inconsistency and variability. Understanding these two distinct but interrelated sources of unfairness, algorithmic and human, is crucial for designing equitable, accountable, and socially aware decision-support systems.

In this study, we investigate fairness in the context of university admission decisions, a domain that exemplifies high-stakes human–machine interaction where outcomes significantly affect individuals' futures, and where text-derived applicant materials such as essays are increasingly processed computationally (Alvero et al. 2020; Liu et al. 2024). We employ a real-world dataset consisting of 1,121 applicant profiles (of which 870 have local qualifications) and evaluate three representative ML models: Extreme Gradient Boosting (XGB), Bidirectional Long Short-Term Memory (Bi-LSTM), and K-Nearest Neighbours (KNN). Textual components of each applicant's profile are encoded using Bidirectional Encoder Representations from Transformers (BERT), enabling the extraction of rich contextual and semantic information for downstream classification

✉ Kwan Hui Lim
kwanhui@acm.org

Junhua Liu
j@forth.ai

Roy Ka-Wei Lee
roy_lee@sutd.edu.sg

[1] Forth AI, Singapore, Singapore

[2] Singapore University of Technology and Design, Singapore, Singapore

[3] Singapore University of Social Sciences, Singapore, Singapore

Published online: 31 July 2026

tasks. The central task examined is the prediction of admission offers, a scenario that demands both predictive accuracy and procedural fairness.

To assess *individual fairness*, we analyze the consistency of decisions across ML models and human experts. Following the definition in Zemel et al. (2013), we operationalize fairness as the degree of agreement in decisions made for applicants with highly similar feature representations. The resulting consistency score quantifies how uniformly comparable candidates are treated by different evaluators. Our empirical findings reveal that ML models, particularly XGB and Bi-LSTM, outperform human evaluators by a margin of more than 14 in fairness consistency, suggesting that well-calibrated ML models can provide more stable and impartial decisions in complex evaluative settings.

We further explore *group fairness* by examining gender-related disparities in classification performance. To this end, we propose a gender-debiasing pipeline that systematically identifies and reduces explicit gender-indicative linguistic cues from applicant features while maintaining the semantic integrity of the data. The pipeline is evaluated by comparing accuracy-parity disparities, summarized through the discrimination reduction score Δ_P, between inferred gender groups before and after debiasing, together with shifts in the distribution of confidently inferred gender labels. These analyses indicate that the proposed debiasing procedure substantially reduces explicit gender-related cues without degrading model performance.

Finally, we assess the downstream impact of debiasing on predictive accuracy. Across all three models, experimental results demonstrate that debiasing does not degrade classification performance, with point-estimate accuracy improving in some cases. These findings indicate that fairness interventions can coexist with strong model accuracy, challenging the common assumption that fairness necessarily entails a trade-off with predictive power. Overall, this study provides evidence supporting the integration of machine learning models into hybrid decision-making frameworks that combine computational precision with human contextual understanding, promoting equitable and transparent admissions practices.

1.1 Contributions

In summary, this work makes the following key contributions[1]:

1. Fairness assessment and debiasing framework. We propose a set of generalizable pipelines for identifying, quantifying, and mitigating both algorithmic and linguistic biases in machine learning models, and empirically evaluate their effectiveness in improving fairness without compromising model interpretability.
2. Empirical evaluation of fairness–accuracy trade-offs. Through extensive experiments, we demonstrate that systematic data debiasing not only preserves but can also enhance model performance, thereby challenging the widely held assumption that improving fairness necessarily reduces predictive accuracy.
3. Comprehensive analysis of human and machine decision-making. We conduct an empirical study on both individual and group fairness using a real-world university admissions dataset. Our findings reveal that ML models consistently outperform human experts in decision consistency, while achieving comparable predictive accuracy.
4. Research questions and insights. This study is guided by six research questions that collectively address key dimensions of fairness and bias in hybrid human–AI decision systems:

RQ1 *Can decision inconsistencies in human evaluations be detected and quantified in terms of reduced individual fairness?*

RQ2 *Are ML models more consistent in decision-making than human evaluators?*

RQ3 *Can ML models outperform human experts in admission decision accuracy?*

RQ4 *How can gender bias be systematically identified in human decision processes?*

RQ5 *Does debiasing input features degrade, preserve, or improve classification performance?*

[1] This journal article is a substantially expanded version of our earlier conference paper (Liu et al. 2025b), with more than 100% new materials. Specifically, this extended version introduces: (i) a new group fairness framework with a complete gender-debiasing pipeline, (ii) comprehensive experiments comparing original versus debiased data, (iii) formalization of discrimination-reduction metrics, (iv) expanded methodological details including hyperparameter search, model configurations, and confidence-interval analysis, and (v) deeper discussions, limitations, and future directions.

RQ6 *Is there a measurable relationship between model complexity and robustness to biased features?*

Together, these contributions provide new empirical and methodological insights into the comparative fairness of human and machine decisions, emphasizing that fairness enhancement and predictive performance can be jointly optimized within well-calibrated ML systems.

2 Related work

2.1 Algorithmic fairness

The concept of *individual fairness*, introduced by Dwork et al. (2012), requires that similar individuals receive similar outcomes. This principle is formalized through a Lipschitz continuity constraint and is closely connected to the mathematical foundations of differential privacy (Dwork 2006). Lahoti et al. (2019) operationalized individual fairness through pairwise fair representations, while (Horesh et al. 2020) proposed paired-consistency as a model-agnostic approach to fairness regularization. However, individual fairness alone does not guarantee *group fairness*, which concerns equitable treatment across demographic groups defined by protected attributes such as gender, race, or age. Hardt et al. (2016) formalized the notion of *equal opportunity*, requiring equalized true positive rates across groups, which has become a foundational group fairness criterion alongside demographic parity and equalized odds. The interpretation of fairness depends heavily on the "worldview" employed (Mahoney et al. 2020), and many practical fairness metrics, particularly those derived from confusion matrices (Castelnovo et al. 2022), reflect these underlying assumptions.

A central question in the fairness literature is whether improving fairness necessarily sacrifices predictive accuracy. Dutta et al. (2020) provided a theoretical perspective on this trade-off through mismatched hypothesis testing, showing that the relationship between fairness and accuracy is more nuanced than commonly assumed.

Recent advances in large language models (LLMs) have expanded the use of automated decision-making in high-stakes contexts. However, emerging evidence shows that LLMs may reproduce or amplify human-like biases (Caliskan et al. 2017), underscoring the need for systematic evaluation of bias in real-world applications and the importance of exploring mitigation strategies. Li et al. (2023a) provide a comprehensive survey of fairness in LLMs, taxonomizing bias evaluation metrics and debiasing techniques for pre-trained language models; our use of BERT-based representations situates this work within that landscape. Beyond lexical interventions such as ours, recent representation-level approaches aim to remove demographic signals directly from learned embeddings: BATED (Li et al. 2025) learns fair representations for pre-trained language models via biased teacher-guided disentanglement, targeting latent bias that term-removal methods may leave behind. Complementing both directions, Robertson et al. (2024) propose a human-in-the-loop fairness-aware model selection framework for navigating complex fairness objective landscapes, reflecting the multi-dimensional nature of fairness objectives that motivates our discussion of metric choice. In related work, Liu et al. (2024) proposed LLM-enhanced hierarchical attention networks for objective decision assessment in similar admissions contexts.

A further practical challenge, directly relevant to our setting, is that protected attributes are often unobserved and must be inferred by proxy. Chen et al. (2019) show that estimating disparities using proxy-inferred protected classes introduces systematic estimation error, and Kallus et al. (2022) characterize the fundamental limits of assessing algorithmic fairness when the protected class is unobserved. These results inform our choice of group fairness criterion in Sect. 5, where gender labels are themselves model-inferred.

Fairness has also been studied extensively in recommender systems and graph-based settings. Shao et al. (2022) proposed FairCF, a fairness-aware collaborative filtering framework, while their subsequent work (Shao et al. 2024) formalized average user-side counterfactual fairness for collaborative filtering. In the domain of social network analysis, Froese et al. (2024) studied modification-fair cluster editing, and Honcharov and Soundarajan (2026) addressed fair dense subgraph discovery. Our work complements these approaches by focusing on fairness in text-derived decision-making rather than recommendation or graph mining settings.

A broad body of prior work has focused on bias mitigation in machine learning pipelines. For instance, Sha et al. (2023) explored robust preprocessing methods to achieve fair and accurate predictive modeling in educational domains, while (Zemel et al. 2013) proposed foundational approaches for fair representation learning. Complementing these computational strategies, Alur et al. (2023) surveyed auditing techniques for assessing fairness and the quality of human expertise. Together, these studies highlight the multifaceted nature of fairness and the importance of principled methodologies for detecting and mitigating bias.

2.2 Fairness in university admissions

Fairness in university admissions has been widely examined across several disciplines. Early studies often relied on small-scale qualitative evaluations (Zimdars 2010) or

focused on high-level policy analyses (Pitman 2016), while others concentrated on specific subcomponents such as standardized testing (Van Busum and Fang 2023). Although informative, the limited scope of these studies constrains the generalizability of their findings. Additional research has evaluated human perceptions of fairness in algorithmic decision-making (Marcinkowski et al. 2020), highlighting differences in how fairness is interpreted when decisions are made by humans versus AI systems.

More quantitative approaches have emerged in recent years. Puranik et al. (2022) proposed the Fair-Greedy policy, balancing score maximization with fair representation by dynamically updating the likelihood of selecting applicants from under-represented groups. Bhattacharya et al. (2017) used a threshold-crossing framework to reveal socioeconomic disparities in admission decisions and to measure differential thresholds across demographic segments.

Despite these advances, there remains comparatively little work that jointly evaluates human and machine decision processes on the same dataset, particularly in contexts where textual features play a substantial role and may encode latent stereotypes or subjective language.

2.3 Differences with earlier work

While prior work has explored both algorithmic fairness and fairness in educational selection processes, relatively few studies have examined biases embedded in text-derived features, even though such features are increasingly used in modern admissions systems (Alvero et al. 2020; Liu et al. 2024). To the best of our knowledge, this study is the first to propose and empirically validate a systematic pipeline for mitigating gender-indicative linguistic bias in applicant text while simultaneously improving individual fairness.

Furthermore, we challenge the common assumption that fairness interventions necessarily compromise model performance. Our empirical results show that debiasing not only preserves predictive accuracy but can in fact enhance it. This finding contributes new evidence to the ongoing discourse on fairness–accuracy trade-offs and demonstrates that well-designed debiasing strategies can improve fairness and utility simultaneously.

3 Data processing

3.1 Dataset

We use the university admissions dataset introduced in Liu et al. (2025a), comprising 1,121 applicant profiles from a single admissions cycle. Of these, 870 profiles correspond to students with local qualifications (i.e., Singapore's GCE A-Level and O-Level examinations), which form the primary subset used in our classification experiments. Table 1 summarizes the original data schema and provides sample entries for each field. The dataset contains four primary textual components: General Certificate of Education Advanced Level (GCEA), General Certificate of Education Ordinary Level (GCEO), Personal Insight Questions (PIQ), and Leadership Experience. For reference and qualitative inspection, these components are also concatenated into a unified `Combined` document; however, embeddings are generated separately for each of the four features (see Sect. 3.3).

Each profile is assigned a binary label `Type`, indicating whether the applicant received an admission offer (*Offered*) or not. For all experiments, we adopt an 80%–10%–10% stratified split for training, validation, and testing, respectively.

3.2 Gender debiasing

The debiasing pipeline aims to reduce explicit gender-indicative language from the PIQ and Leadership features while preserving the original semantic content as much as possible. To achieve this, we curated a comprehensive lexicon of gender-specific terms, including pronouns (e.g., "he", "she"), honorifics (e.g., "Mr.", "Mrs."), and gender-associated nouns (e.g., "son", "daughter"). These terms were systematically filtered out using a hybrid approach that combines spaCy-based tokenization with regular-expression

Table 1 Data format and sample entry

Field	Format	Sample entry
GCEA	[GCEA] School:{school}, UAS:{score}; Grades:{subject-grade pairs}	[GCEA] School:HCI, UAS:90.0; Grades:H1 PROJECT WORK A, H1 GENERAL PAPER A, H2 ECONOMICS A, H2 CHEMISTRY A,..
GCEO	[GCEO] {subject-grade pairs}	[GCEO] HIGHER CHINESE B3, ENGLISH A1, ELEMENTARY MATHEMATICS A2,..
Leadership	[[Leadership] {activity}, Level:{role}, Year:{year}, Category:{type}, Participation:{level}]	[Leadership] Mind Sports Club, Level:Captain, Year:2023, Category:Sports, Participation:Executive Committee
PIQ	[PIQ{n}] {essay text}, where n = 1,2,3,4,5	[PIQ1] Visiting the university for a Learning Journey in Primary 5 left a lasting impression on me..
Type	Offered / Not Offered	Offered

pattern matching. This ensures precise removal of gender cues while minimizing unintended alteration of contextually meaningful text.

To evaluate the effectiveness of the debiasing procedure, we infer gender using a publicly available gender classification model: a `bert-base-uncased` encoder with a classification layer on top, fine-tuned for binary (male/female) gender classification on external gender-labeled data. A prediction is accepted only when its confidence exceeds 0.8; otherwise, the profile is assigned an *Unknown* label. The classifier is not trained or fine-tuned on our admissions data, which contains no gender annotations. The same classifier is applied without modification to both the original and the debiased text. Any shift in the inferred label distribution between the two conditions is therefore attributable to the debiasing intervention on the text, not to changes in the inference model. We note that ground-truth gender labels were unavailable due to privacy constraints, and the classifier output serves as our best available proxy. Using the same pretrained model for both gender inference and debiasing evaluation introduces a potential circularity concern; we discuss this limitation further in Sect. 6.

We emphasize that these confidence intervals do not determine or validate the gender labels themselves; gender assignment is performed solely by the classifier described above. Rather, once profiles have been grouped by their inferred gender label, we compute 95% confidence intervals for the *admission rate within each inferred group*, quantifying the sampling uncertainty of these group-level rates given the finite group sizes. The intervals are computed using the normal approximation to the binomial distribution. For each inferred gender group, the standard error is given by:

$$SE = \sqrt{\frac{p(1-p)}{n}} \tag{1}$$

where p denotes the observed admission rate and n is the corresponding sample size. The 95% confidence interval is then:

$$CI = p \pm (1.96 \times SE) \tag{2}$$

where 1.96 is the critical value of the standard normal distribution. This interval provides an estimate of the range within which the true population admission rate is expected to lie with 95% confidence, accounting for sampling variability within each inferred gender group.

Using this classifier, we compare the distributions of inferred gender labels before and after debiasing, applying the same 0.8 confidence threshold in both conditions.

Table 2 summarizes the gender inference results across all 1,121 profiles in the dataset. The debiasing intervention shows substantial impact, with pronounced shifts in the distribution of inferred gender labels after the removal of gender-specific language.

Specifically, the proportion of confidently inferred gender labels decreases sharply after debiasing: female classifications drop by 83.6% (from 46.7% to 7.7%), and male classifications decline by 75.1% (from 43.6% to 10.9%). Consequently, the *Unknown* category expands from 9.6% to 81.4% of all profiles. These shifts indicate that the debiasing procedure effectively reduces explicit gender-identifying linguistic cues, thereby substantially reducing the model's reliance on gender-associated features. We interpret the sharp rise in *Unknown* labels as an expected consequence of the intervention rather than a measurement failure: because the classifier's evidence for gender consists largely of the very lexical cues that debiasing removes, its inability to make confident predictions on the debiased text is, by construction, direct evidence that those cues have been removed. At the same time, we acknowledge that the large proportion of Unknown-labelled profiles after debiasing limits the statistical power of subsequent group fairness comparisons, and that latent gender signals (e.g., writing style, extracurricular patterns, or semantic associations learned by BERT) may persist despite lexical debiasing (see Sect. 6).

Table 2 Gender inference results before and after debiasing

Category	Before	After	Change (%)
Female	524 (46.7%)	86 (7.7%)	−83.6%
Male	489 (43.6%)	122 (10.9%)	−75.1%
Unknown	108 (9.6%)	913 (81.4%)	+745.4%
Total Profiles	1,121		

3.3 Feature embedding

To prepare the data for fairness assessment and classification, we generate contextual embeddings using the pretrained `bert-base-uncased` model. Each textual feature (GCEA, GCEO, PIQ, Leadership) is tokenized and padded to a maximum sequence length of 512 tokens before being processed through BERT. We extract the 768-dimensional `[CLS]` token embedding as the representation for each feature.

The resulting embeddings are retained both individually and in concatenated form. For the concatenated representation, the feature-level embeddings are combined to produce a 3,072-dimensional vector (4 features × 768 dimensions) that captures holistic profile information. All embedding procedures are applied consistently to both the original and debiased datasets.

3.4 Human decisions

We evaluate the capacity of machine learning models to replicate human judgment by focusing on three decision points used in the admissions workflow: Shortlisting (SL), Admission Recommendation (AR), and Final Offer (OF). These decisions are not annotations produced for this study; they are the actual outcomes recorded at each stage of the real admissions process from which the dataset originates (Liu et al. 2025a). Each decision stage is converted into a binary classification target, where "Shortlisted", "Recommended", and "Offered" are mapped to 1, and all remaining outcomes are mapped to 0.

We treat the Final Offer outcome as the ground-truth label for the classification experiments: the `Type` field (*Offered* / *Not Offered*) is precisely the binarized OF decision. Consequently, classification metrics (precision, recall, F1-score, accuracy) can be computed for the earlier SL and AR stages by comparing them against this final outcome, whereas the OF stage is itself the reference and thus has no independent label to be scored against, so it is evaluated only in terms of decision consistency. We further note that the dataset records one institutional decision per applicant per stage, without identifiers of the individual evaluators involved; inter-rater agreement between specific evaluators therefore cannot be computed, and our consistency metric (Sect. 4) instead measures cross-applicant consistency, i.e., whether similar applicants received similar decisions.

3.5 Data splitting

To ensure reliable and comparable evaluation across experiments, we employ a stratified data-splitting strategy using `sklearn`'s `train_test_split` with a fixed random seed of 42. The dataset is partitioned into 80% training, 10% validation, and 10% test subsets. Identical split indices are applied to both the original and debiased datasets, ensuring that differences in performance arise solely from the debiasing intervention rather than from variations in sample composition.

4 Individual fairness

4.1 Formulation

Individual fairness is grounded in the principle that individuals with similar characteristics should receive similar decisions. Following the formulation in Zemel et al. (2013), we quantify this notion using a consistency metric that measures how similarly a classifier treats applicants with comparable profiles. An overview of our assessment pipeline is shown in Fig. 1.

Formally, given a dataset of N applicants, the consistency score $\mathcal{C}$ for a classifier is defined as:

$$\mathcal{C} = 1 - \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - \frac{1}{k}\sum_{j \in knn(i)} \hat{y}_j| \tag{3}$$

where $\hat{y}_i$ denotes the classifier's predicted label for applicant i, and $\frac{1}{k}\sum_{j\in knn(i)}\hat{y}_j$ represents the average prediction of its k nearest neighbors in the embedding space. The $knn(i)$ set contains the k most similar applicants to i, identified using cosine similarity over their feature embeddings.

The resulting score lies in the interval [0, 1], with higher values indicating greater alignment between the classifier's prediction for an applicant and the predictions made for highly similar applicants. Intuitively, the metric captures the mean absolute deviation between each prediction and the average prediction of its neighbors, thereby operationalizing the degree to which the classifier treats similar profiles similarly. We note that consistency is a necessary but not sufficient condition for individual fairness: a model that consistently assigns unfavourable outcomes to a disadvantaged group would score highly on this metric while remaining substantively unfair. Moreover, if the similarity space itself encodes bias (e.g., through pretrained BERT embeddings), consistent treatment within that space does not guarantee normatively fair outcomes. A further caveat concerns the comparison between human and machine decision-makers: because similarity is defined via nearest neighbours in the BERT embedding space, the metric is computed in a representation that is native to the ML models but not to human evaluators. Elements of an applicant profile such as subject abbreviations and grading codes may carry little signal for a general-purpose embedding model while being highly meaningful to a human expert, so two profiles that appear near-identical in embedding space may differ in ways that are legitimately salient to humans. Human consistency scores under this metric should therefore be read as conservative estimates relative to the humans' own (unobservable) similarity judgments, and part of the human–machine consistency gap may reflect the choice of representation rather than intrinsic differences in decision stability. We therefore interpret the consistency score as a proxy for one important dimension of fairness rather than a complete fairness guarantee.

4.2 Research questions and evaluation overview

This section addresses the first three research questions introduced in Section 1: whether decision inconsistency in

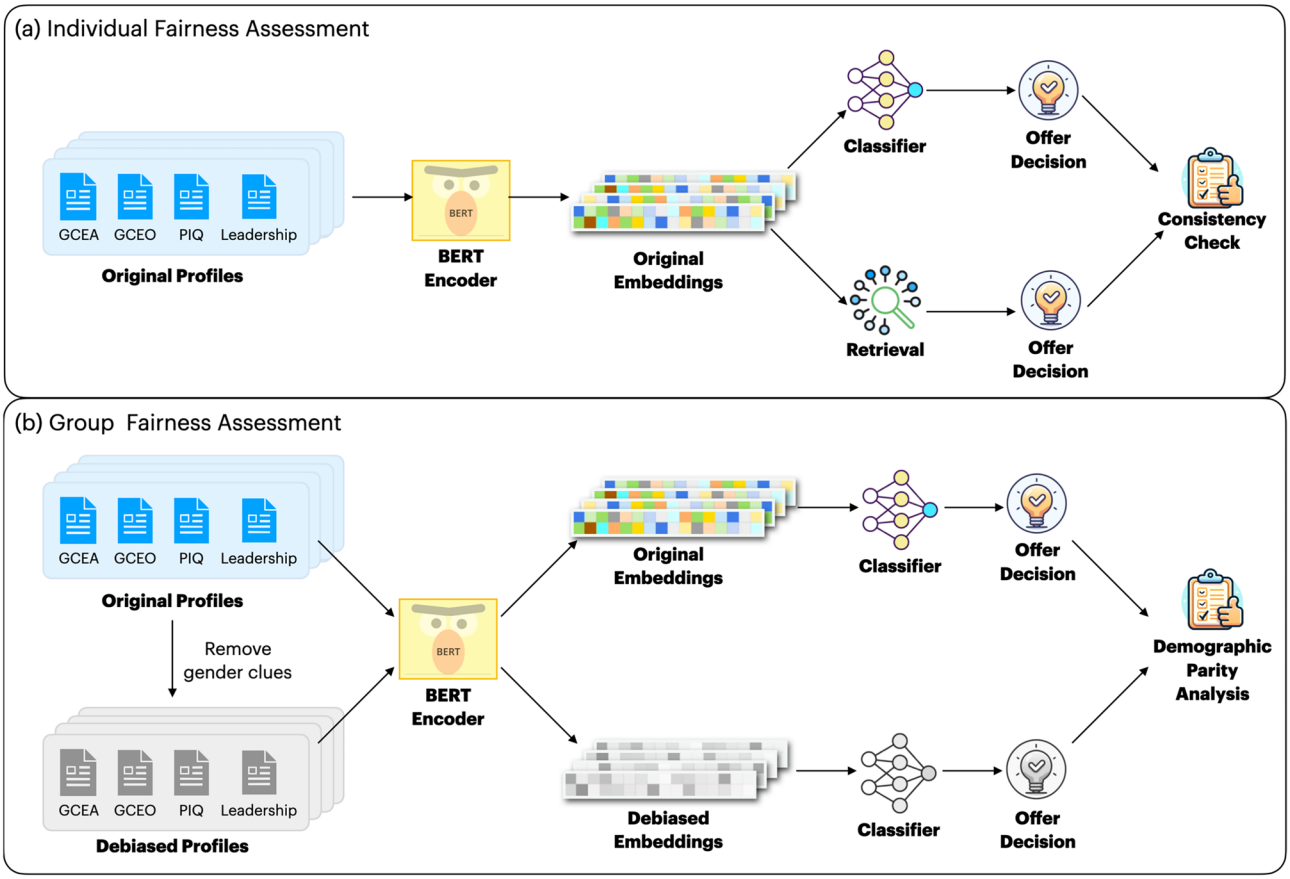


Fig. 1 Proposed pipelines to assess Individual Fairness

human evaluations can be systematically identified (RQ1), whether ML models make more consistent decisions than humans (RQ2), and whether ML models can match or exceed human accuracy in admission decisions (RQ3).

To address these questions, we evaluate both human decisions (Shortlisting, Admission Recommendation, and Offer) and machine learning models using standard classification metrics, precision, recall, F1-score, and accuracy, alongside our consistency-based individual fairness metric (Sect. 4). Table 4 summarizes the results across all models and decision stages.

4.3 Methodology

Our evaluation framework assesses individual fairness across the full admissions pipeline by analyzing three human decision points, Shortlisting (SL), Admission Recommendation (AR), and Offer (OF), alongside two supervised machine learning classifiers (XGB and Bi-LSTM) and a retrieval-based k-Nearest Neighbour (KNN) model. Central to our methodology is the construction of a feature-reranked similarity matrix derived from the BERT embeddings described earlier. This matrix provides a principled measure of similarity between applicant profiles and enables the identification of nearest neighbors required for consistency evaluation.

4.3.1 Implementation

We adopt three representative modelling paradigms: (i) gradient-boosted decision trees, (ii) neural sequence models, and (iii) similarity-based retrieval models. For each paradigm, we evaluate one empirically strong model, XGB, Bi-LSTM, and KNN, respectively.

The XGB classifier is implemented using the `xgboost` library with optimized tree-boosting parameters. The Bi-LSTM model is implemented in PyTorch and consists of a bidirectional LSTM encoder followed by fully connected layers with ReLU activation and a LogSoftmax output layer. A batch size of 1 was selected through the hyperparameter search; this effectively implements stochastic gradient descent, which serves as an implicit regularizer for the relatively small dataset ($N = 870$) and helps reduce overfitting. For the retrieval-based approach, we use the FAISS library to perform efficient similarity search over the embedding

Table 3 Hyperparameter search spaces and optimal values

Model	Hyperparameter	Search space	Optimal
XGB	n_estimators	randint(100, 1000)	415
	max_depth	randint(3, 15)	11
	learning_rate	uniform(0.001, 0.3)	0.014
	subsample	uniform(0.5, 0.3)	0.61
	colsample_bytree	uniform(0.7, 0.3)	0.78
	reg_alpha	uniform(0, 2)	1.13
	reg_lambda	uniform(0, 4)	1.54
Bi-LSTM	hidden_dim	[128, 256, 512, 1024, 2048]	1024
	dropout	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]	0.6
	learning_rate	[1e-4, 1e-3, 3e-4, 3e-3, 5e-4, 5e-3]	1e-3
	batch_size	[1, 4, 8, 16, 32, 64]	1
KNN	k	[3, 4, 5, 6, 7]	5
	similarity	["cosine", "L2"]	L2

`Uniform(loc, scale)` follows the scipy convention, denoting the continuous uniform distribution on $[\mathrm{loc},\ \mathrm{loc} + \mathrm{scale}]$

space, enabling memory-efficient nearest-neighbor retrieval and batch-based prediction.

4.3.2 Hyperparameters

Hyperparameters are selected using randomized search over the validation set. Each training configuration is run for up to 20 epochs with an early-stopping patience of 5 epochs. The best configuration for each model is selected based on average validation accuracy. Table 3 summarizes the search spaces and optimal hyperparameter values for all classifiers. All model selection decisions (hyperparameters and early stopping) are made exclusively on the validation split; the held-out test split (Sect. 3.5) is used only once, for the final evaluation reported here. As this is a single held-out evaluation, the reported test metrics carry the sampling variability quantified in Sect. 4.4.4, and no optimism-correction procedure (e.g., nested cross-validation) is applied.

4.3.3 Metrics

For each decision point and model, we compute standard classification metrics, precision, recall, F1-score, and accuracy, as well as the consistency-based individual fairness metric defined in Sect. 4. The consistency evaluation uses $k = 5$ nearest neighbors, chosen to balance local ranking stability and sensitivity to small variations in profile embeddings. All models operate on the same BERT-based feature space to ensure comparability between human decisions and machine predictions. This unified evaluation pipeline enables a systematic assessment of how closely different decision-making mechanisms adhere to the fairness principle of treating similar applicants similarly, while simultaneously measuring their predictive performance.

4.4 Experimental results

Table 4 summarizes the classification performance and individual fairness consistency for human decision points and machine learning models.

4.4.1 Human decisions

Human shortlisting (SL) decisions achieve the highest precision across all decision points and models, with a value of 84.64%. This indicates strong reliability in correctly identifying qualified candidates during the initial screening stage. The Admission Recommendation (AR) stage shows slightly lower but still robust performance, with an accuracy of 80.87% and an F1-score of 83.21%. These results suggest that human evaluators maintain reasonably consistent and effective judgment during the intermediate phase of the admissions workflow.

Final Offer (OF) decisions do not include traditional classification metrics because the OF outcome is itself the reference label against which the other decision stages and models are scored (Sect. 3.4); OF is therefore evaluated only in terms of fairness consistency.

Table 4 Individual fairness: classification performance and fairness consistency

Model	P	R	F1	A	C(AR)	C(OF)
Human Decisions						
Shortlist(SL)	**0.8464**	**0.8418**	0.8156	0.8155	–	–
Adm. Rec. (AR)	0.8011	0.8193	**0.8321**	0.8087	0.5632	–
Offer (OF)	–	–	–	–	–	0.6023
ML Models						
KNN (k=5)	0.6886	0.6707	0.6897	0.6897	–	–
XGB	0.7902	0.7859	0.7878	0.7931	0.7521	0.7477
Bi-LSTM	0.8291	0.8178	0.8176	**0.8276**	**0.8073**	**0.7797**

Best results are bolded

P precision, *R* Recall, *A* Accuracy, *C* Consistency

4.4.2 Machine learning models

Among the machine learning models, the Bi-LSTM achieves the strongest predictive performance, with an F1-score of 81.76% and an accuracy of 82.76%. These results are comparable to those of human evaluators, with the Bi-LSTM's point-estimate accuracy slightly exceeding that of the human decision points. This indicates the model's ability to replicate human-like decision behaviour at a comparable level of predictive performance.

The XGB classifier also performs competitively, achieving an F1-score of 78.78% and an accuracy of 79.31%, reinforcing the effectiveness of gradient-boosted decision trees in structured text-derived feature settings.

By contrast, the KNN model (with $k = 5$) yields substantially lower performance, with both F1-score and accuracy at 68.97%. This reflects the limitations of non-parametric retrieval-based methods in capturing the nuanced patterns present in applicant profiles and highlights the importance of selecting appropriate model architectures for complex decision-making tasks. Note that the consistency score is not reported for KNN because the model's predictions are deterministically derived from the same nearest-neighbor retrieval used to compute consistency, making the metric trivially perfect and uninformative in this case.

4.4.3 Fairness consistency

Substantial differences emerge between human and machine decision-making when examining individual fairness consistency. For human evaluators, consistency is moderate: the AR stage achieves a score of 56.32%, while the OF stage shows slightly higher consistency at 60.23%. These levels indicate that similar applicants do not always receive similar outcomes, suggesting variability in human judgment.

Machine learning models, however, display markedly higher consistency scores. The XGB classifier reaches 75.21% (AR) and 74.77% (OF), a substantial improvement over human decisions. The Bi-LSTM achieves the highest consistency across all settings, with 80.73% (AR) and 77.97% (OF). These results indicate that ML models, particularly neural architectures, not only approximate human-level predictive accuracy but also outperform human evaluators in delivering more consistent decisions for similar applicants.

4.4.4 Statistical uncertainty

As all experiments use a single fixed random seed and a single held-out test split (Sect. 3.5), we quantify the sampling uncertainty of the reported accuracies analytically. For a model evaluated on the held-out test set of $n = 87$ applicants, the 95% confidence interval of a reported accuracy p under the normal approximation to the binomial distribution is $p \pm 1.96\sqrt{p(1-p)/n}$. This yields 0.8276 ± 0.079 for Bi-LSTM, 0.7931 ± 0.085 for XGB, and 0.6897 ± 0.097 for KNN. These intervals overlap with the corresponding human decision-point accuracies, so the accuracy differences between ML models and human evaluators should be interpreted as indicative rather than statistically established. By contrast, the human–machine gap in consistency scores (approximately 15–24 percentage points) is considerably larger than the accuracy-level sampling variability observed here. Consistency scores are means of absolute deviations rather than simple proportions, and formal paired significance tests (e.g., McNemar's test) require per-instance predictions. We discuss this limitation in Sect. 6.

4.5 Analysis and interpretations

4.5.1 RQ1: identifying decision inconsistency

Although early-stage human decisions, Shortlisting (SL) and Admission Recommendation (AR), exhibit strong classification metrics (both achieving above 80% in F1-score and accuracy), their consistency scores are markedly lower. In particular, AR achieves only 56.32% consistency, while Offer (OF) decisions reach 60.23%. This discrepancy between predictive performance and consistency suggests the presence of subjective variation or inter-rater inconsistency in the evaluation process. We note that such inconsistency may also reflect legitimate contextual judgment, evaluator heterogeneity, or differing weighting of qualitative factors across decision stages, rather than bias per se.

The decrease in consistency from OF to AR (60.23% $\rightarrow$ 56.32%) may indicate that inconsistencies become more pronounced in intermediate decision stages, where evaluation criteria are more complex and less standardized. This points to potential variability in how different reviewers interpret qualitative or borderline cases, highlighting the need for decision-support tools that help maintain uniformity across reviewers and stages.

4.5.2 RQ2: consistency comparison between ML and human

Machine learning models exhibit substantially higher fairness consistency compared to human evaluators. The Bi-LSTM achieves 80.73% (AR) and 77.97% (OF), representing improvements of approximately 24% and 18%, respectively, over human decisions. The XGB model also demonstrates pronounced gains, attaining 75.21% (AR) and 74.77% (OF).

These results show that ML models produce markedly more stable and standardized decision patterns. Unlike human evaluators, whose decisions may fluctuate based on subjective interpretation, contextual factors, or evaluator heterogeneity, ML models apply uniform criteria across similar applicants. This substantial consistency gap underscores the value of ML systems in reducing decision variability, though we reiterate that part of the gap may be attributable to the embedding-based representation in which consistency is measured (Sect. 4).

4.5.3 RQ3: accuracy comparison between ML and human

The accuracy comparison reveals a nuanced trend. Human shortlisting decisions achieve the highest precision (84.64%), reflecting strong ability to identify qualified candidates early in the process. However, the Bi-LSTM model performs competitively, achieving the highest accuracy (82.76%) and an F1-score of 81.76%, closely matching or surpassing human performance across multiple metrics. Given the overlapping confidence intervals reported above, we interpret these accuracy differences as comparable performance rather than a demonstrated advantage. At the same time, Bi-LSTM maintains this level of performance while achieving considerably higher consistency scores.

The XGB model, although slightly weaker in predictive performance (F1-score of 78.78%), still exceeds human evaluators in consistency. This suggests that certain ML architectures may offer a favorable balance between accuracy and fairness consistency, depending on their complexity and representational capacity.

4.5.4 Discussion

These findings yield several implications for improving admissions workflows. First, the high consistency of ML models, especially Bi-LSTM, demonstrates their potential to serve as reliable standardization mechanisms within the admissions pipeline. Second, human evaluators and ML models have complementary strengths: humans achieve the highest precision in shortlisting (84.64%), while the models are more consistent. Hybrid decision systems may therefore be preferable to either alone, with ML models providing consistency checks or secondary scoring alongside human judgment. Third, the decline in human consistency between the OF and AR stages highlights the need for structured decision-support tools, particularly in intermediate stages where evaluation criteria are more complex and less standardized. Finally, the large performance gap between KNN and Bi-LSTM indicates that expressive architectures are important for capturing the nuanced features of applicant profiles: architectural choice affects not only accuracy but also the consistency with which similar applicants are treated.

In summary, while human evaluators demonstrate strong predictive capabilities, their decisions exhibit notable inconsistency, likely stemming from decision variability or subjective interpretation. Machine learning models, particularly neural models, offer both competitive accuracy and substantially higher consistency. These results support the integration of ML-driven consistency checks into admissions workflows and motivate hybrid approaches that leverage the complementary strengths of human judgment and machine learning to enhance reliability in high-stakes decision-making.

5 Group fairness

5.1 Formulation

Group fairness addresses systematic disparities in model behaviour across demographic groups. In this work, we focus on *accuracy parity*, which requires that a classifier achieve comparable accuracy for different groups defined by a protected attribute. Formally, for a binary protected attribute $a \in \{0, 1\}$, accuracy parity is satisfied when

$$\Delta_{Acc} = |P(\hat{Y} = Y|a = 0) - P(\hat{Y} = Y|a = 1)| = 0,$$

where $\hat{Y}$ denotes the classifier's prediction. A non-zero value indicates a violation of accuracy parity and forms one type of *disparate mistreatment* (Zafar et al. 2017).

We note that accuracy parity is one of several group fairness criteria proposed in the literature; others include demographic parity, equal opportunity, and equalized odds (Caton and Haas 2024; Hardt et al. 2016). Our focus on accuracy parity is a deliberate methodological choice given the characteristics of our setting. Because gender labels in our data are inferred by a classifier rather than observed (Sect. 3.2), all group-conditional quantities are estimated with respect to a noisy proxy attribute. Prior work shows that disparity estimates computed on proxy-inferred protected classes can be systematically biased (Chen et al. 2019; Kallus et al. 2022). Criteria that further condition on the true outcome within each group, such as equal opportunity and equalized odds, partition the (already noisy) groups into smaller outcome-conditional cells, compounding this estimation error, particularly at our sample size. Accuracy parity, which conditions only on group membership, is comparatively more robust in this setting, and we therefore adopt it as our primary criterion while acknowledging the evaluation of additional fairness metrics on data with verified protected attributes as an important direction for future work (Sect. 6).

In our setting, we examine gender-related disparities. Unlike settings where protected attributes are explicitly available, gender cues in our dataset appear implicitly through linguistic indicators embedded in textual features (e.g., pronouns, titles, and gendered nouns). These cues are reduced in the debiasing process described in Sect. 3.2, substantially reducing explicit gender signals in the feature space, though latent proxies may remain.

Let $D_{\text{original}} = \{(X_{\text{original}}, Y)\}$ denote the original dataset, where X_{original} represents the unmodified input features and Y is the binary admission label (*Offered* / *Not Offered*). Similarly, let $D_{\text{debiased}} = \{(X_{\text{debiased}}, Y)\}$ denote the debiased dataset, where X_{debiased} contains the features after removal of gender-indicative cues.

To quantify the extent to which debiasing reduces group-level disparities, we define the discrimination reduction score:

$$\Delta_P = |\Delta_{Acc,D_{original}} - \Delta_{Acc,D_{debiased}}|$$

A larger Δ_P indicates a greater decrease in accuracy disparity between gender groups as a result of debiasing.

5.2 Research questions and evaluation overview

The remainder of this section addresses the final three research questions introduced in Section 1: how gender bias in human admission decisions can be systematically identified and quantified (RQ4), whether removing gender-indicative features through debiasing leads to a loss in classification performance (RQ5), and whether there is a relationship between model complexity and a model's ability to handle biased or debiased features (RQ6).

5.3 Methodology

Figure 1b illustrates the proposed pipeline for assessing group fairness. The original dataset first undergoes the debiasing procedure described in Sect. 3.2, which removes gender-indicative linguistic cues from the PIQ and Leadership features. Both the original and debiased datasets are then transformed into contextual embeddings using the pre-trained `bert-base-uncased` model, with each feature encoded as a 768-dimensional vector.

These embeddings serve as input for the various classification approaches. The XGB and Bi-LSTM models are trained on the embedded representations to predict offer decisions, whereas the KNN model operates in a non-parametric manner and performs prediction directly via nearest-neighbour retrieval.

To quantify the effect of debiasing, we compare classification accuracy, consistency, and discrimination scores across the original and debiased datasets. For each model, the exact same architecture and the same hyperparameter configuration (Sect. 4, Table 3) are used in both conditions. The two conditions differ only in their input data (original versus debiased text). This ensures that any performance differences can be attributed to the debiasing intervention rather than model configuration.

Table 5 Group fairness: accuracy and consistency comparisons among ML classifiers using original and debiased data

Model	Data	Accuracy	Consistency	ΔP
XGB	Original	0.8046	0.7303	0.0
	Debiased	0.8046	0.7266	
Bi-LSTM	Original	0.8276	0.7567	0.0195
	Debiased	0.8471	0.7673	
KNN	Original	0.6322	-	0.0345
	Debiased	0.6667		

5.4 Experimental results

Table 5 reports the accuracy, consistency, and discrimination reduction scores for the XGB, Bi-LSTM, and KNN models on both the original and debiased datasets. We note that the group fairness experiments follow their own pipeline (Fig. 1b), configured independently of the individual fairness pipeline (Fig. 1a); accuracies reported here for the same model on the original data may therefore differ slightly from those in Table 4. Accuracy reflects the proportion of correct predictions, consistency measures the stability of predictions across similar instances, and the discrimination score Δ_P captures the change in accuracy disparity between gender groups before and after debiasing.

The Bi-LSTM model exhibits the strongest overall performance. Its accuracy increases from 82.76% to 84.71% after debiasing, accompanied by an improvement in consistency from 75.67 to 76.73%. These gains suggest that removing gender-indicative cues not only preserves but enhances the model's predictive stability. The Bi-LSTM's discrimination score of $\Delta_P = 0.0195$ indicates a modest reduction in group-level disparity.

The XGB model maintains a stable accuracy of 80.46% across both datasets, with only a minor decrease in consistency (73.03% to 72.66%). Its discrimination score of $\Delta_P = 0$ suggests that the model is largely invariant to the removal of gender cues, exhibiting neither improvement nor degradation in group-level fairness.

The KNN model shows the largest relative change, with accuracy improving from 63.22 to 66.67% after debiasing. This improvement, paired with the highest discrimination score ($\Delta_P = 0.0345$), indicates substantial sensitivity to gender-related patterns in the original data. The increase in performance following debiasing suggests that the removal

of such cues enhances the reliability of similarity-based predictions.

Overall, these results demonstrate that debiasing has neutral-to-positive effects across all models, with Bi-LSTM benefiting the most in both accuracy and consistency, and KNN showing the strongest response to bias removal.

As in Sect. 4.4.4, we quantify the sampling uncertainty of these overall accuracies analytically: on the held-out test set ($n = 87$), the 95% binomial confidence intervals are 0.8046 ± 0.083 for XGB (both conditions), 0.8276 ± 0.079 (original) and 0.8471 ± 0.076 (debiased) for Bi-LSTM, and 0.6322 ± 0.101 (original) and 0.6667 ± 0.099 (debiased) for KNN. The original-versus-debiased differences are well within these intervals, so the appropriate reading of Table 5 is that debiasing does *not* harm performance, rather than that it demonstrably improves it. We also note that Table 5 reports the change in accuracy disparity (Δ_P) rather than the per-group accuracies themselves; a small Δ_P therefore indicates that debiasing left the disparity largely unchanged, not that the disparity is small in absolute terms (see Sect. 6).

5.5 Analysis and interpretations

5.5.1 RQ4: assessing gender bias in decision-making patterns

The consistently high performance of the Bi-LSTM and XGB models suggests that the underlying decision-making patterns in the dataset are systematic rather than random. The substantial gap between these models and the KNN baseline indicates that these patterns are complex and require more expressive models to capture effectively. Such complexity may stem from latent biases, such as gender-related cues, embedded within the original textual features. The discrimination scores (Δ_P) across models provide quantitative evidence of these biases: models with lower representational capacity (e.g., KNN) exhibit greater sensitivity to biased signals, whereas more capable models learn to partially compensate for them.

5.5.2 RQ5: impact of debiasing on model performance

Contrary to the common assumption that debiasing may degrade predictive performance, our findings show that performance is preserved, and in some cases enhanced, after debiasing. The Bi-LSTM model, for instance, exhibits improvements in both accuracy (82.76% $\rightarrow$ 84.71%) and consistency (75.67% $\rightarrow$ 76.73%), alongside a moderate discrimination score ($\Delta_P = 0.0195$). These improvements suggest that removing gender-indicative cues reduces noise and allows the model to generalize more effectively. The KNN model shows even more pronounced benefits: its accuracy increases (63.22% $\rightarrow$ 66.67%), and its high Δ_P value (0.0345) reflects substantial sensitivity to bias removal. Overall, these results indicate that debiasing can enhance robustness rather than hinder performance.

5.5.3 RQ6: relationship between model complexity and bias handling

The effects of debiasing differ notably across model architectures, highlighting an important relationship between model complexity and bias sensitivity. The discrimination scores reveal this relationship quantitatively: XGB ($\Delta_P = 0$), Bi-LSTM ($\Delta_P = 0.0195$), and KNN ($\Delta_P = 0.0345$). The KNN model's high sensitivity to debiasing reflects its reliance on direct feature similarity; biased signals therefore have a disproportionately large influence on its predictions. Conversely, XGB's zero discrimination score suggests that its ensemble structure inherently mitigates some forms of bias. The Bi-LSTM, despite its capacity for modeling complex dependencies, still exhibits a moderate Δ_P, indicating that even sophisticated neural models benefit from explicit bias mitigation.

5.5.4 Discussion

Overall, these findings provide valuable insights into how different model architectures interact with biased data. The discrimination scores Δ_P serve as a useful diagnostic for measuring each model's susceptibility to biased features, revealing an inverse trend between model complexity and bias sensitivity. That model performance remains stable or even improves after debiasing demonstrates that fairness-oriented interventions need not compromise predictive accuracy. Instead, debiasing can enhance model generalization by removing spurious correlations.

The differing Δ_P values across models underscore the importance of considering both data bias and model design when developing predictive systems. Practitioners should therefore carefully select model architectures and incorporate appropriate debiasing techniques to ensure both ethical integrity and strong predictive performance. These results support the integration of fairness-aware preprocessing as a standard practice in decision-making pipelines, especially in high-stakes domains such as university admissions.

6 Limitations and future works

While this study offers meaningful insights into fairness in both human and machine decision-making, several limitations present opportunities for future research. First, our assessment of individual fairness focuses primarily on biases

affecting students, without examining potential biases originating from the examiners themselves. The dataset records one institutional decision per applicant per stage, without identifiers of the individual evaluators involved, so classical inter-rater reliability across specific evaluators cannot be computed. Our consistency metric instead captures cross-applicant consistency. Relatedly, because consistency is measured through nearest neighbours in a BERT embedding space, the human–machine consistency comparison is conducted in a representation native to the models, and human consistency scores should be read as conservative estimates under that representation. Future work could analyze variability across individual officers involved in shortlisting and admission recommendations, ideally with rater identifiers and human-elicited similarity judgments, providing a more holistic understanding of human decision-making inconsistencies within the admissions workflow.

Second, our investigation of group fairness centers on gender-related cues inferred from textual features. Gender labels are inferred via a pretrained BERT-based classifier rather than self-reported, introducing potential misclassification and limiting the reliability of group fairness estimates (Chen et al. 2019; Kallus et al. 2022). This is particularly true after debiasing, when the majority of profiles are classified as Unknown. Because ground-truth gender was withheld under the privacy constraints governing the dataset, robustness checks against alternative gender-inference methods or manually verified subsets were not possible, and remain an important validation step for future studies conducted with appropriate consent. Furthermore, our group fairness evaluation focuses on accuracy parity. A more comprehensive audit would include additional metrics such as demographic parity, equal opportunity, equalized odds, FPR/FNR rate parity, and calibration, which in our setting would have compounded the noise introduced by proxy-inferred group labels. Relatedly, we report the change in accuracy disparity (Δ_P) rather than absolute per-group error profiles. A fuller error-type analysis (e.g., confusion-matrix-derived metrics per group), together with visual summaries such as accuracy-versus-fairness trade-off plots, on data with verified protected attributes would strengthen conclusions in high-stakes settings. Extending this analysis to additional demographic attributes, such as nationality, race, or religion, would enable a deeper examination of how different forms of demographic bias manifest and interact.

Third, our lexical debiasing pipeline removes explicit gender markers (pronouns, honorifics, gendered nouns) but does not address proxy variables such as writing style differences, extracurricular patterns, or latent semantics in the BERT embeddings. Recent representation-level approaches, such as fair representation learning for pre-trained language models via disentanglement (Li et al. 2025), directly target such latent signals and constitute natural, complementary baselines for future comparison. Fairness-aware model selection frameworks (Robertson et al. 2024) could likewise support navigating multiple fairness objectives. Evaluating the influence of these residual biases on downstream fairness outcomes remains an important direction for future research, particularly as language models become increasingly integral to decision pipelines (Li et al. 2023a).

Fourth, all experiments are conducted with a single fixed random seed (42) and evaluated once on a single held-out test split, without optimism-correction procedures such as nested cross-validation. While we report analytically derived 95% binomial confidence intervals for all test-set accuracies (Sects. 4.4.4 and 5), these capture only sampling variability of the test split. They do not capture run-to-run training variability, and formal paired significance tests (e.g., McNemar's test) require per-instance predictions. Reporting results over multiple random seeds with mean and standard deviation, together with paired significance testing, would strengthen the robustness of the reported findings.

Finally, this work does not leverage Large Language Models (LLMs) for tasks such as debiasing, feature extraction, or decision support. Future studies could explore the use of LLMs to detect and mitigate bias more effectively, as well as to augment or automate components of the admissions evaluation process. Such approaches may offer improved efficiency, interpretability, and fairness in high-stakes decision-making contexts.

7 Conclusion

Fairness in decision-making is essential in both human and machine-driven evaluation systems. This work investigates individual fairness through a consistency-based metric that evaluates how similarly different decision-makers treat applicants with comparable profiles. Our findings show that machine learning models, particularly XGBoost and Bi-LSTM, achieve substantially higher consistency scores than human evaluators. This indicates that ML models can provide more stable and uniform judgments, reducing the decision inconsistency and subjective variation often observed in human evaluation processes.

Regarding group fairness, we introduce a gender-debiasing pipeline that removes gender-indicative language from applicant features without compromising predictive accuracy. Experimental results demonstrate that all three models, XGBoost, Bi-LSTM, and KNN, either maintain or improve their classification performance on the debiased data. These results challenge the commonly assumed trade-off between fairness and accuracy, showing that appropriate debiasing

can support both equitable decision-making and strong predictive performance.

These findings carry important implications for the design of admissions systems. ML models can serve as reliable decision-support tools, particularly in early screening stages where consistency is critical. They can help identify potential inconsistencies in human decisions and highlight cases that warrant further review. Furthermore, incorporating consistency scores provides an additional evaluative dimension that complements traditional accuracy metrics, enabling more holistic assessments of decision quality.

Future work may expand in several directions. First, examining biases originating from examiners themselves could uncover additional sources of variability in admissions decisions. Second, investigating other demographic attributes, such as nationality, race, or religion, could broaden the scope of group fairness analysis. Third, understanding potential biases introduced by pretrained language models used for embeddings is an important step toward fairer text-based evaluation pipelines. Finally, exploring the use of Large Language Models (LLMs) for debiasing and decision support may offer enhanced efficiency and accuracy in fairness-aware admissions systems.

Acknowledgements This research is supported in part by the Ministry of Education, Singapore (MOE), under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20123-0015), and the Singapore University of Technology and Design (SUTD) under grant RS-MEFAI-00011. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of MOE or SUTD. This work is also supported in part by Forth AI for the use of computational resources, software licenses, in-kind R&D contributions and domain expertise.

Author contributions K.H.L., R.K.L., and J.H. jointly conceived and designed the study. J.H. conducted the experiments, performed the data analysis, and wrote the manuscript. K.H.L. and R.K.L. provided methodological input and critically reviewed the manuscript. All authors read and approved the final version of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Alur R, Laine L, Li DK, Raghavan M, Shah D, Shung D (2023) Auditing for human expertise. Adv Neural Inf Process Syst 36:79439–79468

Alvero A, Arthurs N, Antonio AL, Domingue BW, Gebre-Medhin B, Giebel S, Stevens ML (2020) AI and holistic review: Informing human reading in college admissions. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). pp. 200–206 https://doi.org/10.1145/3375627.3375871

Bhattacharya D, Kanaya S, Stevens M (2017) Are university admissions academically fair? Rev Econ Stat 99(3):449–464

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186. https://doi.org/10.1126/science.aal4230

Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC (2022) A clarification of the nuances in the fairness metrics landscape. Sci Rep 12(1):4209. https://doi.org/10.1038/s41598-022-07939-1

Caton S, Haas C (2024) Fairness in machine learning: a survey. ACM Comput Surv 56(7):1–38

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019) Fairness under unawareness: Assessing disparity when protected class is unobserved. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*). pp. 339–348 https://doi.org/10.1145/3287560.3287594

Dutta S, Wei D, Yueksel H, Chen PY, Liu S, Varshney KR (2020) Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org

Dwork C (2006) Differential privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming Volume Part II. pp. 1–12. ICALP'06, Springer-Verlag, Berlin, Heidelberg

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ITCS '12, Association for Computing Machinery, New York, NY, USA https://doi.org/10.1145/2090236.2090255

Froese V, Kellerhals L, Niedermeier R (2024) Modification-fair cluster editing. Soc Netw Anal Min 14(1):109. https://doi.org/10.1007/s13278-024-01259-0

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Adv Neural Inf Process Syst 29:3323–3331

Honcharov D, Soundarajan S (2026) Fair dense subgraph discovery: Algorithms and analysis. Soc Netw Anal Min 16(1):4. https://doi.org/10.1007/s13278-025-01543-7

Horesh Y, Haas N, Mishraky E, Resheff YS, Meir Lador S (2020) Paired-consistency: An example-based model-agnostic approach to fairness regularization in machine learning. Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, pp 590–604

Kallus N, Mao X, Zhou A (2022) Assessing algorithmic fairness with unobserved protected class using data combination. Manage Sci 68(3):1959–1981. https://doi.org/10.1287/mnsc.2020.3850

Lahoti P, Gummadi KP, Weikum G (2019) Operationalizing individual fairness with pairwise fair representations. Proceedings of the

VLDB Endowment 13(4):506–518. https://doi.org/10.14778/3372716.3372723

Li Y, Du M, Song R, Liu M, Wang Y (2025) BATED: Learning fair representation for pre-trained language models via biased teacher-guided disentanglement. Artif Intell 348:104401. https://doi.org/10.1016/j.artint.2025.104401

Li Y, Du M, Song R, Wang X, Wang Y (2023) A survey on fairness in large language models. arXiv preprint arXiv:2308.10149. https://doi.org/10.48550/arXiv.2308.10149

Li Y, Chen H, Xu S, Ge Y, Tan J, Liu S, Zhang Y (2023) Fairness in recommendation: Foundations, methods, and applications. ACM Transactions on Intelligent Systems and Technology 14(5):1–48

Liu J, Lee RKW, Lim KH (2025) BGM-HAN: A hierarchical attention network for accurate and fair decision assessment on semi-structured profiles. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp 44–58

Liu J, Lee RKW, Lim KH (2025) Understanding fairness-accuracy trade-offs in machine learning models: Does promoting fairness undermine performance? Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp 137–145

Liu J, Lim KH, Lee RKW (2024) Towards objective and unbiased decision assessments with llm-enhanced hierarchical attention networks. arXiv preprint arXiv:2411.08504

Mahoney T, Varshney KR, Hind M (2020) AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning. O'Reilly

Marcinkowski F, Kieslich K, Starke C, Lünich M (2020) Implications of ai (un-) fairness in higher education admissions: the effects of perceived ai (un-) fairness on exit, voice and organizational reputation. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 122–130 https://doi.org/10.1145/3351095.3372867

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607

Pitman T (2016) Understanding 'fairness' in student selection: are there differences and does it make a difference anyway? Stud High Educ 41(7):1203–1216

Puranik B, Madhow U, Pedarsani R (2022) A dynamic decision-making framework promoting long-term fairness. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp 547–556

Robertson J, Schmidt T, Hutter F, Awad N (2024) A human-in-the-loop fairness-aware model selection framework for complex fairness objective landscapes. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). vol. 7, pp. 1231–1242 https://doi.org/10.1609/aies.v7i1.31719

Sha L, Gasevic D, Chen G (2023) Lessons from debiasing data for fair and accurate predictive modeling in education. Expert Syst Appl 231:120914

Shao P, Wu L, Chen L, Zhang K, Wang M (2022) FairCF: Fairness-aware collaborative filtering. SCIENCE CHINA Inf Sci 65(12):222102. https://doi.org/10.1007/s11432-020-3404-y

Shao P, Wu L, Zhang K, Lian D, Hong R, Li Y, Wang M (2024) Average user-side counterfactual fairness for collaborative filtering. ACM Transactions on Information Systems **42**(5), 140:1–140:26 https://doi.org/10.1145/3656639

Van Busum K, Fang S (2023) Analysis of ai models for student admissions: A case study. Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. pp 17–22

Wang Y, Ma W, Zhang M, Liu Y, Ma S (2023) A survey on the fairness of recommender systems. ACM Trans Info Syst 41(3):1–43. https://doi.org/10.1145/3547333

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180. WWW '17, International World Wide Web Conferences Steering Committee https://doi.org/10.1145/3038912.3052660

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: Dasgupta S, McAllester D (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 325–333. PMLR, Atlanta, Georgia, USA 17–19 Jun

Zimdars A (2010) Fairness and undergraduate admission: a qualitative exploration of admissions choices at the university of oxford. Oxf Rev Educ 36(3):307–323

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.