Beyond Predictive Accuracy: Fairness and Bias in Predicting Test Anxiety

Anonymous Author(s)

Anonymous Institute

Abstract. Test anxiety significantly impacts students' academic performance and mental health, with complex interactions influenced by behavioral and demographic factors. This study examines the relationship between metacognitive self-regulation (MSR) behaviors and test anxiety across demographic groups, explores trade-offs between predictive accuracy and fairness in test anxiety prediction models, and investigates how intersecting demographic attributes shape biases. The findings show that specific MSR behaviors, such as classroom distraction and frequent adaptation of study methods, are strongly correlated with test anxiety, highlighting key areas for targeted interventions. Demographic disparities are evident, with females experiencing higher levels of test anxiety and White students reporting more classroom distractions. A trade-off between predictive accuracy and fairness is observed, with highly accurate models not always performing well in terms of fairness, emphasizing the need for balanced model selection. Additionally, the study challenges traditional additive assumptions about fairness, finding that the intersection of demographic attributes produces unexpected compounded effects, such as compounded advantages for Non-White Migrants and mixed outcomes for White Females. We offer insights for designing accurate and equitable predictive models for test anxiety.

Keywords: Intersectional Fairness · Alogrithmic Bias · Test Anxiety.

1 Introduction

Test anxiety is a widespread challenge for students, with well-documented effects on academic performance and mental health [60,7,58,39,47]. It disrupts cognitive functioning, reduces focus, and contributes to cycles of stress and underachievement [10,11]. Among many factors, metacognitive self-regulation (MSR)—the ability to plan, monitor, and adapt learning strategies—has been strongly linked to test anxiety [11,55,54]. However, this relationship is not straightforward. Demographic factors such as race, sex, and migration status may influence how students experience test anxiety and engage in MSR behaviors. For instance, some studies show that female students, despite employing MSR strategies more frequently, still report higher levels of test anxiety than their male counterparts [47,11]. Understanding these variations is crucial for addressing disparities and designing interventions that are equitable and effective [35].

In recent years, predictive modeling has become an invaluable tool for identifying students at risk of test anxiety. These models analyze behavioral, cognitive, and demographic data to predict outcomes and provide early intervention opportunities [16,2]. By leveraging these insights, educators and researchers aim to target support toward students most in need. However, the implementation of predictive models raises important questions about fairness. Predictive systems that perform differently across demographic groups may unintentionally exacerbate inequalities rather than reduce them [4,20,36,19]. For example, if a model underestimates the risk of test anxiety for certain groups—such as non-White students or female migrants—it could result in inadequate support for these populations [57,36]. While fairness in predictive modeling has received attention in educational research, many studies focus on individual demographic attributes, such as race or sex, without considering the compounded effects of intersecting identities [57]. For instance, being both a female and a migrant may create unique vulnerabilities that are not captured by single-attribute fairness evaluations. Furthermore, fairness is often treated as separate from model performance, leaving the trade-offs between predictive accuracy[1] and fairness largely unexplored [56,61]. This gap is particularly relevant for test anxiety prediction, where a lack of attention to fairness could undermine the utility of these systems for vulnerable groups [29].

Despite the progress made in understanding test anxiety and the use of predictive modeling in education, significant gaps remain. The relationship between MSR behaviors and test anxiety, particularly across demographic groups, is not well understood. Additionally, fairness considerations in predictive models for test anxiety require more attention, especially in balancing the trade-offs between predictive accuracy and fairness. Lastly, evaluations of fairness must move beyond isolated attributes to address how intersecting demographic factors shape outcomes. To address these challenges, this study explores the following research questions (RQs):

RQ1a: What is the relationship between MSR behaviors and test anxiety?
RQ1b: How do MSR behaviors and test anxiety differ across demographic groups such as race, sex, and migrant status?
RQ2: Are the most accurate predictive models for test anxiety also the most fair, or do trade-offs exist between predictive accuracy and fairness?.
RQ3: How does the intersection of demographic attributes compound or mitigate biases in predictive models for test anxiety?

To answer these RQs, we use Spearman's rank correlation and Mann-Whitney U tests to explore how students' demographic attributes relate to their MSR behaviors and test anxiety. We then train five predictive models and evaluate them based on their predictive accuracies using four metrics, as well as their fairness across sex, migrant status, and race by measuring differences in the same metrics. Lastly, we investigate how intersecting demographic attributes affect fairness using a new diagnostic metric we introduce.

[1] In this paper, *predictive accuracy* refers to the model's performance evaluated using one or more metrics such as precision and accuracy. In contrast, *accuracy* specifically denotes the ratio of correct predictions to the total number of predictions.

Our **contributions** are threefold: (1) we introduce the Residual Fairness Gap (RFG), a novel metric for assessing intersectional fairness in predictive models; (2) we demonstrate the trade-offs between predictive accuracy and fairness, emphasizing the importance of informed model selection; and (3) we show that the combined effects of intersecting demographic attributes are often complex, going beyond simple additive assumptions.

2 Related Works

This section briefly outlines related works that inform the research questions, without providing a comprehensive review.

2.1 MSR and Test Anxiety

The relationship between test anxiety and MSR is a topic of ongoing discussion in educational research. Test anxiety, known for disrupting cognitive processes and negatively impacting academic performance, has been studied extensively [60,10,11]. On the other hand, MSR is often associated with better academic outcomes, though its connection to test anxiety is less clear-cut [43,33]. Some research suggests that students who actively engage in MSR tend to experience lower levels of test anxiety, likely due to feeling more prepared and in control of their learning [48]. However, not all findings align with this view. In some cases, frequent use of MSR strategies has been linked to heightened stress, especially among students who are acutely aware of their academic challenges or feel external pressure to succeed [27]. Demographics further complicate this relationship [47,11,6]. Female students, for example, often report higher levels of test anxiety than males, even though they tend to use MSR strategies more effectively [28,14]. Despite these insights, the literature remains inconclusive on how generalizable these patterns are across different populations. This study aims to contribute to this ongoing debate.

2.2 Predictive Modeling of Anxiety Disorders

Machine learning has become a valuable tool for identifying individuals at risk of of various anxiety disorders, enabling early and targeted interventions [16,2,51,31,1]. For example, Almadhor et al. [1] trained several models to predict anxiety levels, finding that Random Forest achieved the highest predictive accuracy. Similarly, Priya et al. [51] also applied machine learning to anxiety prediction, demonstrating strong performance in identifying negative cases. While these studies showcase the potential of predictive modeling, they tend to prioritize accuracy over fairness. Little attention is paid to whether predictions work consistently across diverse demographic groups, highlighting the need for research that considers both predictive accuracy and fairness to ensure that these tools serve all students effectively.

2.3 Fairness of Predictive Models in Education

Fairness in predictive modeling is an important issue in education [24,53,40,36,18]. Research shows that models optimized for accuracy may often perform worse for underrepresented groups [36,4,24]. Nonetheless there is a lack of consensus on the trade-off between fairness and predictive accuracy predictive even in the general fair machine learning community [44]. Certain studies show that fairness and predictive accuracy can co-exist without a strict trade-off [24,25]. However, there are other studies that show that optimizing for fairness comes at cost to predictive accuracy [20,44,61]. These conflicting findings emphasize the need for further exploration, particularly in educational contexts where fairness is as important as predictive accuracy.

Furthermore, fairness evaluations in educational predictive models often neglect intersectionality according to Verger and colleagues [57]. To the best of our knowledge, only Gardner and colleagues [25] evaluate the fairness of predictive models in education along the intersection of multiple demographic attributes using AUC Gap. The AUC Gap performs well in highlighting intersectional subgroup disparity but it does not explicitly show whether there is a compounded (dis)advantage for a particular subgroup or not. The Residual Fairness Gap which we propose in Section 3.3, pinpoints intersectional subgroups with compounded (dis)advantages.

3 Methods

3.1 Data

We used survey data from the Motivated Strategies for Learning Questionnaire (MSLQ) [50], consisting of 81 items grouped into 15 subscales, collected over an 8-week period (April–June 2024) via Prolific with 672 consenting participants. Responses were rated on a 7-point Likert scale. Demographic data included race (54% White, 46% Non-White), migrant status, and sex (50.35% Male, 49.08% Female). For statistical power, race was categorized as White vs. Non-White, and sex analysis excluded the <1% who selected "Prefer not to say." This study focuses on the Metacognitive Self-Regulation (MSR) and Test Anxiety (TA) subscales from the MSLQ, comprising twelve and five items respectively. The survey items have been rephrased for brevity, and we will use their aliases throughout this study (e.g., *"During class time I often miss important points because I'm thinking of other things"* is rephrased as `distracted_during_class`). See the Appendix 5 for the complete list.

Latent Test Anxiety Score Derivation: To represent test anxiety, we created a composite score from the five TA items. Internal consistency was verified with Cronbach's Alpha ($\alpha = 0.80$) and McDonald's Omega ($\omega = 0.81$), indicating strong reliability [49,52,22]. Using exploratory factor analysis (EFA) [46], we generated the composite score, confirming suitability with a Kaiser-Meyer-Olkin (KMO) test score of 0.80. The factor loadings (FLs) showed varying correlations of the five TA items with test anxiety, with `fear_of_failure` (FL$= 0.81$) being the most correlated (See Table .2). The factor scores were normalized between 0 and 1 to reflect increasing test anxiety. We then analyzed the relationship between the composite score (i.e., test anxiety), MSR behaviours, and the various demographic identities.

3.2 Statistical Analysis

To investigate the relationship between test anxiety, metacognitive self-regulation (MSR), and student demographics, we conducted two statistical analyses. First, we used Spearman's rank correlation to analyze the relationship between twelve MSR behaviors and test anxiety. Despite some MSR features showing zero or near-zero correlation (e.g., `assess_concept_mastery`, $\rho = 0.00$), permutation importance (PMI) revealed that even these features could have predictive utility. Second, we used the Mann-Whitney U test to examine how race, sex, and migrant status affect test anxiety and MSR behaviors, calculating Cliff's Delta (δ) to measure effect size and direction. Following these analyses, we trained five machine learning models to predict test anxiety, evaluating their predictive accuracy and fairness.

3.3 Predicting Test Anxiety

Predictive Models: To predict test anxiety, we selected five machine learning (ML) models commonly used in classification tasks within learning analytics: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), XGBoost (XGB), and Multilayer Perceptron (MLP) [30]. These models have been effectively applied in similar contexts, such as predicting anxiety disorders and related mental health conditions [2].

Predictive Accuracy and Fairness Metrics: To evaluate the predictive accuracy of our models for identifying test anxiety, we used commonly applied metrics in learning analytics—accuracy, F1 score, area under the receiver operating characteristic curve (AUC-ROC), and area under the precision-recall curve (AUC-PR) [24,53,36]. These metrics provide a comprehensive view, addressing the nuances of predicting test anxiety, where missing true cases or over-predicting can have serious consequences [8]. To assess fairness, we compared these metrics across demographic groups following established conventions [36,53,4], examining consistency and potential biases. Significant disparities in metrics between groups would indicate unfair outcomes, as the model may work better for some groups than others. Additionally, while fairness analysis in learning analytics often focuses on single demographic attributes, the intersection of multiple attributes remains underexplored [57]. In this work, we propose a diagnostic metric for evaluating algorithmic fairness at the intersectional level.

Proposed Metric Residual Fairness Gap (RFG) RFG compares the *actual* predictive accuracy of an intersectional subgroup to the *expected* predictive accuracy, which is computed as the average marginal predictive accuracies of its constituent groups. This approach reveals whether the intersection of demographic attributes introduces compounded effects—either positive or negative—on subgroup predictive accuracy. Given multiple sensitive attributes $A, B, \ldots, K$, where a, b, etc., represent specific groups (e.g., a could represent "female" in sex and b could represent "Black" in race), RFG is defined as follows: $\text{RFG}_{a,b,\ldots,k} = \text{Metric}_{a,b,\ldots,k} - \frac{\text{Metric}_a + \text{Metric}_b + \cdots + \text{Metric}_k}{n}$. *Where*: Metric represents any predictive accuracy metric, such as precision or F1-score. $\text{Metric}_{a,b,\ldots,k}$ is the *actual* predictive accuracy of the intersectional subgroup (e.g., Black females) w.r.t the chosen metric, while $\text{Metric}_a, \text{Metric}_b, \ldots, \text{Metric}_k$ are the

Table 1: *Ground truth* distribution showing the prevalence (or base rates) of test anxiety at different thresholds ($\tau = 0.4$, $\tau = 0.5$, and $\tau = 0.6$) across the overall population and different demographic groups in the dataset.

			Migrant Status		Race		Sex	
		Overall	Migrant	Non-Migrant	White	Non-White	Male	Female
Total Sample		672	107	565	363	309	338	334
Thresholds	$\tau = 0.4$	75.3%	**80.4%**	74.3%	75.8%	74.8%	74.6%	76.0%
	$\tau = 0.5$	62.9%	**69.2%**	61.8%	65.0%	60.5%	60.9%	65.0%
	$\tau = 0.6$	47.0%	**50.5%**	46.4%	49.0%	44.7%	45.3%	48.8%

marginal predictive accuracy scores of constituent groups w.r.t the chosen metric, and n is the number of sensitive attributes. The RFG score evaluates whether the intersectional subgroup performs better ($RFG > 0$, compounded advantage), worse ($RFG < 0$, compounded disadvantage), or as expected ($RFG = 0$, no compounded effects) compared to the average of its marginal groups.

Model Training and Evaluation: We built the test anxiety prediction models using the demographic variables and the metacognitive self-regulation items as input features, with binarized test anxiety scores as the target. Three binarization thresholds—0.4, 0.5, and 0.6—were used: 0.4 assumed false negatives to be costlier, 0.5 followed standard practice in classification, and 0.6 was derived from an ad-hoc ROC analysis for optimal sensitivity-specificity balance. We found that 0.4 overestimated test anxiety prevalence, while 0.6 underestimated it. This led us to choose 0.5 as the most balanced option as per Table 1. Nonetheless, we trained and tested all models using each threshold. In this paper, we will focus on 0.5 threshold, however, results for 0.4 and 0.6 thresholds are included in the Appendix 5.

The five models were trained as follows. We split the dataset into an 80% training set and a 20% test set, using stratification to maintain the distribution of test anxiety. We determined optimal hyper-parameters through grid search with 5-fold cross-validation. After training, we evaluated the models on the test set by performing bootstrap sampling for 100 times to ensure robustness [23], ensuring that each bootstrap iteration contained the same *number* of observations as the original test set. For each iteration, we calculated the predictive accuracy and fairness metrics, then we averaged the results and calculated their standard deviations.

We assessed predictive accuracy using accuracy, F1 score, AUC-ROC, and AUC-PR. We evaluated fairness by computing differences in predictive accuracy across demographic groups and tested their significance using independent samples t-tests. Finally, to capture the compounded effect of bias across intersectional subgroups, we applied the RFG metric to bootstrap-averaged results for pairwise intersectional subgroups.

Table 2: Spearman's rank correlation (ρ) between MSR features and test anxiety (TA), with PMI (permutation importance). Asterisks denote significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ and is consistent throughout this study.

MSR	ρ	PMI
`distracted_during_class`	0.45***	0.78
`formulate_guiding_questions`	0.03	0.16
`clarify_confusing_content`	0.00	0.08
`adjust_reading_strategy`	-0.03	0.09
`preview_course_material`	-0.00	0.11
`self_check_understanding`	-0.01	0.12
`adapt_study_methods`	0.34***	0.35
`mindless_class_reading`	0.04	0.10
`identify_learning_objectives`	0.07	0.11
`assess_concept_mastery`	-0.00	0.11
`set_study_goals`	-0.05	0.11
`review_unclear_notes`	-0.01	0.07

4 Results Discussion

4.1 RQ1: Dynamics of MSR, Test Anxiety, and Demographics

Recall that RQ1 explores the link between MSR, test anxiety, and their variation by demographics. As shown in Table 2, interestingly, we observed that out of the twelve MSR behaviors, only two—`distracted_during_class` ($\rho = 0.45$) and `adapt_study_methods` ($\rho = 0.34$)—are significantly ($p < 0.001$) correlated with test anxiety. However, it is not surprising to find that distraction during class was the most correlated MSR behaviour with test anxiety. We speculate that this could be due to the fact that students who are distracted during class may have missed important points that are crucial to understanding the study material, for example, due to mind wandering [21], thus feeling unprepared and anxious before tests [7]. Another interesting finding was that adapting study methods to fit course requirements or an instructor's teaching style may not always reduce test anxiety, as previously found in studies such as [58]. Rather, we found that students who adapt their study methods may sometimes have high test anxiety. A probable reason for this could be that the increased pressure to adapt study methods or the frequency at which students keep changing study methods could destabilize their study routines [45,39]. This aligns with other research suggesting that frequent cognitive adjustments, especially when tied to metacognitive strategies, can increase test anxiety by adding to the mental load and making it harder for students to regulate their emotions effectively [26,27].

Across all demographic groups that we considered, i.e., race, sex, and migrant status, it was only in terms of sex that we observed significant difference in test anxiety. For instance, as shown in Table 3, females were found to be 10% more likely than males to experience test anxiety. There are several existing studies that corroborates to this finding [47,39]. Focusing on the MSR behaviours that we found to be significantly correlated with test anxiety, i.e.,

Table 3: Mann-Whitney U test results for TA and MSR behaviours by Race (W: White, NW: Non-White), Sex (M: Male, F: Female), and Migrant Status (M: Migrant, NM: Non-Migrant). δ represents Cliff's delta effect sizes. A negative δ signifies higher value for Females, Non-White, or Migrants and vice versa.

Variable	**Race**		**Sex**		**Migrant Status**	
	δ	Higher	δ	Higher	δ	Higher
Test Anxiety (TA)	0.03	W	-0.1*	F	-0.09	M
`adapt_study_methods`	0.04	W	-0.07	F	0.01	NM
`adjust_reading_strategy`	-0.21***	NW	-0.05	F	-0.02	M
`assess_concept_mastery`	-0.17***	NW	-0.11*	F	-0.07	M
`clarify_confusing_content`	-0.21***	NW	-0.13**	F	-0.03	M
`distracted_during_class`	0.15***	W	0.04	M	-0.07	M
`formulate_guiding_questions`	-0.29***	NW	-0.06	F	0.02	NM
`identify_learning_objectives`	-0.21***	NW	-0.11**	F	0.01	NM
`mindless_class_reading`	-0.19***	NW	0.03	M	-0.04	M
`preview_course_material`	-0.2***	NW	-0.1*	F	0.01	NM
`review_unclear_notes`	-0.18***	NW	-0.04	F	0.11	NM
`self_check_understanding`	-0.19***	NW	-0.08	F	0.05	NM
`set_study_goals`	-0.14**	NW	-0.16***	F	0.01	NM

Table 4: Average Predictive Accuracy of all models. The values are the the average ± standard deviation. Boldened and red scores are the highest and least overall averages respectively

Model	AUC-ROC	AUC-PR	Accuracy	F1 Score	Overall Average
LR	0.66 ± 0.05	0.76 ± 0.05	0.67 ± 0.04	0.75 ± 0.03	0.71 ± 0.04
MLP	0.54 ± 0.05	0.67 ± 0.05	0.61 ± 0.04	0.71 ± 0.03	0.63 ± 0.04
RF	0.66 ± 0.05	0.75 ± 0.05	0.69 ± 0.04	0.79 ± 0.03	$\mathbf{0.72 \pm 0.04}$
SVM	0.66 ± 0.05	0.75 ± 0.05	0.66 ± 0.04	0.75 ± 0.03	0.7 ± 0.05
XGB	0.60 ± 0.05	0.72 ± 0.05	0.64 ± 0.04	0.76 ± 0.03	0.68 ± 0.04

`distracted_during_class` and `adapt_study_methods`, we did not find any significant difference across the various demographic groups except *race*. Specifically, we found that the White students are 15% more likely to be distracted during class compared to their Non-White counterparts. We conjecture that cultural differences can influence how students report being distracted. For example, White students might be more open about mentioning (even) minor distractions, while Non-White students, who are aware of biases and stereotypes [17,38], might downplay their distractions [59,32]. This is in line with studies showing that sociocultural norms affect emotional self-awareness and self-assessment [32]. Further future studies are needed to investigate this finding in detail.

4.2 RQ2: Trade-offs Between Predictive Accuracy and Fairness

RQ2 examines whether the most accurate models for predicting test anxiety are also the fairest or involve trade-offs. Firstly, in terms of predictive accuracy, no model consistently outperformed others across all metrics as shown in Table 4. However, averaging across metrics, the RF model performed best, while the deep learning model (i.e., MLP) performed worst—contrary to prior studies where deep learning models excel [53]. This may be due to our dataset size, as neural networks often underperform on smaller datasets [9]. Nonetheless, similar to our results, several related studies have often found RF to outperform other models whenever such comparative analysis are done [34,42,12]. ***Ancillary Note**: It is worth mentioning that the RF model was not the best-performing model on thresholds $\tau = 0.4$ and $\tau = 0.6$ although the MLP remained the worst performing model (see Appendix 5)*

Table 5: Predictive accuracy disparities favor *historically* advantaged groups (Whites, Males, Non-Migrants) with negative values, and disadvantaged groups (Non-Whites, Females, Migrants) with positive values [4,36].

	Model	AUC-PR	AUC-ROC	Accuracy	F1 Score
Race	LR	0.02 ± 0.07 *	0.05 ± 0.08 ***	-0.07 ± 0.06 ***	-0.08 ± 0.06 ***
	MLP	0.06 ± 0.08 ***	0.13 ± 0.08 ***	0.02 ± 0.06 **	-0.01 ± 0.06
	RF	0.12 ± 0.06 ***	0.17 ± 0.07 ***	-0.02 ± 0.06	-0.02 ± 0.04 **
	SVM	0.09 ± 0.07 ***	0.11 ± 0.08 ***	-0.06 ± 0.06 ***	-0.07 ± 0.06 ***
	XGB	0.02 ± 0.08	0.08 ± 0.08 ***	0.03 ± 0.06 **	0.0 ± 0.06
Sex	LR	-0.02 ± 0.07	-0.12 ± 0.07 ***	-0.17 ± 0.06 ***	-0.13 ± 0.05 ***
	MLP	-0.09 ± 0.07 ***	-0.16 ± 0.08 ***	-0.11 ± 0.06 ***	-0.08 ± 0.06 ***
	RF	-0.03 ± 0.07 **	-0.1 ± 0.08 ***	0.0 ± 0.06	0.01 ± 0.04 *
	SVM	-0.06 ± 0.08 ***	-0.17 ± 0.08 ***	-0.14 ± 0.06 ***	-0.1 ± 0.06 ***
	XGB	-0.05 ± 0.06 ***	-0.15 ± 0.07 ***	0.03 ± 0.06 ***	0.04 ± 0.05 ***
MS	LR	0.05 ± 0.08 ***	-0.05 ± 0.1 **	0.02 ± 0.08	0.03 ± 0.07 ***
	MLP	0.04 ± 0.1 **	-0.11 ± 0.11 ***	-0.03 ± 0.08 *	0.0 ± 0.08
	RF	0.11 ± 0.08 ***	0.01 ± 0.1	0.0 ± 0.08	0.01 ± 0.06
	SVM	0.1 ± 0.08 ***	0.02 ± 0.1	0.04 ± 0.08 **	0.04 ± 0.07 ***
	XGB	0.06 ± 0.09 ***	-0.07 ± 0.1 ***	0.05 ± 0.08 ***	0.04 ± 0.06 ***

In terms of fairness, no model was consistently fair across all metrics and demographics as shown in Table 5. For instance, in terms of race, the LR model was *less* favorable to Whites by 2% (AUC-PR) and 5% (AUC-ROC) but *more* favorable to the same Whites by 7% (accuracy) and 8% (F1-score), supporting the fairness *"impossibility theorem"* which posits the mutual exclusivity of certain metrics [5,13,37].

We also observed a phenomenon that we call "phantom unfairness". Specifically, we found that some disparities which we would otherwise call unfairness quickly disappear when subjected to statistical test. For example, consider the fairness of RF in terms of race in Table 5. We observed that disparities in AUC-PR (0.12±0.06***), AUC-ROC(0.17 ± 0.07***) and F1-score (-0.02 ± 0.04 **)

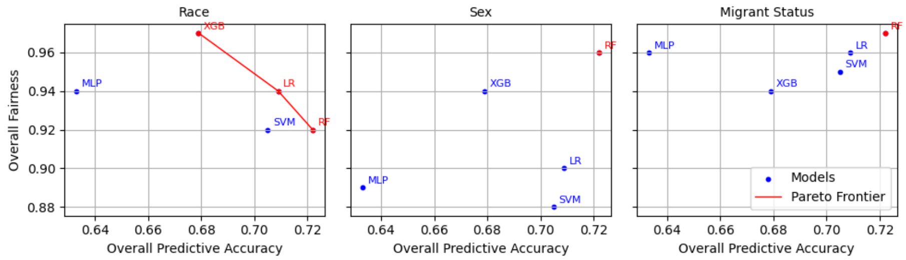

Fig. 1: Pareto frontier illustrating the trade-off between fairness and predictive accuracy. Red and blue points mark optimal and suboptimal models respectively.

are significant. However, the disparity in accuracy (-0.02 $\pm$ 0.06) is not significant statistically. Most fairness analysis in the educational domain and the general fairness community do not subject measured group disparities to statistical rigor. A noteworthy exception, however, is the study by Gardner and colleagues [24].

Moving on to the crux of RQ2, we look at the trade-off between **overall fairness** and **overall predictive accuracy** across the three sensitive attributes: race, sex, and migrant status. Overall predictive accuracy of each model is operationalized as the mean predictive accuracy of that model across all evaluation metrics. Overall fairness of each model is operationalized using the absolute mean disparity in predictive accuracy between groups, normalized as $1 - |mean\ disparity|$. Using absolute values ensured that positive and negative differences did not cancel each other out, allowing us to capture the extent of unfairness regardless of its direction.

The results, as shown in Figure 1, highlight a clear trade-off between fairness and predictive accuracy. RF consistently achieved the highest predictive accuracy across all three attributes but did not always perform best in terms of fairness. For example, in the race analysis, XGB achieved the highest fairness but at the cost of slightly lower predictive accuracy. For sex and migrant status, RF was the only model on the Pareto frontier, meaning it offered the best balance between fairness and predictive accuracy, while other models like MLP and SVM were suboptimal, underperforming in both desiderata. Models below the frontier are less effective, as better-performing alternatives exist. Overall, the findings indicate that the most accurate models are not necessarily the most fair. Hence, improving fairness may often come at the expense of predictive accuracy. This finding partially contradicts studies [24,53] that reported no strict trade-off between predictive accuracy and fairness. However, numerous other studies, including ours, demonstrate that such a trade-off does exist [3,41,44].

4.3 RQ3: Intersectional Bias in Test Anxiety Models

This RQ aims to explore the compounded effect of the intersection demographic attributes. From the results in Table 6, we found that the intersection of demographic attributes can result in (1) compounded disadvantage, (2) compounded advantage, or (3) indifference, albeit, mostly in unexpected ways. For instance, *let us* ***focus*** *on the RF model* which we found to be the pareto optimal as per

Table 6: **RFG** for metrics: PR (AUC-PR), ROC (AUC-ROC), Acc (Accuracy), and F1 (F1 Score). Column initials indicate demographic intersections: NWM (Non-White Migrant), WM (White Migrant), NWNM (Non-White Non-Migrant), WNM (White Non-Migrant), NWF (Non-White Female), NWM (Non-White Male), WF (White Female), WM (White Male), FM (Female Migrant), MM (Male Migrant), FNM (Female Non-Migrant), MNM (Male Non-Migrant).

		Race-Migrant Status				Race-Sex				Sex-Migrant Status			
		NWM	WM	NWNM	WNM	NWF	NWM	WF	WM	FM	MM	FNM	MNM
LR	PR	0.15	-0.05	-0.01	0.02	-0.01	0.04	-0.01	-0.00	-0.04	0.13	0.02	-0.00
	ROC	0.18	-0.24	0.00	-0.01	-0.03	0.04	-0.12	0.00	-0.11	0.08	-0.01	0.02
	Acc	0.07	-0.09	-0.02	0.01	-0.11	0.07	0.04	-0.01	0.04	0.03	-0.03	0.04
	F1	0.06	-0.05	-0.03	0.02	-0.12	0.07	0.06	-0.02	0.02	0.04	-0.02	0.03
MLP	PR	0.13	-0.07	-0.00	0.03	-0.01	0.07	-0.02	-0.01	-0.01	0.13	0.01	0.02
	ROC	0.07	-0.27	0.02	-0.02	-0.01	0.07	-0.16	-0.01	-0.05	0.04	-0.04	0.04
	Acc	0.06	-0.17	-0.00	0.01	-0.06	0.07	-0.01	-0.02	-0.01	0.01	-0.03	0.03
	F1	0.06	-0.15	-0.02	0.02	-0.08	0.06	0.02	-0.03	-0.04	0.04	-0.02	0.01
RF	PR	0.11	-0.03	0.01	-0.00	0.00	0.07	-0.01	-0.04	0.02	0.07	0.00	0.00
	ROC	0.13	-0.23	0.02	-0.04	-0.04	0.09	-0.12	-0.06	-0.02	-0.03	-0.03	0.02
	Acc	0.07	-0.08	-0.01	0.01	-0.07	0.05	0.09	-0.07	-0.03	0.06	0.01	-0.01
	F1	0.05	-0.06	-0.02	0.01	-0.05	0.04	0.07	-0.05	-0.01	0.04	0.01	-0.01
SVM	PR	0.14	0.02	0.01	-0.02	-0.02	0.08	-0.04	-0.02	0.08	0.04	-0.02	0.02
	ROC	0.15	-0.11	0.01	-0.04	-0.05	0.09	-0.15	-0.03	0.06	-0.09	-0.05	0.05
	Acc	0.08	-0.08	-0.03	0.02	-0.07	0.04	0.02	0.01	0.04	0.05	-0.03	0.03
	F1	0.07	-0.04	-0.03	0.03	-0.09	0.04	0.04	-0.00	0.02	0.05	-0.02	0.02
XGB	PR	0.16	-0.13	-0.02	0.04	-0.02	0.06	0.03	-0.02	0.01	0.13	0.03	-0.00
	ROC	0.15	-0.36	-0.00	-0.01	-0.02	0.05	-0.10	0.01	-0.04	0.07	-0.02	0.03
	Acc	0.06	-0.05	-0.01	-0.00	-0.00	0.02	0.02	-0.03	0.10	-0.01	0.00	-0.00
	F1	0.05	-0.04	-0.02	0.01	-0.01	0.02	0.02	-0.03	0.06	0.01	0.01	-0.01

RQ2. From Table 5, for each demographic attribute in isolation, we found that the fairness of the RF model in terms of AUC-PR is: 0.12±0.06*** for race (advantage Non-Whites), -0.03±0.07** for sex (advantage Males), and 0.11±0.08*** for migrant status (advantage migrants). With this in mind, one might expect the RF to have compounded advantage for Non-White Males, Non-White Migrants, and Male Migrants, for example. Similarly, one might expect compounded disadvantage for White Females, White Non-Migrants, and Female Non-Migrants. Our results in Table 6 sometimes agree with this hypothesis and at other times, disagree. For example, we observed that being Non-White Migrant resulted in a 11% improvement in predictive accuracy in terms of AUC-PR as compared to the average of the marginal AUC-PRs for Non-Whites and Migrants in isolation. In fact, across all models and all metrics, we observed a compounded advantage for Non-White Migrants. On the reverse, we expected compounded disadvantage for White Females. Truly, we observed a decline in predictive accuracy for White Females in terms of AUC-PR although it was a mealsy 1% decline. Yet, in terms of some other metrics such AUC-ROC for the same RF model, we observed as high as a 12% decline in AUC-ROC for White Females relative the average

of the marginal AUC-ROCs for Whites and Females in isolation. Nonetheless, we observed instances where there was actually an “unexpected” improvement in predictive accuracy for White Females for the same RF model in terms of accuracy (9%) and F1-Score (7%). Overall, our results suggest that the combination of two supposed disadvantages as it were, may not necessarily result in compounded disadvantage as prior research [15] suggests and vice versa.

5 Concluding Discussions and Implications

Demographic Inequities in Test Anxiety and the Need for Targeted Interventions: Classroom distractions and inconsistent study methods have a significant negative impact on test anxiety. This is particularly concerning for White students, who are more prone to distractions, and for females, who experience higher anxiety levels compared to males. These patterns suggest potential inequities in how test anxiety manifests across demographics, emphasizing the need for targeted strategies to create focused learning environments and stabilize study routines. Moreover, the disproportionate effects of distractions and anxiety among specific groups imply that interventions should be tailored to address these demographic differences, such as designing inclusive classroom practices and offering gender-sensitive support programs.

Balancing Accuracy and Fairness in Predictive Models Through Thoughtful and Rigorous Evaluation: The trade-off between accuracy and fairness means that practitioners have to think carefully about what matters most for their specific goals. Is predictive accuracy the priority? Is fairness more important? Or is there a need to strike a balance between the two? For example, RF models were the most accurate overall, but they were not always the fairest—especially when looking at race, where XGB performed better in fairness but at a slight cost to accuracy. Additionally, the idea of “phantom unfairness”, where disparities disappear after proper statistical testing, reminds us how critical it is to validate fairness metrics rigorously. Without this step, we risk overreacting to perceived unfairness and making decisions that do not actually solve the problem. These insights make it clear that balancing accuracy and fairness requires thoughtful decisions backed by robust and transparent methods.

Rethinking Intersectional Bias: Beyond Additive Assumptions in Predictive Fairness: Our findings indicate that the intersection of demographic attributes does not always lead to predictable compounded effects, challenging the assumption that combining disadvantages consistently exacerbates bias. This highlights the importance of moving beyond simple additive assumptions about fairness and adopting a more nuanced approach to understanding how intersectional attributes interact in predictive models. The variation in compounded effects further emphasizes the need for context-specific fairness evaluations, as the impact of intersectionality often depends on the metric or model, making tailored interventions essential to effectively addressing biases.

Limitation: We used binary demographic groupings due to the small dataset to ensure statistical power, but this limits the detection of nuanced group differences. Future work with larger samples will enable finer categorization.

References

1. Almadhor, A., Abbas, S., Sampedro, G.A., Alsubai, S., Ojo, S., Al Hejaili, A., Strazovska, L.: Multi-class adaptive active learning for predicting student anxiety. IEEE Access (2024)
2. Arif, M., Basri, A., Melibari, G., Sindi, T., Alghamdi, N., Altalhi, N., Arif, M.: Classification of anxiety disorders using machine learning methods: a literature review. Insights Biomed Res **4**(1), 95–110 (2020)
3. Badar, M., Sikdar, S., Nejdl, W., Fisichella, M.: Fairtrade: Achieving pareto-optimal trade-offs between balanced accuracy and fairness in federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10962–10970 (2024)
4. Baker, R.S., Hawn, A.: Algorithmic bias in education. International Journal of Artificial Intelligence in Education pp. 1–41 (2021)
5. Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., Stoyanovich, J.: The possibility of fairness: Revisiting the impossibility theorem in practice. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 400–422 (2023)
6. Bidjerano, T.: Gender differences in self-regulated learning. Online Submission (2005)
7. Blankstein, K.R., Toner, B.B., Flett, G.L.: Test anxiety and the contents of consciousness: Thought-listing and endorsement measures. Journal of Research in Personality **23**(3), 269–286 (1989)
8. Bradford, A., Meyer, A.N., Khan, S., Giardina, T.D., Singh, H.: Diagnostic error in mental health: a review. BMJ Quality & Safety (2024)
9. Brigato, L., Iocchi, L.: A close look at deep learning with small data. In: 2020 25th international conference on pattern recognition (ICPR). pp. 2490–2497. IEEE (2021)
10. Cassady, J.C., Johnson, R.E.: Cognitive test anxiety and academic performance. Contemporary educational psychology **27**(2), 270–295 (2002)
11. Chapell, M.S., Blanding, Z.B., Silverstein, M.E., Takahashi, M., Newman, B., Gubi, A., McCann, N.: Test anxiety and academic performance in undergraduate and graduate students. Journal of educational Psychology **97**(2), 268 (2005)
12. Chen, J., Zhou, X., Yao, J., Tang, S.K.: Evaluation of student performance based on learning behavior with random forest model. In: 2024 13th International Conference on Educational and Information Technology (ICEIT). pp. 266–272. IEEE (2024)
13. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
14. Cleary, T.J., Velardi, B., Schnaidman, B.: Effects of the self-regulation empowerment program (srep) on middle school students' strategic skills, self-efficacy, and mathematics achievement. Journal of school psychology **64**, 28–42 (2017)
15. Crenshaw, K.: Women of color at the center: Selections from the third national conference on women of color and the law: Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review **43**(6), 1241–1279 (1991)
16. Daza, A., Saboya, N., Necochea-Chamorro, J.I., Ramos, K.Z., Valencia, Y.d.R.V.: Systematic review of machine learning techniques to predict anxiety and stress in college students. Informatics in medicine unlocked p. 101391 (2023)
17. Deckman, S.L.: Managing race and race-ing management: Teachers' stories of race and classroom conflict. Teachers College Record **119**(11), 1–40 (2017)

18. Deho, O.B., Joksimovic, S., Li, J., Zhan, C., Liu, J., Liu, L.: Should learning analytics models include sensitive attributes? explaining the why. IEEE Transactions on Learning Technologies (2022)
19. Deho, O.B., Joksimovic, S., Liu, L., Li, J., Zhan, C., Liu, J.: Assessing the fairness of course success prediction models in the face of (un)equal demographic group distribution. In: Proceedings of the Tenth ACM Conference on Learning @ Scale. p. 48–58. L@S '23, Association for Computing Machinery, New York, NY, USA (2023)
20. Deho, O.B., Zhan, C., Li, J., Liu, J., Liu, L., Duy Le, T.: How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? British Journal of Educational Technology (2022)
21. Desideri, L., Ottaviani, C., Cecchetto, C., Bonifacci, P.: Mind wandering, together with test anxiety and self-efficacy, predicts student's academic self-concept but not reading comprehension skills. British Journal of Educational Psychology **89**(2), 307–323 (2019)
22. Edwards, A.A., Joyner, K.J., Schatschneider, C.: A simulation study on the performance of different reliability estimation methods. Educational and Psychological Measurement **81**(6), 1089–1117 (2021)
23. Efron, B., Rogosa, D., Tibshirani, R.: Resampling methods of estimation. In: Smelser, N.J., Baltes, P.B. (eds.) International Encyclopedia of the Social & Behavioral Sciences, pp. 13216–13220. Elsevier, New York, NY (2004)
24. Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. p. 225–234. LAK19, Association for Computing Machinery, New York, NY, USA (2019)
25. Gardner, J., Yu, R., Nguyen, Q., Brooks, C., Kizilcec, R.: Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1664–1684 (2023)
26. Garnefski, N., Kraaij, V., Spinhoven, P.: Negative life events, cognitive emotion regulation and emotional problems. Personality and Individual differences **30**(8), 1311–1327 (2001)
27. Ghribnavaz, S., Nouri, R., Moghadasin, M.: Relationship between metacognition believes and exam anxiety: Mediating role of cognitive emotion regulation. Journal of Cognitive Psychology **5**(4), 1–10 (2018)
28. Gustems-Carnicer, J., Calderón, C., Calderón-Garrido, D.: Stress, coping strategies and academic achievement in teacher education students. European Journal of Teacher Education **42**(3), 375–390 (2019)
29. Häuselmann, A., Custers, B.: Substantive fairness in the gdpr: Fairness elements for article 5.1 a gdpr. Computer Law & Security Review **52**, 105942 (2024)
30. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N.: Predicting academic performance: a systematic literature review. In: Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education. pp. 175–199 (2018)
31. Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K., Hilbert, K.: Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. Digital Health **7**, 20552076211060659 (2021)
32. Huggins, C.F., Williams, J.H., Sato, W.: Cross-cultural differences in self-reported and behavioural emotional self-awareness between japan and the uk. BMC Research Notes **16**(1), 380 (2023)

33. Huntley, C.D., Young, B., Tudur Smith, C., Fisher, P.L.: Metacognitive beliefs predict test anxiety and examination performance. In: Frontiers in Education. vol. 8, p. 1051304. Frontiers Media SA (2023)
34. Hutt, S., Gardener, M., Kamentz, D., Duckworth, A.L., D'Mello, S.K.: Prospectively predicting 4-year college graduation from student applications. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge. pp. 280–289 (2018)
35. Hyseni Duraku, Z., Hoxha, L.: Self-esteem, study skills, self-concept, social support, psychological distress, and coping mechanism effects on test anxiety and academic performance. Health psychology open **5**(2), 2055102918799963 (2018)
36. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: The Ethics of Artificial Intelligence in Education, pp. 174–202. Routledge (2022)
37. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
38. Kunesh, C.E., Noltemeyer, A.: Understanding disciplinary disproportionality: Stereotypes shape pre-service teachers' beliefs about black boys' behavior. Urban Education **54**(4), 471–498 (2019)
39. Lawrence, A.A.: Relationship between study habits and test anxiety of higher secondary students. Online Submission **3**(6), 1–9 (2014)
40. Lele, S., Gasevic, D., Chen, G.: Lessons from debiasing data for fair and accurate predictive modeling in education (2022)
41. Li, J., Li, G.: The triangular trade-off between robustness, accuracy and fairness in deep neural networks: A survey. ACM Computing Surveys (2024). https://doi.org/10.1145/3645088
42. Lingjun, H., Levine, R.A., Fan, J., Beemer, J., Stronach, J.: Random forest as a predictive analytics alternative to regression in institutional research. Practical Assessment, Research, and Evaluation **23**(1), 1 (2019)
43. Mega, C., Ronconi, L., De Beni, R.: What makes a good student? how emotions, self-regulated learning, and motivation contribute to academic achievement. Journal of educational psychology **106**(1), 121 (2014)
44. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
45. Nicolas, T.M., Arambulo, R.: Test anxiety, readiness, and intervention strategies for enhancing board exam performance among psychology students. The Quest: Journal of Multidisciplinary Research and Development **2**(3) (2023)
46. Norris, M., Lecavalier, L.: Evaluating the use of exploratory factor analysis in developmental disability psychological research. Journal of autism and developmental disorders **40**, 8–20 (2010)
47. Núñez-Peña, M.I., Suárez-Pellicioni, M., Bono, R.: Gender differences in test anxiety and their impact on higher education students' academic achievement. Procedia-Social and Behavioral Sciences **228**, 154–160 (2016)
48. Onwunyili, F.C., Onwunyili, M.C.: Effect of self-regulated learning on test anxiety: Academic achievement and metacognition among secondary school students in anambra state. South Eastern Journal of Research and Sustainable Development (SEJRSD) **3**(2), 90–104 (2020)
49. Orçan, F.: Comparison of cronbach's alpha and mcdonald's omega for ordinal data: Are they different? International Journal of Assessment Tools in Education **10**(4), 709–722 (2023)

50. Pintrich, P.: A manual for the use of the motivated strategies for learning questionnaire (mslq). National Center for Research to Improve Postsecondary Teaching and Learning (1991)
51. Priya, A., Garg, S., Tigga, N.P.: Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science **167**, 1258–1267 (2020)
52. Santos, J.R.A.: Cronbach's alpha: A tool for assessing the reliability of scales. The journal of Extension **37**(2), 15 (1999)
53. Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V.M., Gasevic, D., Chen, G.: Assessing algorithmic fairness in automatic classifiers of educational forum posts. In: International Conference on Artificial Intelligence in Education. pp. 381–394. Springer (2021)
54. Silaj, K.M., Schwartz, S.T., Siegel, A.L.M., Castel, A.: Test anxiety and metacognitive performance in the classroom. Educational Psychology Review **33**, 1809 – 1834 (2021). https://doi.org/10.1007/s10648-021-09598-6
55. Spada, M., Nikcevic, A., Moneta, G., Ireson, J.: Metacognition as a mediator of the effect of test anxiety on a surface approach to studying. Educational Psychology **26**, 615 – 624 (2006). https://doi.org/10.1080/01443410500390673
56. Valdivia, A., Sánchez-Monedero, J., Casillas, J.: How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. International Journal of Intelligent Systems **36**(4), 1619–1643 (2021)
57. Verger, M., Fan, C., Lallé, S., Bouchet, F., Luengo, V.: A comprehensive study on evaluating and mitigating algorithmic unfairness with the madd metric. Journal of Educational Data Mining **16**(1), 365–409 (2024)
58. Wittmaier, B.C.: Test anxiety and study habits. The Journal of Educational Research **65**(8), 352–354 (1972)
59. Xie, J.L., Roy, J.P., Chen, Z.: Cultural and individual differences in self-rating behavior: an extension and refinement of the cultural relativity hypothesis. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior **27**(3), 341–364 (2006)
60. Zeidner, M.: Test anxiety in educational contexts: Concepts, findings, and future directions. In: Emotion in education, pp. 165–184. Elsevier (2007)
61. Zhao, H., Gordon, G.J.: Inherent tradeoffs in learning fair representations. Journal of Machine Learning Research **23**(57), 1–26 (2022)

APPENDIX A

Table .1: Survey Questions and their aliases. *Hereinafter, we will refer to specific survey items by their aliases.MSR items (first twelve rows) are highlighted in green and test anxiety items (last five rows) are highlighted in gray

Survey Item	Alias*
1: During class time I often miss important points because I'm thinking of other things.	`distracted_during_class`
2: When reading for this course, I make up questions to help focus my reading.	`formulate_guiding_questions`
3: When I become confused about something I'm reading for this class, I go back and try to figure it out.	`clarify_confusing_content`
4: If course readings are difficult to understand, I change the way I read the material.	`adjust_reading_strategy`
5: Before I study new course material thoroughly, I often skim it to see how it is organized.	`preview_course_material`
6: I ask myself questions to make sure I understand the material I have been studying in this class.	`self_check_understanding`
7: I try to change the way I study in order to fit the course requirements and the instructor's teaching style.	`adapt_study_methods`
8: I often find that I have been reading for this class but don't know what it was all about.	`mindless_class_reading`
9: I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over.	`identify_learning_objectives`
10: When studying for this course I try to determine which concepts I don't understand well.	`assess_concept_mastery`
11: When I study for this class, I set goals for myself in order to direct my activities in each study period.	`set_study_goals`
12: If I get confused taking notes in class, I make sure I sort it out afterwards.	`review_unclear_notes`
1: When I take a test, I think about how poorly I am doing compared with other students.	`comparison_with_others`
2: When I take a test, I think about items on other parts of the test I can't answer.	`difficult_questions_fixation`
3: When I take tests, I think of the consequences of failing.	`fear_of_failure`
4: I have an uneasy, upset feeling when I take an exam.	`exam-induced_uneasiness`
5: I feel my heart beating fast when I take an exam.	`exam-induced_heart_racing`

Table .2: Factor Loadings showing the correlation (coefficient) of each of the five survey items to the test anxiety. CWO: `comparison_with_others`, DQF: `difficult_questions_fixation`, FOF: `fear_of_failure`, EIU: `exam-induced_uneasiness`, EHR: `exam-induced_heart_racing`.

CWO	DQF	FOF	EIU	EHR
0.59	0.54	**0.81**	0.72	0.70

Table .3: Average Predictive Accuracy of all models for threshold $\tau = 0.4$. The values are the average $\pm$ standard deviation. Boldened and red scores are the highest and least overall averages, respectively.

Model	AUC-ROC	AUC-PR	Accuracy	F1 Score	Overall Average
LR	0.70 ± 0.06	0.84 ± 0.04	0.79 ± 0.03	0.87 ± 0.02	**0.80 ± 0.04**
MLP	0.61 ± 0.06	0.82 ± 0.05	0.67 ± 0.04	0.78 ± 0.03	0.72 ± 0.04
RF	0.66 ± 0.05	0.83 ± 0.04	0.78 ± 0.04	0.87 ± 0.03	0.79 ± 0.04
SVM	0.71 ± 0.06	0.85 ± 0.04	0.75 ± 0.04	0.86 ± 0.02	0.79 ± 0.04
XGB	0.61 ± 0.07	0.80 ± 0.05	0.75 ± 0.04	0.84 ± 0.03	0.75 ± 0.05

Table .4: Threshold $\tau = 0.4$ fairness results. Predictive accuracies across Race, Sex, and Migrant Status (MS) groups, highlighting mean disparities, standard deviations, and statistical significance.

	Model	AUC-PR	AUC-ROC	Accuracy	F1 Score
Race	LR	-0.19 ± 0.06 ***	-0.28 ± 0.08 ***	-0.12 ± 0.06 ***	-0.07 ± 0.04 ***
	MLP	-0.11 ± 0.06 ***	-0.10 ± 0.08 ***	0.00 ± 0.06	0.00 ± 0.04
	RF	-0.06 ± 0.06 ***	-0.13 ± 0.09 ***	-0.01 ± 0.06	0.00 ± 0.04
	SVM	0.00 ± 0.06	-0.04 ± 0.09 **	-0.02 ± 0.06 **	-0.01 ± 0.04 **
	XGB	-0.09 ± 0.06 ***	-0.11 ± 0.09 ***	-0.04 ± 0.06 ***	-0.02 ± 0.04 ***
Sex	LR	0.07 ± 0.06 ***	0.10 ± 0.09 ***	0.04 ± 0.05 ***	0.04 ± 0.03 ***
	MLP	0.05 ± 0.06 ***	0.10 ± 0.08 ***	0.14 ± 0.06 ***	0.10 ± 0.05 ***
	RF	-0.01 ± 0.06	0.03 ± 0.08 **	0.03 ± 0.05 ***	0.02 ± 0.03 ***
	SVM	-0.03 ± 0.06 ***	-0.04 ± 0.08 **	-0.01 ± 0.05	0.00 ± 0.03
	XGB	0.04 ± 0.06 ***	0.05 ± 0.08 ***	0.01 ± 0.05	0.02 ± 0.04 **
MS	LR	0.14 ± 0.04 ***	0.22 ± 0.08 ***	0.07 ± 0.06 ***	0.05 ± 0.04 ***
	MLP	0.15 ± 0.05 ***	0.25 ± 0.08 ***	0.08 ± 0.06 ***	0.05 ± 0.05 ***
	RF	0.15 ± 0.04 ***	0.26 ± 0.08 ***	0.08 ± 0.06 ***	0.05 ± 0.04 ***
	SVM	0.12 ± 0.04 ***	0.21 ± 0.08 ***	0.01 ± 0.06	0.00 ± 0.04
	XGB	0.10 ± 0.06 ***	0.15 ± 0.12 ***	0.14 ± 0.06 ***	0.09 ± 0.04 ***

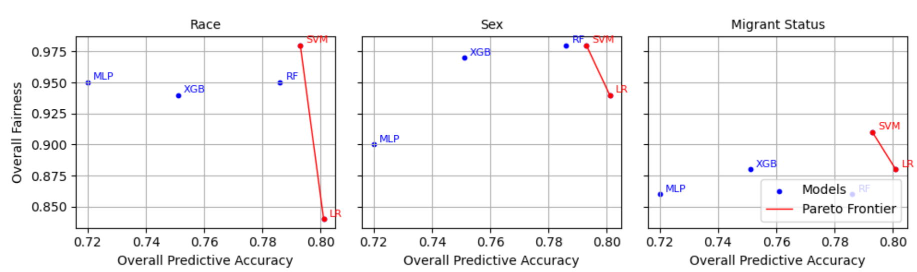

Fig. .1: Pareto frontier illustrating the trade-off between fairness and predictive accuracy. Red and blue points mark optimal and suboptimal models respectively for $\tau = 0.4$

Table .5: Threshold $\tau = 0.4$. RFG for various predictive accuracy metrics: PR(AUC-PR), ROC (AUC-ROC), Acc(Accuracy), and F1(F1 Score). Column initials represent demographic intersections: NWM (Non-White Migrant), WM (White Migrant), NWNM (Non-White Non-Migrant), WNM (White Non-Migrant), NWF (Non-White Female), NWM (Non-White Male), WF (White Female), WM (White Male), FM (Female Migrant), MM (Male Migrant), FNM (Female Non-Migrant), and MNM (Male Non-Migrant)

		Race-Migrant Status				Race-Sex				Migrant-Sex			
		NWM	WM	NWNM	WNM	NWF	NWM	WF	WM	FM	MM	FNM	MNM
LR	AUC-PR	0.10	0.04	-0.07	0.05	0.01	-0.07	0.04	0.08	0.02	0.10	0.02	-0.02
	AUC-ROC	0.17	0.08	-0.14	0.07	-0.02	-0.14	0.08	0.11	0.04	0.19	0.03	-0.04
	Accuracy	0.03	0.05	-0.05	0.03	0.03	-0.12	-0.00	0.06	-0.05	0.10	0.02	-0.03
	F1 Score	0.01	0.03	-0.03	0.02	0.02	-0.09	-0.00	0.04	-0.04	0.07	0.01	-0.03
MLP	AUC-PR	0.11	0.03	-0.06	0.03	0.02	-0.05	0.02	0.05	0.05	0.09	0.01	-0.02
	AUC-ROC	0.19	0.01	-0.10	0.02	-0.01	-0.04	0.07	0.02	0.12	0.13	0.01	-0.03
	Accuracy	0.09	-0.09	-0.04	0.01	0.02	-0.04	0.05	-0.01	0.05	-0.01	0.04	-0.03
	F1 Score	0.06	-0.08	-0.03	0.01	0.01	-0.03	0.03	-0.01	0.02	-0.01	0.03	-0.02
RF	AUC-PR	0.06	0.07	-0.05	0.02	0.03	-0.05	-0.02	0.09	0.01	0.09	-0.01	-0.01
	AUC-ROC	0.11	0.13	-0.09	0.02	0.02	-0.08	-0.02	0.12	0.03	0.19	0.02	-0.04
	Accuracy	-0.00	0.07	-0.02	-0.00	0.01	-0.03	-0.00	0.03	-0.04	0.10	0.01	-0.02
	F1 Score	-0.01	0.04	-0.01	-0.00	0.01	-0.03	-0.01	0.02	-0.04	0.06	0.01	-0.01
SVM	AUC-PR	0.03	0.08	-0.01	0.00	0.05	-0.02	-0.03	0.07	0.01	0.07	-0.01	0.01
	AUC-ROC	0.06	0.13	-0.05	0.00	0.01	-0.04	-0.02	0.09	0.02	0.15	-0.01	-0.01
	Accuracy	-0.03	0.04	-0.01	0.00	0.02	-0.04	-0.03	0.04	-0.07	0.05	-0.00	-0.00
	F1 Score	-0.02	0.03	-0.00	0.00	0.01	-0.03	-0.02	0.03	-0.05	0.03	-0.00	-0.00
XGB	AUC-PR	0.07	0.05	-0.03	0.02	0.01	-0.03	0.02	0.06	-0.05	0.15	0.04	-0.02
	AUC-ROC	0.15	-0.00	-0.06	0.03	-0.05	-0.02	0.06	0.05	-0.12	0.34	0.07	-0.06
	Accuracy	0.12	-0.01	-0.05	0.02	-0.01	-0.03	0.00	0.03	-0.13	0.20	0.02	-0.03
	F1 Score	0.07	-0.01	-0.03	0.01	-0.00	-0.03	-0.01	0.02	-0.10	0.13	0.02	-0.02

Table .6: Average Predictive Accuracy of all models for threshold $\tau = 0.6$. The values are the average $\pm$ standard deviation. Boldened and red scores are the highest and least overall averages, respectively.

Model	AUC-ROC	AUC-PR	Accuracy	F1 Score	Overall Average
LR	0.75 ± 0.05	0.74 ± 0.06	0.70 ± 0.04	0.67 ± 0.06	$\mathbf{0.71 \pm 0.05}$
MLP	0.57 ± 0.05	0.58 ± 0.07	0.58 ± 0.04	0.51 ± 0.06	0.56 ± 0.05
RF	0.72 ± 0.04	0.71 ± 0.06	0.66 ± 0.03	0.57 ± 0.05	0.66 ± 0.05
SVM	0.74 ± 0.05	0.73 ± 0.06	0.70 ± 0.04	0.67 ± 0.06	$\mathbf{0.71 \pm 0.05}$
XGB	0.64 ± 0.05	0.65 ± 0.06	0.63 ± 0.05	0.58 ± 0.06	0.62 ± 0.05

Table .7: Threshold $\tau = 0.6$. fairness results. Predictive accuracies across Race, Sex, and Migrant Status (MS) groups, highlighting mean disparities, standard deviations, and statistical significance.

	Model	AUC-PR	AUC-ROC	Accuracy	F1 Score
Race	LR	-0.09 ± 0.08 ***	-0.09 ± 0.06 ***	-0.15 ± 0.06 ***	-0.17 ± 0.08 ***
	MLP	-0.04 ± 0.08 **	-0.08 ± 0.07 ***	-0.07 ± 0.06 ***	-0.08 ± 0.08 ***
	RF	-0.10 ± 0.08 ***	-0.05 ± 0.06 ***	-0.14 ± 0.06 ***	-0.20 ± 0.08 ***
	SVM	-0.14 ± 0.08 ***	-0.11 ± 0.06 ***	-0.12 ± 0.06 ***	-0.13 ± 0.08 ***
	XGB	-0.12 ± 0.08 ***	-0.10 ± 0.06 ***	-0.11 ± 0.06 ***	-0.16 ± 0.08 ***
Sex	LR	-0.07 ± 0.08 ***	-0.10 ± 0.06 ***	-0.07 ± 0.06 ***	0.02 ± 0.08 *
	MLP	-0.09 ± 0.08 ***	-0.15 ± 0.07 ***	-0.14 ± 0.06 ***	-0.11 ± 0.08 ***
	RF	-0.07 ± 0.08 ***	-0.10 ± 0.06 ***	-0.12 ± 0.06 ***	-0.06 ± 0.08 ***
	SVM	-0.06 ± 0.07 ***	-0.09 ± 0.06 ***	-0.03 ± 0.06 ***	0.05 ± 0.07 ***
	XGB	-0.01 ± 0.08	-0.06 ± 0.06 ***	0.03 ± 0.06 ***	0.08 ± 0.08 ***
MS	LR	0.04 ± 0.08 **	-0.02 ± 0.06 **	-0.08 ± 0.06 ***	-0.04 ± 0.08 **
	MLP	-0.10 ± 0.10 ***	-0.11 ± 0.08 ***	-0.05 ± 0.06 ***	0.00 ± 0.09
	RF	0.00 ± 0.08	-0.04 ± 0.07 ***	-0.12 ± 0.06 ***	-0.10 ± 0.08 ***
	SVM	0.07 ± 0.08 ***	0.00 ± 0.07	-0.07 ± 0.06 ***	-0.04 ± 0.08 ***
	XGB	0.02 ± 0.08 *	-0.03 ± 0.08 **	-0.04 ± 0.07 ***	-0.05 ± 0.09 ***

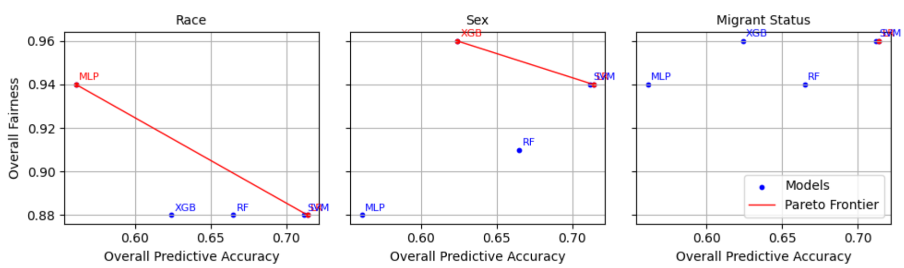

Fig. .2: Pareto frontier illustrating the trade-off between fairness and predictive accuracy. Red and blue points mark optimal and suboptimal models respectively for $\tau = 0.6$

Table .8: Threshold $\tau = 0.6$. RFG for various predictive accuracy metrics: PR(AUC-PR), ROC (AUC-ROC), Acc(Accuracy), and F1(F1 Score). Column initials represent demographic intersections: NWM (Non-White Migrant), WM (White Migrant), NWNM (Non-White Non-Migrant), WNM (White Non-Migrant), NWF (Non-White Female), NWM (Non-White Male), WF (White Female), WM (White Male), FM (Female Migrant), MM (Male Migrant), FNM (Female Non-Migrant), and MNM (Male Non-Migrant)

		Race-Migrant Status				Race-Sex				Sex-Migrant Status			
		NWM	WM	NWNM	WNM	NWF	NWM	WF	WM	FM	MM	FNM	MNM
LR	PR	-0.10	0.18	0.03	0.01	-0.03	0.02	0.05	0.07	0.07	0.08	0.01	0.03
	ROC	-0.08	0.12	0.00	0.01	-0.08	0.03	0.03	0.04	-0.07	0.05	-0.00	0.03
	Acc	-0.11	0.19	0.00	0.02	-0.05	-0.04	0.01	0.07	-0.07	-0.01	0.00	0.03
	F1	-0.11	0.20	-0.02	0.03	-0.05	-0.08	0.03	0.07	0.02	-0.03	0.01	0.01
MLP	PR	-0.13	0.16	0.05	0.01	-0.01	0.06	0.03	0.04	0.12	0.03	-0.02	0.07
	ROC	-0.18	0.03	0.04	0.00	-0.05	0.05	-0.03	0.05	-0.05	0.02	-0.03	0.06
	Acc	-0.04	0.07	-0.01	0.02	-0.07	0.05	0.01	0.03	-0.06	-0.01	-0.02	0.04
	F1	-0.20	0.19	0.01	-0.00	-0.05	0.03	0.01	0.05	0.06	-0.05	-0.04	0.04
RF	PR	-0.11	0.21	0.05	0.01	-0.06	0.01	0.04	0.04	0.05	0.06	-0.00	0.02
	ROC	-0.12	0.17	0.03	0.01	-0.10	0.05	0.03	0.00	-0.08	0.04	0.00	0.03
	Acc	-0.11	0.13	0.00	0.02	-0.09	0.02	0.02	0.04	-0.10	-0.04	-0.00	0.05
	F1	-0.15	0.18	-0.03	0.04	-0.16	0.03	0.07	0.02	-0.05	-0.05	0.01	0.04
SVM	PR	-0.05	0.15	0.01	0.03	-0.04	0.04	0.07	0.06	0.05	0.10	0.01	0.03
	ROC	-0.06	0.10	-0.01	0.02	-0.08	0.04	0.04	0.05	-0.08	0.09	-0.00	0.03
	Acc	-0.11	0.20	0.01	0.02	-0.04	-0.04	0.01	0.06	-0.07	-0.00	0.01	0.02
	F1	-0.12	0.21	-0.01	0.02	-0.03	-0.09	0.02	0.05	0.01	-0.02	0.02	-0.00
XGB	PR	-0.16	0.23	0.02	0.02	-0.00	-0.00	0.07	0.02	0.07	0.07	0.02	0.00
	ROC	-0.15	0.17	0.02	0.02	-0.05	0.01	0.02	0.02	-0.06	-0.00	0.00	0.03
	Acc	-0.11	0.26	0.01	0.01	-0.02	-0.04	0.02	0.03	-0.03	-0.01	0.02	-0.01
	F1	-0.21	0.30	0.00	0.02	-0.05	-0.08	0.04	0.04	0.03	-0.06	0.02	-0.01