MAAT:

A Novel Ensemble Approach to Addressing Fairness and Performance Bugs for Machine Learning Software

Claudia Farkas,
Amir Haeri,
Ahmed Radwan

Supervsied By: Maleknaz Nayebi

FAIRNESS

IN ML

Defining fairness can be tricky, as it's highly dependent on the specific application domain and the ML task being performed

- Machine learning (ML) is being increasingly used across industries—finance, healthcare, hiring, and more—but concerns about fairness and bias are growing.
- Fairness ensures that ML models do not produce biased outcomes that could lead to discrimination against certain groups (e.g., by race, gender, or socioeconomic status).
- Historical data often contains inherent biases, which are learned by ML models and can result in unfair decisions

Age: sensitive for loans, not for movies.

Location: sensitive for dating, not for weather.

Income: sensitive for insurance, not for shopping

Addressing Bias in Machine Leanring

Late-Stage Bias Mitigation

Early-Stage Bias Mitigation

Early-Stage Bias Mitigation

Early-Stage Bias Mitigation

- Instead of later in the development, why not during the requirements phase?
- Identifying sensitive features (e.g., race, gender) can help mitigate the bias.

Late-Stage Bias Mitigation

Late-Stage Bias Mitigation

- Most current approaches to addressing fairness issues are applied late **in the development cycle**, usually during **data preprocessing** or **model training**.

- Data Rebalancing.
- Ensemble Learning.
- Fairness Constraint.
- Knowledge Distillation.

EXISTING WORK

Current Methodologies

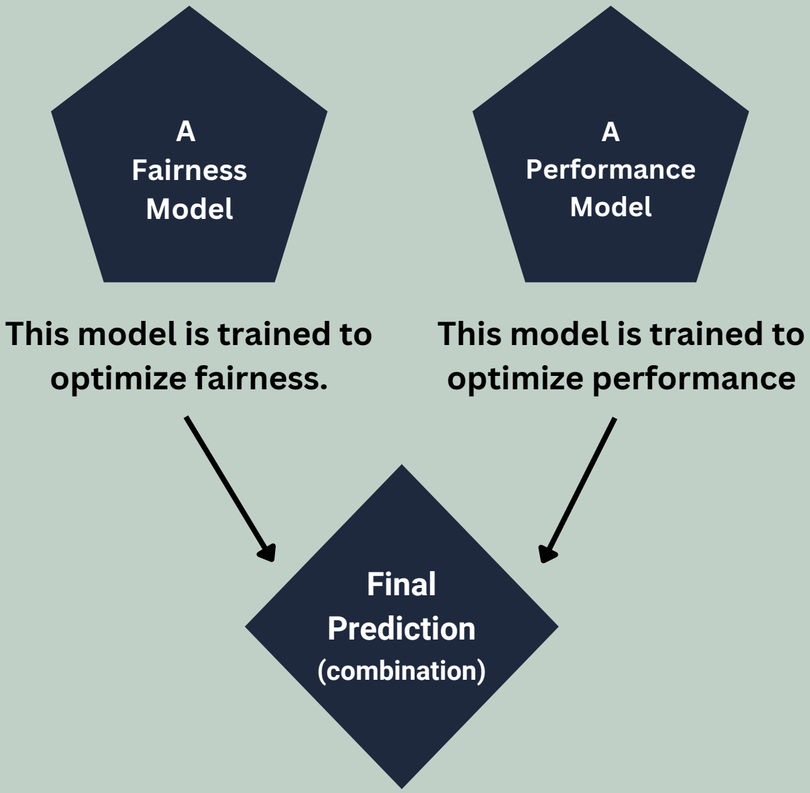
Model	Category	Details	Disadvantages
Reweighing (REW):	Pre-Processing	Adjusts the weights of training samples to reduce bias.	it can be sensitive to small changes in the training data.
Adversarial Debiasing (ADV)	In-Proessing	Uses adversarial techniques to minimize evidence of the protected attribute in predictions while maximizing performance.	an be computationally expensive to train
Reject Option Classification (ROC):	Post-Processing	Focuses on predictions with high uncertainty and adjusts them to improve fairness.	may not be effective for all types of data.
Fairway:	Combined (Pre- and In-Processing)	Combines pre-processing and in-processing techniques. It uses situation testing to remove ambiguous data and then uses multi-objective optimization.	can be complex to implement and computationally expensive.
Fair-SMOTE	Ensemble	Combines two pre-processing techniques. It first generates new data points to equalize the number of training samples in different subgroups. Then, it removes ambiguous data points like Fairway	the data generation process may not always create realistic or helpful data.

MAAT

MAAT Overview

A novel fairness-performance ensemble approach designed to improve this trade-off

MAAT works by training two different models:



What? Whether an ensemble method can effectively balance fairness and performance by combining models with different goals.

So? Existing methods typically focus on either fairness or performance, making it difficult to balance both goals. MAAT, on the other hand, attempts to overcome this limitation by combining models specifically optimized for each objective.

MAAT Overview

A novel fairness-performance ensemble approach designed to improve this trade-off

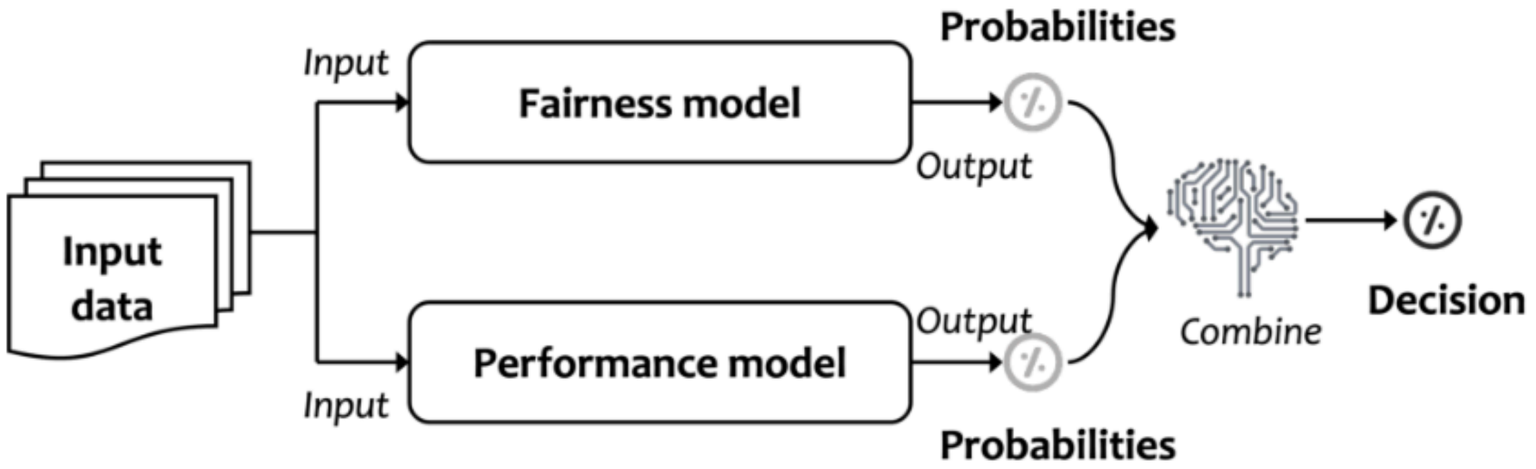


Figure 1: Overview of MAAT.

MAAT Methodology

MAAT Framework

As mentioned earlier, the MAAT framework consists of two models, i.e. fairness and performance model. First we go through the fairness model.

FAIRNESS MODEL

1. Bias in the **training data** is the root cause of ML software bias. (e.g. adult dataset)

2. Use **data debugging** to locate and modify the data that causes program bugs.

3. Encoding the **WAE** (we're all equal) worldview. No statistical association between the outcome decision and the protected attribute.

How to encode WAE?

First divide the training data into four subgroups:

1. Favorable (F)
2. Unfavorable (U)

} Outcome labels

3. Privileged (P)
4. Unprivileged (U)

} Protected attributes

Use PF, PU, UF, and UU to denote the # of samples in each subgroup.

Now what?

Make the favorable rates of privileged and unprivileged groups equal:

$$\frac{PF}{PF + PU} = \frac{UF}{UF + UU}$$

But is that enough?

Spoiler alert: No!

Types of bias

1. Selection bias
2. Label bias

The formula on the previous page is enough to mitigate selection bias only. In order to mitigate label bias as well, we need to do more.

Why does label bias happen?
ML software tends to falsely produce the favorable label for the privileged and the unfavorable label for the unprivileged, label bias mainly exists in the Privileged & Favorable and Unprivileged & Unfavorable subgroups.

$$\frac{PF-a}{PF-a+PU}=\frac{UF}{UF+UU-b}$$

$$\frac{PF+PU}{UF+UU}=\frac{PF-a+PU}{UF+UU-b}$$

- Calculate a and b
- Randomly remove a samples from the Privileged & Favorable subgroup, and
- b samples from the Unprivileged & Unfavorable subgroup .
- Train the fairness model.

PERFORMANCE MODEL

Performance Model

Using traditional ML algorithms on the original training data, as the performance model.

Ensemble Learning

Combination

Average the produced probabilities of the two models

E.g. binary classification:

- unfavorable label: 0
- favorable label: 1

Probability vectors:

- fairness model: $[p_{0f}, p_{1f}]$
- performance model: $[p_{0p}, p_{1p}]$

- predict 0 if: $\frac{p_{0f} + p_{0p}}{2} \geqslant \frac{p_{1f} + p_{1p}}{2}$
- predict 1 otherwise

Evaluation

Datasets

Name	Protected attribute(s)	#Features	Favorable label	Majority label	Size
Adult	Sex, Race	14	1 (income > 50K)	0 (75.2%)	45,222
Compas	Sex, Race	10	0 (no recidivism)	0 (54.5%)	6,167
German	Sex	20	1 (good credit)	1 (70.0%)	1,000
Bank	Age	20	1 (subscriber)	0 (87.3%)	30,488
Mep	Race	41	1 (utilizer)	0 (82.8%)	15,830

- **Protected Attribute:** attributes of individuals that are legally or ethically safeguarded against discrimination.
- **Diversity:** Covers financial, social, and medical application domains.
- **Benchmark Tasks:** Datasets are used for both single and multiple protected attribute tasks.

- **Favourable Outcome:** Desired prediction ((Y_hat = 1)
- **Unfavourable Outcome:** Undesired prediction (Y_hat = 0)
- **Privileged Group:** Historically advantaged (A = 1)
- **Unprivileged Group:** Historically disadvantaged (A = 0).

General Metrics and Measurements

Evaluation Criteria:

- **Two key aspects evaluated:**
  - Fairness and Performance.
- **Combined into 15 fairness-performance measurements:**
  - 3 fairness metrics × 5 performance metrics.
- **Goal:**
  - Improve fairness while minimizing performance degradation.
  - Measure effectiveness using the **Fairea tool** to classify trade-offs.
- **Trade-Off Classification Levels:**
  - **Win-Win:** Improves both fairness and performance.
  - **Good Trade-Off:** Improves fairness with acceptable performance loss.
  - **Other Levels:** Poor, Lose-Lose, and Inverted Trade-Offs.

Fairea tool

A benchmarking framework for fairness-performance trade-offs.

How It Works:

1. Creates a baseline by simulating naive models with controlled predictions (mutations).
2. Compares bias mitigation methods against this baseline.

In MAAT they repeated the mutation 50 times.

Fairness Metrics

1. Statistical Parity Difference (SPD)

- **What It Measures:** Difference in favourable outcome probabilities between unprivileged and privileged groups.
- **Formula:** $$SPD = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)$$
- **Ideal Value:** 0 (indicates fairness).
- **Example:** SPD highlights if one group receives disproportionately more offers in hiring.

2. Equal Opportunity Difference (EOD)

- **What It Measures:** Difference in true-positive rates (TPR) between unprivileged and privileged groups.
- **Formula:** $$EOD = P(\hat{Y} = 1 \mid A = 0, Y = 1) - P(\hat{Y} = 1 \mid A = 1, Y = 1)$$
- **Ideal Value:** 0 (equal chances for positive outcomes).
- **Example:** In healthcare, it ensures equal detection rates for diseases across groups.

Fairness Metrics

Average Odds Difference (AOD)

- **W**hat It Measures:
  - AOD evaluates fairness by measuring the average of the false-positive rate difference and the true-positive rate difference between privileged and unprivileged groups. It ensures the model treats both groups equally in terms of errors and successes.

$$\frac{1}{2}\left[\left|P(\hat{Y}=1 \mid A=0, Y=0) - P(\hat{Y}=1 \mid A=1, Y=0)\right| + \left|P(\hat{Y}=1 \mid A=0, Y=1) - P(\hat{Y}=1 \mid A=1, Y=1)\right|\right]$$

Example:

- **In a criminal justice setting:**
  - Scenario: A model predicts recidivism (likelihood of re-offending).
- **Impact:** AOD ensures the false-positive rates (incorrectly predicting recidivism) and true-positive rates (correctly predicting recidivism) are equal for different racial or gender groups.

Performance Metrics

Accuracy	Measures the percentage of correct predictions out of the total predictions made.
Precision	Proportion of correctly predicted positive instances among all instances predicted as positive.
Recall	Proportion of correctly predicted positive instances out of all actual positive instances.
F1-Score	Harmonic mean of precision and recall, balancing both metrics.
MCC (Matthews Correlation Coefficient)	Evaluates predictions' quality for imbalanced datasets, ranging from -1 to 1.

Performance Metrics

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Numerator: Quantifies Agreement and Disagreement

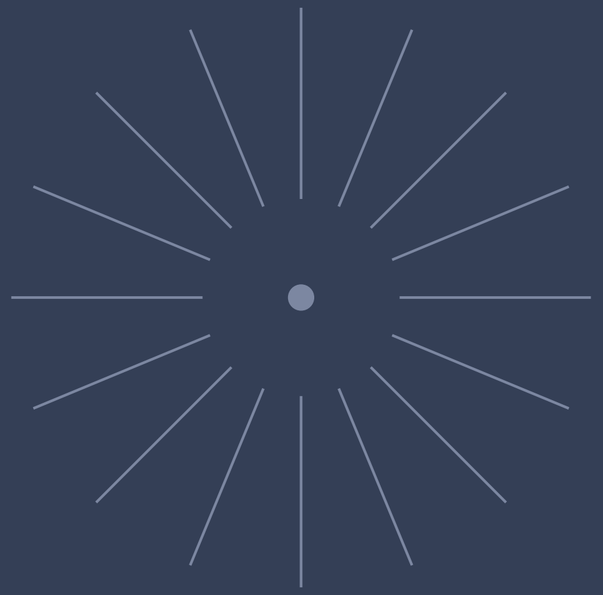
- TP×TN: Reflects correct predictions (true positives and true negatives).
- FP×FN: Reflects incorrect predictions.

Denominator: Normalizes for Dataset Imbalance

- The denominator ensures MCC is normalized by multiplying the sums of rows and columns of the confusion matrix

Combining Terms:

- The numerator provides the raw agreement-disagreement measure.
- The denominator adjusts for scale and balances the metric across datasets with varying proportions of classes.



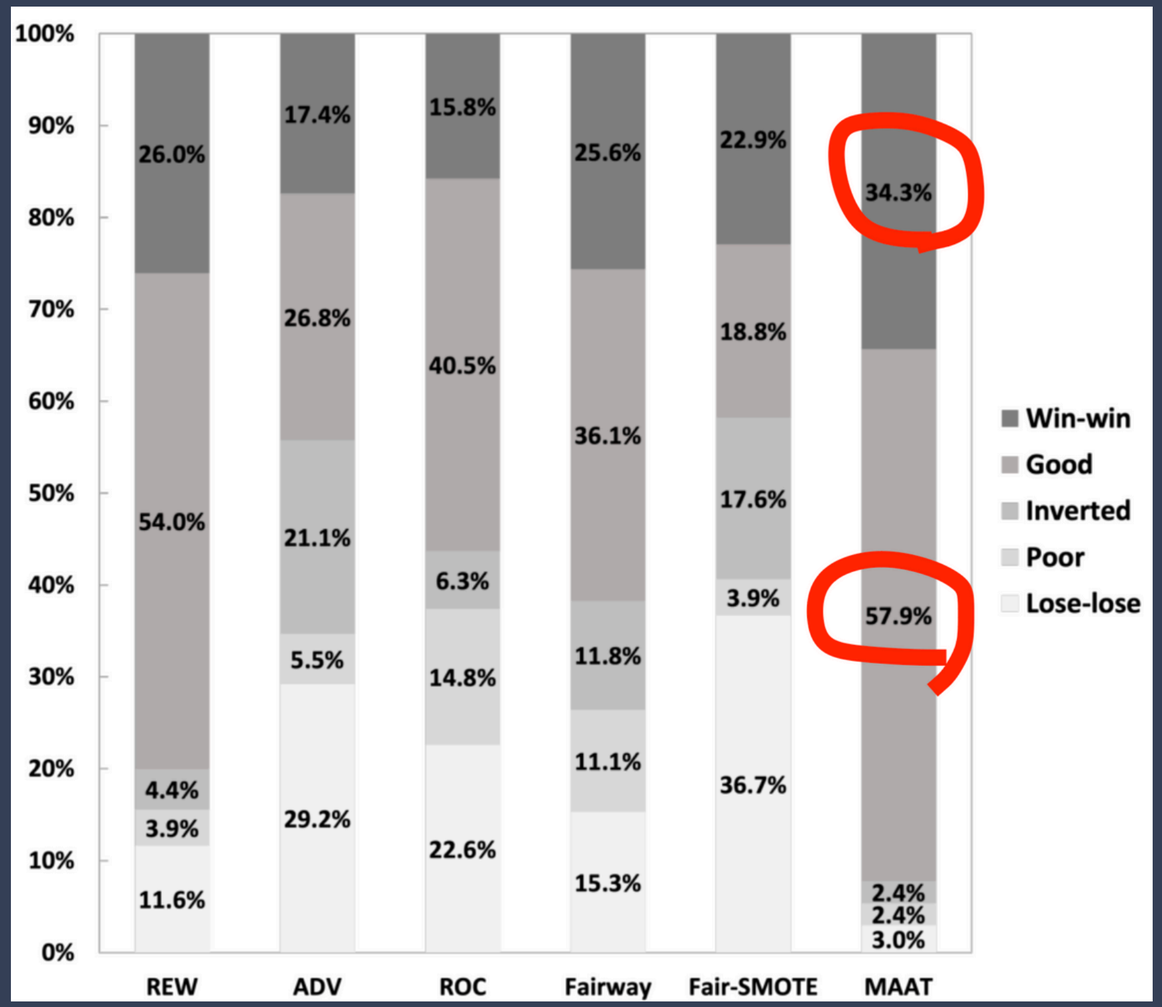
Results

Was it good?

Research Questions

1. *Trade-off Effectiveness*
2. *Applicability*
3. *Influence of Fairness and Performance Models*
4. *Influence of Combination Strategies*
5. *Multiple Protected Attributes*

RQ 1 - What fairness-performancet rade- off does MAAT achieve?



- MAAT achieves the best trade-off, with 92.2% of the mitigation cases falling in **good** or **win-win** trade-off.
- MAAT achieves **poor** or **lose-lose** tradeoff in much fewer cases (only 5.4%) than existing methods.

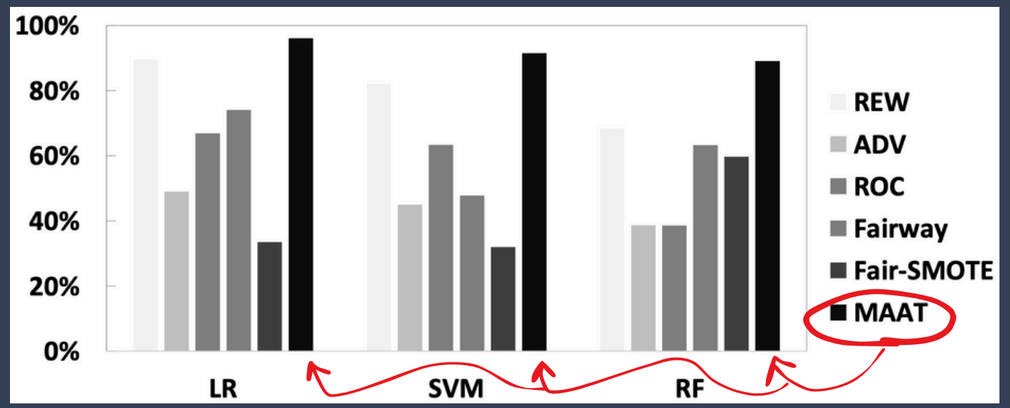
RQ 1 - Answer

	REW	ADV	ROC	Fairway	Fair-SMOTE	MAAT
Fairness ↑	87.3%	33.3%	66.7%	69.8%	33.3%	96.8%
Performance ↓	42.9%	51.4%	76.2%	61.9%	48.6%	44.8%

- Use the Mann Whitney U-test to test whether the fairness/performance is significantly improved/decreased.
  - MAAT improves fairness in 96.8% of the scenarios.
  - Existing methods between 33.3% - 87.3%
  - MAAT decreases ML performance in 44.8% of the scenarios
  - Current best alternative is 42.9%

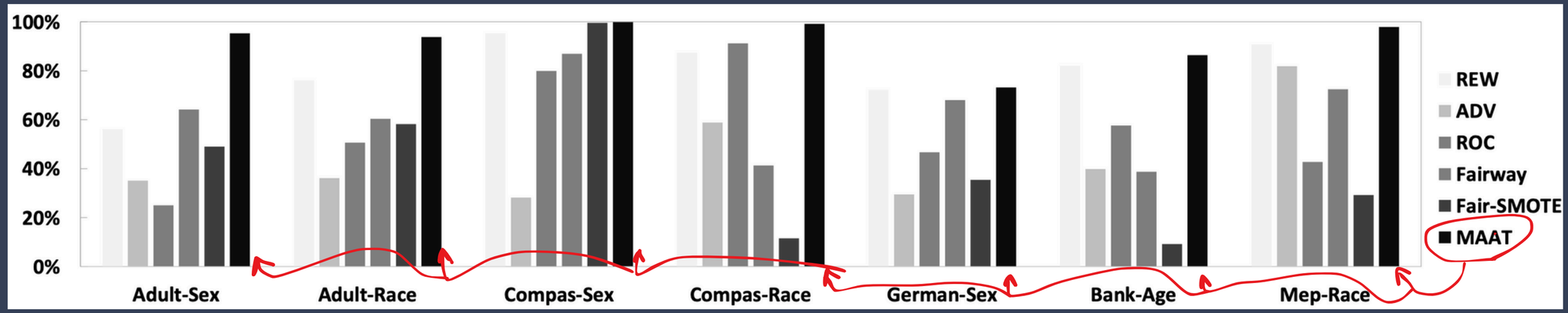
MAAT can significantly improve fairness in 96.8% of the scenarios, without decreasing ML performance too much.

RQ 2 - How well does MAAT apply to different ML algorithms, decision tasks, and fairness-performance measurements?



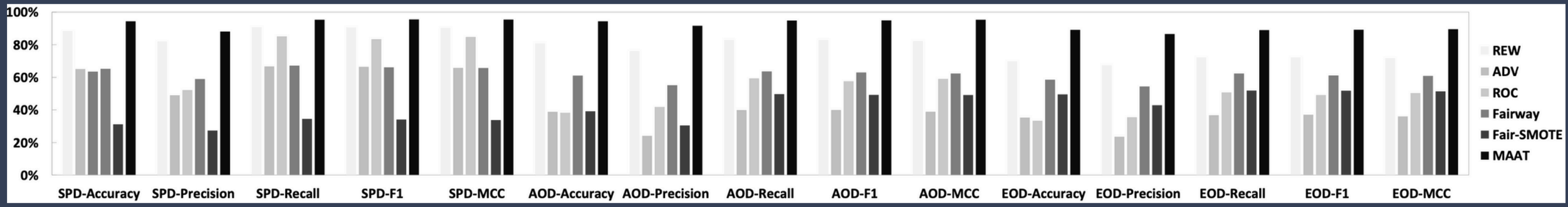
Proportions of cases beating the baseline for different ML algorithms:

- Linear Regression (LR)
- Support Vector Machine (SVM)
- Random Forest (RF)



Proportions of cases beating the baseline for different tasks.

RQ 2 - Answer



Proportions of cases beating the baseline for different fairness-performance measurements

The superiority of MAAT over the state-of-the- art holds on all the ML algorithms, decision tasks, and fairness-performance measurements that we study.

RQ3: Influence of Fairness and Performance Models

-Fairness Models

Aims to understand the impact of using different fairness models and performance models on MAAT's effectiveness.

Table 3: (RQ3) Proportions of cases beating the trade-off baseline, achieved by existing bias mitigation methods, their combinations with MAAT, and the default setting of MAAT. The results show that the ensemble approach of MAAT can improve the trade-off for each method, but the default setting of MAAT still performs the best.

eg.

REW	Fairway	Fair-SMOTE	M-REW	M-Fairway	M-Fair-SMOTE	MAAT
80.0%	61.7%	41.7%	84.0%	65.8%	44.0%	92.2%

cases where the standalone fairness models (without MAAT) exceeded the trade-off baseline.

same fairness models but integrated into MAAT as the fairness component "M-"

cases where MAAT (using its default training data & debugging technique as the fairness model) beat the baseline

How to read?

Compare all models and variants against MAAT's default setting, MAAT with its training data debugging technique is the most effective, achieving 92.2% of cases beating the baseline.

RQ3: Influence of Fairness and Performance Models

-Performance Models

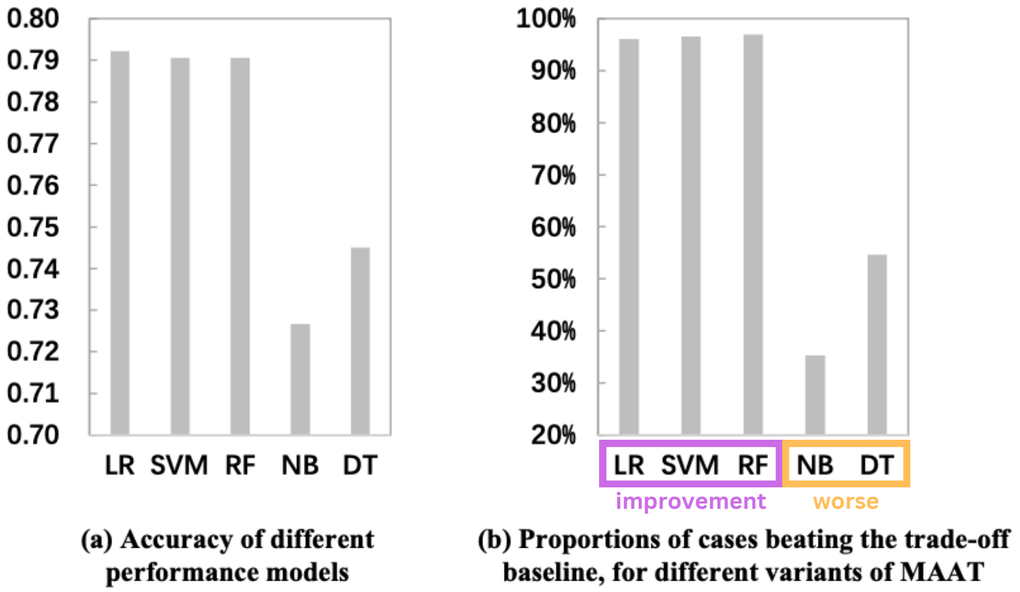
Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF) NB (Naive Bayes), DT (Decision Tree)

Looking at the Accuracy Graph (4a):

- Tested five different ML algorithms: LR, SVM, RF, NB, and DT
- The more sophisticated algorithms (LR, SVM, RF) achieved accuracy around 0.77-0.79
- While simpler models like NB and DT performed notably lower at around 0.71-0.73

Looking at the Trade-off Effectiveness Graph (4b):

- The same pattern emerged in MAAT's effectiveness
- LR, SVM, and RF variants achieved ~96-97% success in beating the trade-off baseline
- NB and DT variants only achieved 35.3% and 54.6% respectively



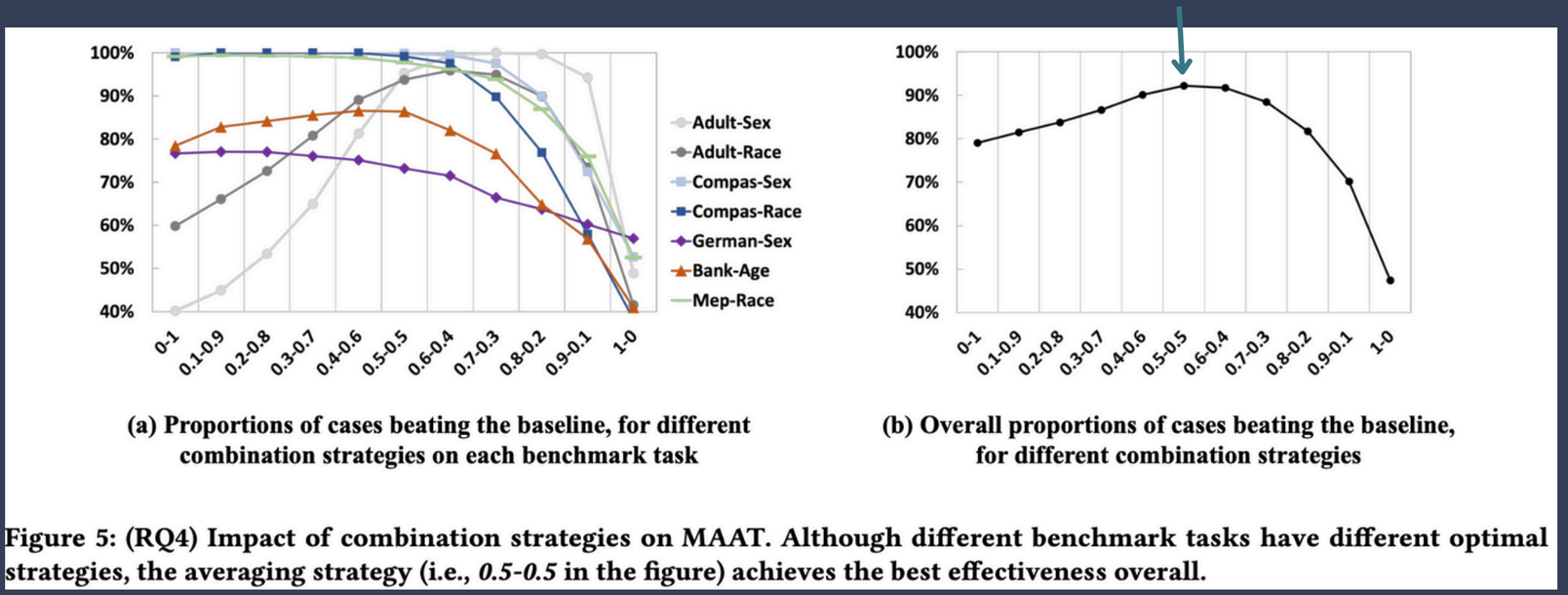
(a) Accuracy of different performance models

(b) Proportions of cases beating the trade-off baseline, for different variants of MAAT

Figure 4: (RQ3) Impact of the performance model on MAAT. MAAT tends to have better effectiveness with more accurate performance models.

RQ4: Influence of Combination Strategies

Explores the influence of different combination strategies on MAAT's effectiveness.



(a) Proportions of cases beating the baseline, for different combination strategies on each benchmark task

(b) Overall proportions of cases beating the baseline, for different combination strategies

Figure 5: (RQ4) Impact of combination strategies on MAAT. Although different benchmark tasks have different optimal strategies, the averaging strategy (i.e., *0.5-0.5* in the figure) achieves the best effectiveness overall.

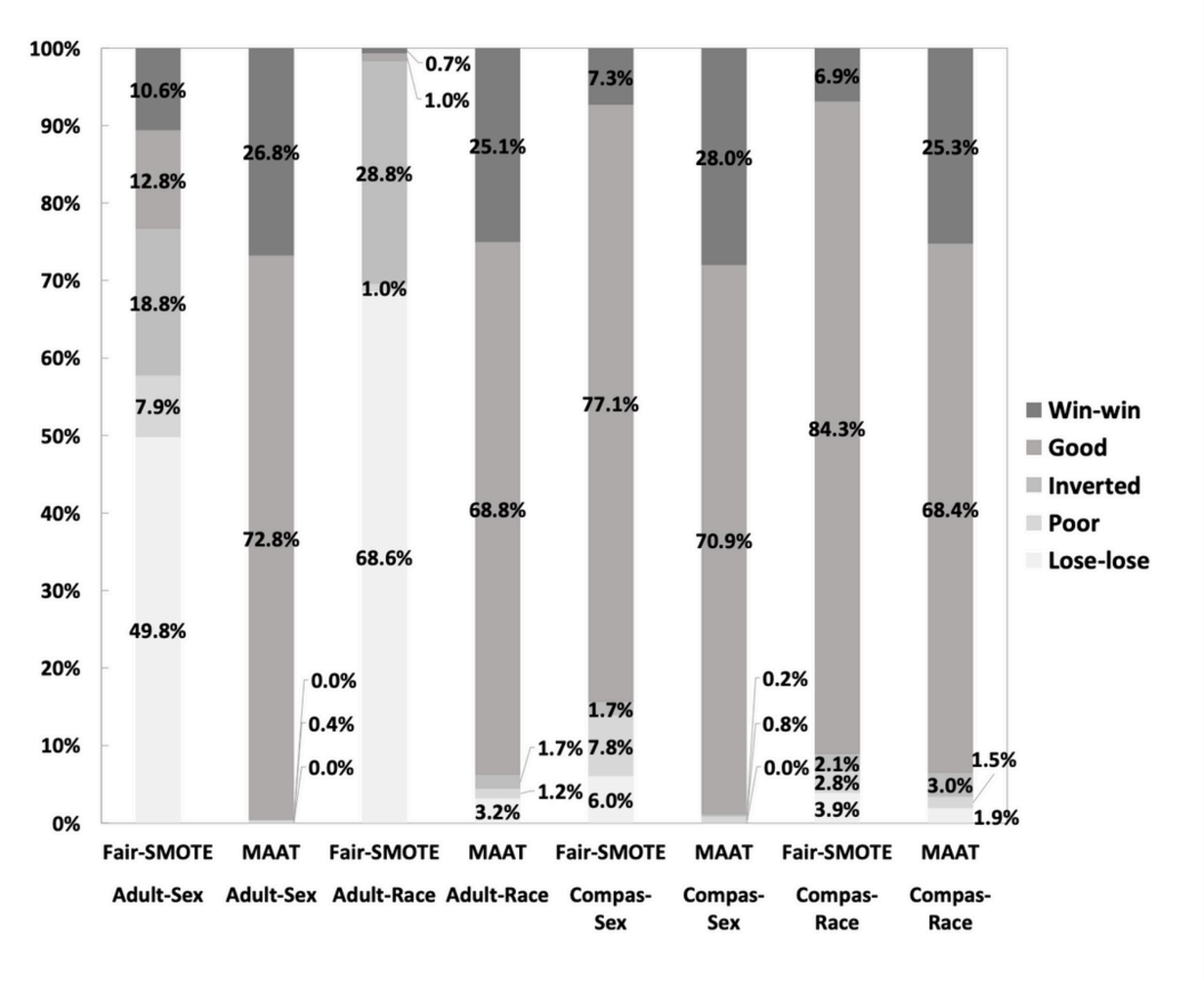
A) E.g: "Adult-Sex" shows the performance of each combination strategy when applied to the Adult dataset with "Sex" as the protected attribute.
Its best performance (almost 100% of cases beating the baseline) when using the "0.8-0.2" combination strategy. Meaning; giving more weight to the performance model leads to a better balance between fairness and accuracy

The Process: 11 different combination strategies, denoted as "0-1," "0.1-0.9," "0.2-0.8," and so on, up to "1-0." (represent the weights assigned to the output probability vectors of the 2 models.

Where, the performance model: ($[p0p\ ,\ p1p\ ]$) and the fairness model: ($[p0f\ ,\ p1f\ ]$)

Example, the 0.1-0.9 strategy calculates the final probability vector as $0.1 * [p0p\ ,\ p1p\ ] + 0.9 * [p0f\ ,\ p1f\ ]$. = *probability vector which guides the final prediction made by MAAT.*

RQ 5 - Is MAAT effective for handling multiple protected attributes simultaneously?



- **MAAT**: Separate fairness model per protected attribute + performance model
- **Fair-SMOTE**: Data balancing for class and protected attributes
- **Setup**: 3 Models × 2 Methods × 2 Datasets (Adult/Compas) × 50 Runs
- **Results**:
  - MAAT: 96.5% good/win-win trade-off vs Fair-SMOTE's 50.2%
  - MAAT: 0.4-4.4% poor/lose-lose vs Fair-SMOTE's 7.7-69.6%
- Training separate fairness models proved effective for multiple attributes

Challenges and Limitations

- **Fairness-Performance Trade-off:**
  - Balancing fairness improvements often comes at the cost of some reduction in ML performance, even with the MAAT ensemble.
- **Algorithm Dependence:**
  - While MAAT is widely applicable, its effectiveness may vary across ML algorithms, especially when using less accurate performance models.
- **Protected Attribute Access:**
  - Studying fairness requires access to protected attributes, which might not always be available due to regulations like GDPR.
- **Task-Specific Optimization:**
  - The ensemble strategy for combining models (fairness vs. performance) might need manual adjustment depending on the application.
- **Scalability with Data Size:**
  - While MAAT is computationally efficient compared to alternatives, handling very large datasets or high-dimensional data could still present challenges.
- **Limited Scope in Evaluation:**
  - MAAT's evaluation is primarily based on classical ML algorithms and structured data. Its performance on deep learning models or unstructured data (e.g., images, text) needs further exploration.
- **Trade-off Baseline Variability:**
  - The results and effectiveness levels heavily depend on the trade-off baseline defined by Fairea, which may vary with dataset and metrics selection.

THANK YOU