FairRF: Multi-Objective Search for Single and Intersectional Software Fairness

Giordano d'Aloisio*, Max Hort, Rebecca Moussa***, Federica Sarro***

Università degli Studi dell'Aquila, Italy*

Simula Research Laboratory, Norway**

University College London, UK***

Motivation

AI and ML-based systems in sensitive domains must be fair

Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness

Stefano Canali, Viola Schiaffonati, Andrea Aliverti

Fairness in credit scoring: Assessment, implementation and profit implications

Nikita Kozodoi [a], Johannes Jacob [a], Stefan Lessmann [a]

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU is regulated by the AI Act, the world's first comprehensive AI law. Find out how it protects you.

Sources of bias

Biased Training Data

Wrong Algorithm Implementation

Human Interaction

Sources of bias

Biased Training Data

Wrong Algorithm Implementation

Human Interaction

Problem

More than 300 bias mitigation methods available

Issue: Most of them are black-box - they do not allow stakeholders to control the trade-off between fairness and effectiveness

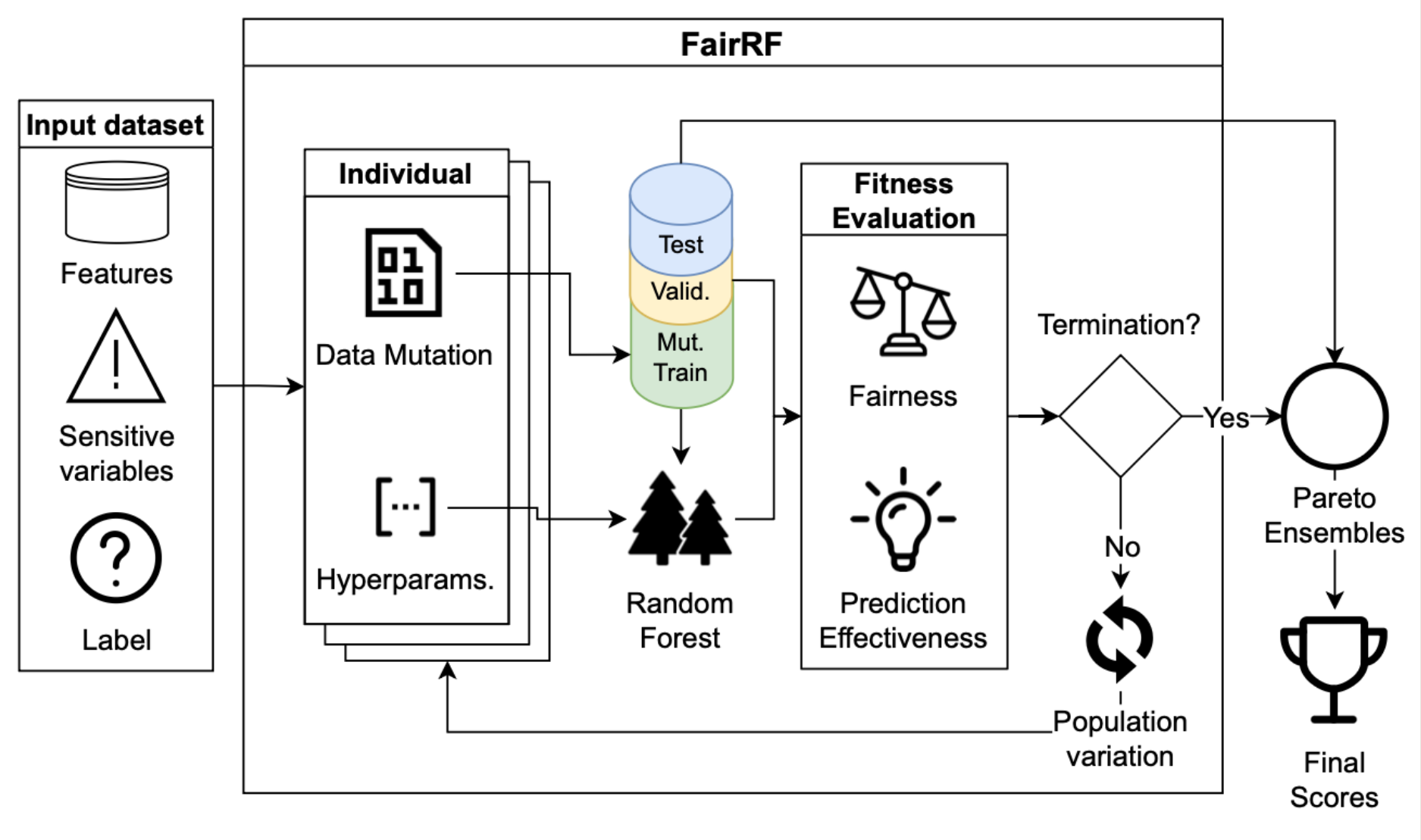
Insight: Frame bias mitigation as a multi-objective optimisation problem between fairness and effectiveness — return a Pareto front of solutions, letting stakeholders choose based on their priorities.

Background

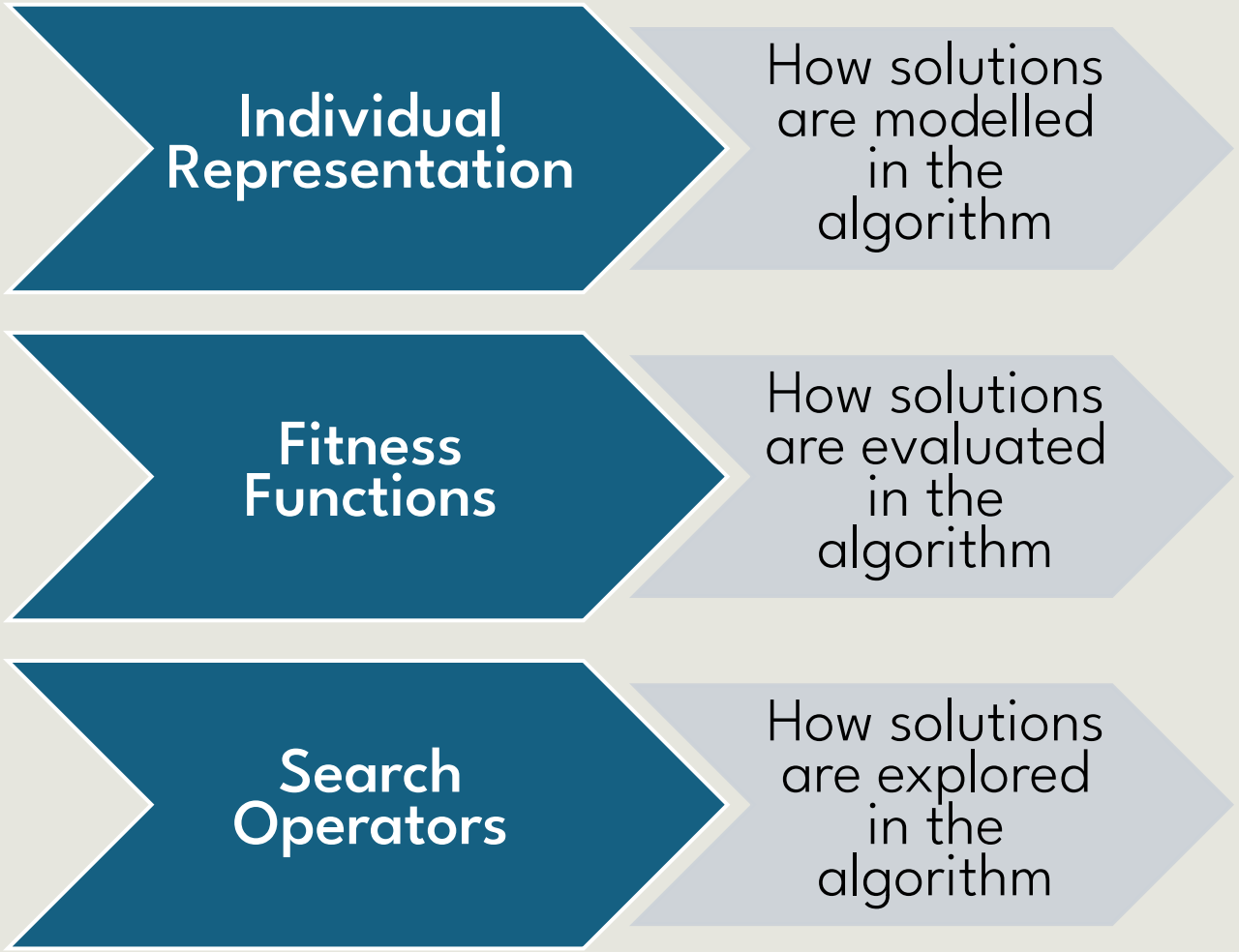
Multi-Objective Evolutionary Search

1. Initialise random population P
   a) Evaluate individuals' fitness
   b) Select the best non-dominated individuals P^*
   c) Create new population P' applying Crossover operators on P^*
   d) Mutate individuals on P'
   e) $P = P^* + P'$
   f) Repeat until termination criteria is reached (e.g., max generations are explored)
2. Return non-dominated individuals

FairRF Approach



Main Ingredients for Search-Based Algorithms



Individual Representation

RF Hyperparameters

- Estimators
- Quality Criterion
- Max Depth
- Min Samples Split
- Max Features

Data Mutation Value

- $n \in [0.1, 1.0]$
- Represents the fraction of sensitive feature values to flip (bit-flip mutation: $f(x) = 1 - x$)
- Applied only on training set – test and validation are untouched to avoid evaluation bias

Fitness Functions

Fairness: Statistical Parity Difference (SPD)

$$|P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)|$$

- 0 = perfect fairness
- 1 = maximum bias
- Minimised by FairRF

Effectiveness: Accuracy

$$\frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- Measures proportion of correct predictions
- Maximised by FairRF

Search Operators

- **Search algorithm:** NSGA2
- **Crossover:** Single Point Crossover with 60% probability
- **Mutation:** Random Mutation with 20% probability
- Explore a population of 50 individuals for 25 generations

Evaluation

- **RQ1 – Algorithm Variations:** *How does FairRF compare against algorithm variations using different base classifiers? (LR, KNN, CART, SVM)*
- **RQ2 – Base Algorithm:** *To what extent is FairRF able to improve fairness and effectiveness in predictions compared to base approaches?*
- **RQ3 – SOTA Single:** *How does FairRF compare against state-of-the-art bias mitigation methods for bias mitigation with single sensitive attributes?*
- **RQ4 – SOTA Intersec.:** *How does FairRF compare against state-of-the-art bias mitigation methods for intersectional bias mitigation?*

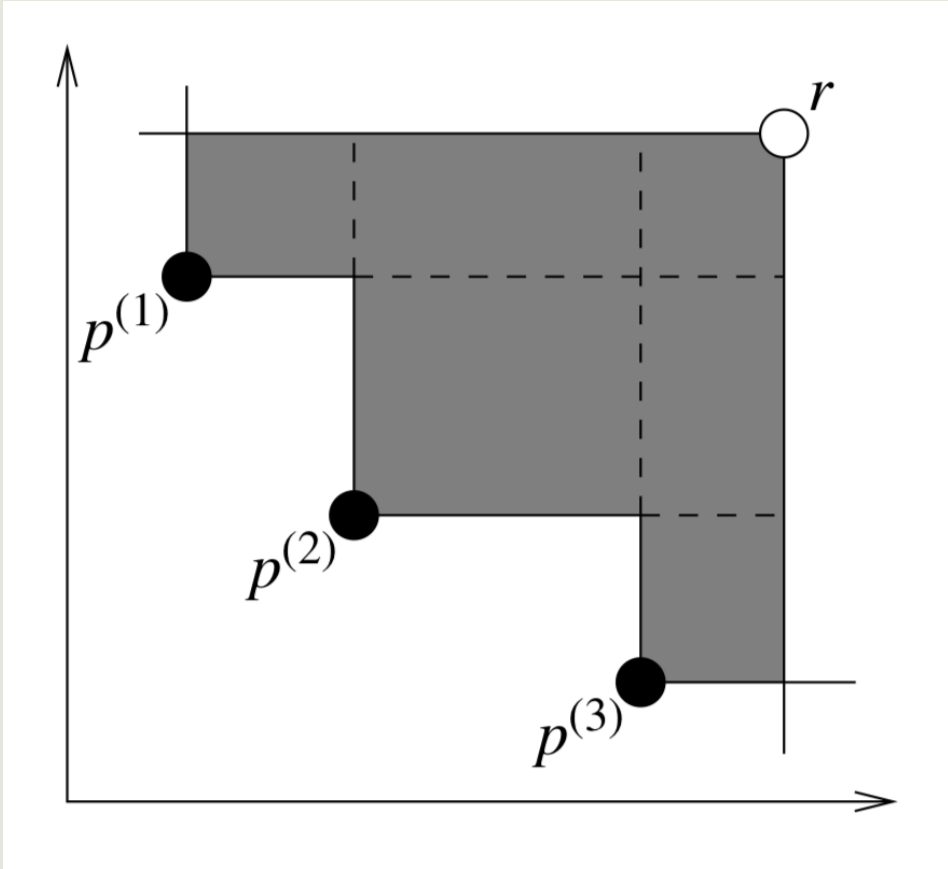
Datasets

Dataset	Description	Positive Label	Sensitive variable(s)	(Un)privileged groups
Adult Income	Predict the yearly income of US people	Income > 50.000	*Sex* *Race*	*Men vs Women* *White vs Black*
COMPAS	Predict the recidivism of condemned people	No-recidivism	*Race* *Sex*	*White vs Black* *Women vs Men*
German Credit	Predict the credit goodness	Good credit	*Age* *Sex*	*< 25 y.o. vs >= 25 y.o.* *Men vs Women*
Bank Telemarketing	Predict the subscription to a bank program	Subscription	*Age*	*< 25 y.o. vs >= 25 y.o.*
MEPS	Predict the usage of medical furniture	Usage	*Race*	*White vs Black*

Evaluation Metrics – RQ1

Hypervolume

- Multi-objective metric to measure the quality of a Pareto front in the objective space
- Higher = better coverage



Evaluation Metrics – RQ2, RQ3

Fairness

- Statistical Parity Difference
- Equal Opportunity Difference (EOD)

$$TPR_{unpriv} - TPR_{priv}$$

- Average Odds Difference (AOD)

$$\frac{1}{2}\left[\left(TPR_{unpriv} - TPR_{priv}\right) + \left(FPR_{unpriv} - FPR_{priv}\right)\right]$$

Effectiveness

- Accuracy
- Precision
- Recall
- F1 Score
- Matthews Correlation Coefficient (MCC)

Evaluation Metric – RQ4

Intersectional Fairness

- Worst-Case Scenario (WCS)

$$\mathrm{WCS-SPD} = \max_{s \in S}[P(\hat{Y}=1|S=s) - P(\hat{Y}=1|S \neq s)] - \min_{s \in S}[P(\hat{Y}=1|S=s) - P(\hat{Y}=1|S \neq s)]$$

- Average (AVG)

$$\mathrm{AVG-SPD} = \frac{\sum_{s \in S} P(\hat{Y}=1|S=s) - P(\hat{Y}=1|S \neq s)}{|S|}$$

RQ1: FairRF vs Algorithm Variations

FairRF (Random Forest)	FairLR (Logistic Regression)	FairSVM (Support Vector Machine)	FairKNN (K-Nearest Neighbours)	FairCART (Decision Tree)
Best HV in 5 / 8 scenarios (62.5%)	Competitive; wins in 2 scenarios	Wins in 2 scenarios; drops effectiveness	Entirely dominated by FairRF	Never beats FairRF

RQ1: FairRF vs Algorithm Variations

FairRF (Random Forest)	FairLR (Logistic Regression)	FairSVM (Support Vector Machine)	FairKNN (K-Nearest Neighbours)	FairCART (Decision Tree)
Best HV in 5 / 8 scenarios (62.5%)	Competitive; wins in 2 scenarios	Wins in 2 scenarios; drops effectiveness	Entirely dominated by FairRF	Never beats FairRF

Answer to RQ1: Random Forests work best as base classifier — FairRF wins in 62.5% of scenarios (dataset + sensitive variable), followed by FairLR and FairSVM.

RQ2: FairRF vs Base Classifiers

Baselines: RF, LR, SVM (hyperparameter-tuned via Grid Search without data mutation) + Random Search (RS)

Fairness Improvement

- SPD: 84.7% of scenarios
- EOD: 65.6% of scenarios
- AOD: 75.0% of scenarios

Effectiveness

- Accuracy improves in 59.3% of scenarios
- Precision/Recall/F1 slightly lower (fairness/effectiveness trade-off)
- MCC significantly better in 56.2% of cases

RQ2: FairRF vs Base Classifiers

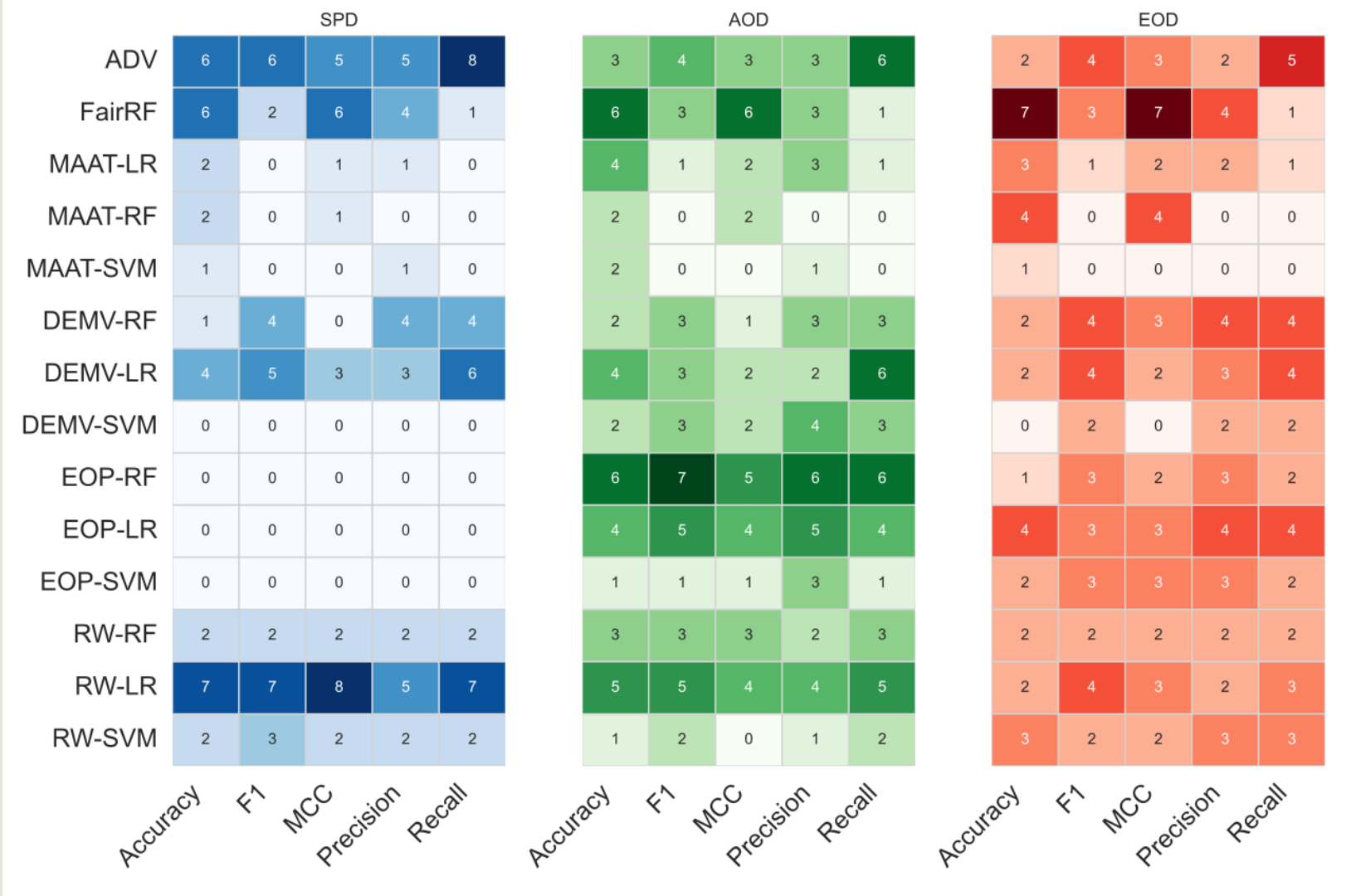
Baselines: RF, LR, SVM (hyperparameter-tuned via Grid Search without data mutation) + Random Search (RS)

Fairness Improvement	Effectiveness
• SPD: 84.7% of scenarios	• Accuracy improves in 59.3% of scenarios • MCC significantly better in 56.2% of cases

Answer to RQ2: FairRF significantly improves fairness without significantly impacting prediction effectiveness

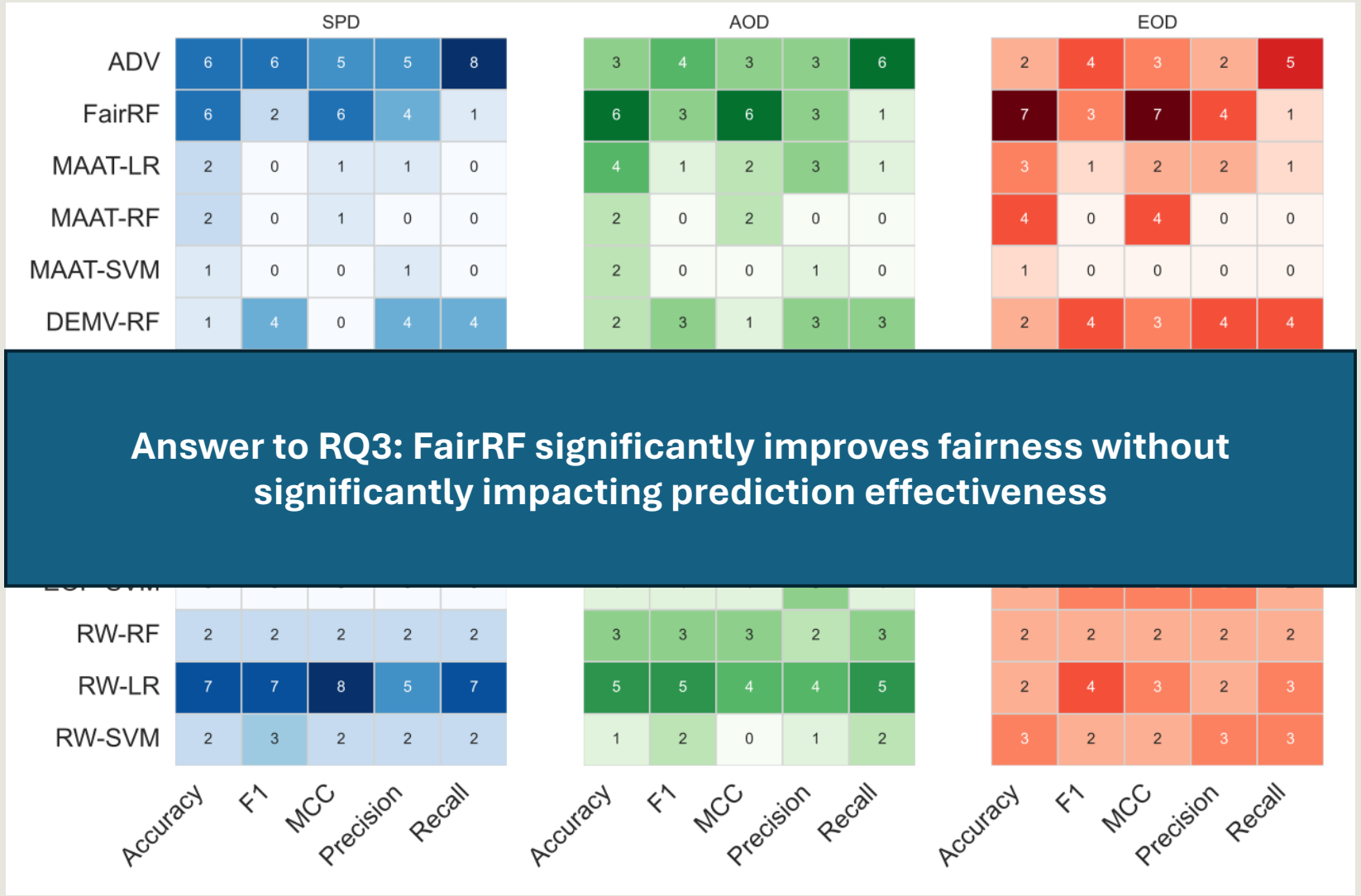
RQ3: FairRF vs SOTA

- Baselines: ADV (Adversarial Debiasing), MAAT, DEMV, EOP, RW across LR, RF, SVM variants
- Metric: Pareto optimality count — how many times a method appears on the global Pareto front across all scenarios



RQ3: FairRF vs SOTA

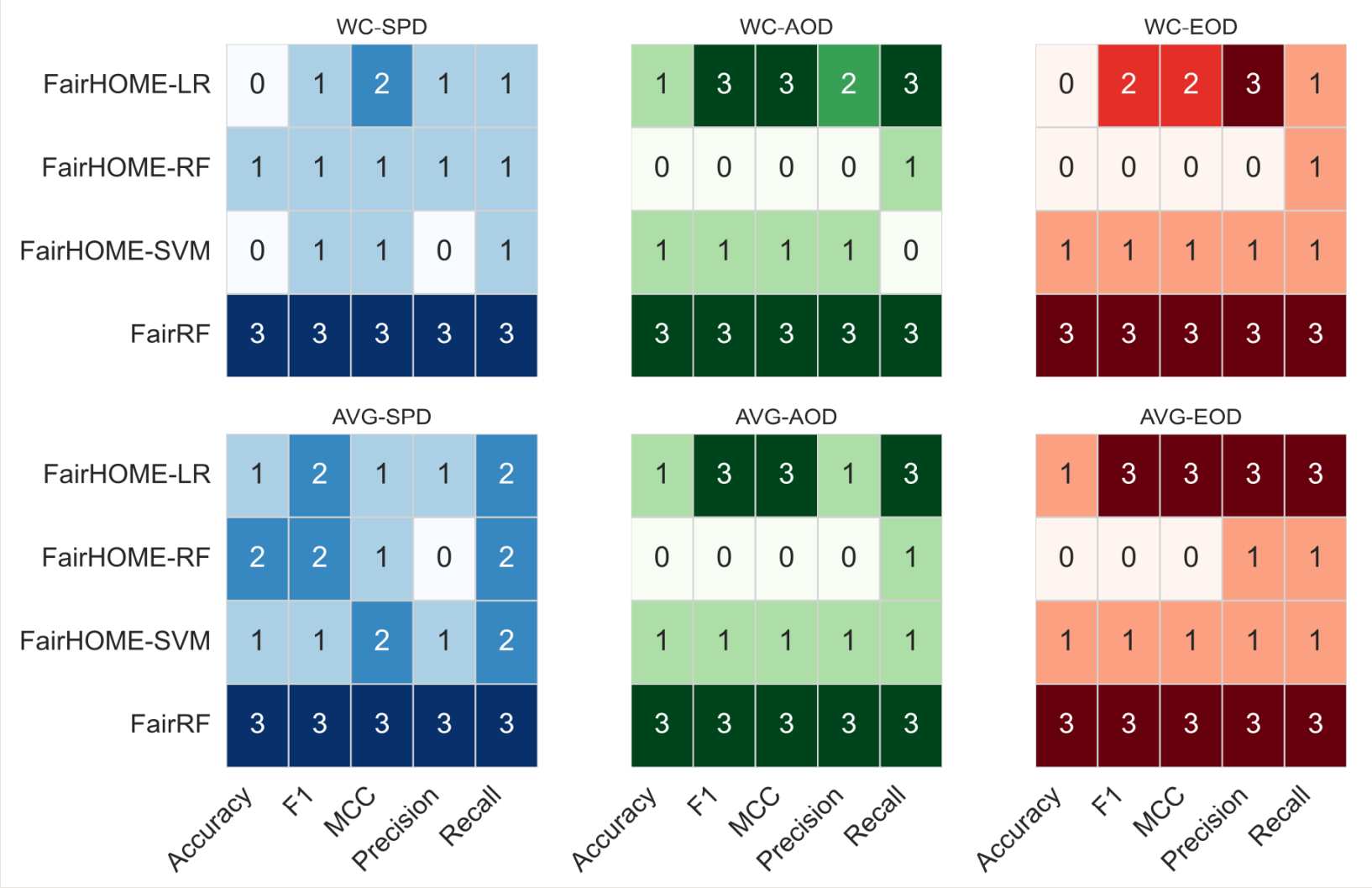
- Baselines: ADV (Adversarial Debiasing), MAAT, DEMV, EOP, RW across LR, RF, SVM variants
- Metric: Pareto optimality count — how many times a method appears on the global Pareto front across all scenarios



Answer to RQ3: FairRF significantly improves fairness without significantly impacting prediction effectiveness

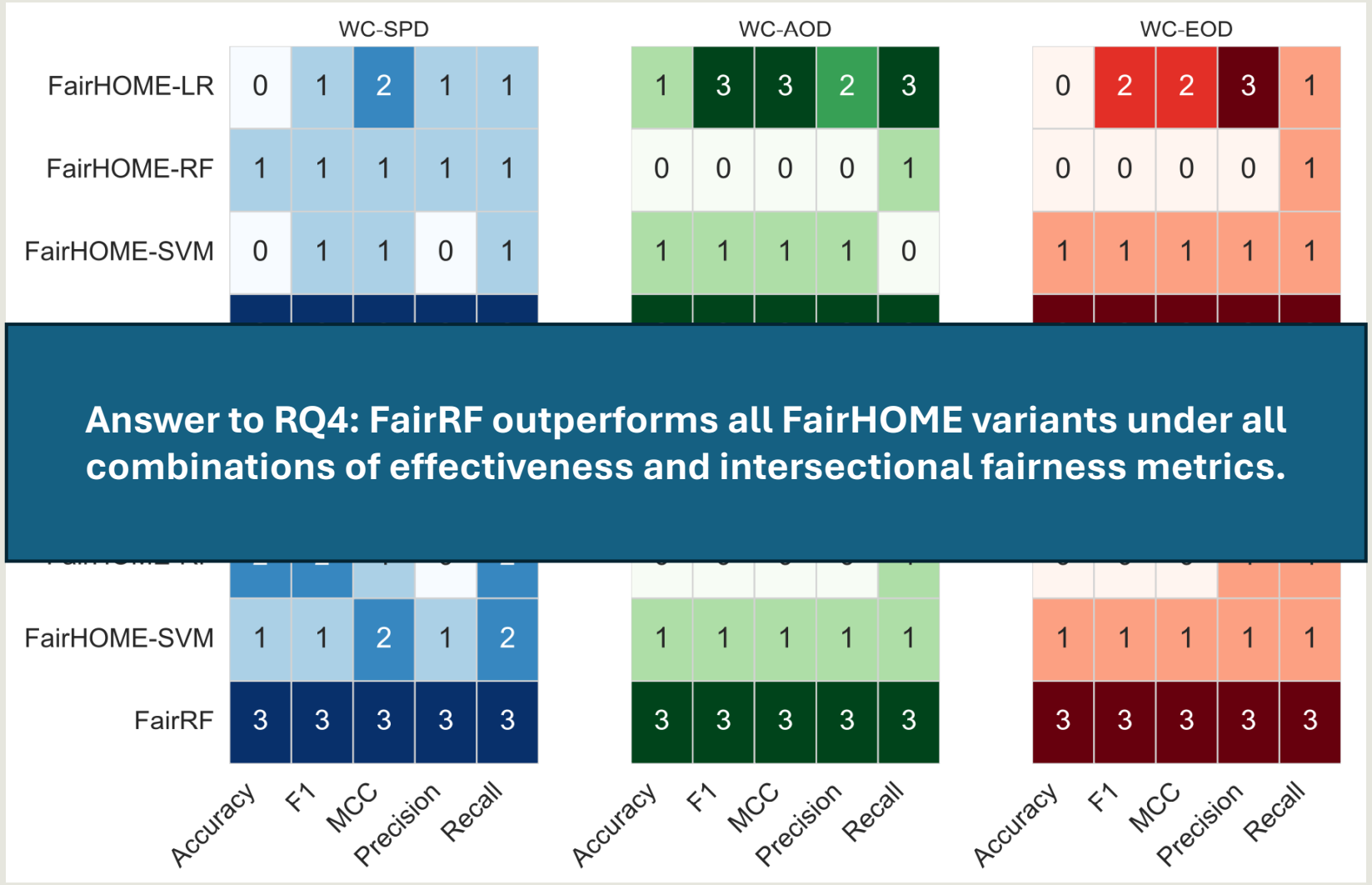
RQ4: FairRF vs SOTA Intersectional Bias

Baseline: FairHOME with LR, RF, SVM base classifiers



RQ4: FairRF vs SOTA Intersectional Bias

Baseline: FairHOME with LR, RF, SVM base classifiers



Answer to RQ4: FairRF outperforms all FairHOME variants under all combinations of effectiveness and intersectional fairness metrics.

Discussion

- Data mutation is essential, as shown in the answer to RQ2
- SPD fitness generalises across other fairness metrics
- Accuracy fitness may impact TPR
- FairRF overcomes the SOTA for intersectional bias in all scenarios analysed
- **Limitation:** Pareto front may contain many solutions — guiding stakeholder selection is an open challenge. Evaluation limited to binary classification with tabular data.

Future Work

- Extend FairRF to fairness in multi-class classification
- Develop tools to help stakeholders navigate the Pareto front
- Test other fitness functions
- Select automatically the optimal base classifier for a given dataset (AutoML)

Thank you for your attention!

giordano.daloisio@univaq.it

Replication Package