A Human-in-the-Loop Fairness-Aware Model Selection Framework for Complex Fairness Objective Landscapes

Jake Robertson[1,2], Thorsten Schmidt[1], Frank Hutter[1,3], Noor Awad[1]

[1] University of Freiburg, [2] Zuse School ELIZA, [3] ELLIS Institute Tübingen

In a Nutshell

Despite the recognized trade-offs among various algorithmic fairness concepts, existing fairness-aware ML methods typically focus on optimizing a single, user-specified fairness measure. This approach is problematic because:

1. Real-world FairML scenarios often involve intricate and varied stakeholder concerns, encompassing multiple fairness criteria
2. Concentrating on one fairness notion may not only compromise other pertinent fairness metrics, but also potentially result in adverse downstream effects

ManyFairHPO is a human-centered, optimization-driven framework that allows fairness practitioners to specify, rank, and optimize for multiple fairness metrics. ManyFairHPO facilitates fairness modeling decisions that effectively balance fairness objectives and reduce conflict-associated risks

Background

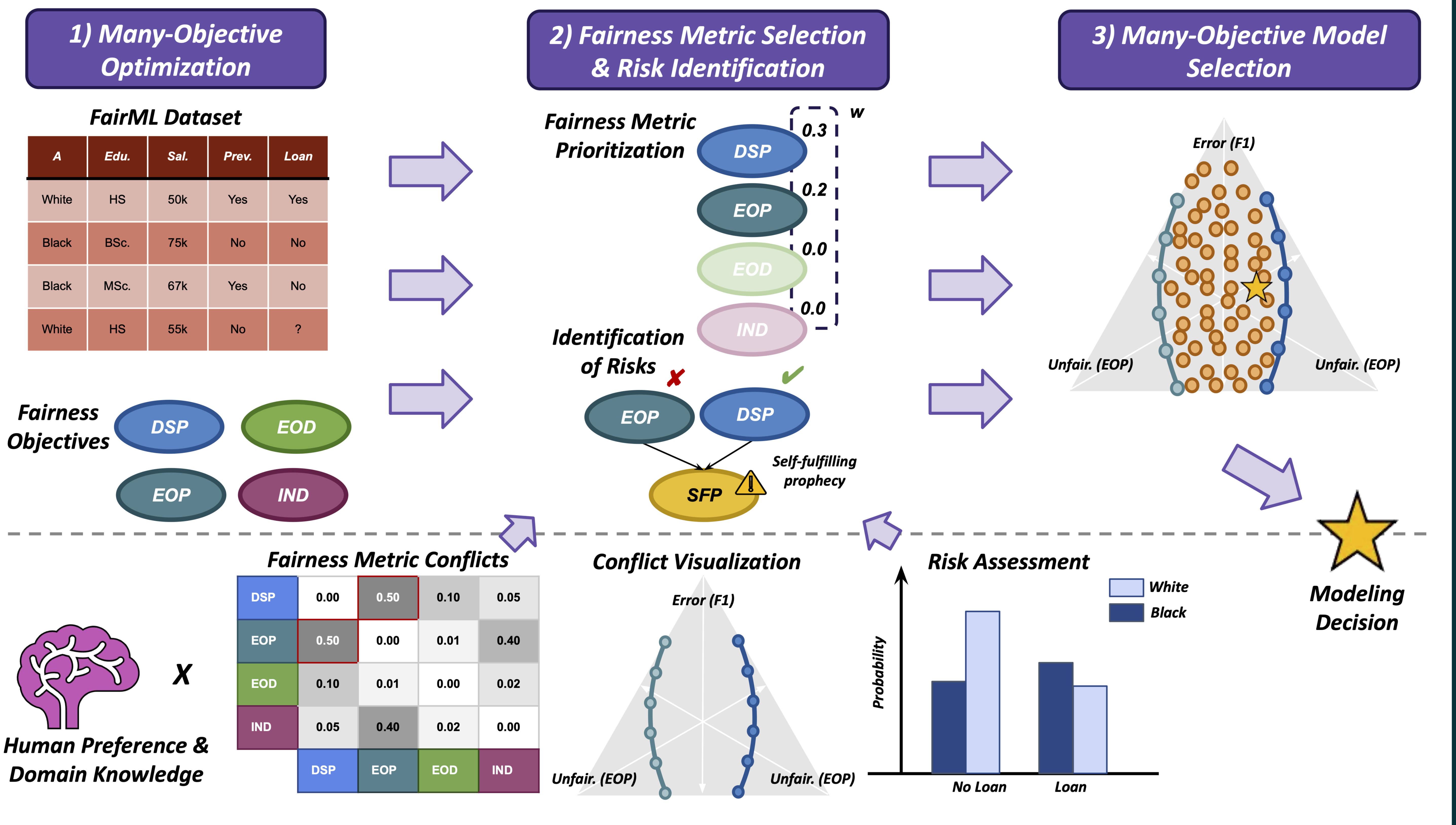
Fairness Metrics	*Social Objectives*
Statistical Parity (DSP)	*Equality*
Equal Opportunity (EOP)	*Equity*
Equalized Odds (EOD)	*Equity*
Inverse Distance (IND)	*Individual Justice*

		Violated Metric			
		DSP	EOP	EOD	IND
Satisfied Metric	DSP		*SFP*	*SFP*	*SFP*
	EOP	*?*		*?*	*?*
	EOD	*?*	*?*		*?*
	IND	*?*	*?*	*?*	

- Multi-objective Hyperparameter Optimization (MOHPO) involves adjusting typical ML design parameters (e.g. neural network structure) to approximate the Pareto Front of conflicting ML goals (e.g. accuracy and energy consumption)
- In fairness applications, MOHPO has been used to balance accuracy with a single, user-specified fairness criterion
- However, the established Impossibility Theorem shows that optimizing one fairness notion can unintentionally violate other relevant concepts
- This results in 1) a compromise between related social objectives and potentially 2) undesirable downstream effects (e.g. Self-Fulfilling Prophecy)

Many-Objective Fairness-Aware Hyperparameter Optimization (ManyFairHPO)

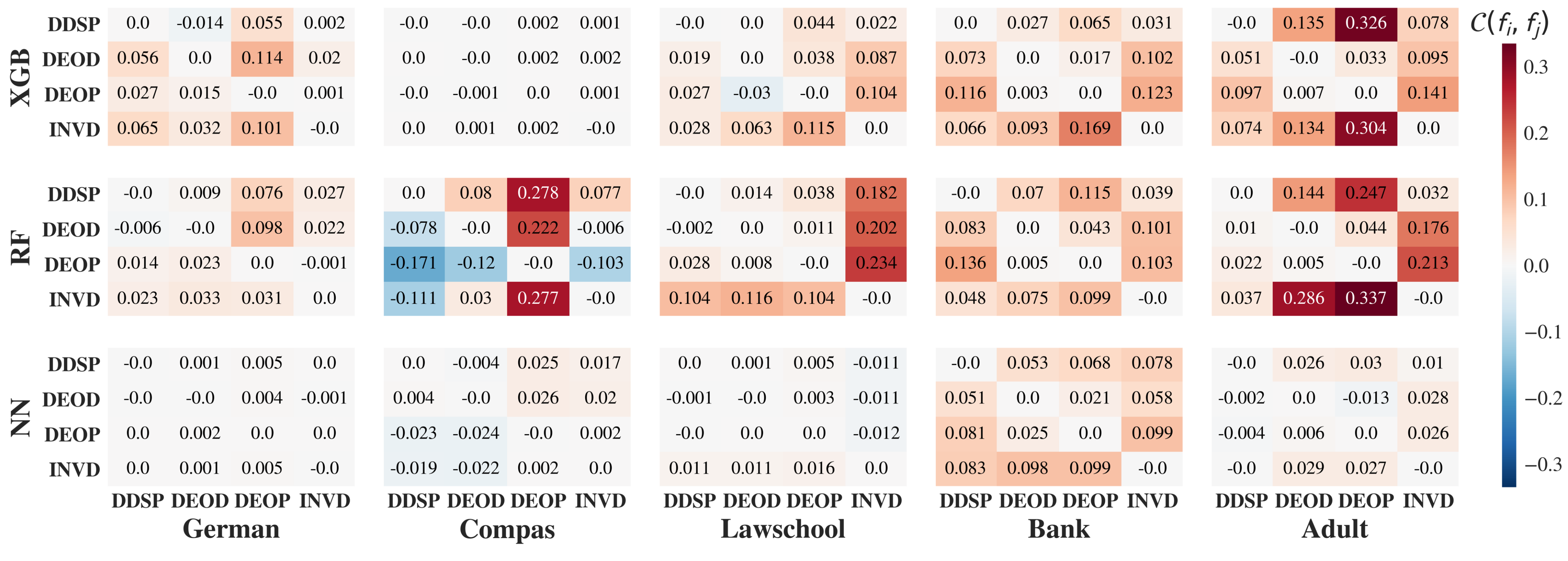
1) **Many-Objective Optimization for a set of fairness metrics**
   - ML dataset
   - fairness metrics
2) **Fairness Metric Selection and Risk Identification**
   - Human preferences and domain knowledge: i.e. *which set of fairness metrics are important for my task?*
   - Identified fairness metric conflicts and their associated risks: *which fairness metrics are in conflict and what might be the implications?*
3) **Many-Objective Model Selection**
   - Fairness metric weights are translated by single-objective scalarization into a single model selection decision


1) Many-Objective Optimization
2) Fairness Metric Selection & Risk Identification
3) Many-Objective Model Selection
FairML Dataset
Fairness Objectives
DSP
EOD
EOP
IND
Fairness Metric Prioritization
w
0.3
0.2
0.0
0.0
DSP
EOP
EOD
IND
Identification of Risks
EOP
DSP
SFP
Self-fulfilling prophecy
Error (F1)
Unfair. (EOP)
Unfair. (EOP)
Modeling Decision
Human Preference & Domain Knowledge
X
Fairness Metric Conflicts
Conflict Visualization
Error (F1)
Unfair. (EOP)
Unfair. (EOP)
Risk Assessment
White
Black
Probability
No Loan
Loan

A	*Edu.*	*Sal.*	*Prev.*	*Loan*
White	HS	50k	Yes	Yes
Black	BSc.	75k	No	No
Black	MSc.	67k	Yes	No
White	HS	55k	No	?

DSP	0.00	0.50	0.10	0.05
EOP	0.50	0.00	0.01	0.40
EOD	0.10	0.01	0.00	0.02
IND	0.05	0.40	0.02	0.00
	DSP	EOP	EOD	IND

Fairness Metric Conflicts (Fairness Datasets)



German

		DDSP	DEOD	DEOP	INVD
XGB	DDSP	0.0	-0.014	0.055	0.002
	DEOD	0.056	0.0	0.114	0.02
	DEOP	0.027	0.015	-0.0	0.001
	INVD	0.065	0.032	0.101	-0.0
RF	DDSP	-0.0	0.009	0.076	0.027
	DEOD	-0.006	-0.0	0.098	0.022
	DEOP	0.014	0.023	0.0	-0.001
	INVD	0.023	0.033	0.031	0.0
NN	DDSP	-0.0	0.001	0.005	0.0
	DEOD	-0.0	-0.0	0.004	-0.001
	DEOP	0.0	0.002	0.0	0.0
	INVD	0.0	0.001	0.005	-0.0

Compas

		DDSP	DEOD	DEOP	INVD
XGB	DDSP	-0.0	-0.0	0.002	0.001
	DEOD	0.0	-0.0	0.002	0.002
	DEOP	-0.0	-0.001	0.0	0.001
	INVD	0.0	0.001	0.002	-0.0
RF	DDSP	0.0	0.08	0.278	0.077
	DEOD	-0.078	-0.0	0.222	-0.006
	DEOP	-0.171	-0.12	-0.0	-0.103
	INVD	-0.111	0.03	0.277	-0.0
NN	DDSP	0.0	-0.004	0.025	0.017
	DEOD	0.004	-0.0	0.026	0.02
	DEOP	-0.023	-0.024	-0.0	0.002
	INVD	-0.019	-0.022	0.002	0.0

Lawschool

		DDSP	DEOD	DEOP	INVD
XGB	DDSP	-0.0	0.0	0.044	0.022
	DEOD	0.019	0.0	0.038	0.087
	DEOP	0.027	-0.03	-0.0	0.104
	INVD	0.028	0.063	0.115	0.0
RF	DDSP	-0.0	0.014	0.038	0.182
	DEOD	-0.002	0.0	0.011	0.202
	DEOP	0.028	0.008	-0.0	0.234
	INVD	0.104	0.116	0.104	-0.0
NN	DDSP	0.0	0.001	0.005	-0.011
	DEOD	-0.001	-0.0	0.003	-0.011
	DEOP	-0.0	0.0	0.0	-0.012
	INVD	0.011	0.011	0.016	0.0

Bank

		DDSP	DEOD	DEOP	INVD
XGB	DDSP	0.0	0.027	0.065	0.031
	DEOD	0.073	0.0	0.017	0.102
	DEOP	0.116	0.003	0.0	0.123
	INVD	0.066	0.093	0.169	0.0
RF	DDSP	-0.0	0.07	0.115	0.039
	DEOD	0.083	0.0	0.043	0.101
	DEOP	0.136	0.005	0.0	0.103
	INVD	0.048	0.075	0.099	-0.0
NN	DDSP	-0.0	0.053	0.068	0.078
	DEOD	0.051	0.0	0.021	0.058
	DEOP	0.081	0.025	0.0	0.099
	INVD	0.083	0.098	0.099	-0.0

Adult

		DDSP	DEOD	DEOP	INVD
XGB	DDSP	-0.0	0.135	0.326	0.078
	DEOD	0.051	-0.0	0.033	0.095
	DEOP	0.097	0.007	0.0	0.141
	INVD	0.074	0.134	0.304	0.0
RF	DDSP	0.0	0.144	0.247	0.032
	DEOD	0.01	-0.0	0.044	0.176
	DEOP	0.022	0.005	-0.0	0.213
	INVD	0.037	0.286	0.337	-0.0
NN	DDSP	-0.0	0.026	0.03	0.01
	DEOD	-0.002	0.0	-0.013	0.028
	DEOP	-0.004	0.006	0.0	0.026
	INVD	-0.0	0.029	0.027	-0.0

Figure 1. Fairness metric conflicts discovered by ManyFairHPO on common fairness datasets. Problem specific conflicts can guide practitioners in selecting and prioritizing fairness metrics and identifying and assessing fairness metric conflict related risks

Stakeholder Compromise

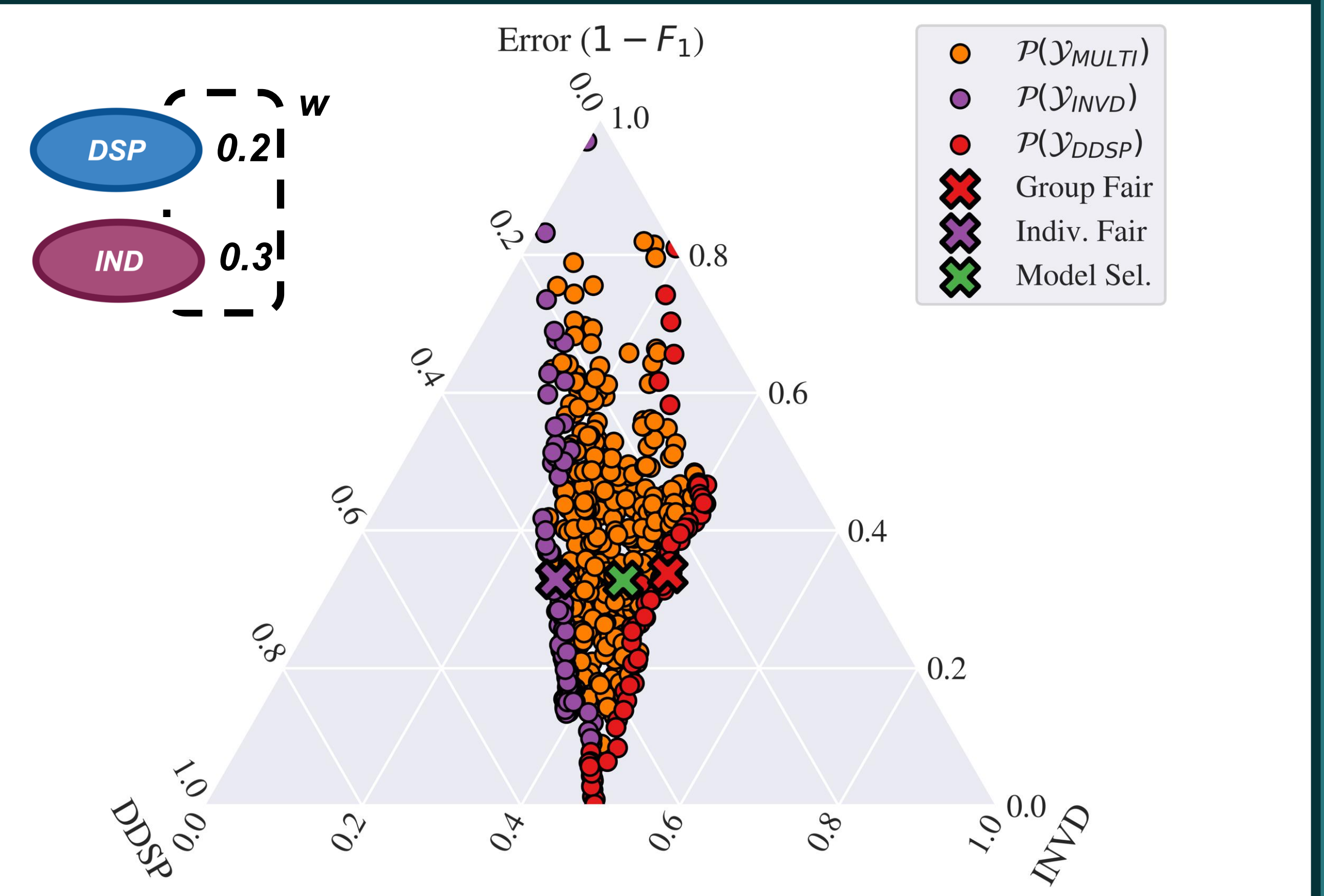


Figure 2. Given a set of fairness metric weights, ⟨0.2, 0.3⟩ for DDSP and INVD on a , ManyFairHPO selects a model (green) that balances this conflict