Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification

Maliha Tashfia Islam
mtislam@cs.umass.edu
University of Massachusetts Amherst

Anna Fariha
annafariha@microsoft.com
Microsoft

Alexandra Meliou
ameli@cs.umass.edu
University of Massachusetts Amherst

Babak Salimi
bsalimi@ucsd.edu
University of California, San Diego

Classifiers can exhibit discriminatory behavior

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Historical biases in training data often cause **discrimination**!

Two new challenges in fair classification

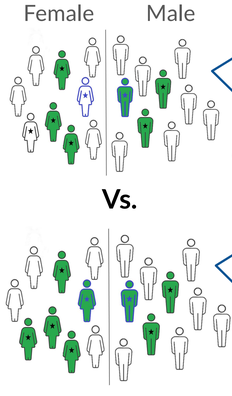
Classification algorithms can be modified to behave fairly, but present **two challenges**:

- ❑ How to choose an appropriate **definition of fairness**?
- ❑ Which **technique** to use to incorporate fairness into a classifier?

👉 **Our work addresses both challenges!**

Fairness is subjective and application dependent

★: similar individuals
green: good outcome



Group fairness ensures equal gender representation, but similar individuals may not be treated similarly

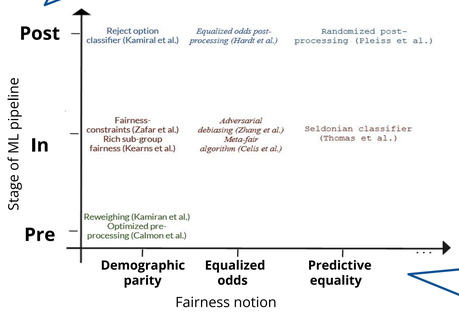
Individual fairness treats all similar individuals similarly, but no fairness guarantee for groups

👉 Informative categorization of 34 fairness notions

	Fairness notion	Granularity: group: demography-aware	Granularity: group: error-aware	Granularity: individual	Causal hierarchy: observation	Causal hierarchy: intervention	Causal hierarchy: counterfactual	Additional requirements: prediction probability	Additional requirements: causality model	Additional requirements: resolving attribute	Additional requirements: similarity metric
non-causal	conditional statistical parity [24]	✓			✓						
non-causal	demographic parity† [26]	✓			✓						
non-causal	intersectional fairness [32]	✓			✓						
non-causal	overall accuracy equality [10]		✓		✓						
non-causal	treatment equality [10]		✓		✓						
non-causal	equalized odds [41]		✓		✓						
non-causal	equal opportunity‡ [41]		✓		✓						
non-causal	resilience to random bias [30]		✓		✓						
non-causal	preference-based fairness [96]		✓		✓						
non-causal	calibration [23]		✓		✓			✓			
non-causal	calibration within groups [53]		✓		✓			✓			
non-causal	positive class balance [53]		✓		✓			✓			
non-causal	negative class balance [53]		✓		✓			✓			
non-causal	individual discrimination†† [34]			✓	✓						
non-causal	metric multifairness [52]			✓	✓						✓
causal	proxy fairness [50]	✓				✓			✓		
causal	total causal effect [72]	✓				✓			✓		
causal	path-specific fairness [105]	✓				✓			✓		
causal	unresolved discrimination [50]	✓				✓				✓	
causal	interventional/justifiable fairness [80]	✓				✓				✓	
causal	fair on average causal effect [49]	✓				✓			✓		
causal	non-discrimination criterion [104]	✓				✓			✓		
causal	equality of effort [42]	✓				✓			✓		
causal	individual direct discrimination [103]			✓		✓			✓		

Fair classifiers vary across two dimensions

Where to apply fairness notion?



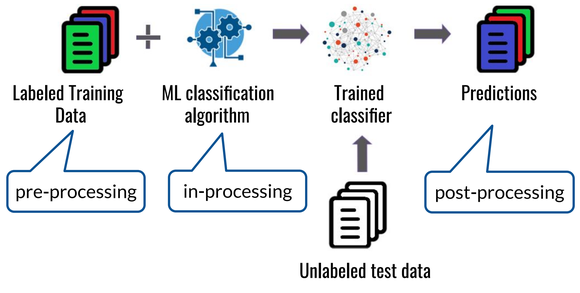
Which fairness notion?

Three stages of ML pipeline to enforce fairness

Pre-processing modifies data ***before*** training

In-processing constrains classifier objective ***during*** training

Post- processing modifies prediction ***after*** training



👉 Empirical investigation of fair classifiers

Approaches: 18 fair techniques

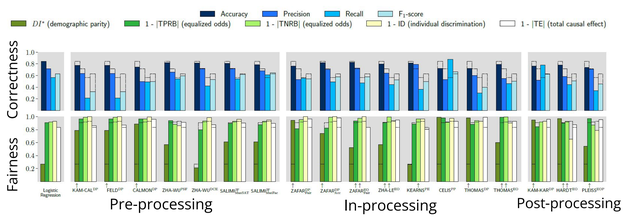
Datasets: **3** real-world datasets

Biases: gender and **racial**

Evaluation axes: **4** areas
- ☛ Correctness- fairness **tradeoff**
- ☛ **Scalability** issues
- ☛ Effect of training **data errors**
- ☛ Sensitivity to the choice of **ML model**

Correctness-fairness tradeoff

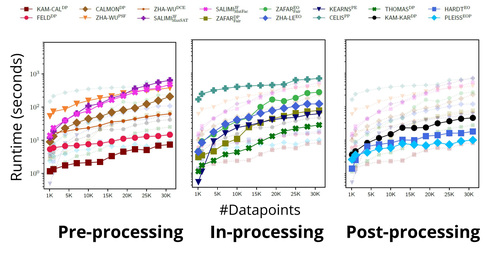
Fair approaches **trade** accuracy for fairness



Bigger compromise in accuracy when target fairness is **low** in fairness-unaware setting

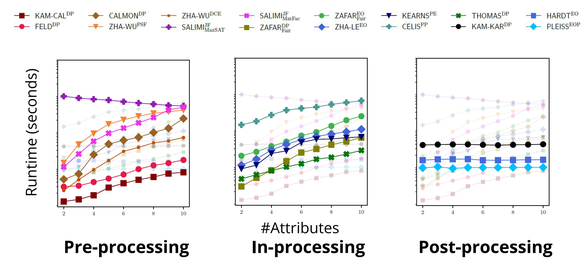
Efficiency gap

Post- processing is most **efficient** due to inherent simplicity



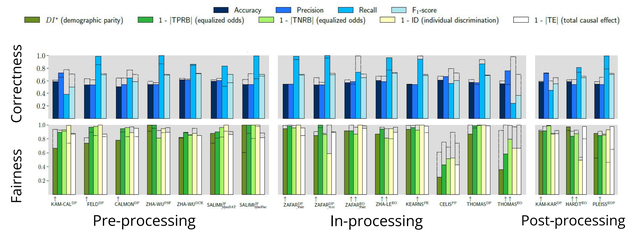
Complex mechanisms can **lower efficiency**: common in **pre-** and **in-**processing

Scalability issues



Pre-processing **scales worse** with increasing #attributes than with increasing #datapoints

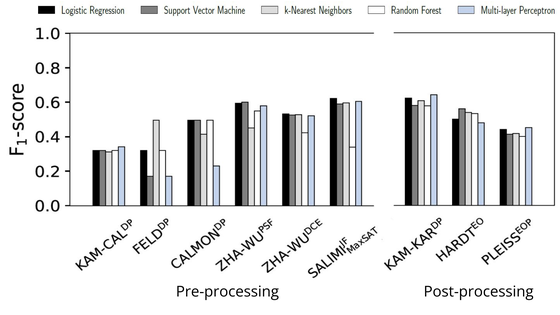
What happens if training data has errors?



Pre- and **in-**processing can **fail** to build fair models when training data contains **errors**

Sensitivity to the choice of ML model

Post-processing is more **stable**



Pre-processing can **fluctuate** depending on classifier model