Human evaluation of learned causal graph

Hi! Would you mind taking 30 minutes to complete this form? It would be great if you can submit your response by 5/10/2023. Thank you!

Introduction

To conduct a comprehensive study on the trade-offs among model fairness, performance, and robustness, we use causal discovery techniques to construct causal graphs from data, which can facilitate the understanding of the intricate relationships among various kinds of metrics. In this evaluation, several subgraph of the learned causal graph will be presented. Then, you are requested to mark any edges on the subgraph that you disagree with and note any edges not present in the subgraph that should be included.

Nodes in the causal graph

Method Type:

1. **Pre-processing** methods manipulate the training data;
2. **In-processing** methods modify the model during training;
3. **Post-processing** methods modify the trained model's predictions.

Most methods can be found
in https://aif360.readthedocs.io/en/stable/modules/algorithms.html

Fairness Type

1. **Individual fairness** requires that all individuals be treated equally.
2. **Group fairness** is concerned with the fairness of the given model's predictions for different groups, where the groups are defined by one or more sensitive attributes.

Most fairness metrics can be found
in https://aif360.readthedocs.io/en/stable/modules/metrics.html

Other Metrics:
Except for the Causal Discrimination Score (CDS), all fairness metrics are computed by the widely-used AIF360. We implement the CDS metric by ourselves. For robustness, we evaluate the model from two perspectives: the adversarial attack and the membership inference attack. For the adversarial attack, we employ the Fast Gradient Sign Method (FGSM) and the Projected Gradient (PGD). Here, the success rate is defined as the proportion of the number of adversarial examples that are capable of fooling the model to the total number of adversarial examples. In our experiments, Torchattacks implementation is used. For the membership inference attack, we use two simple but effective methods: the rule-based method and the black-box method, for the sake of practicality. The rule-based method assumes that a sample is a member if the model correctly predicts its label; otherwise, the sample is the non-member. The black-box method trains a model to predict whether a sample is a member or not. Both methods are implemented by ART here.

Metrics				
Data	Data_Cons	Data_sens_DP	Data_sens_SPD	Data_other_DP
Train_sensitive	Train_sens_DP	Train_sens_SPD	Train_sens_AOD	
Train_other	Train_other_DP	Train_other_SPD	Train_other_AOD	
Test_sensitive	Test_sens_DP	Test_sens_SPD	Test_sens_AOD	
Test_other	Test_other_DP	Test_other_SPD	Test_other_AOD	
Individual fair	Train_Cons	Train_TI	Train_CDS	Test_Cons
Performance	Train_Acc	Train_Loss	Train_F1	Test_Acc
Robustness	AE_FGSM	AE_PGD	MI_BlackBox	MI_Rule
Pre-processing	FairMask	Fairway	FairSmote	LTDD
	DIR	RW		
In-processing	AdDebias	EGR	PR	
Post-processing	EO	CEO	ROC	

1

Metrics categories

Category	Name
Performance (2)	Accuracy (Acc) [37]
	F1 score (F1) [37]
Group Fairness (3)	Disparate Impact (DI) [37]
	Statistical Parity Difference (SPD) [37]
	Average Odds Difference (AOD) [37]
Individual Fairness (3)	Consistency (Cons) [38]
	Theil Index (TI) [38]
	Causal Discrimination Score (CDS) [13]
Robustness (4)	FGSM's Success Rate (FGSM) [39]
	PGD's Success Rate (PGD) [40]
	Rule-Based Membership Inference's Accuracy (Rule) [41]
	Black-Box Membership Inference's Accuracy (Bbox) [41]

2

Fairness methods

Category	Name
Pre-processing (6)	Reweighing [1]
	Disparate Impact Remover (DIR) [4]
	FairWay [8]
	FairSmote [9]
	FairMask [10]
	LTDD [11]
In-processing (3)	Adversarial Debiasing (AD) [6]
	Prejudice Remover (PR) [2]
	Exponentiated Gradient Reduction (EGR) [7]
Post-processing (3)	Reject Option Classification (ROC) [3]
	Equalized Odds (EO) [35]
	Calibrated Equalized Odds (CEO) [5]

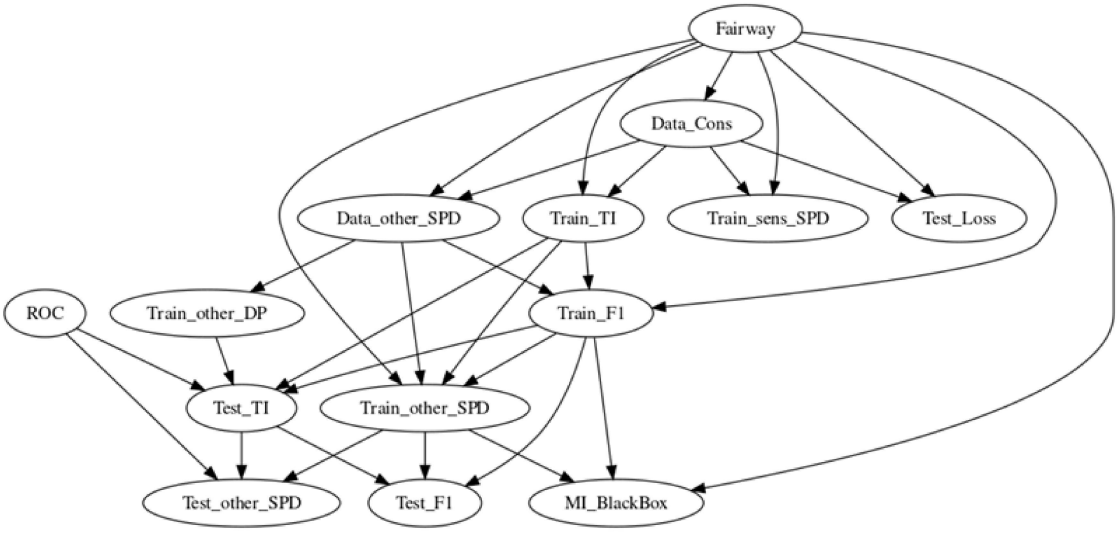
Dataset

Our experiments are conducted on three real-world datasets: Adult, COMPAS, and German. Note that they all
have two sensitive attributes, and we set the target sensitive attribute of all aforementioned fairness-improving methods to one attribute. Then, there are six scenarios in total (e.g., Adult_sex, Adult_race).

TABLE I: Dataset Information.

Dataset	Size	Favorable Class	Sensitive Attribute	Privileged Group
Adult [31]	48,842×12	income>50K	sex race	sex=Male race=White
COMPAS [32]	7,214×11	no recidivism	sex race	sex=Female race=Caucasian
German [33]	1,000×21	good credit	sex age	sex=male age>30

Adult_sex Subgraph 1



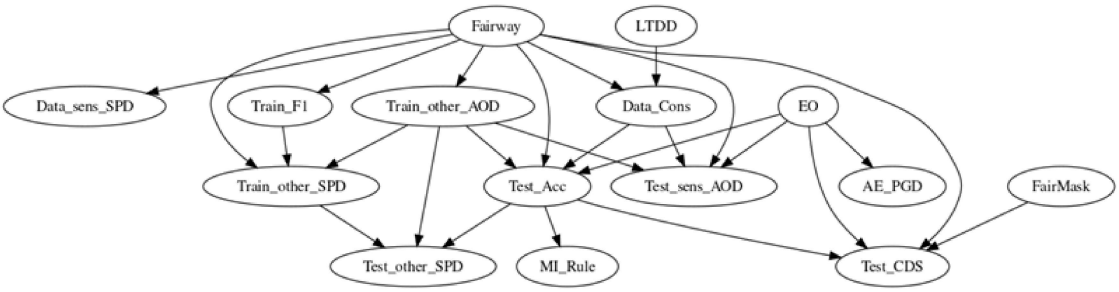
3

Error edges

4

Missed edges

Adult_sex Subgraph 2



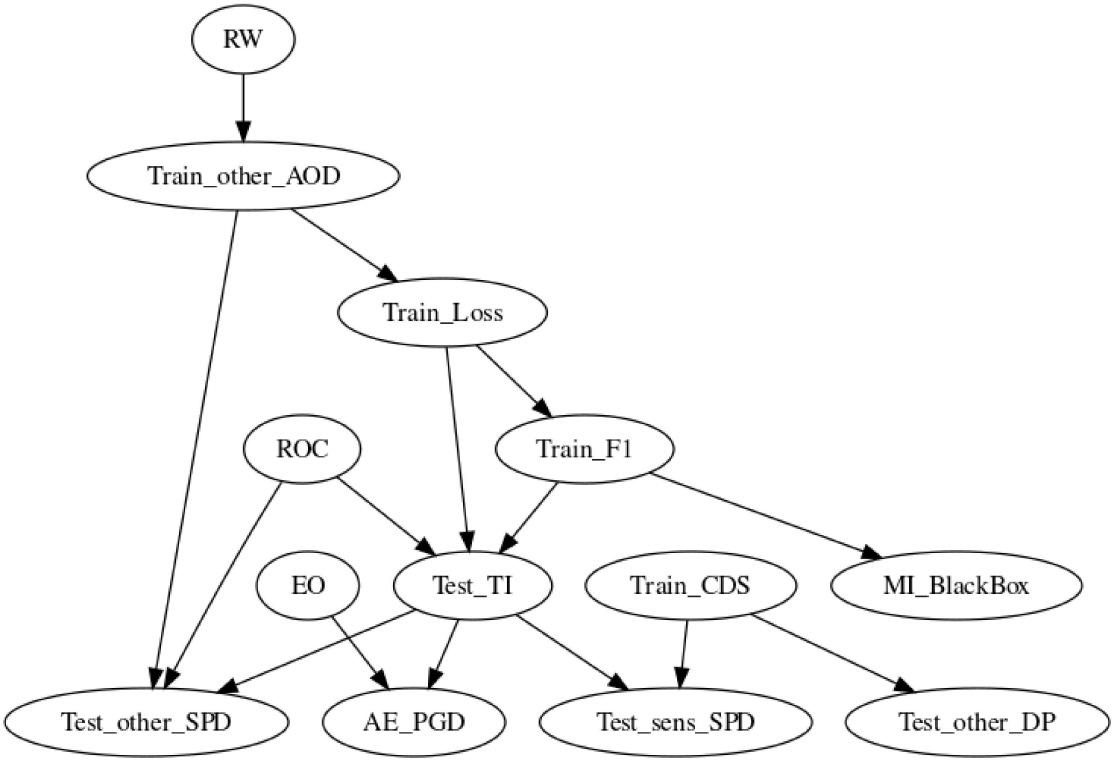
5

Error edges

6

Missed Edges

Adult_sex Subgraph 3



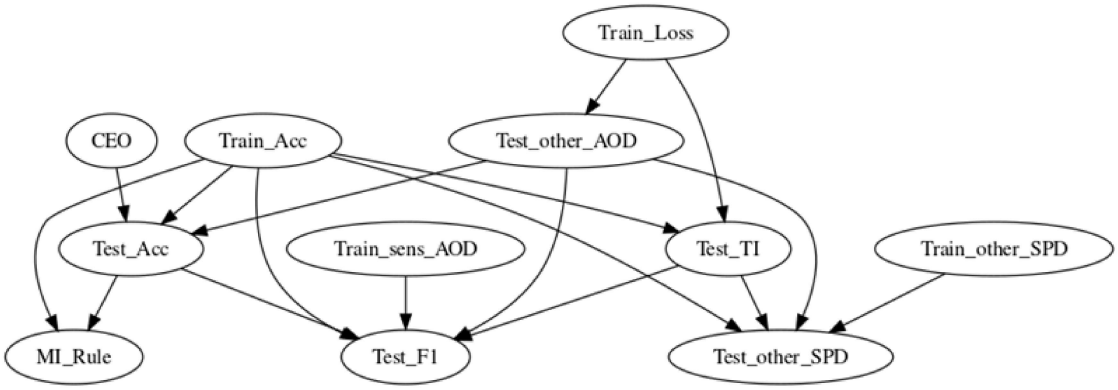
7

Error edges

8

Missed edges

Adult_race Subgraph 1



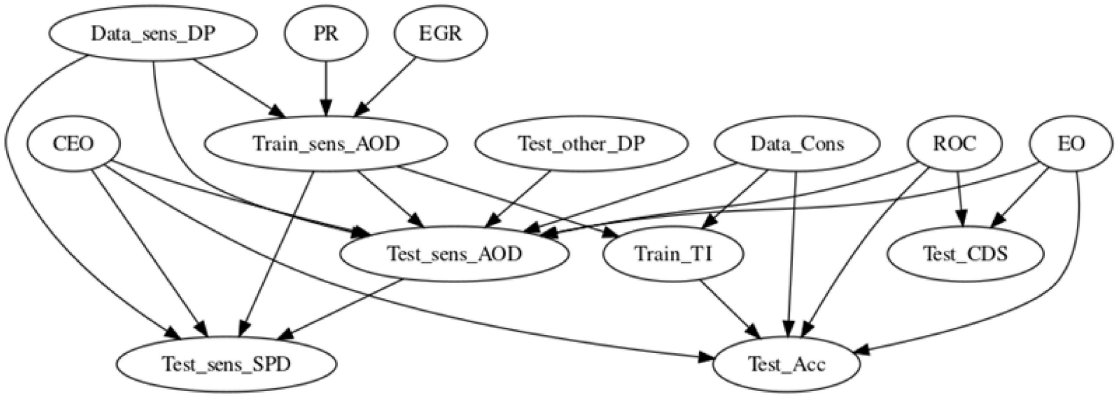
9

Error edges

10

Missed Edges

Adult_race, Subgraph 2



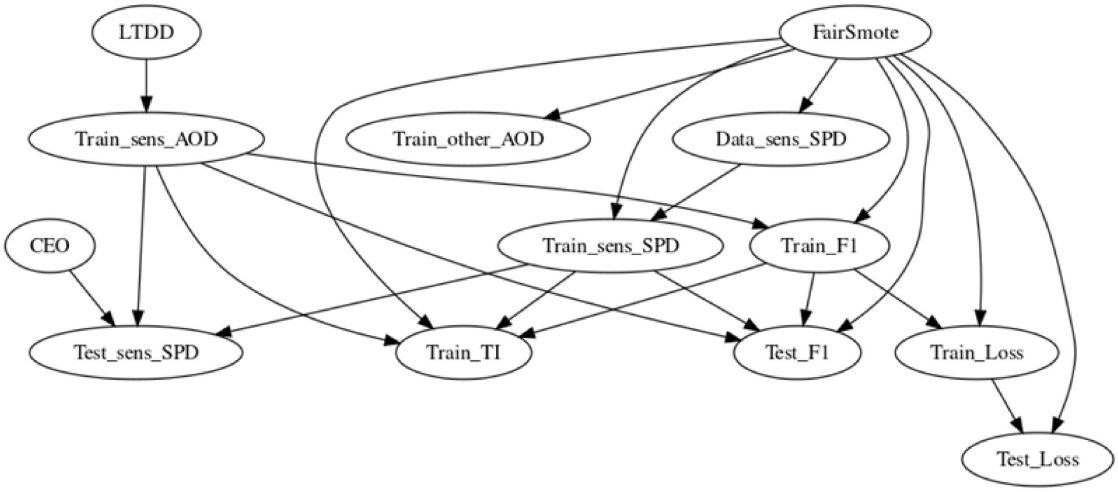
11

Error edges

12

Missed edges

Adult_race Subgraph 3



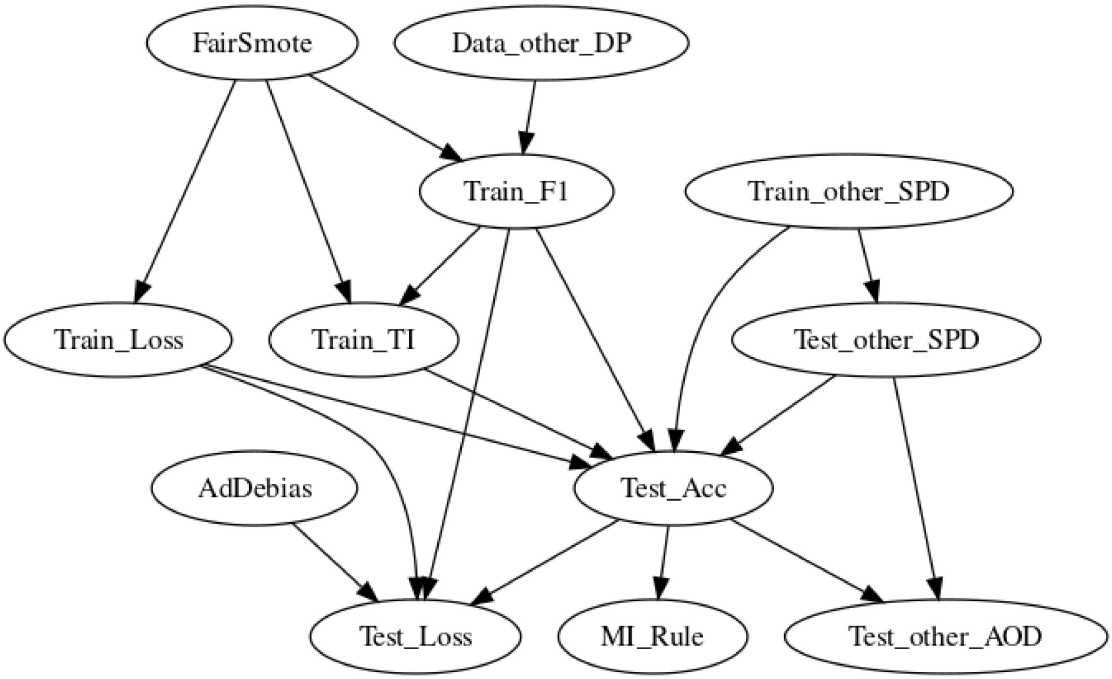
13

Error edges

14

Missed Edges

Compas_sex Subgraph 1



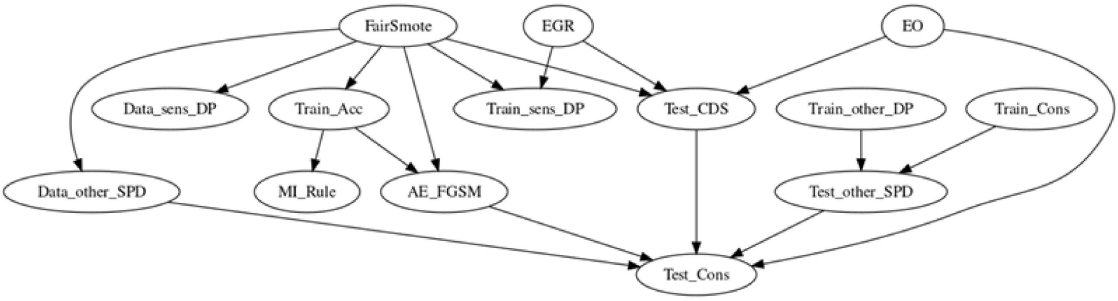
15

Error edges

16

Missed edges

Compas_sex Subgraph 2



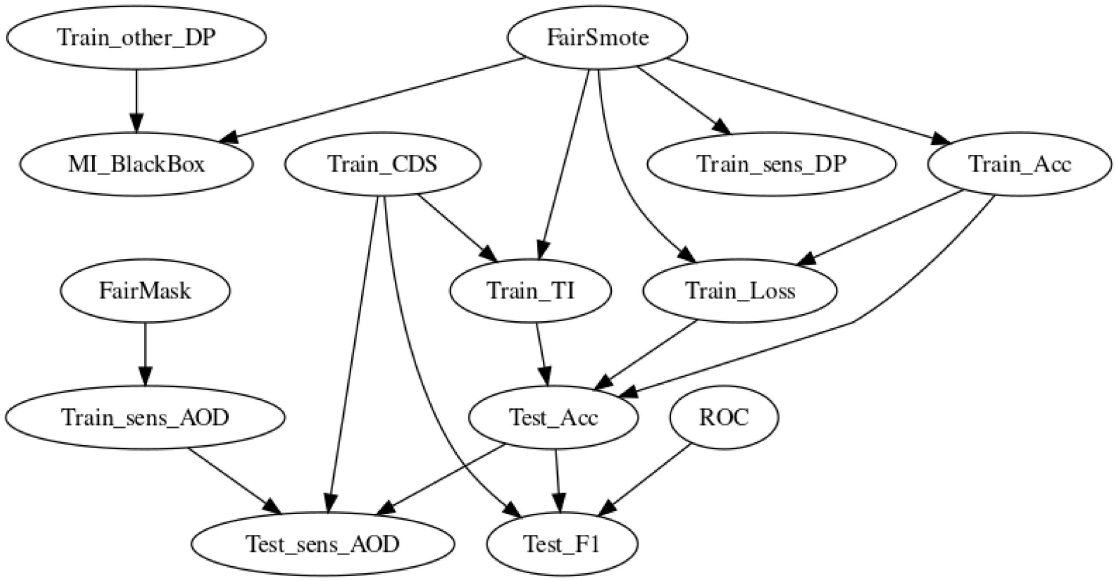
17

Error edges

18

Missed edges

Compas_sex Subgraph 3



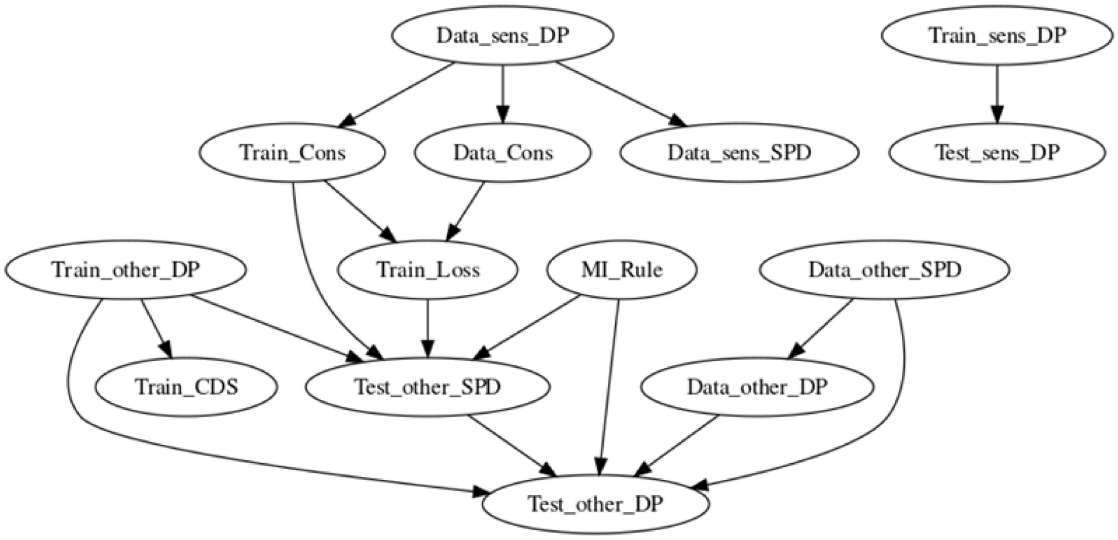
19

Error edges

20

Missed edges

Compas_race Subgraph 1



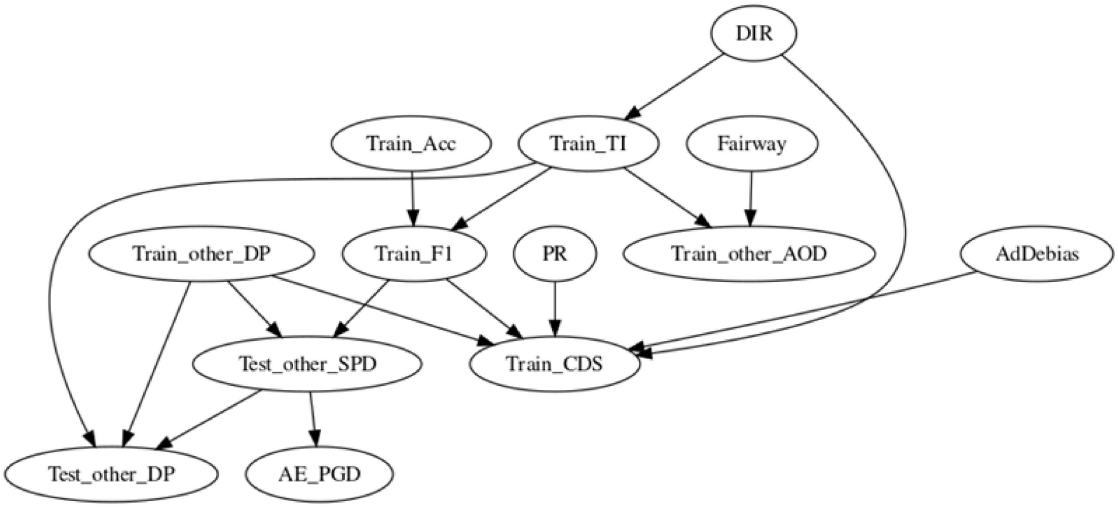
21

Error edges

22

Missed edges

Compas_race Subgraph 2



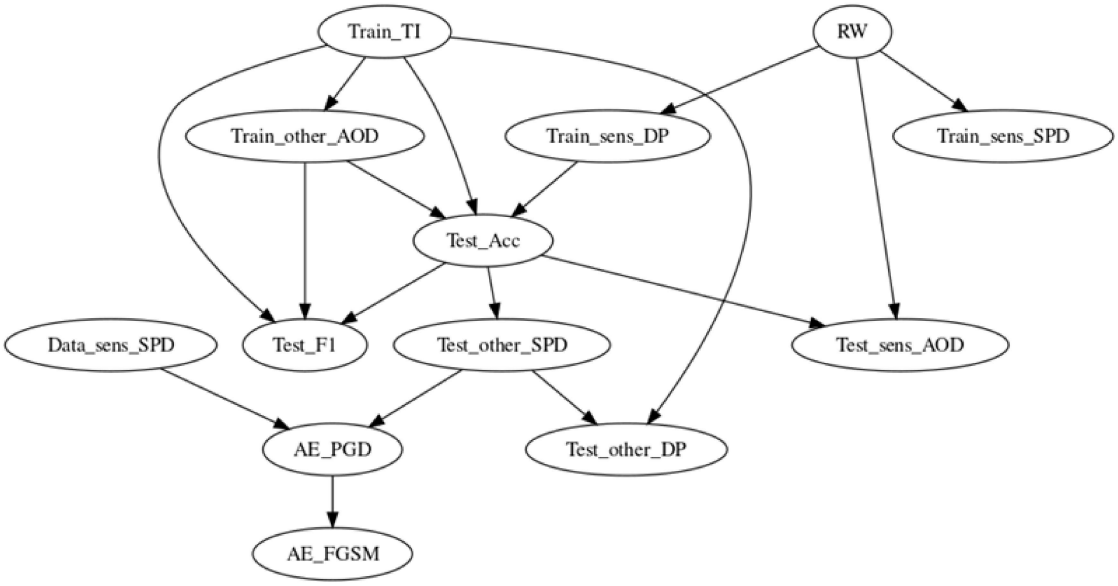
23

Error edges

24

Missed edges

Compas_race Subgraph 3



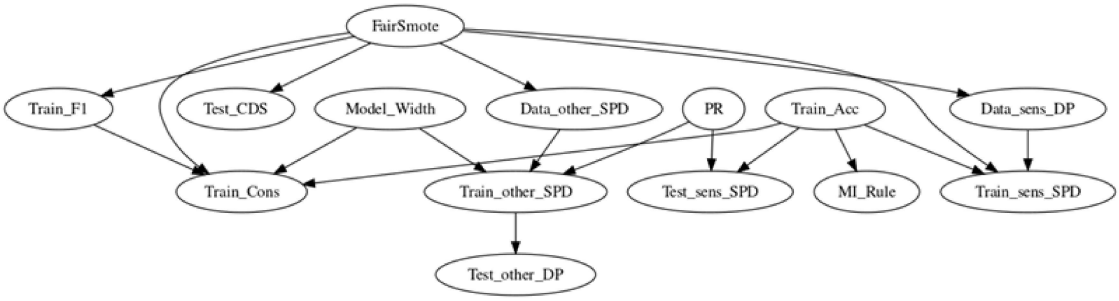
25

Error edges

26

Missed edges

German_sex Subgraph 1



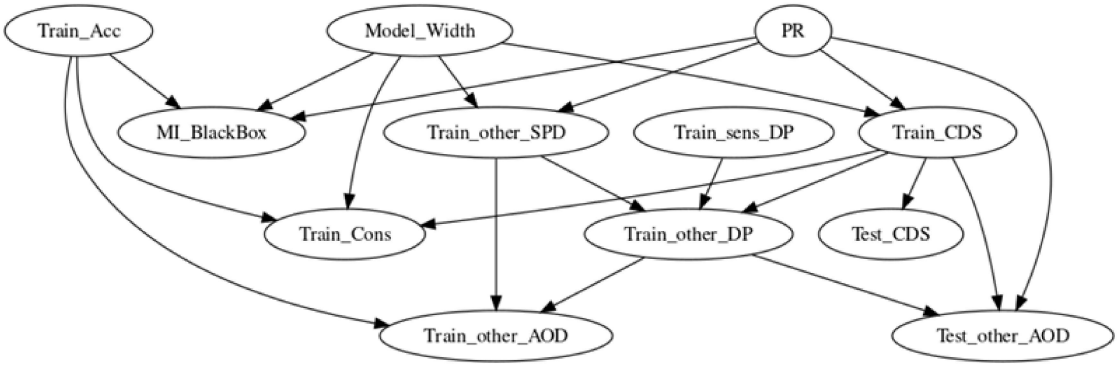
27

Error edges

28

Missed edges

German_sex Subgraph 2



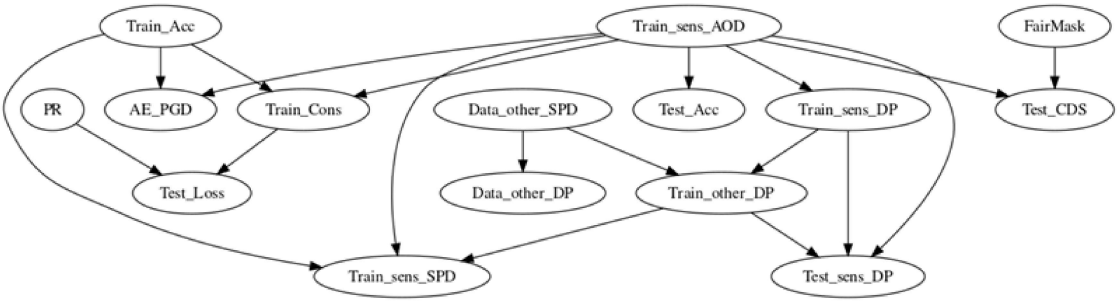
29

Error edges

30

Missed edges

German_sex Subgraph 3



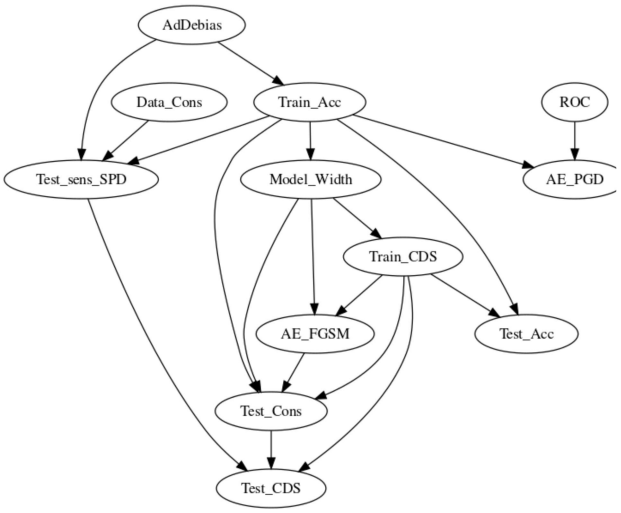
31

Error edges

32

Missed edges

German_age Subgraph 1



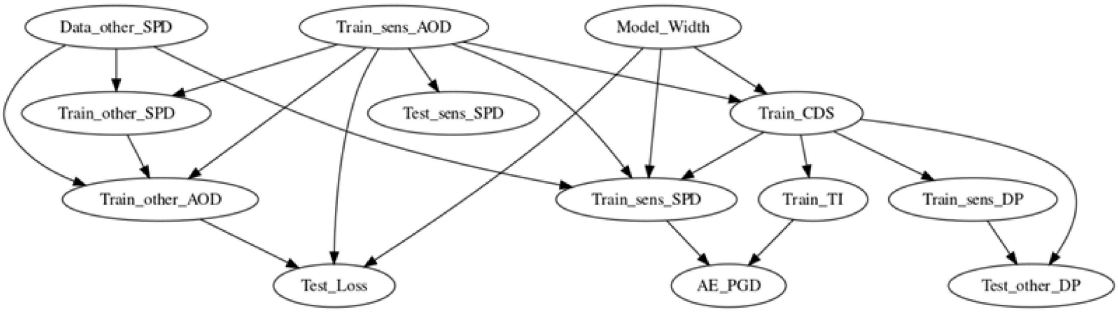
33

Error edges

34

Missed edges

German_age Subgraph 2



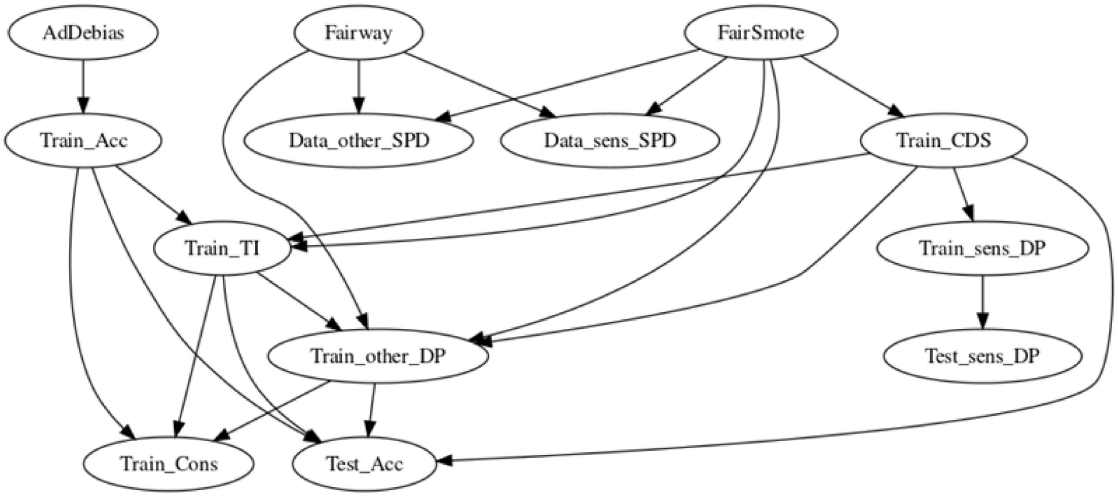
35

Error edges

36

Missed edges

German_age Subgraph 3



37

Error edges

38

Missed edges