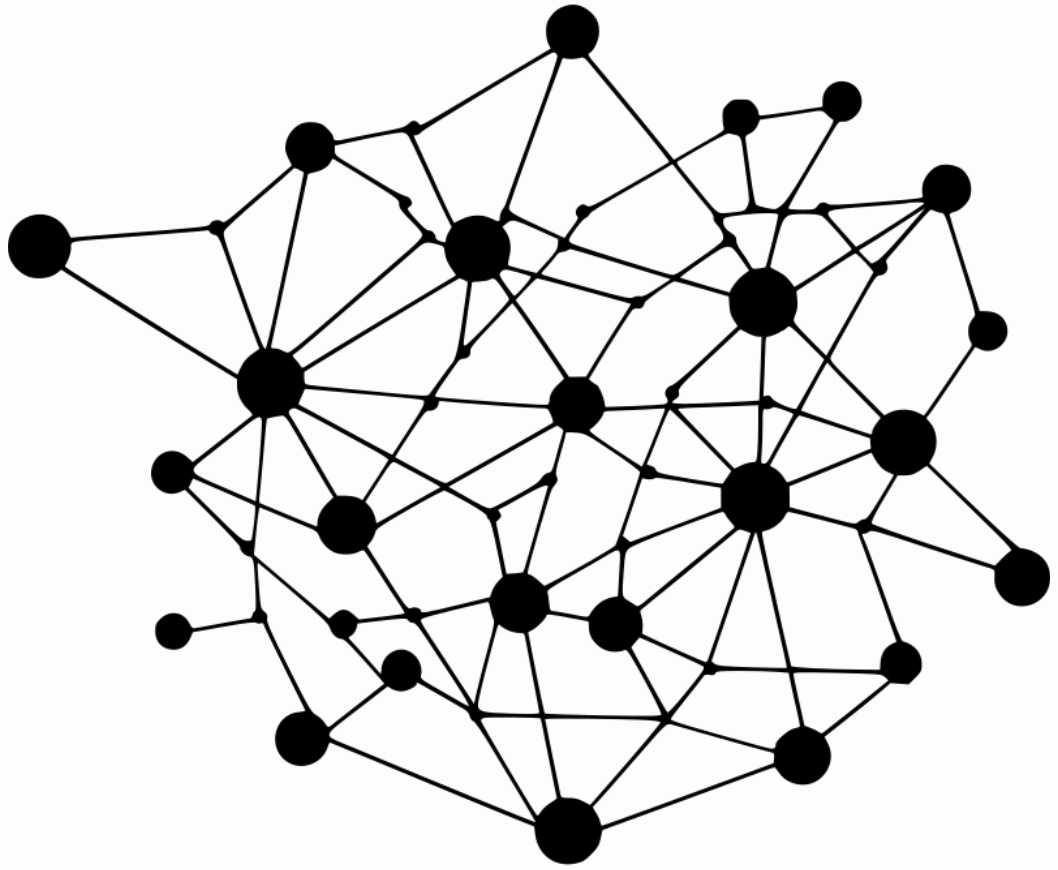
ALGORITHMIC FAIRNESS IN X-RAY DIAGNOSIS ⚖️

Elle McFarlane, Andrea Matamoros Alonso, Alicia García Sanz, Jesús Díaz Pereira

ARCHITECTURE & DATA INFO

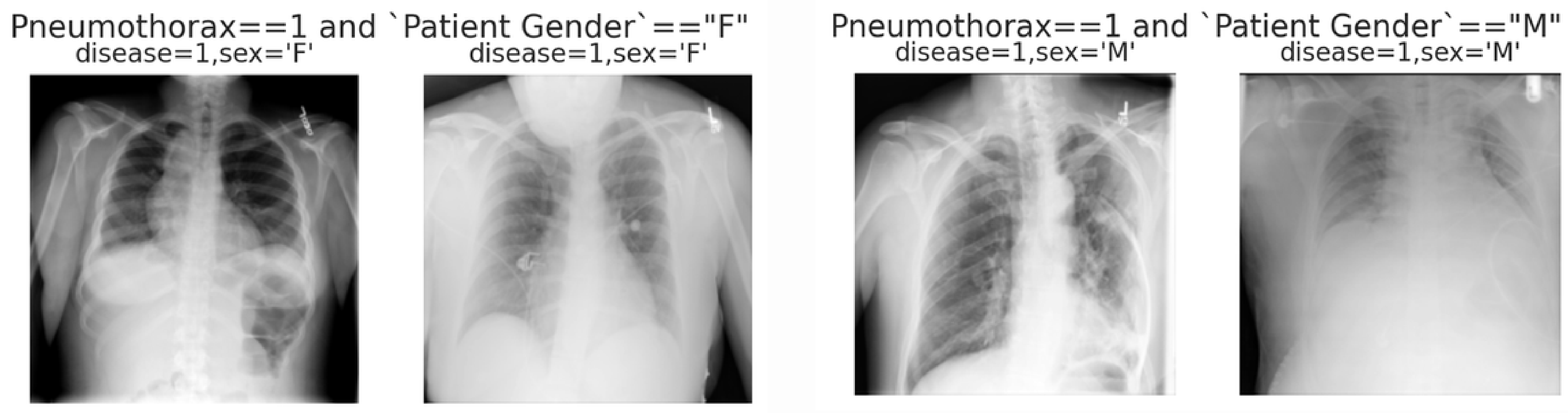


2 **X-ray images**:

Image size: 224x224

Dataset **Validation**: 2k images

Dataset **Test**: 8k images



1 **Model:**

Already pre-trained ResNet model

Deep residual network with skip connections

44 layers

1 input channel (grayscale)

Binary output

3 **Data splits nomenclature**

Training set - train the model

Validation set - train the threshold optimizer

Test set - test the mitigated model

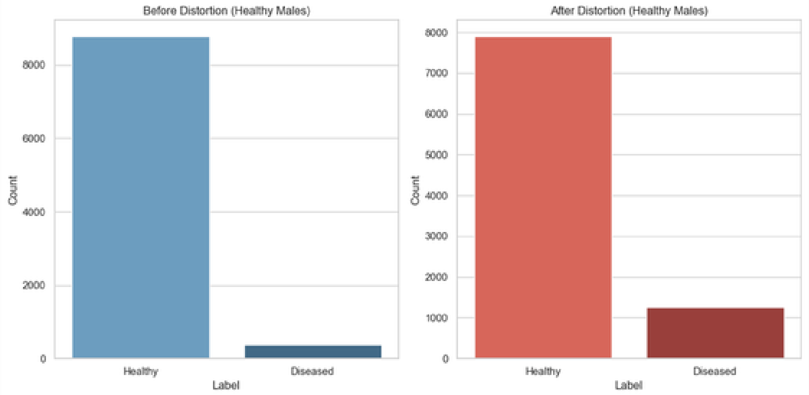
METHOD

1 Modify fairlearn library (& model.py)

Threshold optimizer fit/predict: add batch_size and device options for speed/memory efficiency #1416

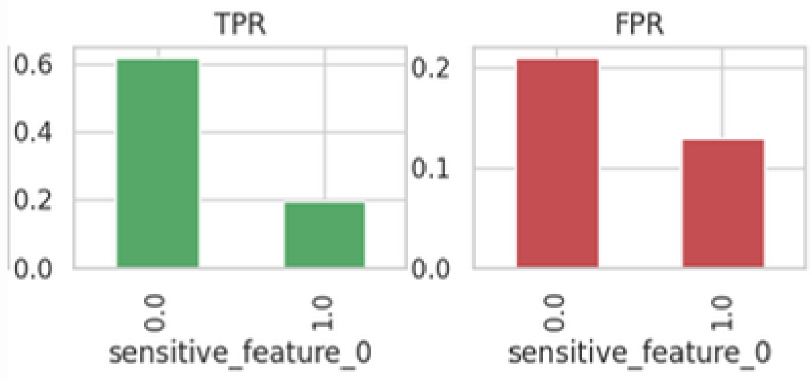
Open ellemcfarlane wants to merge 3 commits into fairlearn:main from ellemcfarlane:batch_thresh_op

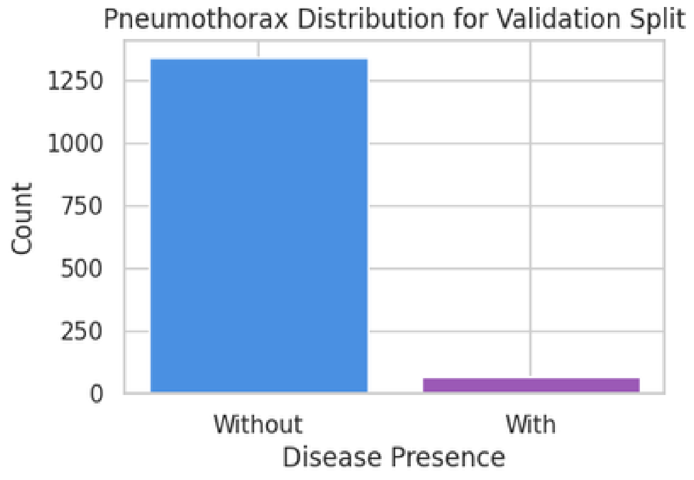
2 flip 10% of males without pneumothorax



3 diagnose bias (& choose one to mitigate)

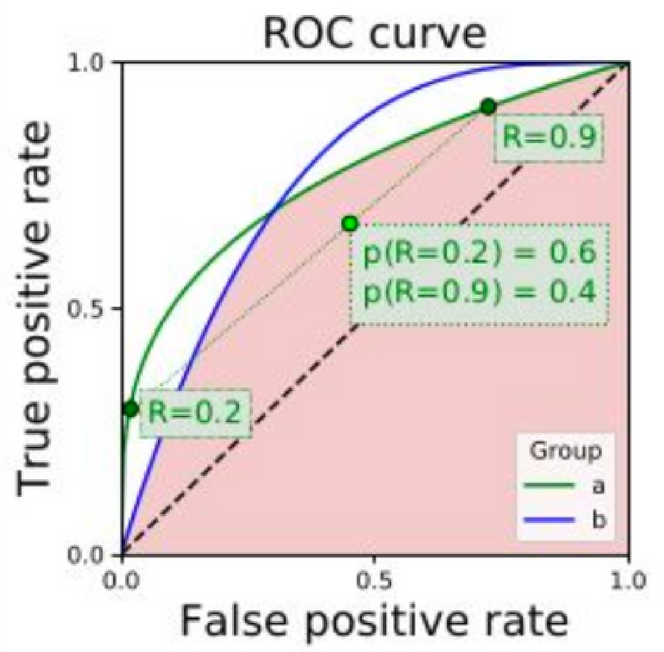
4 Post-process mitigation with threshold optimizer





optimize equalized-odds

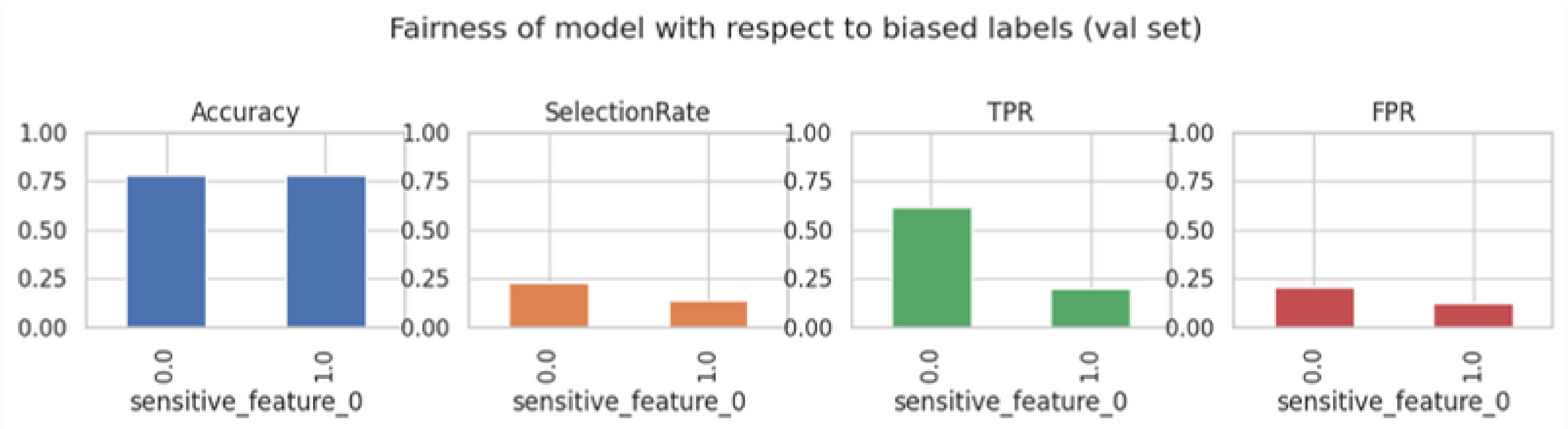
subject to balanced-accuracy



5 Problems and solutions

- Accuracy vs balanced accuracy
- Hard labels vs “predict_proba”

BIASED LABELS CAN AFFECT FAIRNESS DIAGNOSIS



VALIDATION SET

In true labels (left) there is no 20% discrepancy in TPR -> with biased lables (right) there is a false TPR discrepancy

Key takeaway: biased labels can lead to false diagnosis of unfairness

BIASED LABELS CAN AFFECT FAIRNESS EVALUATION

TEST SET

True equalized odds ratio before and after mitigation: **.62 -> .71**
Biased equalized odds ratio before and after mitigation: **.43 -> .89**

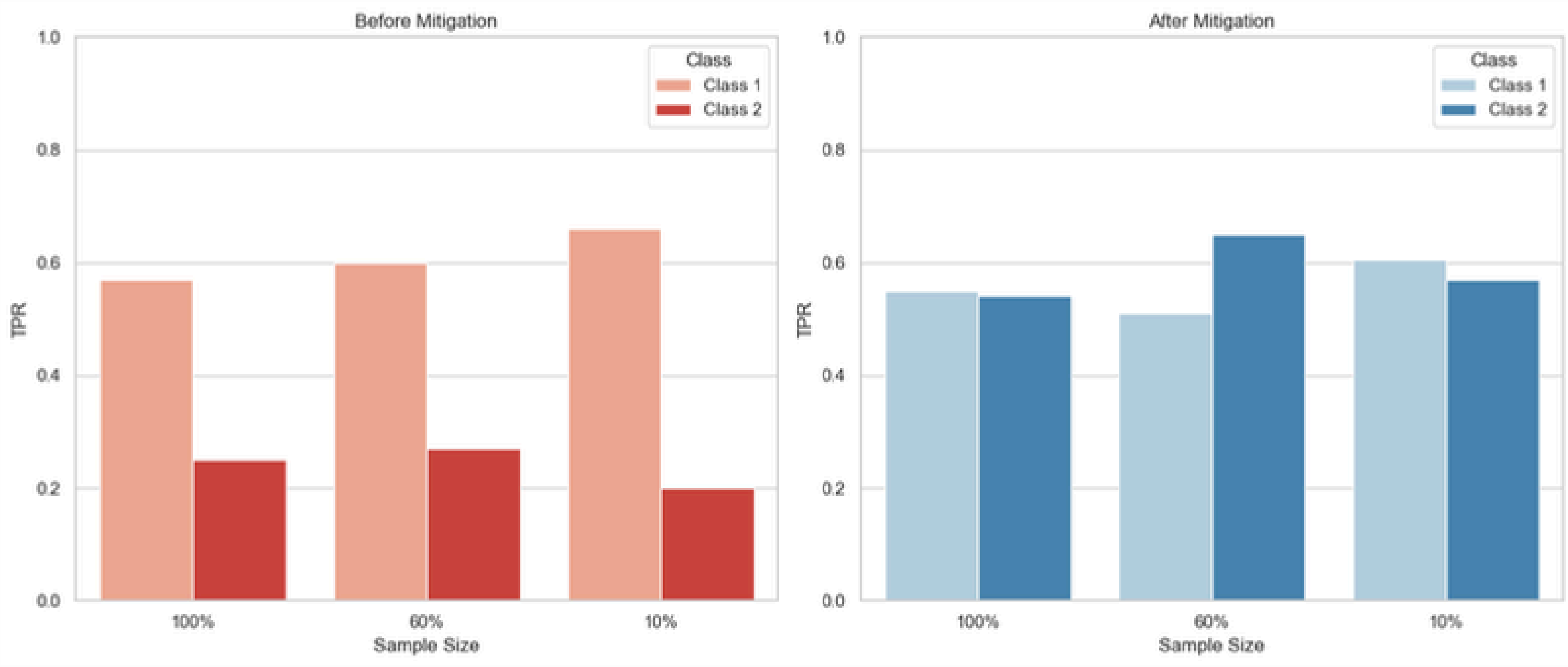
Key takeaway: evaluation can look fair, but not be in reality

EFFECTS OF SAMPLE SIZE

We explored how different sample sizes impact the final results.

Motivation: Small sample sizes may lead to suboptimal thresholds and less reliable outcomes.

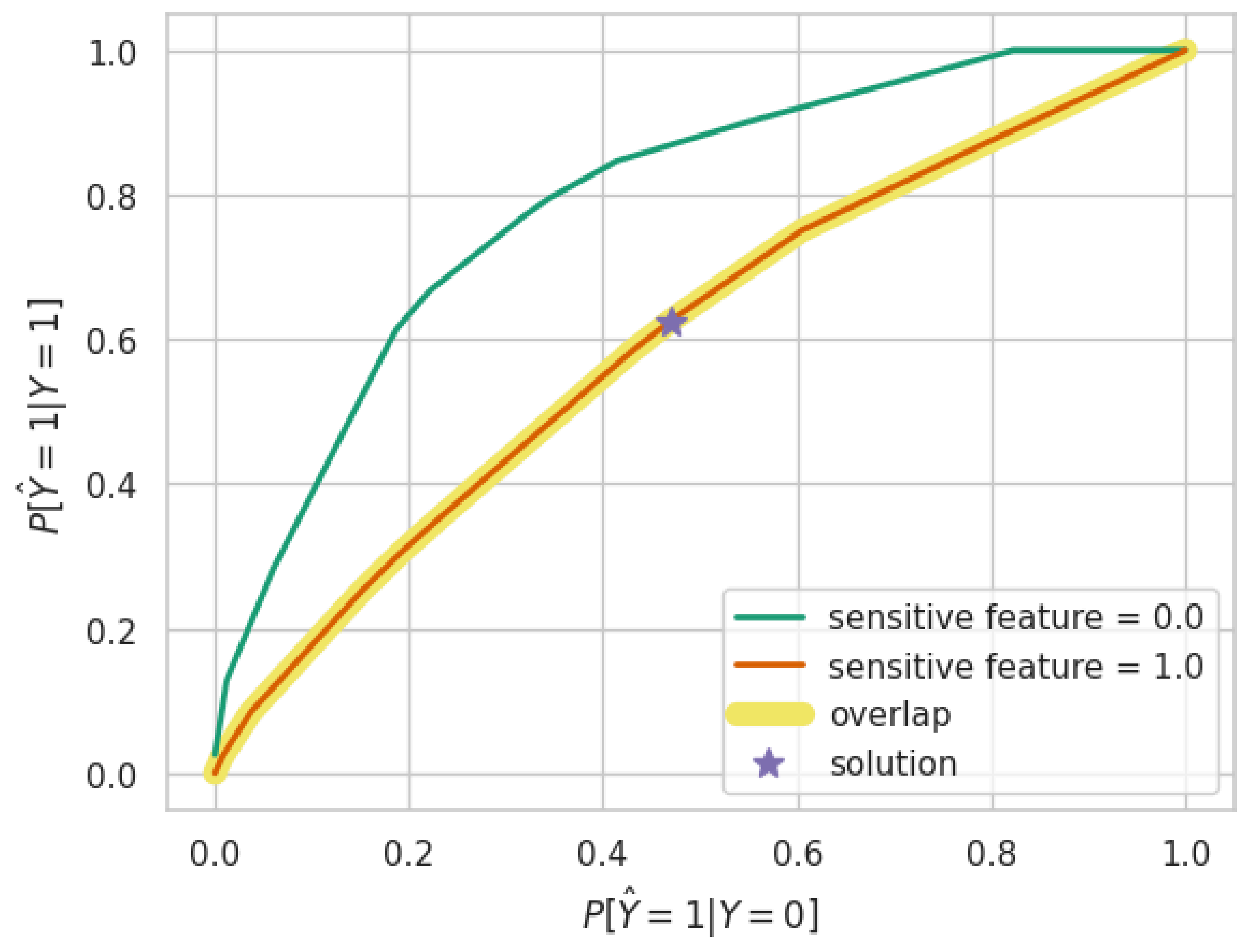
10%, 60% and 100% of the Validation Data were used to fit the TO for each of the cases



This indicates that mitigation is less effective when we **subsample the dataset, disparaties in the metrics being optimized for each class increase**

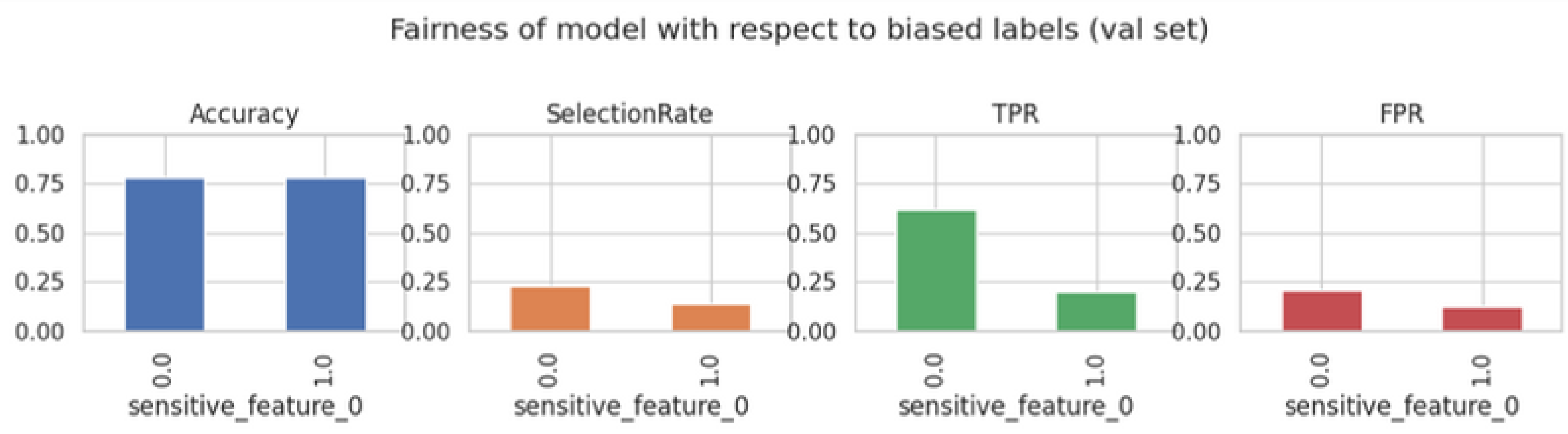
SUPPLEMENTARY SLIDES

ROC CURVE




```
{
    "0.0": {
        "p_ignore": 0.6093801911899192,
        "prediction_constant": 0.46900000000000003,
        "p0": 0.5553095238095237,
        "operation0": "[>0.1940901130437851]",
        "p1": 0.4446904761904763,
        "operation1": "[>0.1109326183795929]"
    },
    "1.0": {
        "p_ignore": 0.0,
        "prediction_constant": 0.46900000000000003,
        "p0": 0.9981463414634144,
        "operation0": "[>0.06031285971403122]",
        "p1": 0.0018536585365855895,
        "operation1": "[>0.04268590360879898]"
    }
}
```

FULL VALIDATION MITGIATION RESULTS



FULL VALIDATION DIAGNOSIS NUMBERS

```
BIASED VAL
#### BEFORE VAL mitigation ####
                     Accuracy  SelectionRate       TPR       FPR
sensitive_feature_0
0.0                   0.78156       0.231206  0.615385  0.208709
1.0                   0.77983       0.137784  0.197917  0.128289

TRUE VAL
#### BEFORE VAL mitigation ####
                     Accuracy  SelectionRate       TPR       FPR
sensitive_feature_0
0.0                  0.781560       0.231206  0.615385  0.208709
1.0                  0.860795       0.137784  0.482759  0.122963
```