UNIVERSITY OF AMSTERDAM

[Re] Exacerbating Algorithmic Bias through Fairness Attacks

Matteo Tafuro*, Andrea Lombardo*, Tin Hadži Veljković, Lasse Becker-Czarnetzki

Date: December 2022 | NeurIPS 2022

*Today's presenters

Outline

Outline

Motivation

ML Reproducibility Challenge:

Encourage the publishing and sharing of scientific results that are reliable and reproducible.

Reproducibility study:

Verify the empirical results and claims in the paper by reproducing the computational experiments

Motivation

ML Reproducibility Challenge:

Encourage the publishing and sharing of scientific results that are reliable and reproducible.

Reproducibility study:

Verify the empirical results and claims in the paper by reproducing the computational experiments

Reproducibility study

N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan,
"Exacerbating Algorithmic Bias through Fairness Attacks"
AAAI, vol. 35, no. 10, pp. 8930-8938, May 2021

Exacerbating Algorithmic Bias through Fairness Attacks

Ninareh Mehrabi[1,2], Muhammad Naveed[1], Fred Morstatter[1,2], Aram Galstyan[1,2]
[1]University of Southern California - [2]Information Sciences Institute
{ninarehm, mnaveed}@usc.edu, {fredmors, galstyan}@isi.edu

Abstract

Algorithmic fairness has attracted significant attention in recent years, with many quantitative measures suggested for characterizing the fairness of different machine learning algorithms. Despite this interest, the robustness of those fairness measures with respect to an intentional adversarial attack has

appear unfair in order to depreciate their value and credibility. Some adversaries can even profit from such attacks by biasing decisions for their benefit, e.g., in credit or loan applications. Thus, one should consider fairness when assessing the robustness of ML systems.

Our contributions. In this work, we propose data poison-

Outline

Introduction

Two families of ***poisoning attacks*** that inject malicious points into the models' training sets and intentionally target the ***fairness*** of a classification model.

Influence Attack on Fairness (IAF)

$$L_{adv}(\hat{\theta}; \mathcal{D}_{test}) = \ell_{acc} + \lambda \ell_{fairness}$$

(An attacker can hence harm both accuracy and fairness simultaneously)

Anchoring Attack

Introduction

Two families of ***poisoning attacks*** that inject malicious points into the models' training sets and intentionally target the ***fairness*** of a classification model.

Influence Attack on Fairness (IAF)

$$L_{adv}(\hat{\theta}; \mathcal{D}_{test}) = \ell_{acc} + \lambda \ell_{fairness}$$

(An attacker can hence harm both accuracy and fairness simultaneously)

Anchoring Attack

Introduction

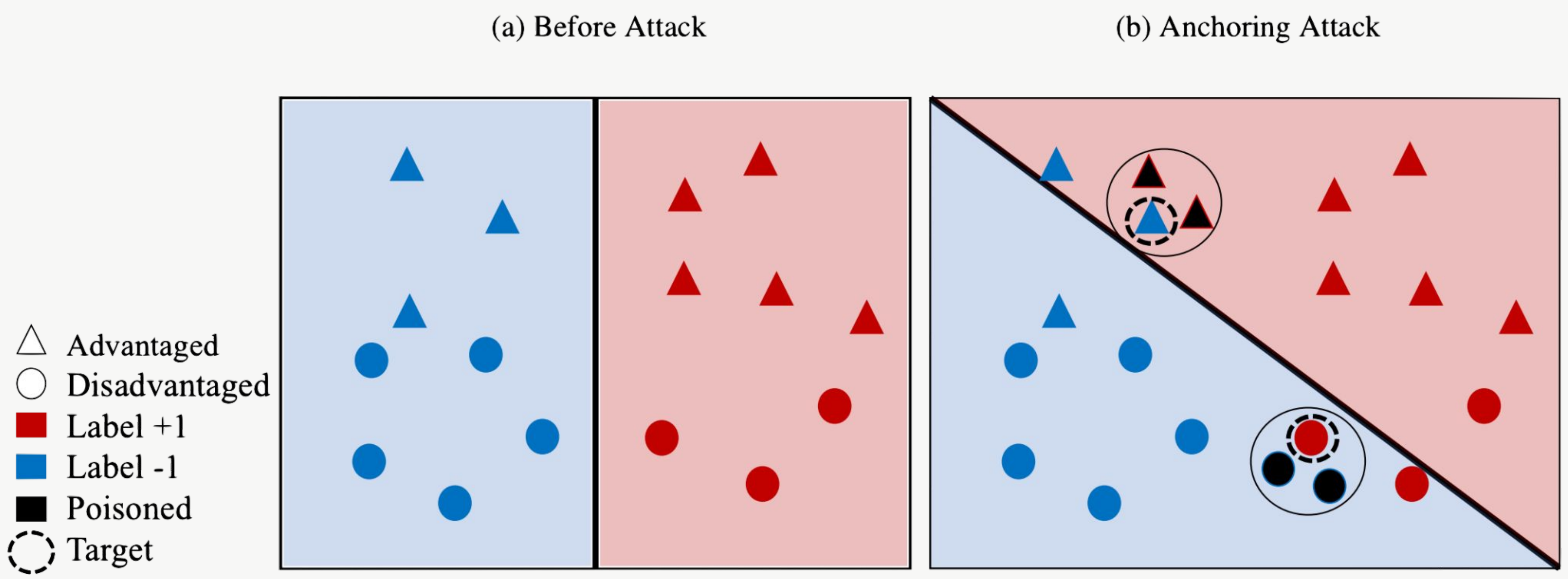
Two families of ***poisoning attacks*** that inject malicious points into the models' training sets and intentionally target the ***fairness*** of a classification model.

Influence Attack on Fairness (IAF)

$$L_{adv}(\hat{\theta}; \mathcal{D}_{test}) = \ell_{acc} + \lambda \ell_{fairness}$$

(An attacker can hence harm both accuracy and fairness simultaneously)

Anchoring Attack



Scope of reproducibility

Claim 1 Increasing the parameter λ results in stronger attacks against fairness.

Claim 2 The proposed IAF outperforms the attack of Koh et al. *[1]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 3 The proposed IAF outperforms the attack of Solans et al. *[2]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 4 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. *[1]* in degrading the SPD and EOD of the classification model, on all three datasets.

Claim 5 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. *[2]* in degrading the SPD and EOD of the classification model, on all three datasets.

[1] Stronger data poisoning attacks break data sanitization defenses
Koh et al, Mach Learn 111, 1–47 (2022)

[2] Poisoning Attacks on Algorithmic Fairness
Solans et al, ECML PKDD 2020

Scope of reproducibility

Claim 1 Increasing the parameter λ results in stronger attacks against fairness.

Claim 2 The proposed IAF outperforms the attack of Koh et al. *[1]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 3 The proposed IAF outperforms the attack of Solans et al. *[2]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 4 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. *[1]* in degrading the SPD and EOD of the classification model, on all three datasets.

Claim 5 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. *[2]* in degrading the SPD and EOD of the classification model, on all three datasets.

[1] Stronger data poisoning attacks break data sanitization defenses
Koh et al, Mach Learn 111, 1–47 (2022)

[2] Poisoning Attacks on Algorithmic Fairness
Solans et al, ECML PKDD 2020

Scope of reproducibility

Claim 1 Increasing the parameter λ results in stronger attacks against fairness.

Claim 2 The proposed IAF outperforms the attack of Koh et al. *[1]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 3 The proposed IAF outperforms the attack of Solans et al. *[2]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 4 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. *[1]* in degrading the SPD and EOD of the classification model, on all three datasets.

Claim 5 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. *[2]* in degrading the SPD and EOD of the classification model, on all three datasets.

[1] Stronger data poisoning attacks break data sanitization defenses
Koh et al, Mach Learn 111, 1–47 (2022)

[2] Poisoning Attacks on Algorithmic Fairness
Solans et al, ECML PKDD 2020

Scope of reproducibility

Claim 1 Increasing the parameter λ results in stronger attacks against fairness.

Claim 2 The proposed IAF outperforms the attack of Koh et al. *[1]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 3 The proposed IAF outperforms the attack of Solans et al. *[2]* in affecting both fairness metrics (SPD and EOD), on all three datasets.

Claim 4 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. *[1]* in degrading the SPD and EOD of the classification model, on all three datasets.

Claim 5 Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. *[2]* in degrading the SPD and EOD of the classification model, on all three datasets.

[1] Stronger data poisoning attacks break data sanitization defenses
Koh et al, Mach Learn 111, 1–47 (2022)

[2] Poisoning Attacks on Algorithmic Fairness
Solans et al, ECML PKDD 2020

Outline

- Motivation
- Introduction
- Methodology
- Results
- Discussion

Methodology

- **2 baselines:** Koh et al., Solans et al.
- **3 datasets:** German, COMPAS, Drug Consumption

Setup

- **Existing code implementation:** Missing parts
- **Model description:** SVM with SH loss, L2 regularization
- **Fairness metrics:** SPD and EOD

Methodology

- **2 baselines:** Koh et al., Solans et al.
- **3 datasets:** German, COMPAS, Drug Consumption

Setup

- **Existing code implementation:** Missing parts
- **Model description:** SVM with SH loss, L2 regularization
- **Fairness metrics:** SPD and EOD

Methodology

- **2 baselines:** Koh et al., Solans et al.
- **3 datasets:** German, COMPAS, Drug Consumption

Setup

- **Existing code implementation:** Missing parts
- **Model description:** SVM with SH loss, L2 regularization
- **Fairness metrics:** SPD and EOD

Methodology

- **2 baselines:** Koh et al., Solans et al.
- **3 datasets:** German, COMPAS, Drug Consumption

Setup

- **Existing code implementation:** Missing parts
- **Model description:** SVM with SH loss, L2 regularization
- **Fairness metrics:** SPD and EOD

Methodology

- **2 baselines:** Koh et al., Solans et al.
- **3 datasets:** German, COMPAS, Drug Consumption

Setup

- **Existing code implementation:** Missing parts
- **Model description:** SVM with SH loss, L2 regularization
- **Fairness metrics:** SPD and EOD

Outline

Results

- Effect of λ on different metrics
- Comparison between novel attacks and the baselines
- Effects of different stopping metrics (beyond original paper)

Effects of λ on different metrics

- ***Claim 1:*** Larger values of λ results in stronger attacks against fairness

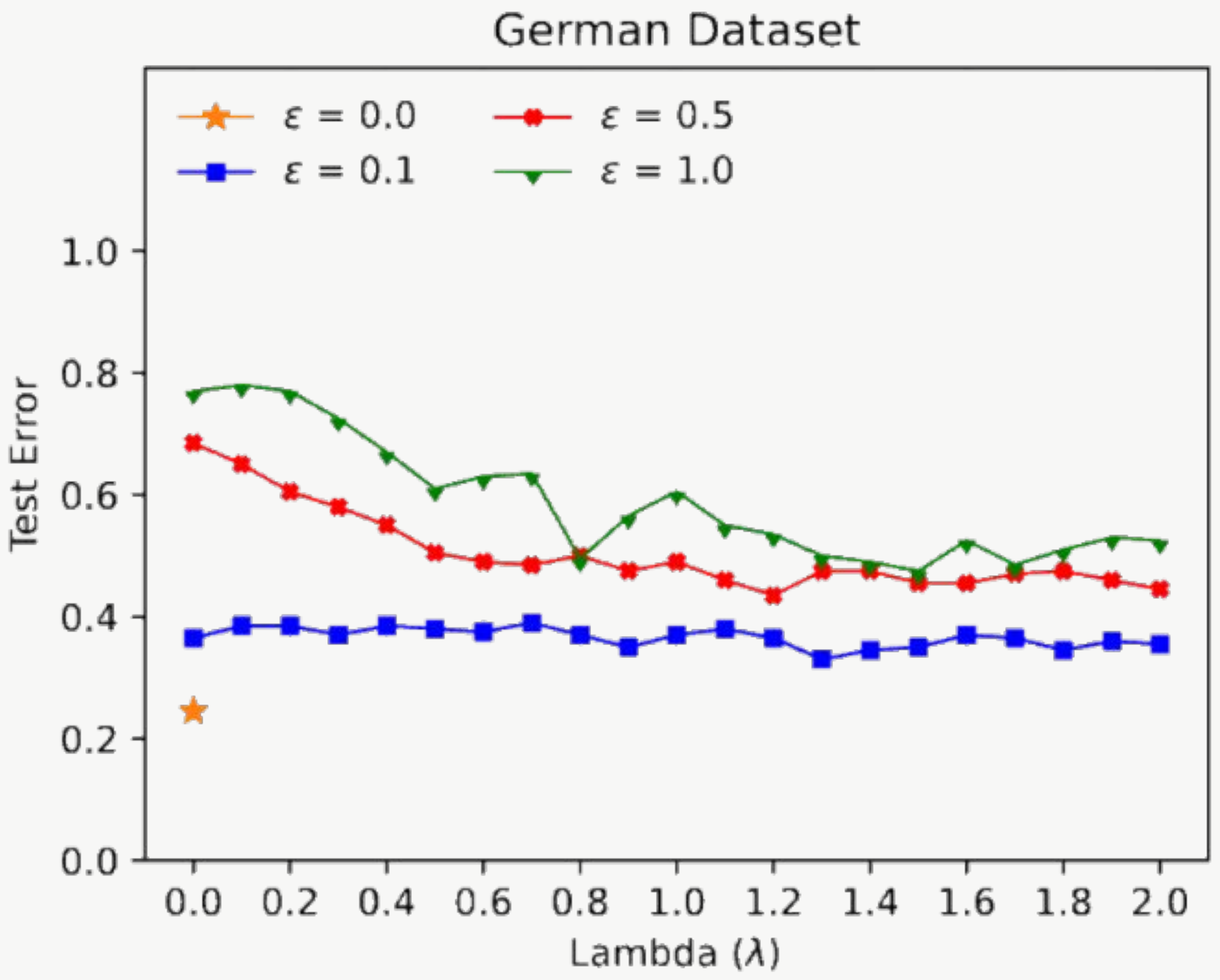
(Test error)

(SPD)

(EOD)

Effects of λ on different metrics

- ***Claim 1:*** Larger values of λ results in stronger attacks against fairness



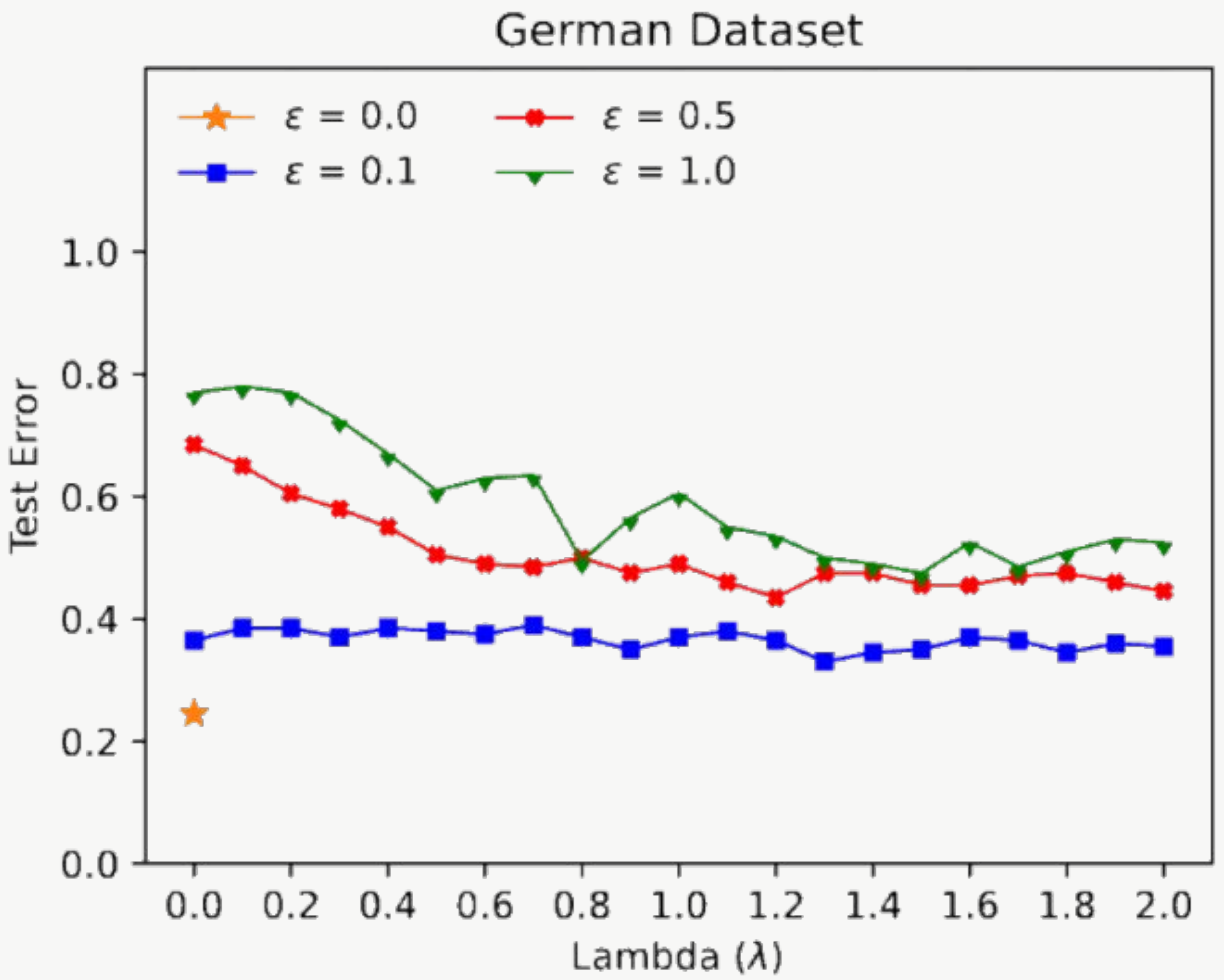
(Test error)

(SPD)

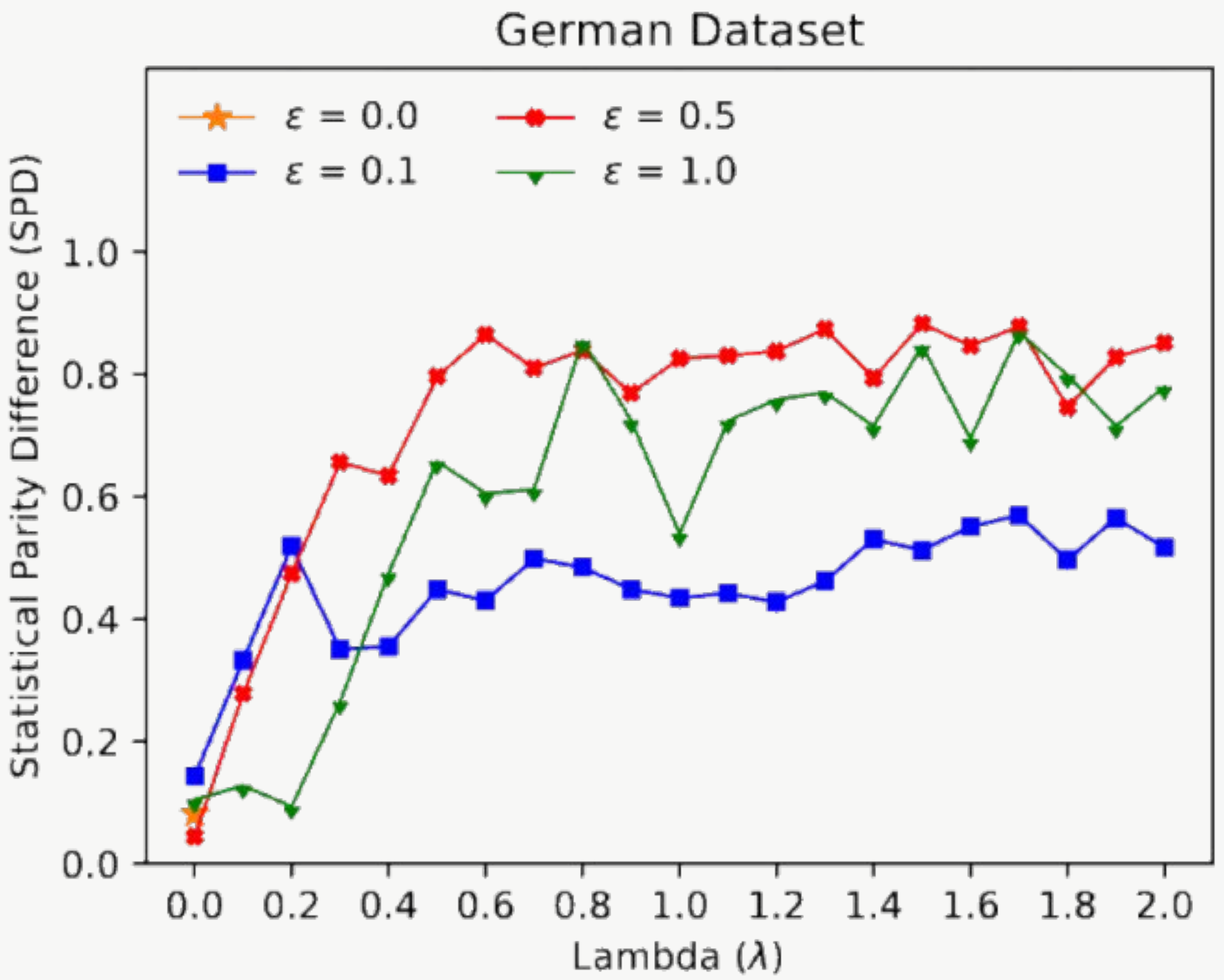
(EOD)

Effects of λ on different metrics

- ***Claim 1:*** Larger values of λ results in stronger attacks against fairness



(Test error)

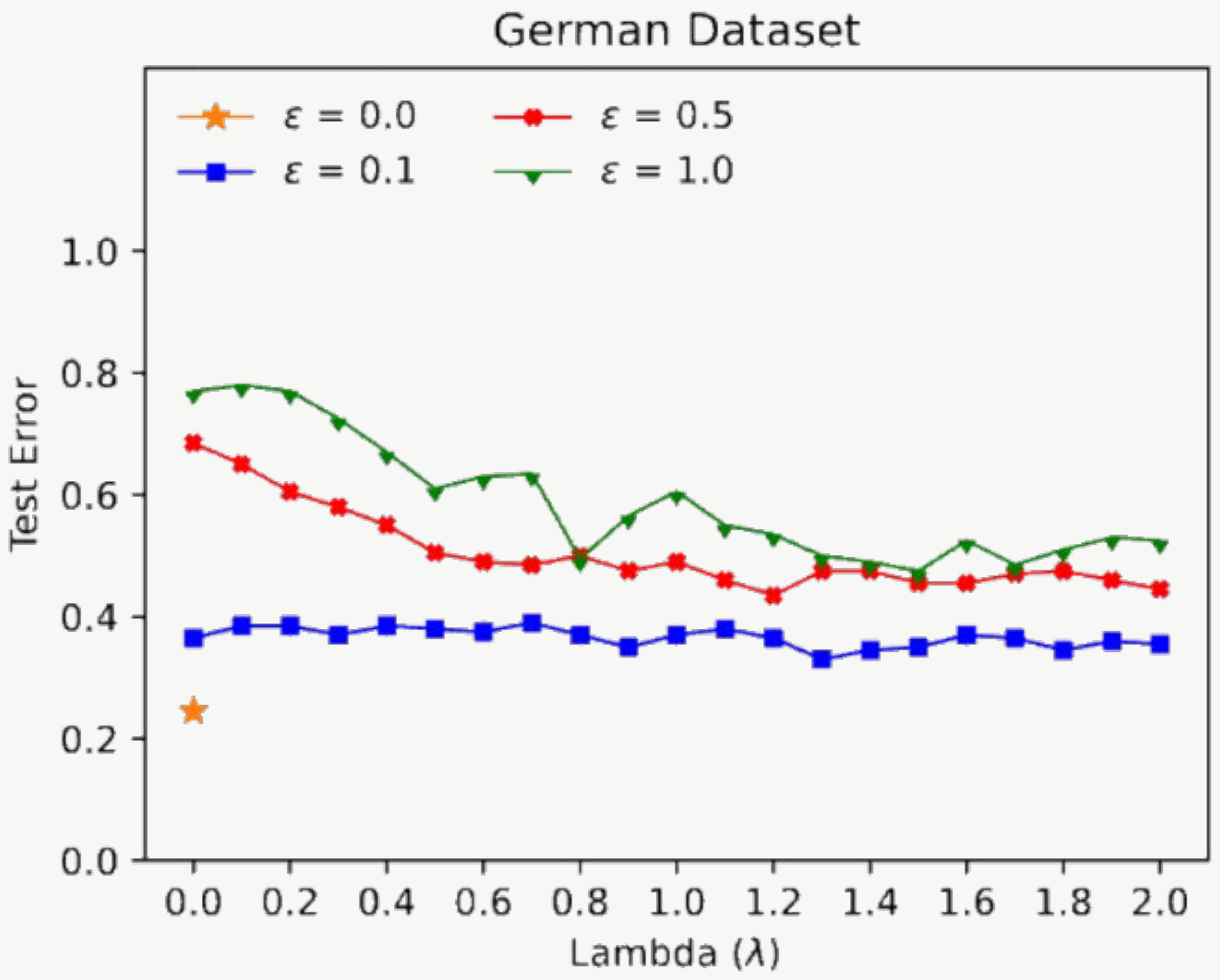


(SPD)

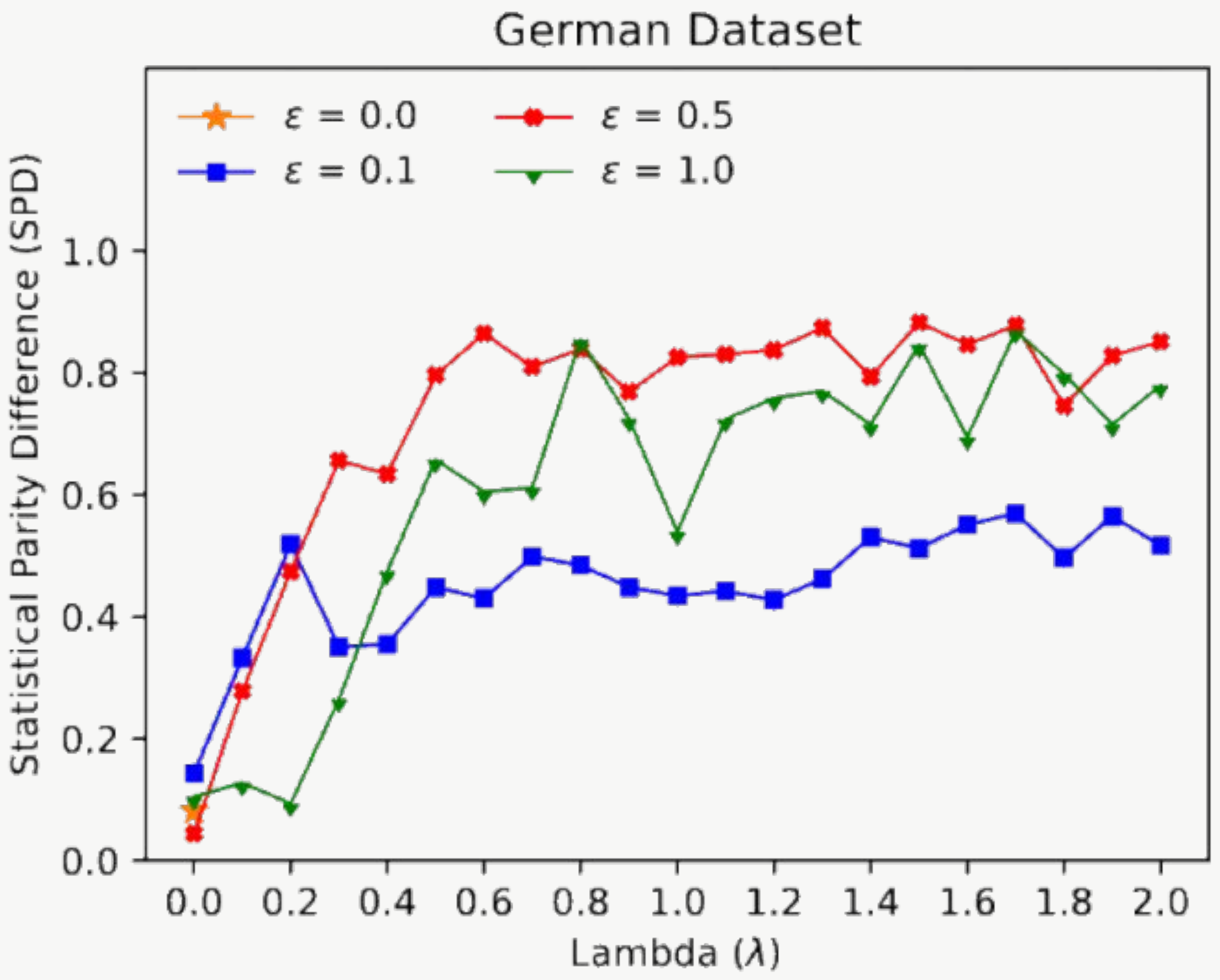
(EOD)

Effects of λ on different metrics

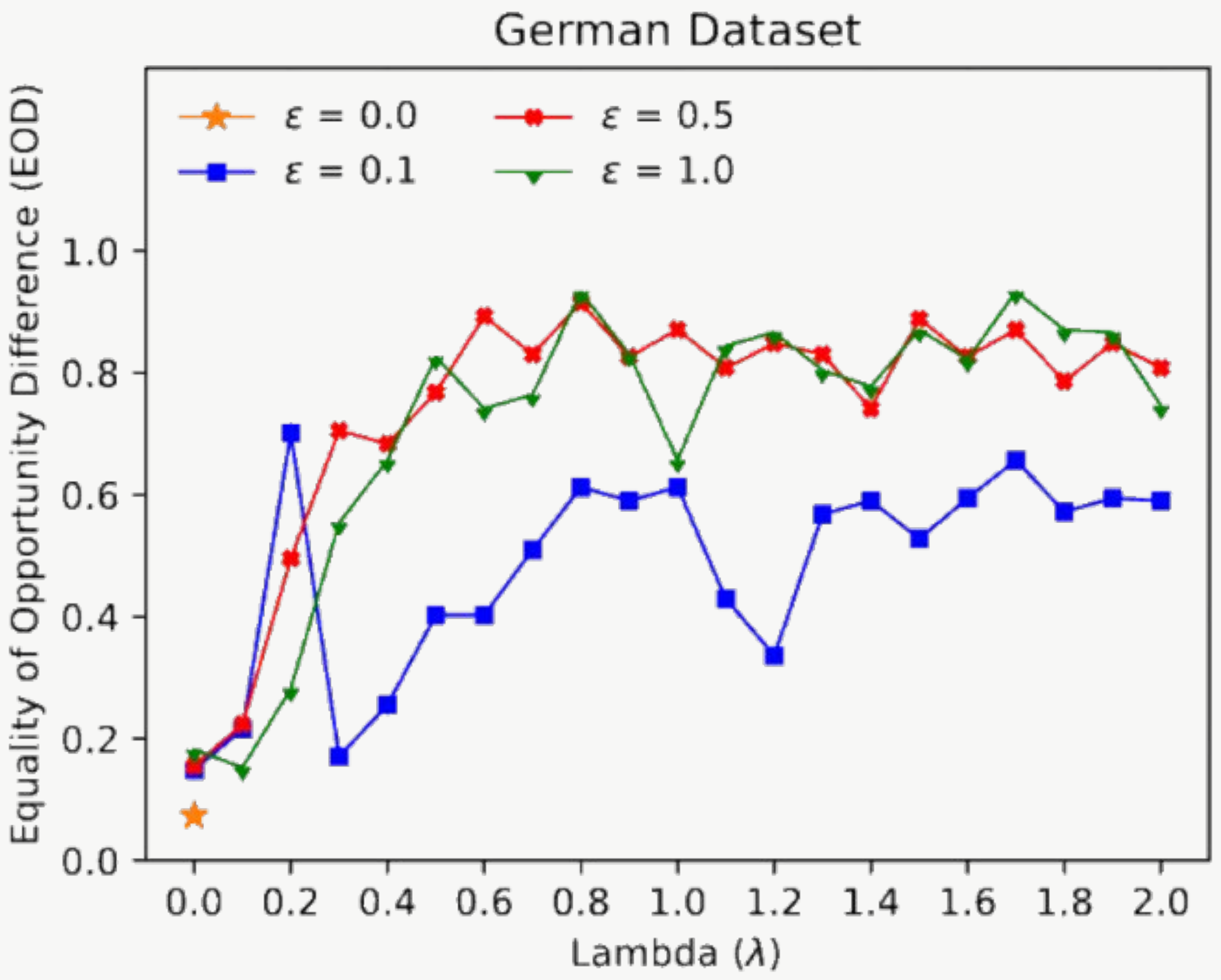
- ***Claim 1:*** Larger values of λ result in stronger attacks against fairness



(Test error)



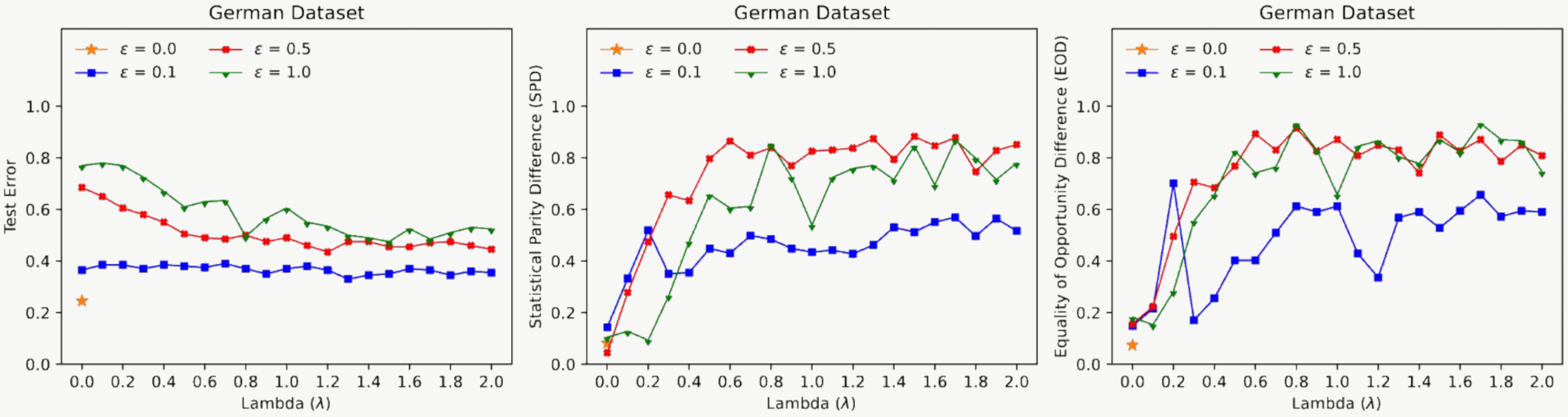
(SPD)



(EOD)

Effects of λ on different metrics

- ***Claim 1:*** Larger values of λ results in stronger attacks against fairness



(Test error) **(SPD)** **(EOD)**

Influence attack vs baselines

- ***Claim 2:*** The proposed IAF outperforms the attack of Koh et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.

- ***Claim 3:*** The proposed IAF outperforms the attack of Solans et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.

(Test error) (SPD) (EOD)

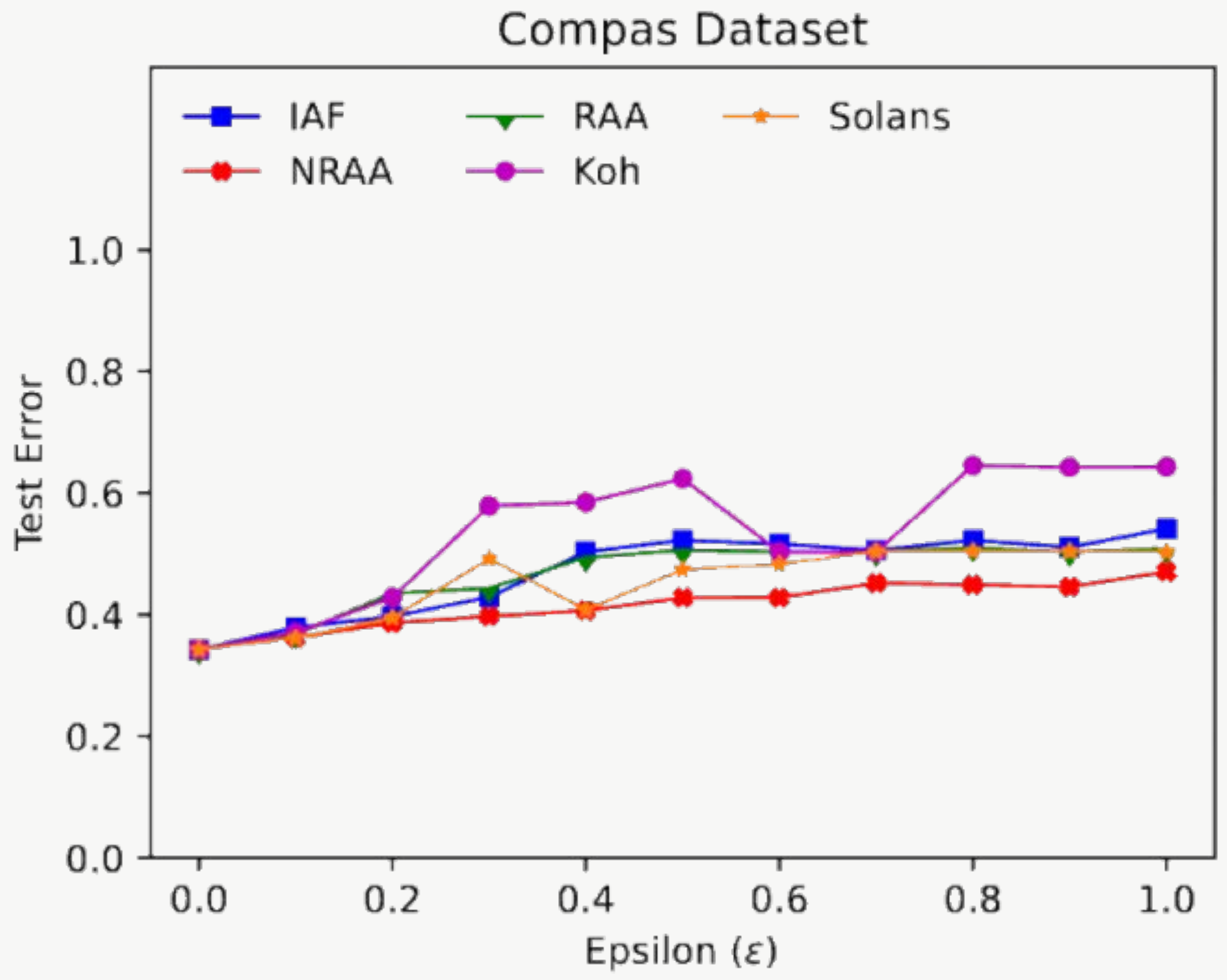
Influence attack vs baselines

- ***Claim 2:*** The proposed IAF outperforms the attack of Koh et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.
- ***Claim 3:*** The proposed IAF outperforms the attack of Solans et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.

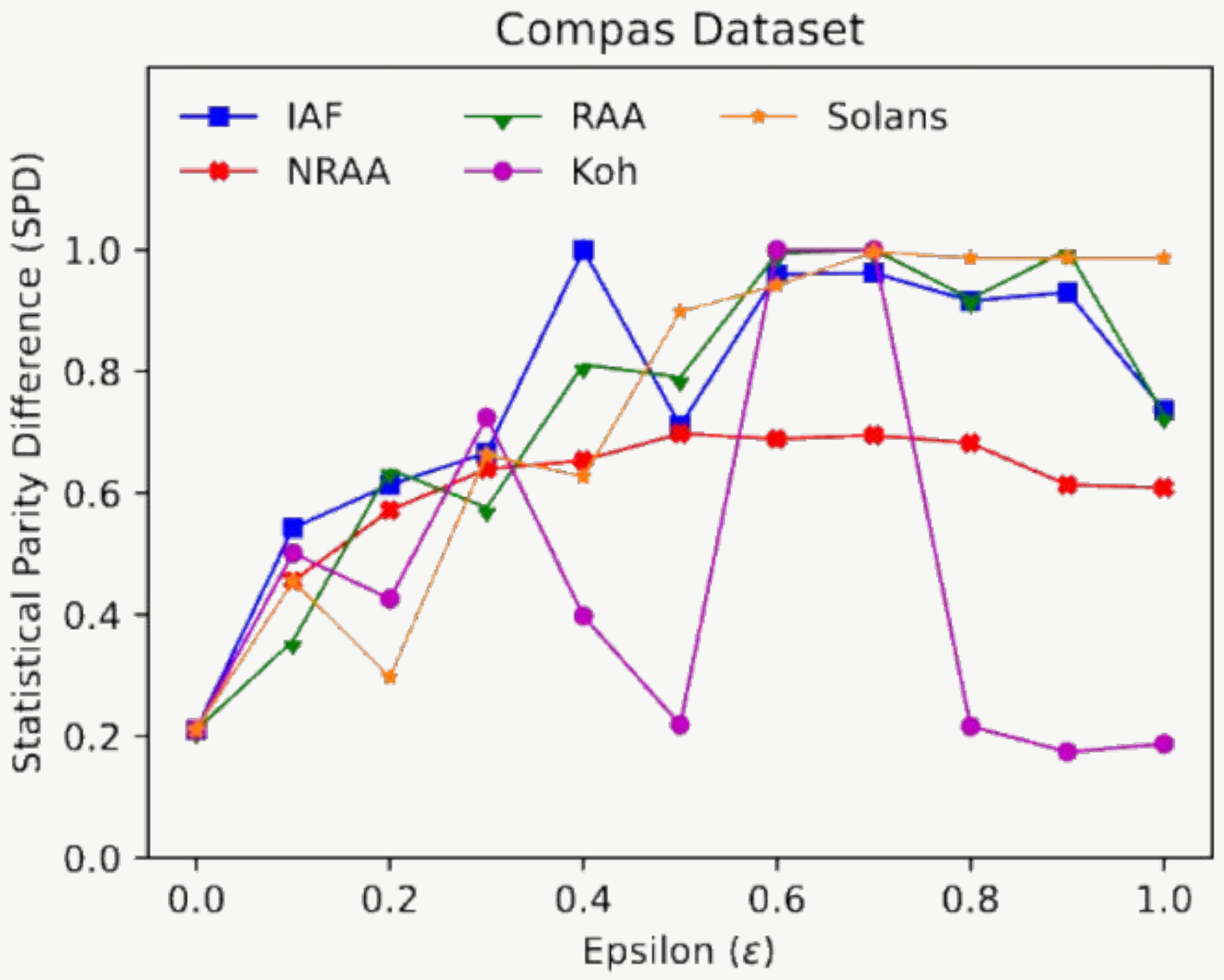
(Test error) (SPD) (EOD)

Influence attack vs baselines

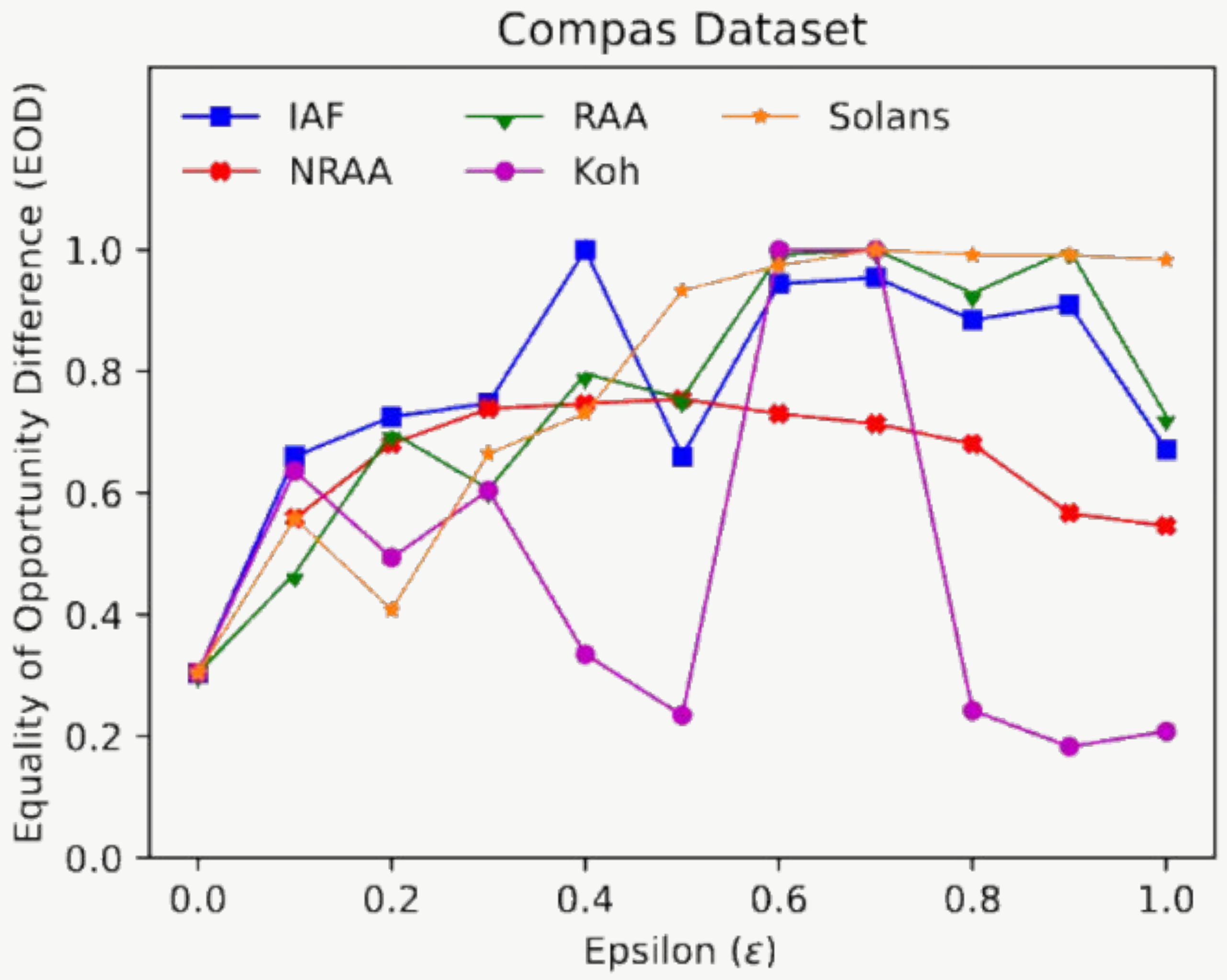
- ***Claim 2:*** The proposed IAF outperforms the attack of Koh et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.
- ***Claim 3:*** The proposed IAF outperforms the attack of Solans et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.



(Test error)



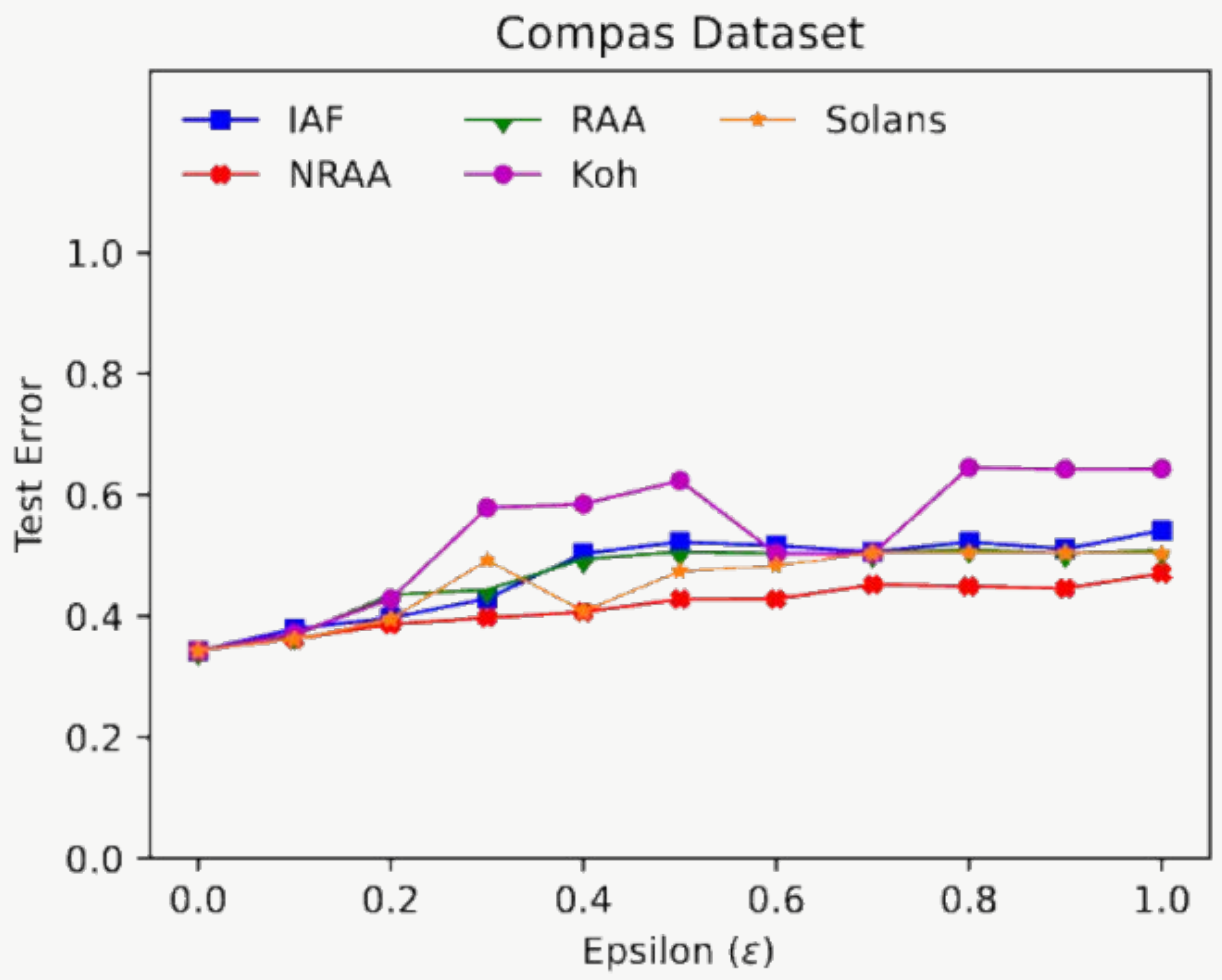
(SPD)



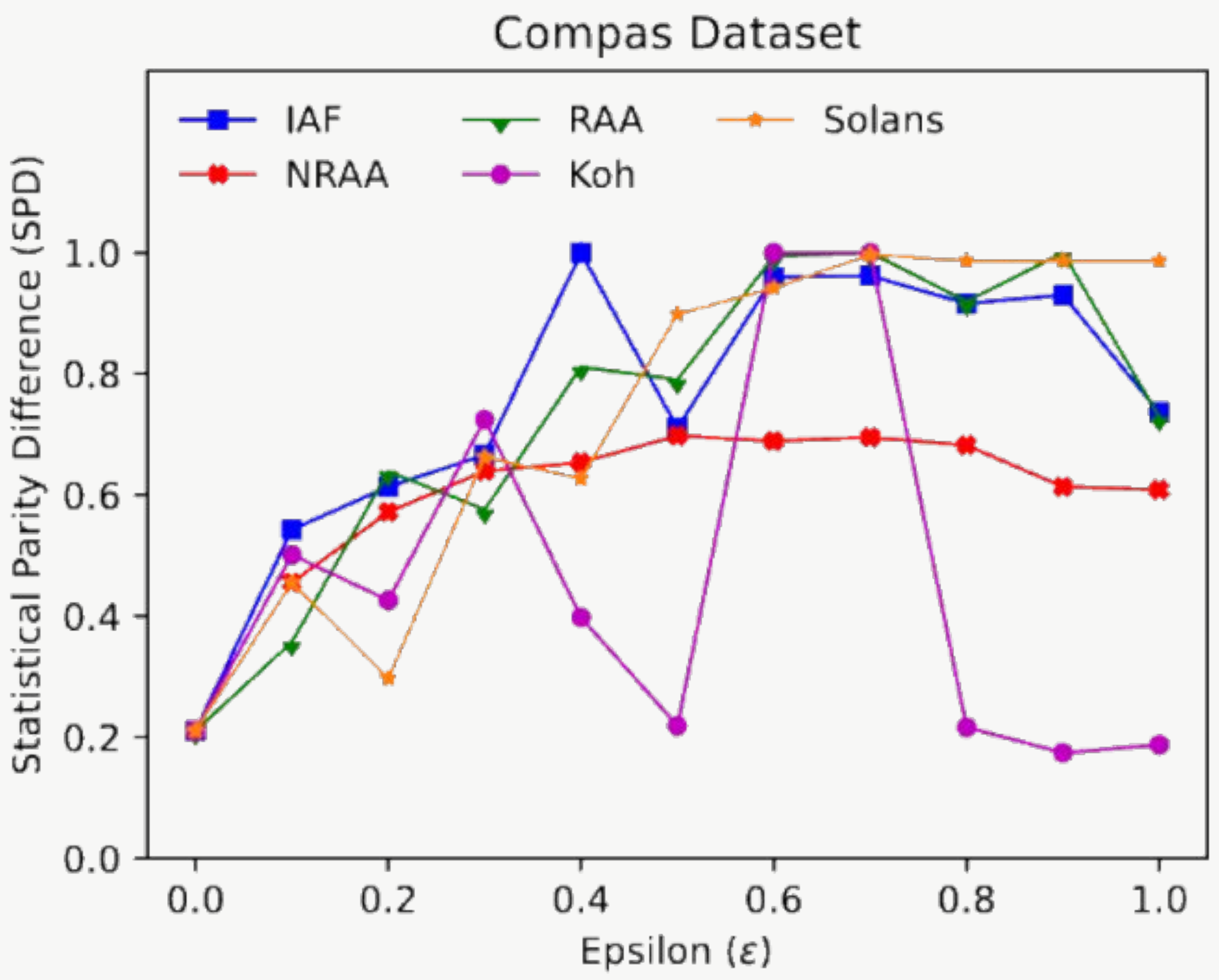
(EOD)

Influence attack vs baselines

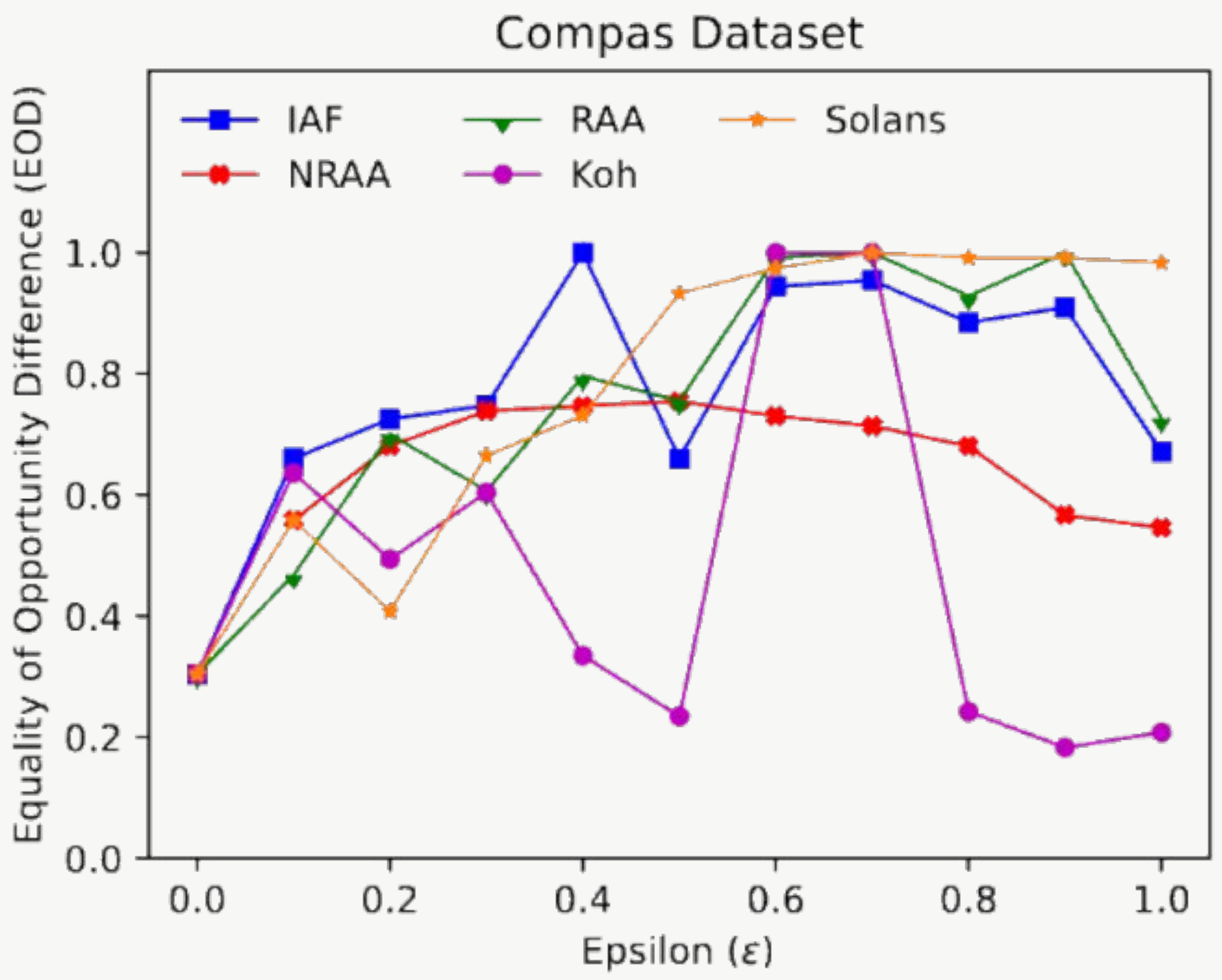
- ***Claim 2:*** The proposed IAF outperforms the attack of Koh et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.
- ***Claim 3:*** The proposed IAF outperforms the attack of Solans et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.



(Test error)



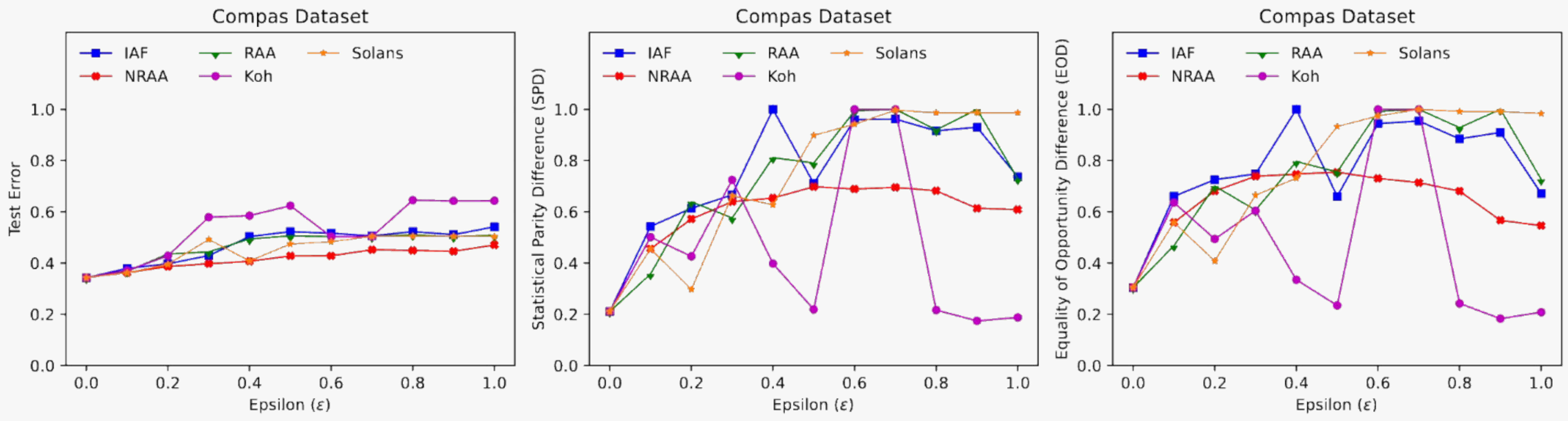
(SPD)



(EOD)

Influence attack vs baselines

- ***Claim 2:*** The proposed IAF outperforms the attack of Koh et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.
- ***Claim 3:*** The proposed IAF outperforms the attack of Solans et al. in affecting both fairness metrics (SPD and EOD), on all three datasets.



(Test error) **(SPD)** **(EOD)**

Anchoring attack vs baselines

- ***Claim 4:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. in degrading the SPD and EOD of the classification model, on all three datasets.

- ***Claim 5:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. in degrading the SPD and EOD of the classification model, on all three datasets.

(Test error) (SPD) (EOD)

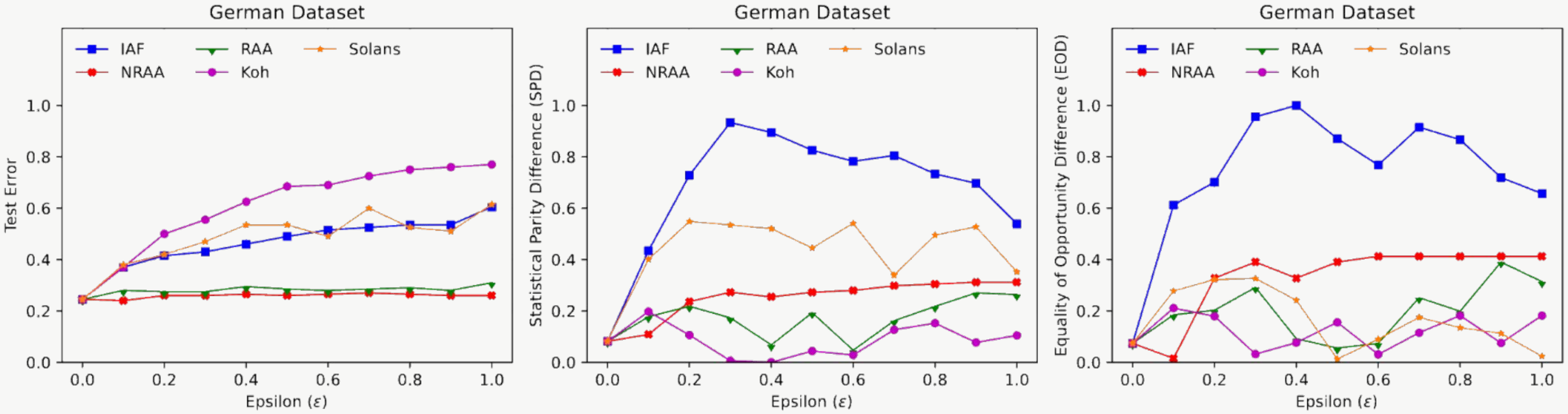
Anchoring attack vs baselines

- ***Claim 4:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. in degrading the SPD and EOD of the classification model, on all three datasets.
- ***Claim 5:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. in degrading the SPD and EOD of the classification model, on all three datasets.

(Test error) (SPD) (EOD)

Anchoring attack vs baselines

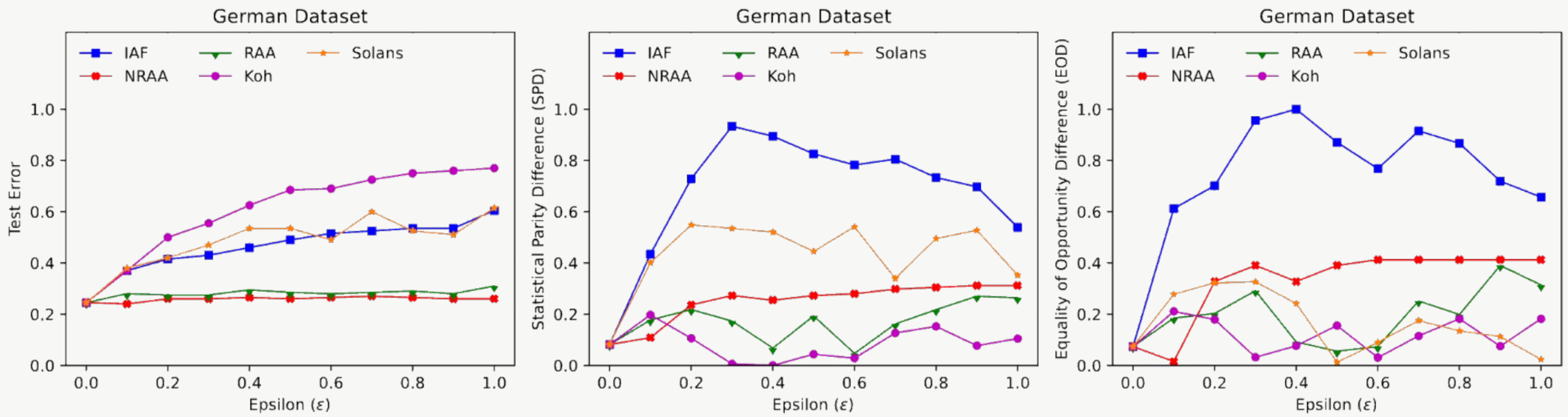
- ***Claim 4:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. in degrading the SPD and EOD of the classification model, on all three datasets.
- ***Claim 5:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. in degrading the SPD and EOD of the classification model, on all three datasets.



(Test error) **(SPD)** **(EOD)**

Anchoring attack vs baselines

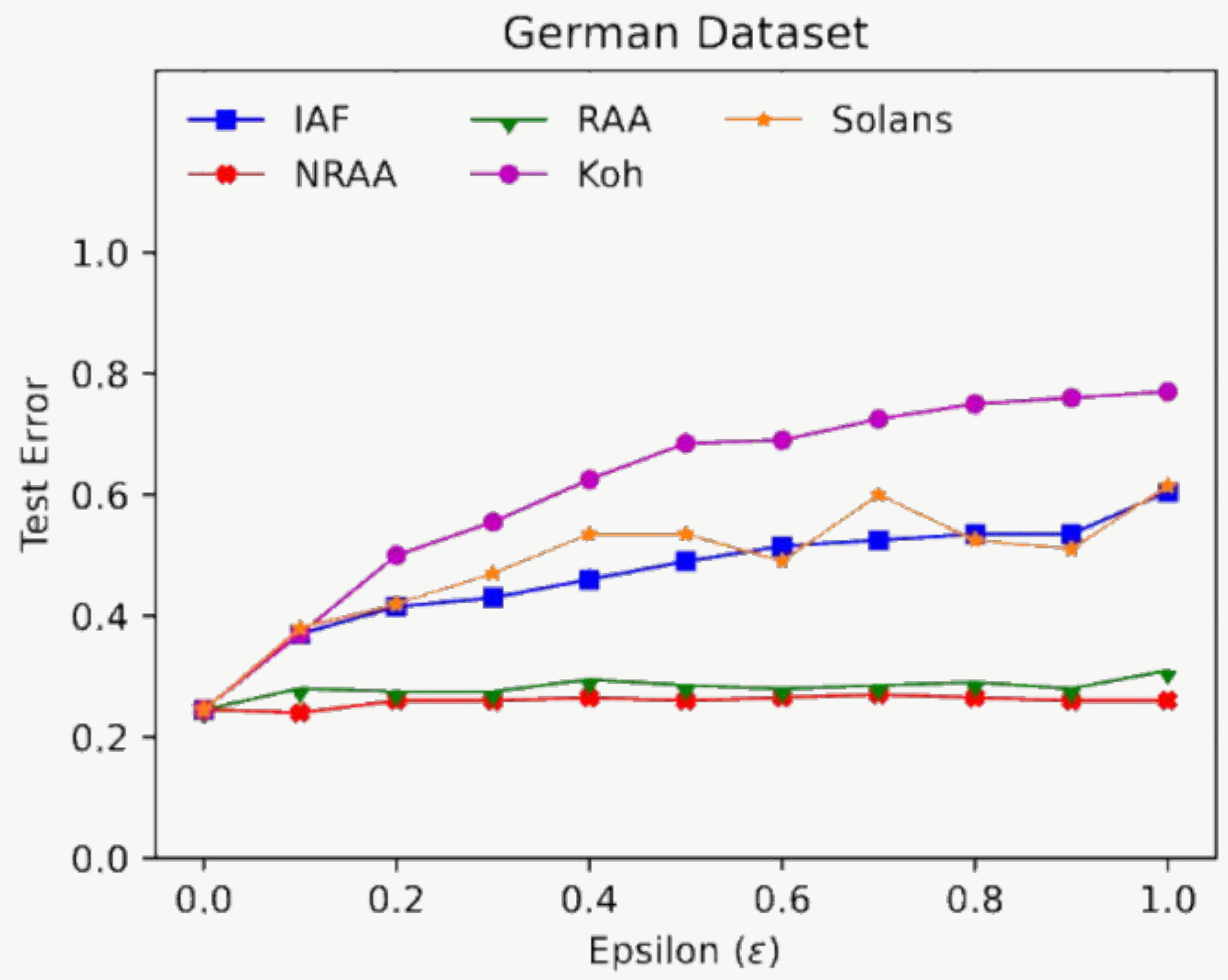
- ***Claim 4:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. in degrading the SPD and EOD of the classification model, on all three datasets.
- ***Claim 5:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. in degrading the SPD and EOD of the classification model, on all three datasets.



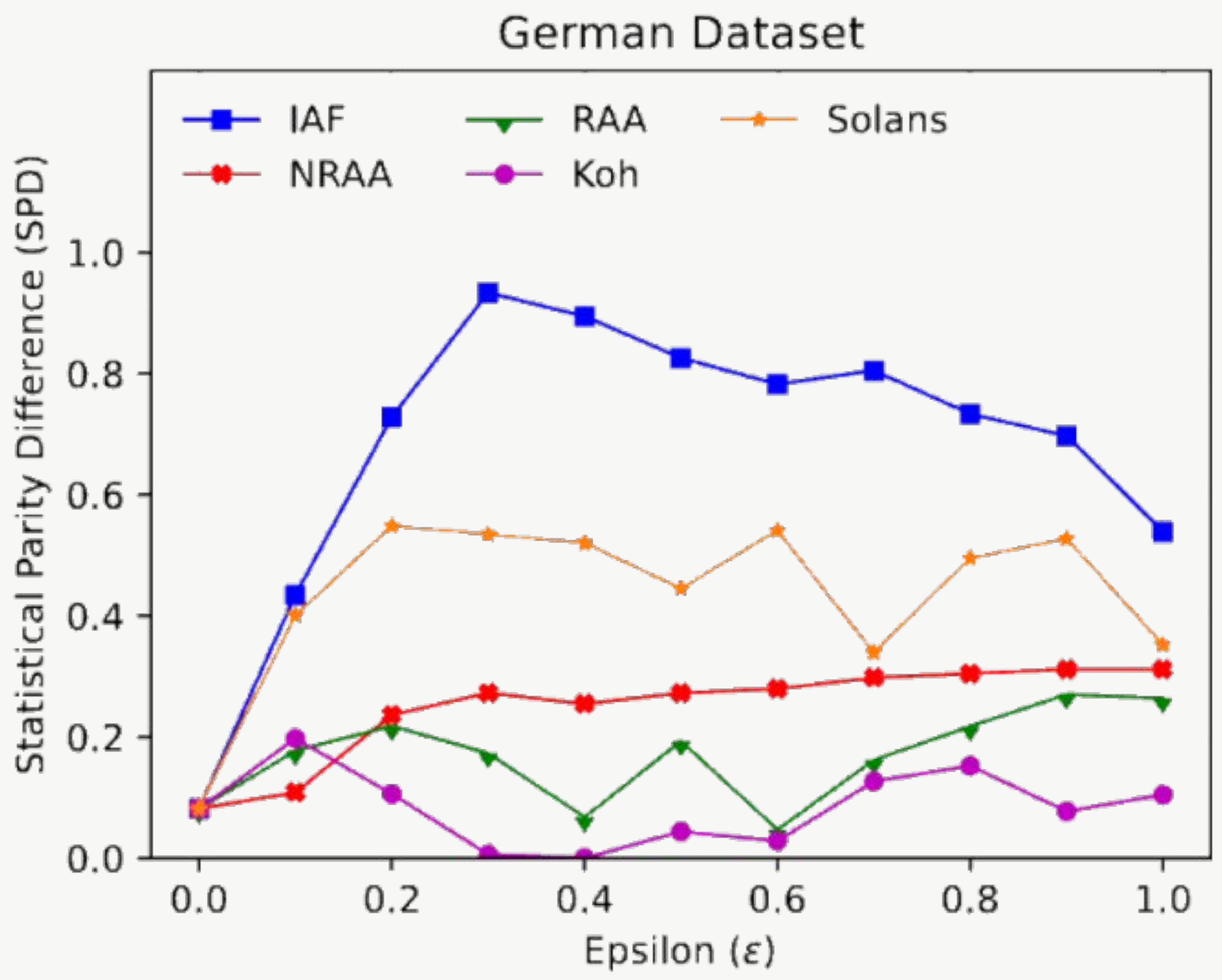
(Test error) **(SPD)** **(EOD)**

Anchoring attack vs baselines

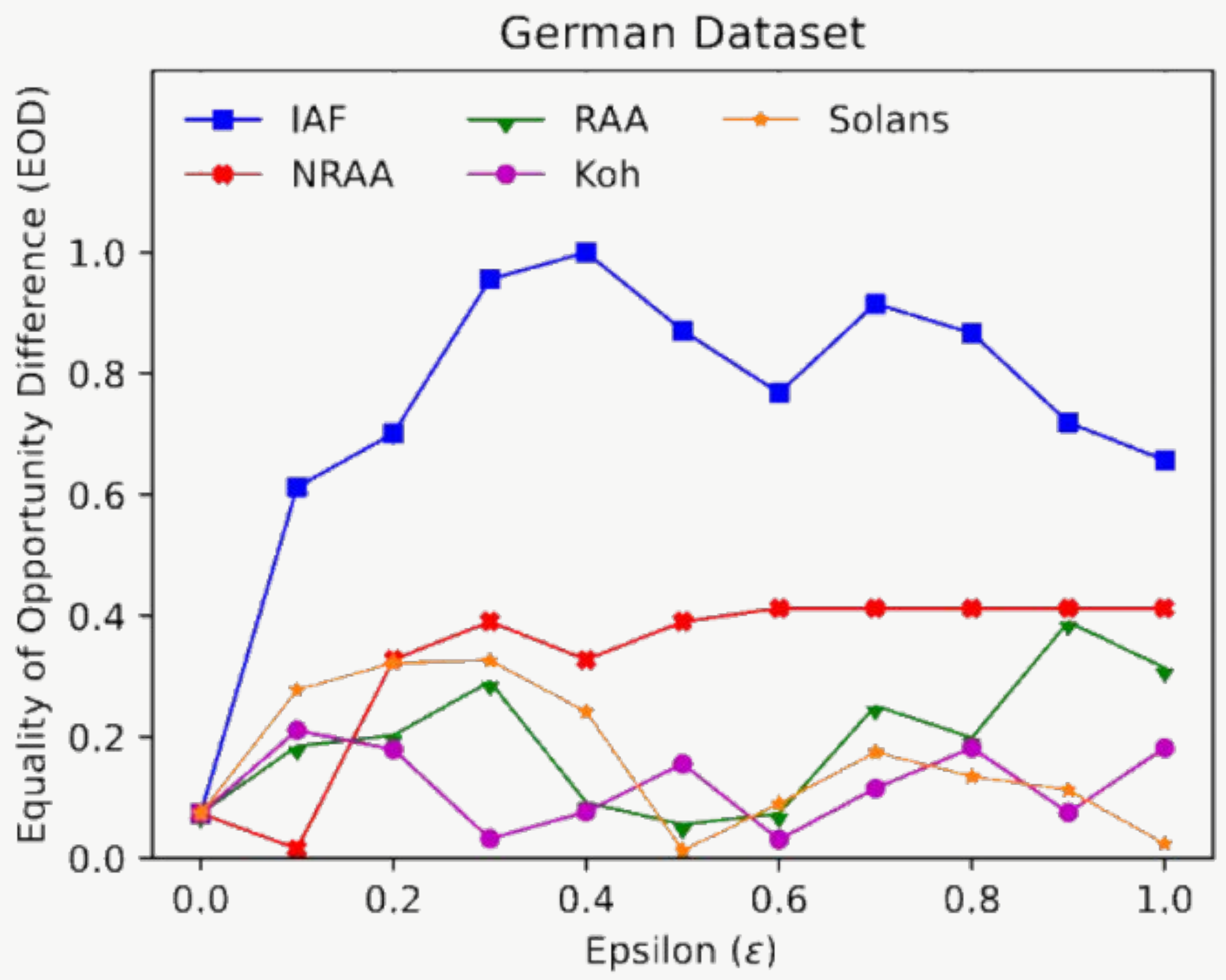
- ***Claim 4:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Koh et al. in degrading the SPD and EOD of the classification model, on all three datasets.
- ***Claim 5:*** Both random and non-random anchoring attacks (RAA and NRAA, respectively) outperform Solans et al. in degrading the SPD and EOD of the classification model, on all three datasets.



(Test error)



(SPD)



(EOD)

Results beyond the original paper:

Effects of different stopping metrics

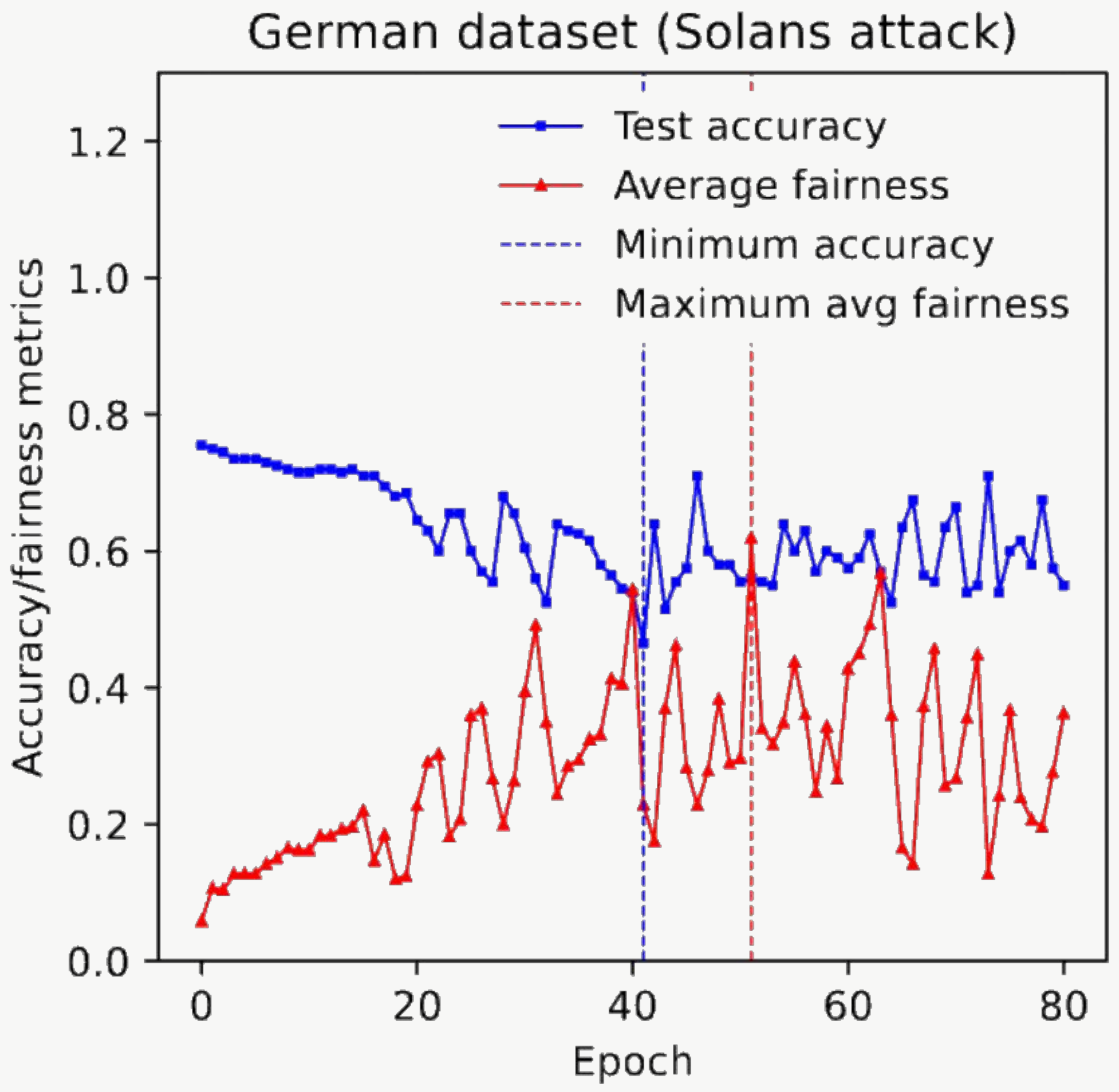
- Early stopping metric: **accuracy** or **average fairness**?

Value	German (Solans)	Drug (IAF)
Min. test accuracy	0.465	0.506
Avg. fairness at the point of min. accuracy	0.229	0.822
Actual max. average fairness	0.619	1.000

Results beyond the original paper:

Effects of different stopping metrics

- Early stopping metric: **accuracy** or **average fairness**?

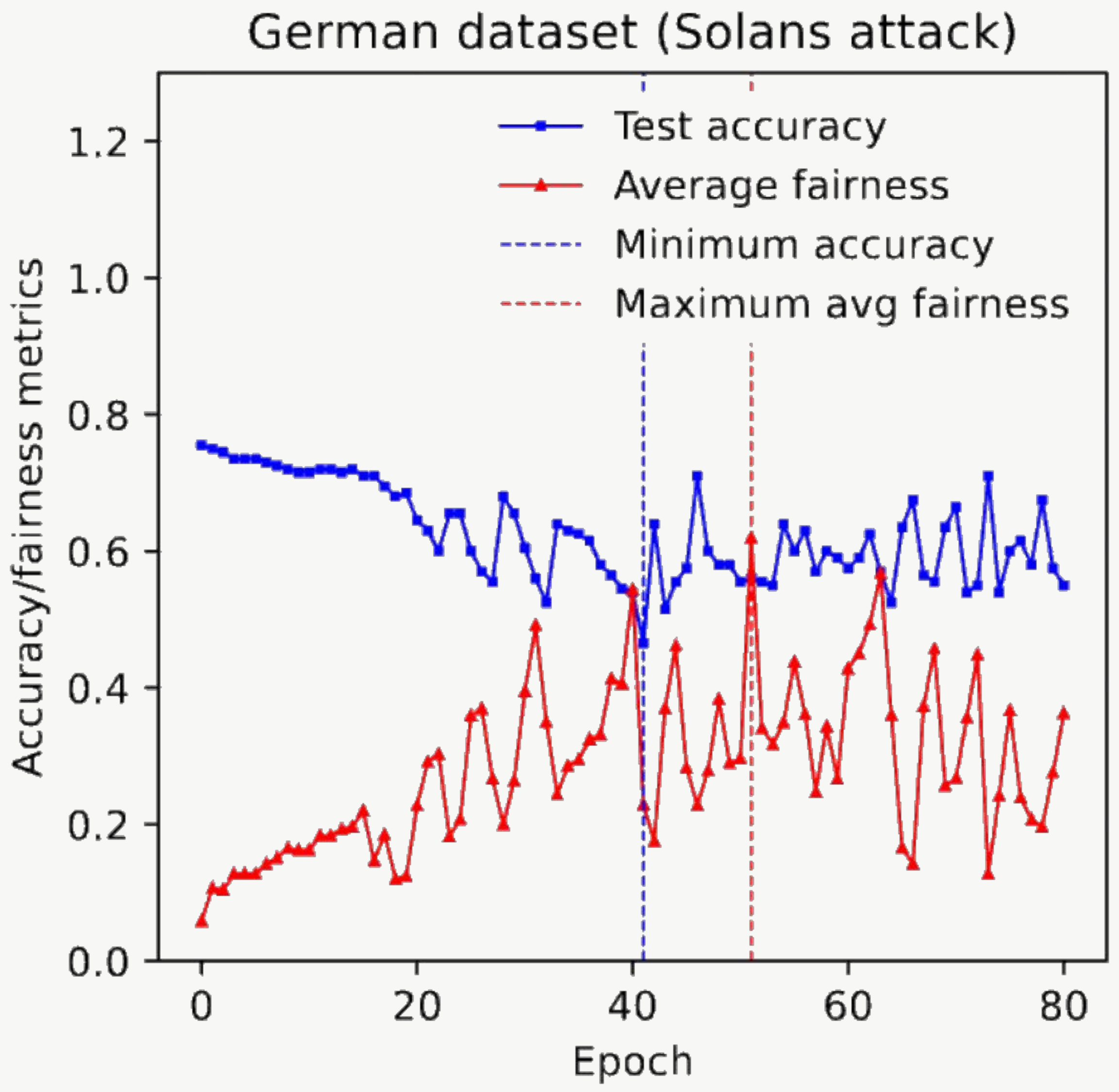


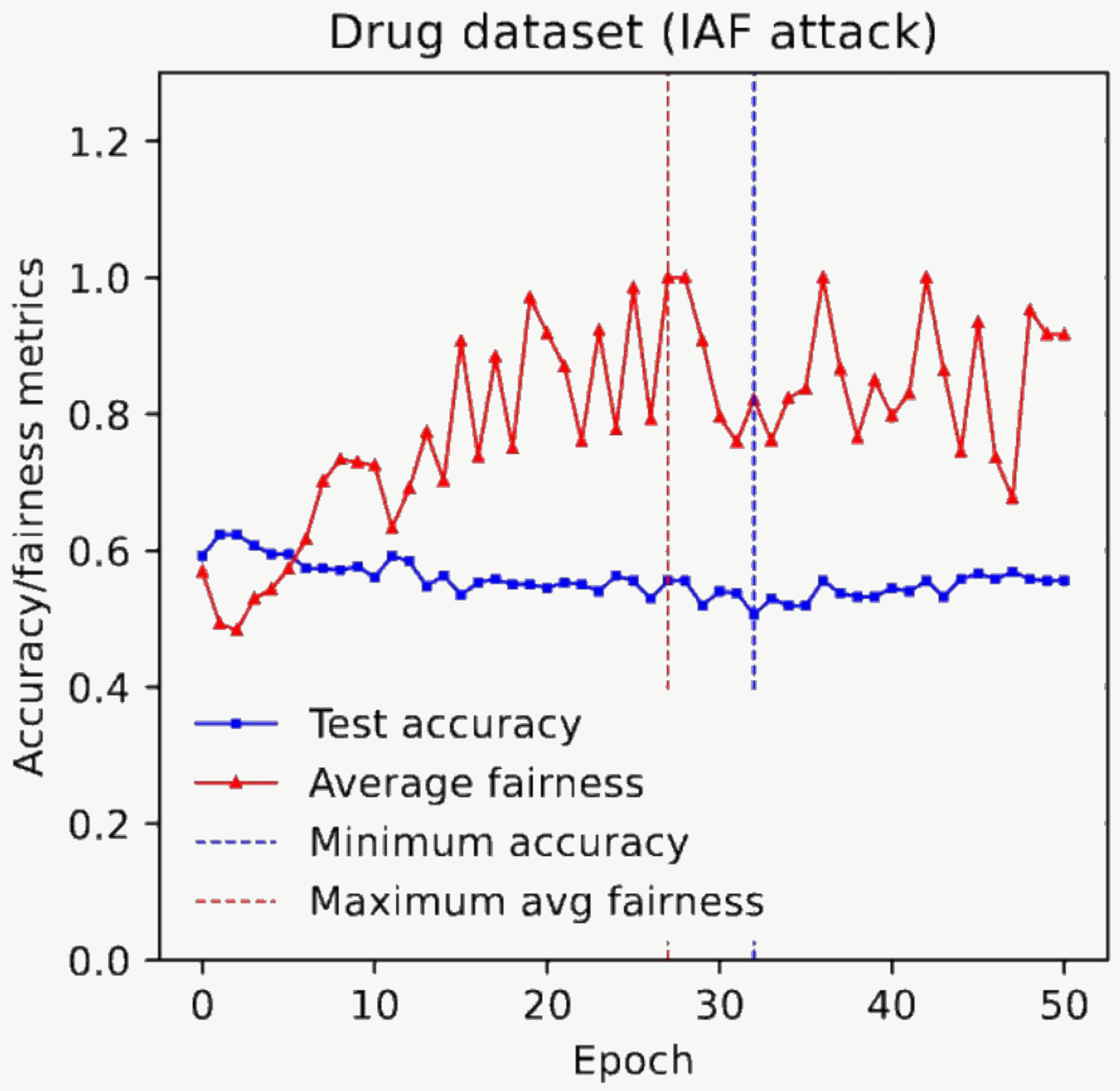
Value	German (Solans)	Drug (IAF)
Min. test accuracy	0.465	0.506
Avg. fairness at the point of min. accuracy	0.229	0.822
Actual max. average fairness	0.619	1.000

Results beyond the original paper:

Effects of different stopping metrics

- Early stopping metric: **accuracy** or **average fairness**?



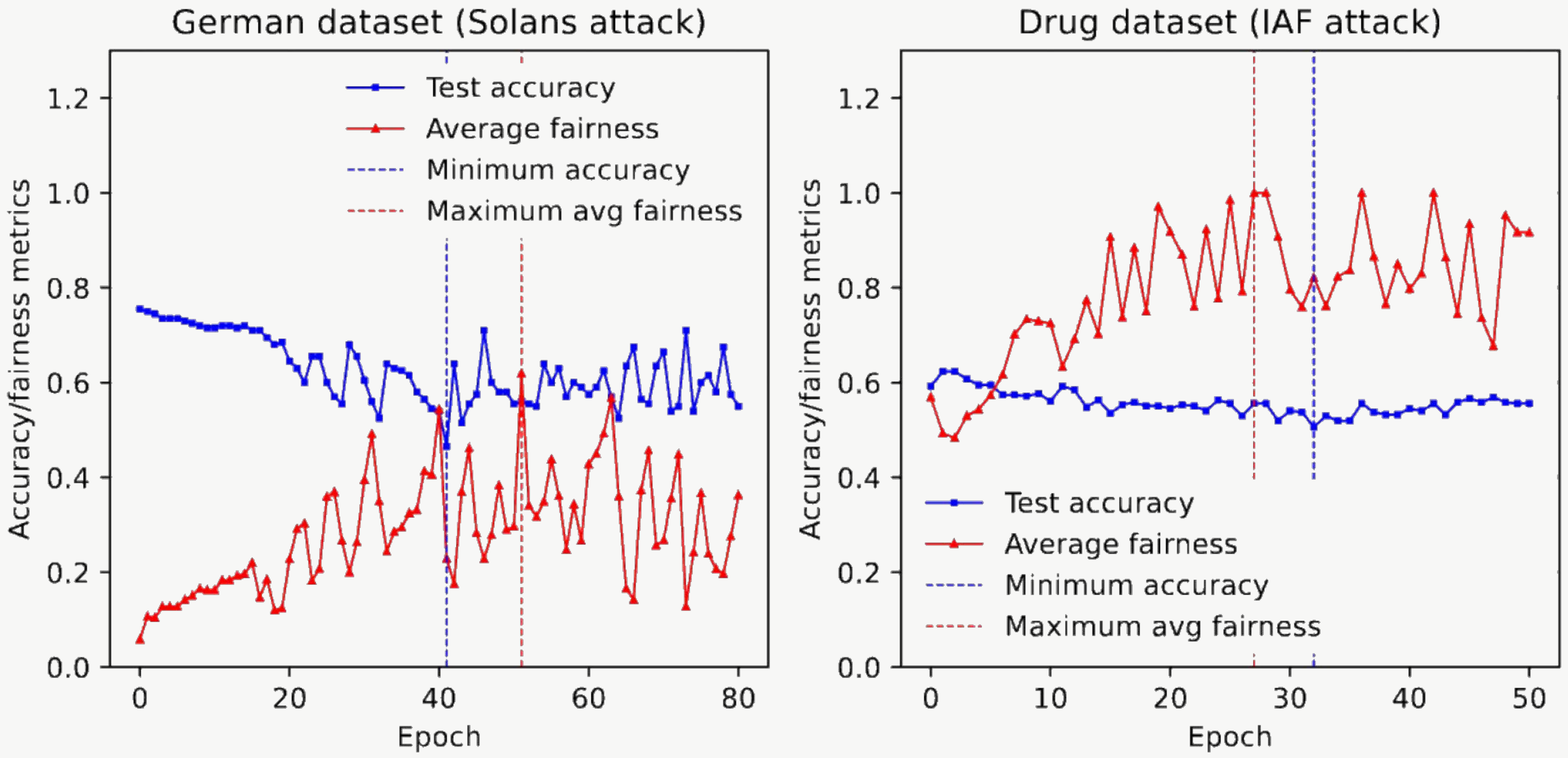


Value	German (Solans)	Drug (IAF)
Min. test accuracy	0.465	0.506
Avg. fairness at the point of min. accuracy	0.229	0.822
Actual max. average fairness	0.619	1.000

Results beyond the original paper:

Effects of different stopping metrics

- Early stopping metric: **accuracy** or **average fairness**?



Value	German (Solans)	Drug (IAF)
Min. test accuracy	0.465	0.506
Avg. fairness at the point of min. accuracy	0.229	0.822
Actual max. average fairness	0.619	1.000

Results beyond the original paper:

Effects of different stopping metrics

- Early stopping metric: **accuracy** or **average fairness**?

Value	German (Solans)	Drug (IAF)
Min. test accuracy	0.465	0.506
Avg. fairness at the point of min. accuracy	0.229	0.822
Actual max. average fairness	0.619	1.000

Outline

Summary of the results

- **Average the metrics over the ε values**
- Base our results on quantifiable measures instead of solely relying on visual inspection

Attack	German Dataset: Test error	German Dataset: SPD	German Dataset: EOD	Compas Dataset: Test Error	Compas Dataset: SPD	Compas Dataset: EOD	Drug Dataset: Test error	Drug Dataset: SPD	Drug Dataset: EOD
	(Stopping metric: Fairness / Accuracy)			*(Stopping metric: Fairness / Accuracy)*			*(Stopping metric: Fairness / Accuracy)*		
IAF	0.40/**0.47**	**0.84**/0.68	**0.88**/0.74	0.46/**0.47**	**0.83**/0.75	**0.87**/0.77	0.43/**0.45**	**0.89**/0.75	**0.90**/0.76
NRAA	0.26/0.26	**0.26**/0.25	**0.36**/0.33	0.41/**0.42**	0.59/0.59	0.64/0.64	0.39/0.39	0.53/0.53	0.53/0.53
RAA	0.27/**0.28**	**0.24**/0.17	**0.36**/0.19	0.47/0.47	**0.84**/0.73	**0.87**/0.75	0.42/**0.44**	**0.66**/0.55	**0.68**/0.57
Koh	0.27/**0.61**	**0.17**/0.08	**0.13**/0.12	0.45/**0.53**	**0.81**/0.46	**0.85**/0.48	0.40/**0.56**	**0.56**/0.26	**0.56**/0.29
Solans	0.40/**0.48**	**0.65**/0.44	**0.49**/0.16	0.44/**0.45**	**0.76**/0.73	**0.83**/0.78	0.40/**0.56**	**0.53**/0.28	**0.55**/0.32

Discussion

Better statistics could give clearer insights

- Multiple runs with **different seeds**

Results depend on the chosen stopping metric

- ***Claim 1*** (supported) invariant to the stopping metric
- ***Claims 2-5*** - dependence on the stopping metric

In general, the paper presents novel methods intuitively and clearly, <u>but</u>:

- Missing **implementation details** of attacks
- **Incomplete** code, **incompatible** dependencies
- Data **preprocessing** not specified

Discussion

Better statistics could give clearer insights

- Multiple runs with **different seeds**

Results depend on the chosen stopping metric

- ***Claim 1*** (supported) invariant to the stopping metric
- ***Claims 2-5*** - dependence on the stopping metric

In general, the paper presents novel methods intuitively and clearly, <u>but</u>:

- Missing **implementation details** of attacks
- **Incomplete** code, **incompatible** dependencies
- Data **preprocessing** not specified

Discussion

Better statistics could give clearer insights

- Multiple runs with **different seeds**

Results depend on the chosen stopping metric

- ***Claim 1*** (supported) invariant to the stopping metric
- ***Claims 2-5*** - dependence on the stopping metric

In general, the paper presents novel methods intuitively and clearly, but:

- Missing **implementation details** of attacks
- **Incomplete** code, **incompatible** dependencies
- Data **preprocessing** not specified

Discussion

Better statistics could give clearer insights

- Multiple runs with **different seeds**

Results depend on the chosen stopping metric

- ***Claim 1*** (supported) invariant to the stopping metric
- ***Claims 2-5*** - dependence on the stopping metric

In general, the paper presents novel methods intuitively and clearly, <u>but</u>:

- Missing **implementation details** of attacks
- **Incomplete** code, **incompatible** dependencies
- Data **preprocessing** not specified

Conclusion

Reproduction required many **educated guesses**

Obtained similar findings that support **3 out of 5** claims

Too many missing details: **paper not reproducible**

Conclusion

Reproduction required many **educated guesses**

Obtained similar findings that support **3 out of 5** claims

Too many missing details: **paper not reproducible**

Conclusion

Reproduction required many **educated guesses**

Obtained similar findings that support **3 out of 5** claims

Too many missing details: **paper not reproducible**

UNIVERSITY OF AMSTERDAM

Thank you!

Matteo Tafuro*, Andrea Lombardo*, Tin Hadži Veljković, Lasse Becker-Czarnetzki

Contact information:

- Matteo Tafuro
  tafuromatteo00@gmail.com
- Andrea Lombardo
  andrealombardo2m@gmail.com
- Lasse Becker-Czarnetzki
  lasse.becza@gmail.com
- Tin Hadži Veljković
  tin.hadzi@gmail.com

Acknowledgements:

We would like to acknowledge the help of ***ELLIS*** and ***Qualcomm*** for funding our travel expenses to NeurIPS 2022.

e l l i s
European Laboratory for Learning and Intelligent Systems

Qualcomm