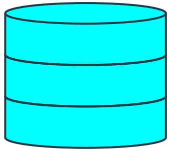
An Empirical Investigation into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification

Prakhar Ganesh

What is Model Multiplicity?

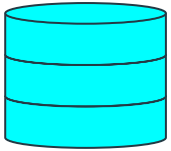
What is Model Multiplicity?

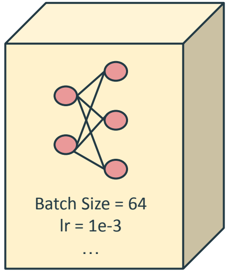


Train & Val Dataset

What is Model Multiplicity?

Train & Val Dataset

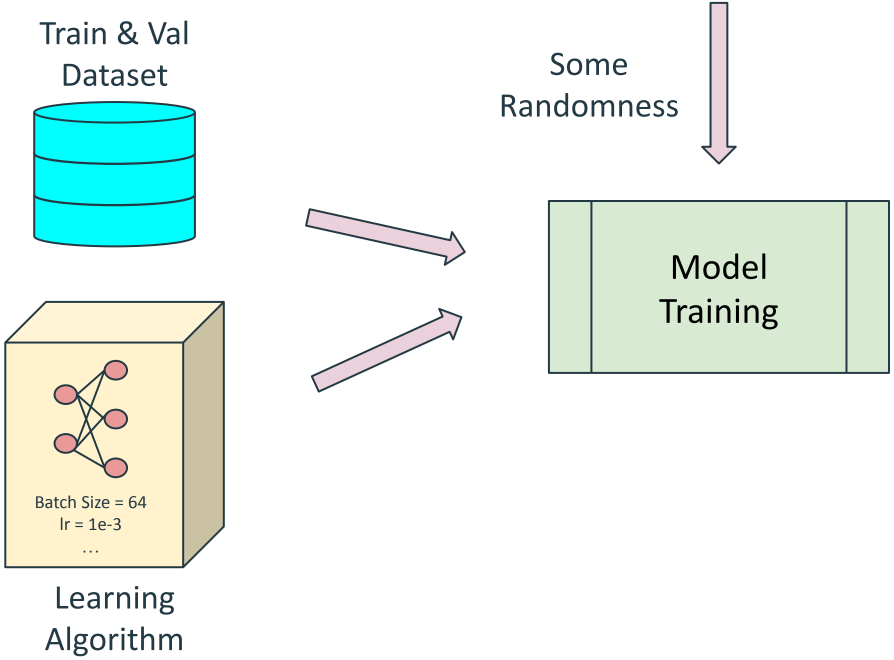




Batch Size = 64
lr = 1e-3
…

Learning Algorithm

What is Model Multiplicity?



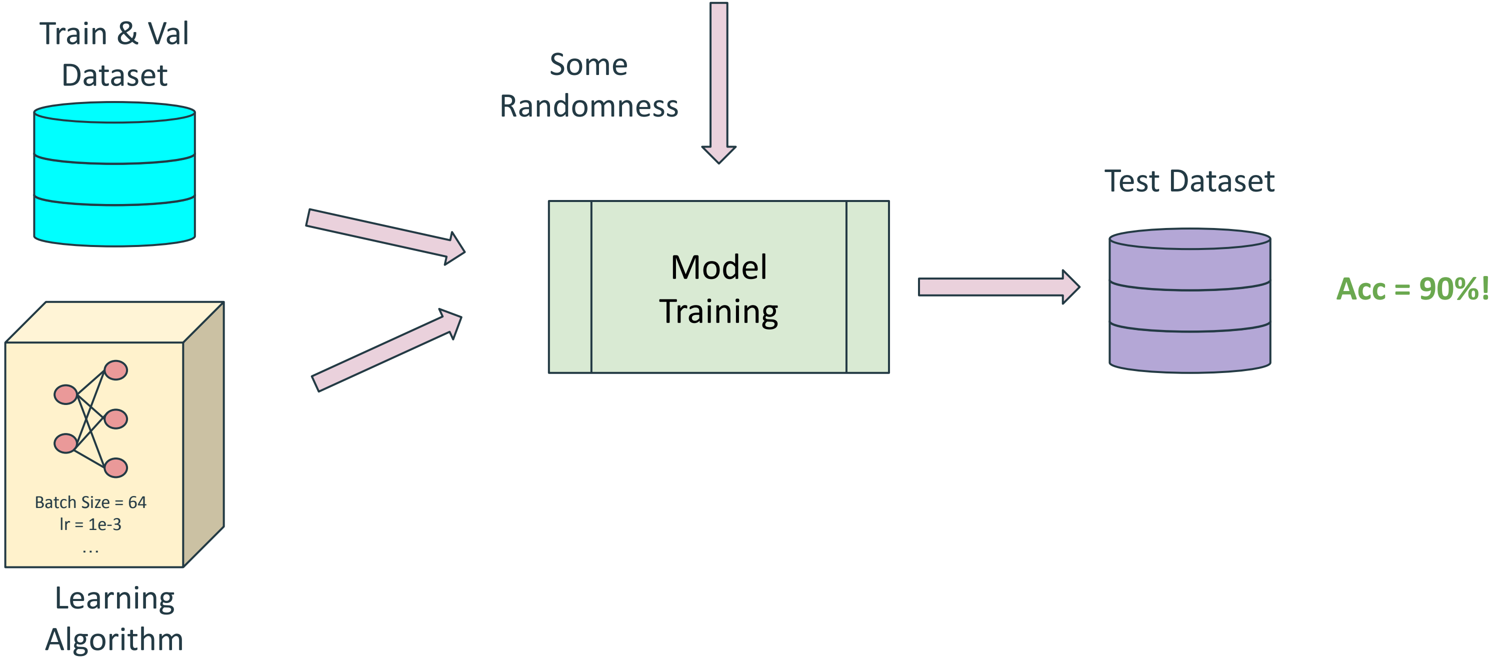
Train & Val Dataset

Some Randomness

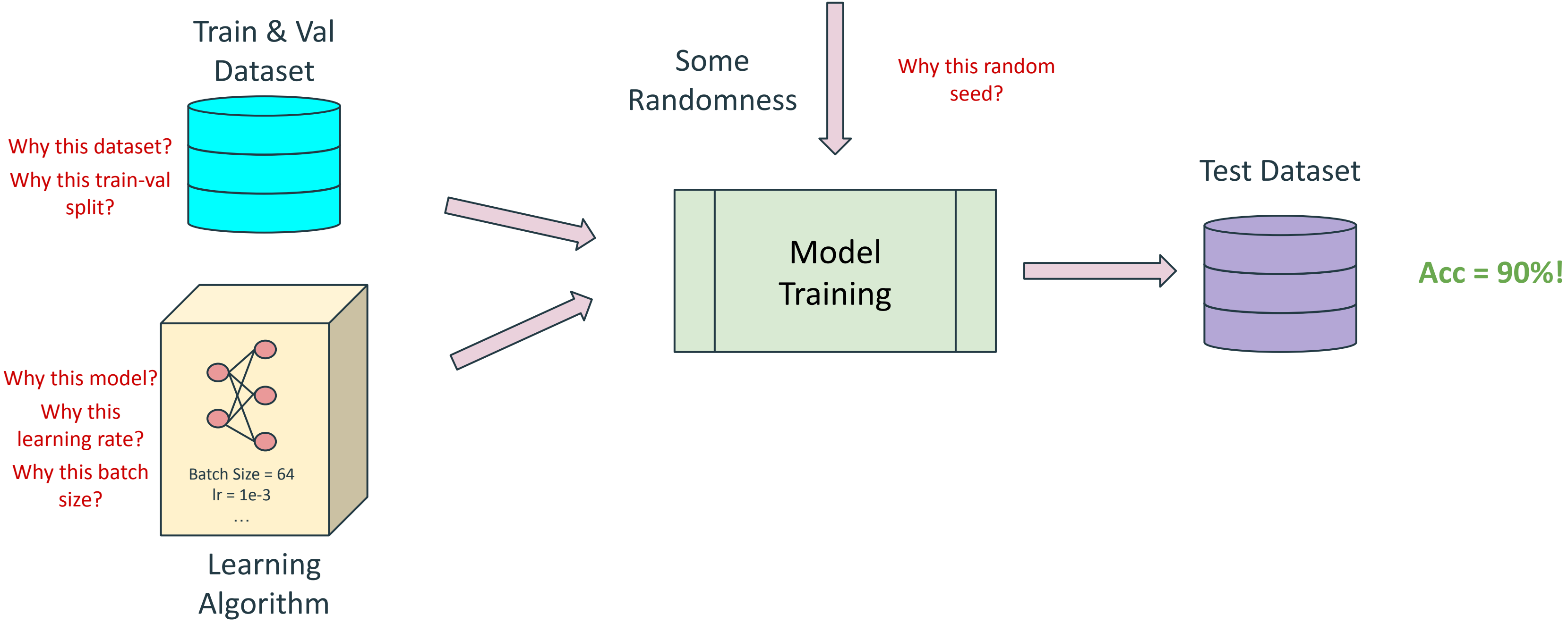
Model Training

Learning Algorithm

What is Model Multiplicity?



What is Model Multiplicity?



Train & Val Dataset

Why this dataset?

Why this train-val split?

Batch Size = 64
lr = 1e-3
…

Learning Algorithm

Why this model?

Why this learning rate?

Why this batch size?

Some Randomness

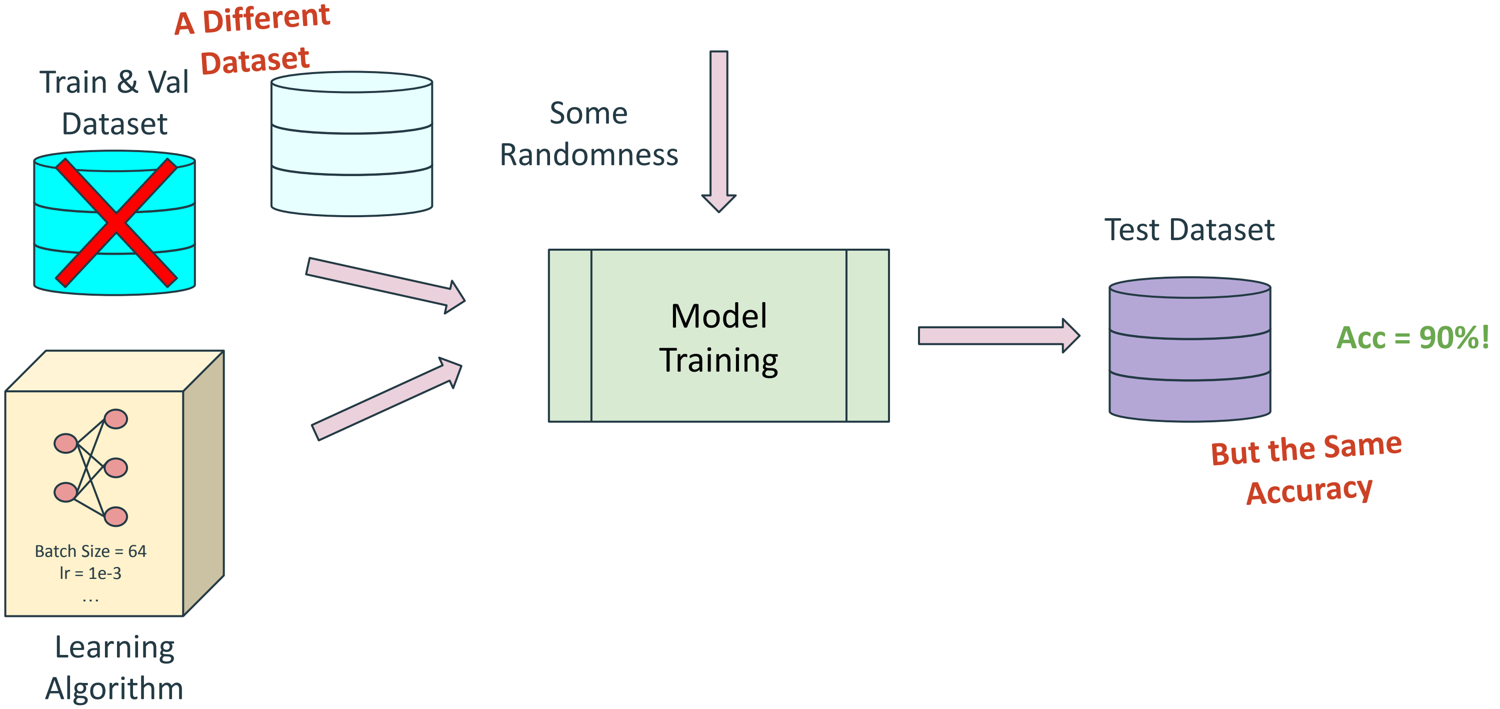
Why this random seed?

Model Training

Test Dataset

Acc = 90%!

What is Model Multiplicity?



A Different Dataset

Train & Val Dataset

Some Randomness

Model Training

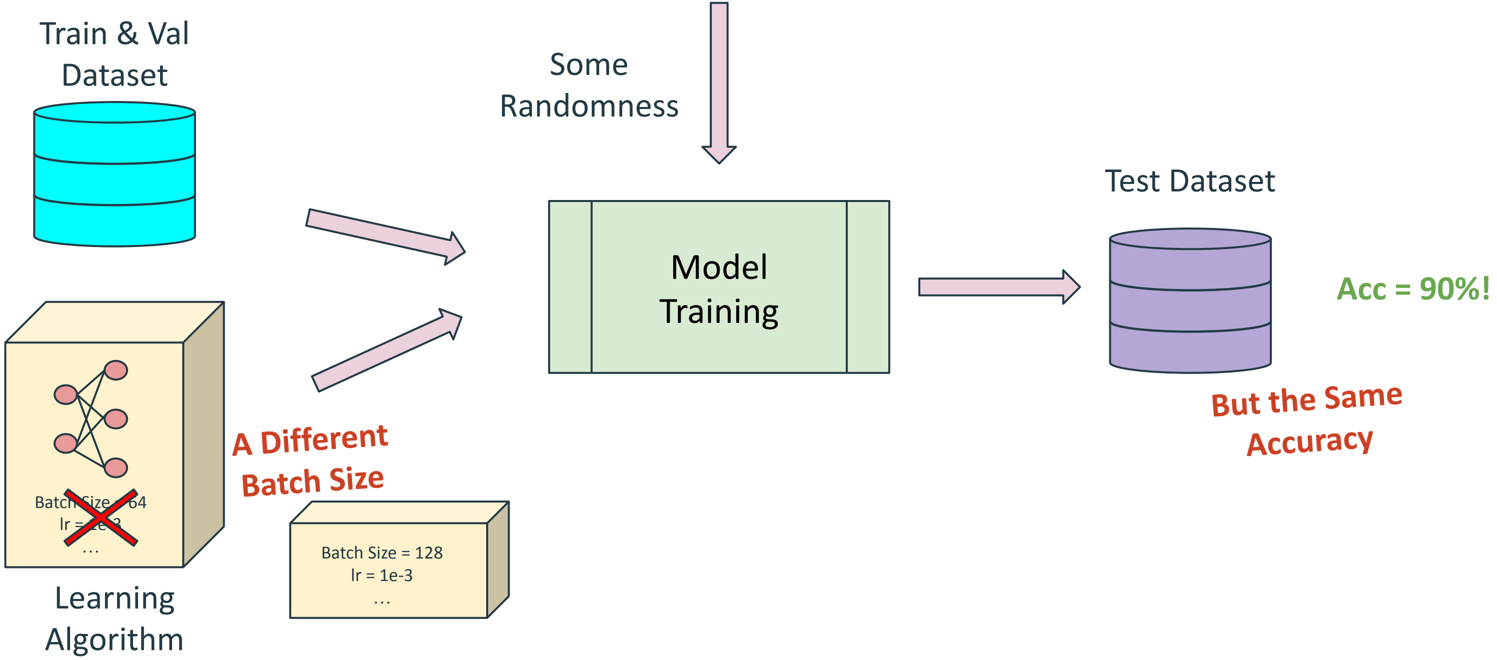
Test Dataset

Acc = 90%!

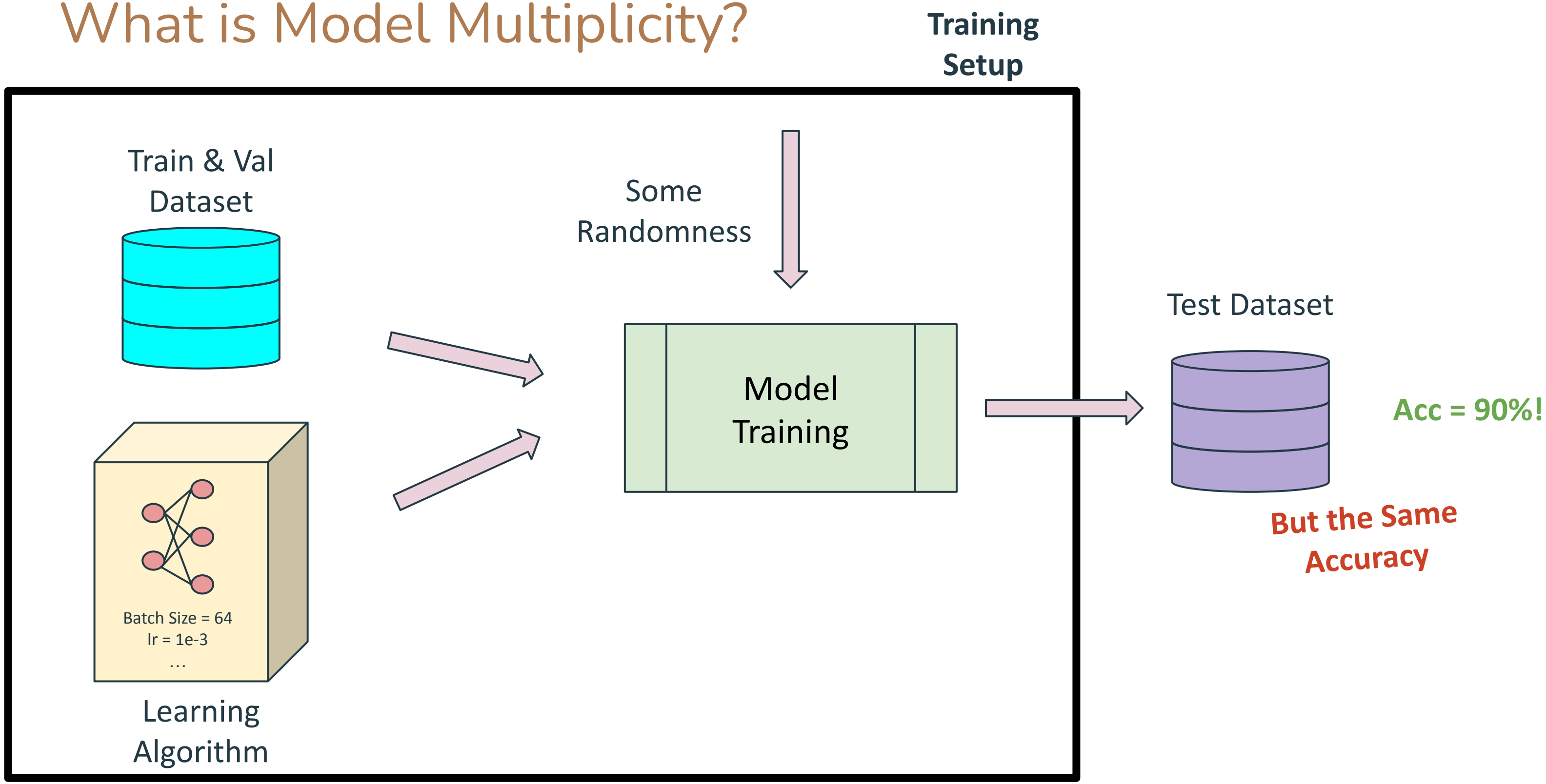
But the Same Accuracy

Learning Algorithm

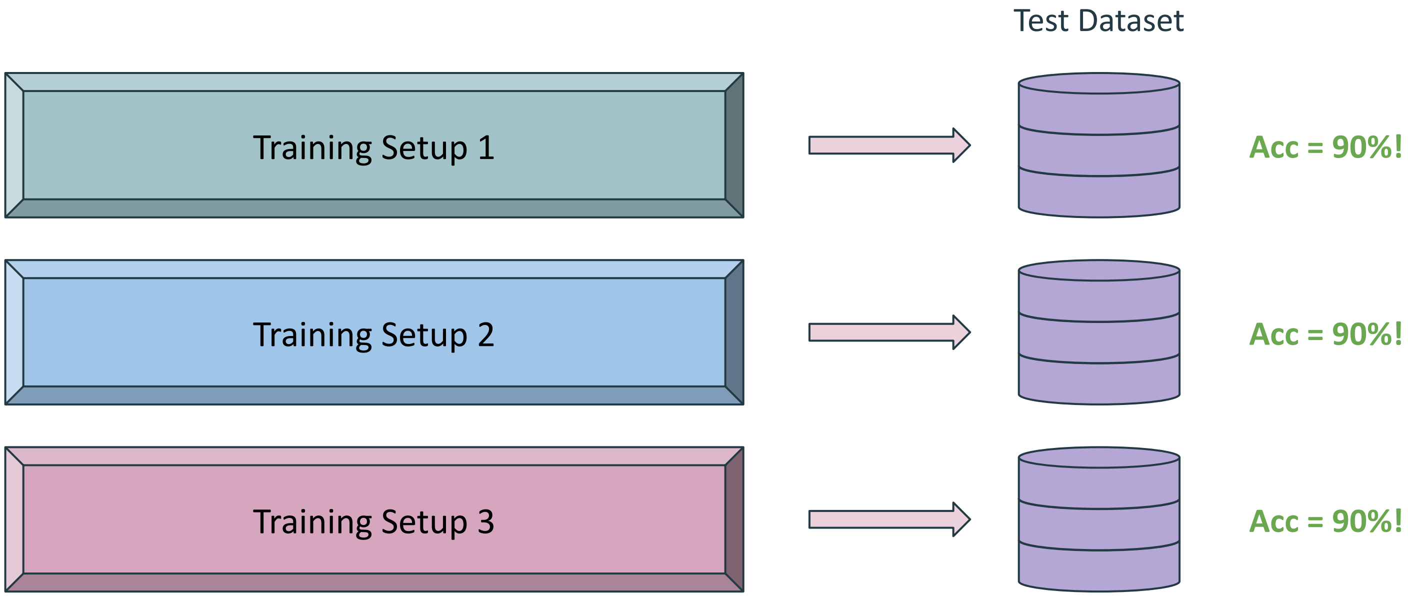
What is Model Multiplicity?



What is Model Multiplicity?



What is Model Multiplicity?



Test Dataset

Training Setup 1

Acc = 90%!

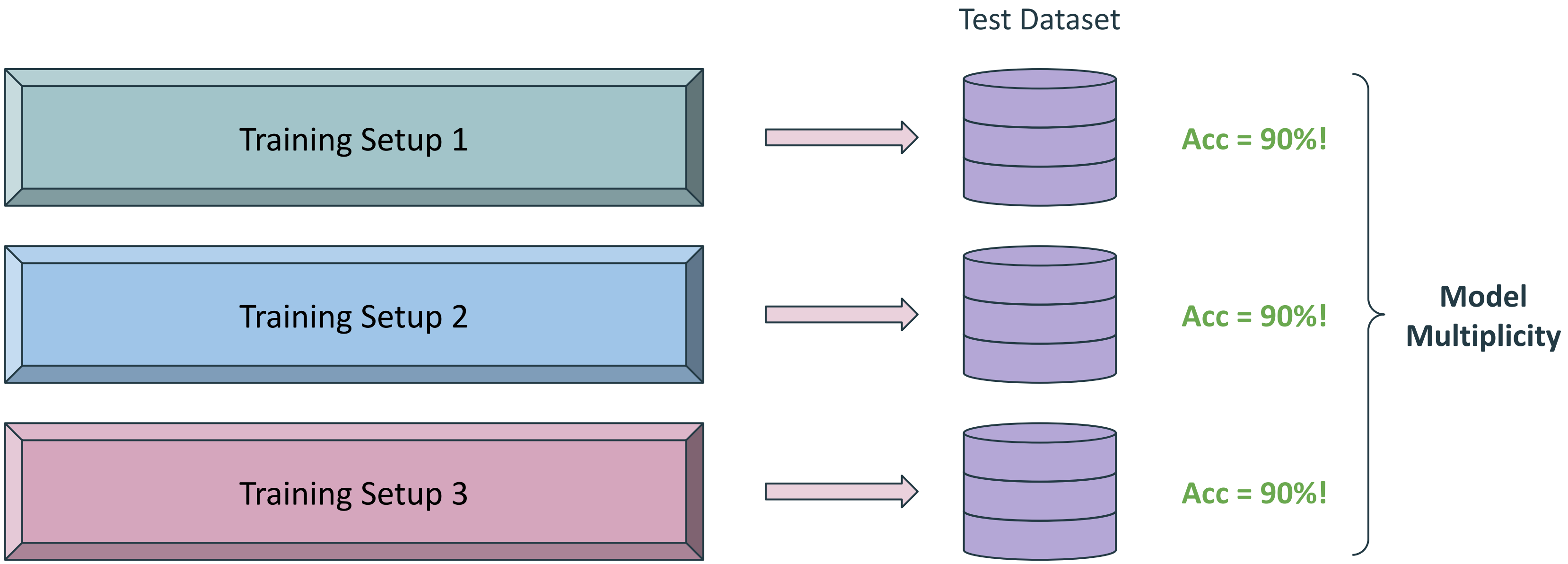
Training Setup 2

Acc = 90%!

Training Setup 3

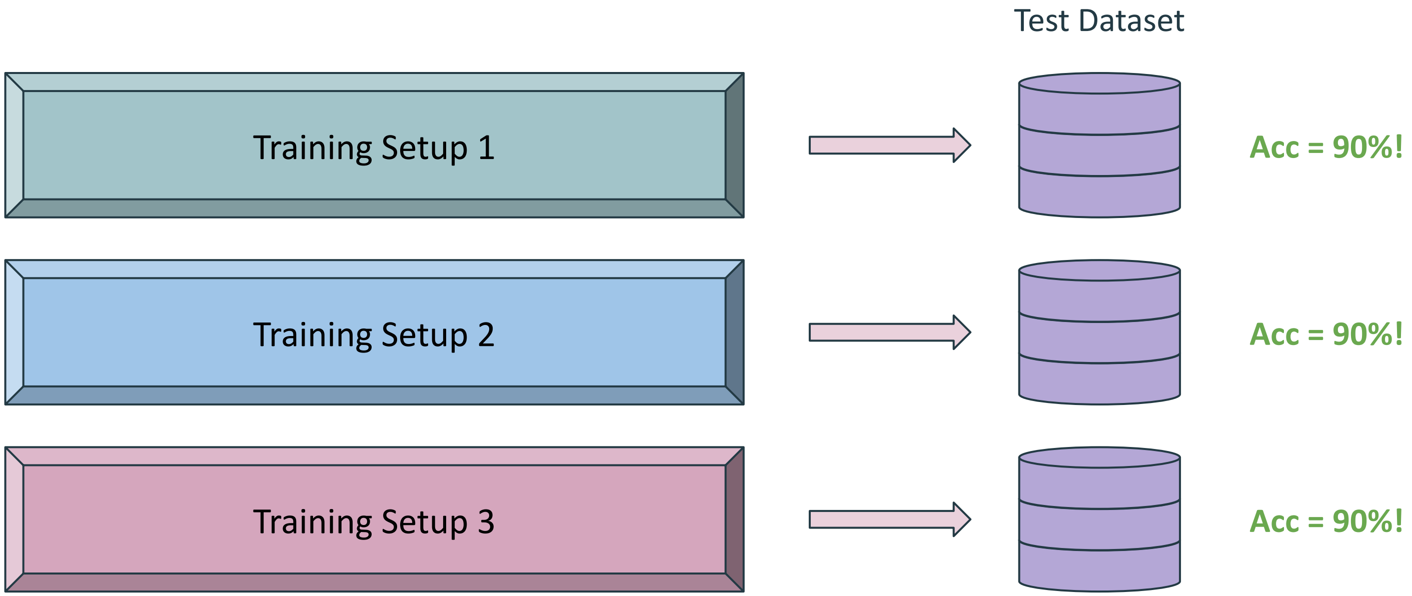
Acc = 90%!

What is Model Multiplicity?

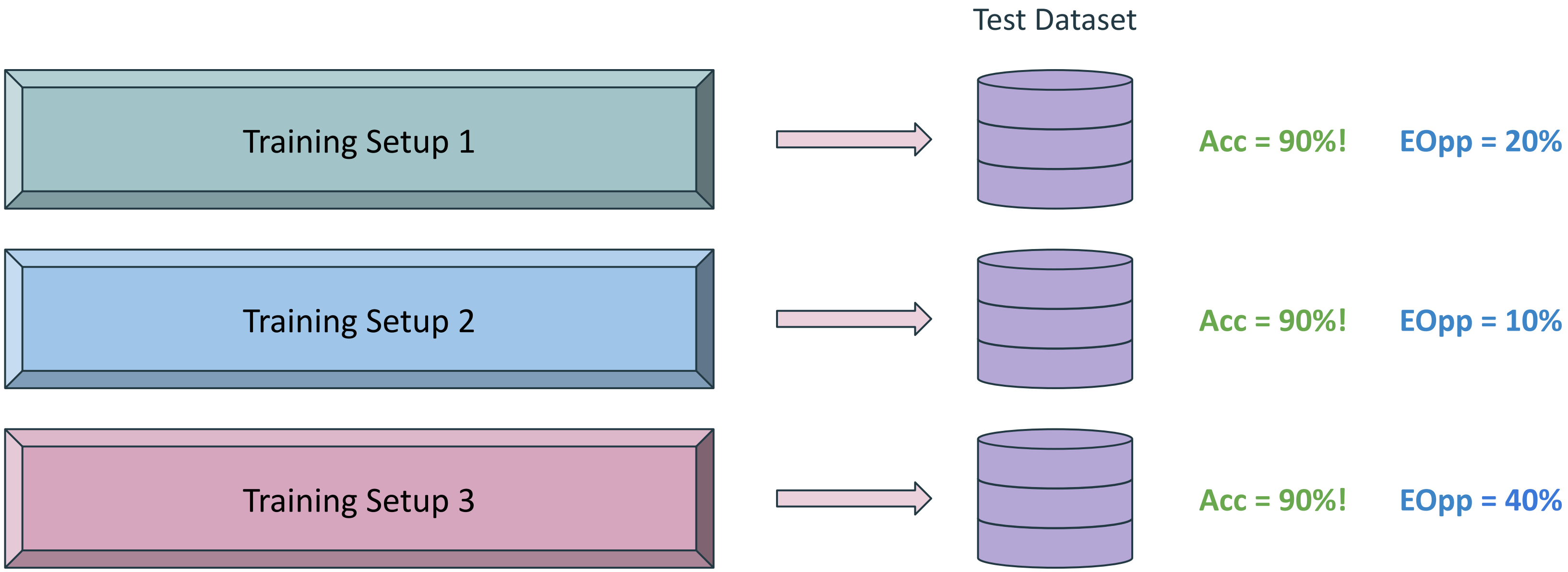


How does Model Multiplicity Impact Trustworthy ML?

How does Model Multiplicity Impact Trustworthy ML?



How does Model Multiplicity Impact Trustworthy ML?



Test Dataset

Training Setup 1

Acc = 90%! EOpp = 20%

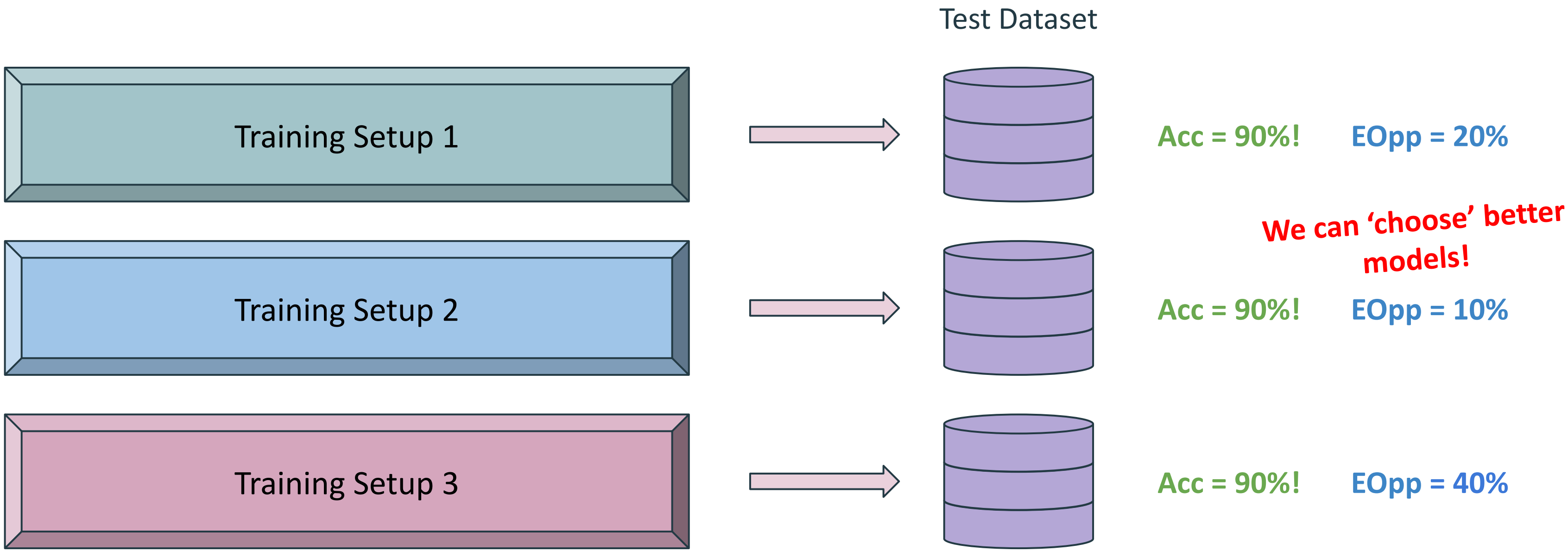
Training Setup 2

Acc = 90%! EOpp = 10%

Training Setup 3

Acc = 90%! EOpp = 40%

How does Model Multiplicity Impact Trustworthy ML?



Test Dataset

Training Setup 1 → Acc = 90%! EOpp = 20%

We can ‘choose’ better models!

Training Setup 2 → Acc = 90%! EOpp = 10%

Training Setup 3 → Acc = 90%! EOpp = 40%

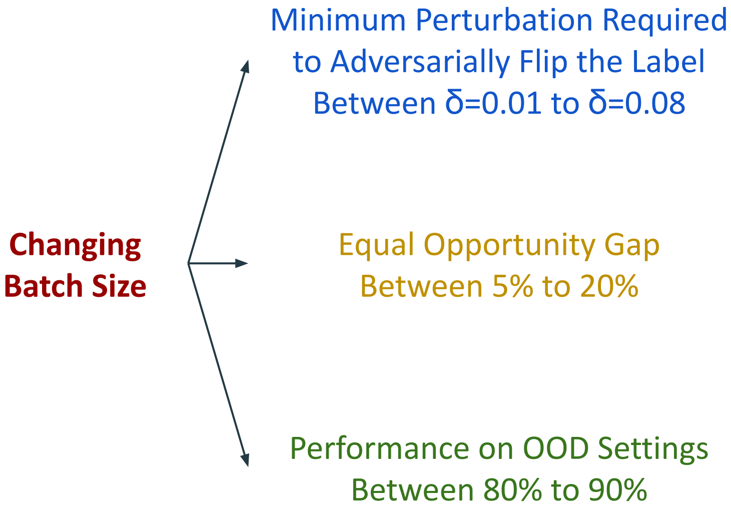
Contributions

- A standardized framework to measure and study model multiplicity.
- An empirical benchmark study of model multiplicity in image classification.
- Investigating the impact of model selection on unseen failure cases.

Contributions

- **A standardized framework to measure and study model multiplicity.**
- An empirical benchmark study of model multiplicity in image classification.
- Investigating the impact of model selection on unseen failure cases.

Accuracy Under Intervention



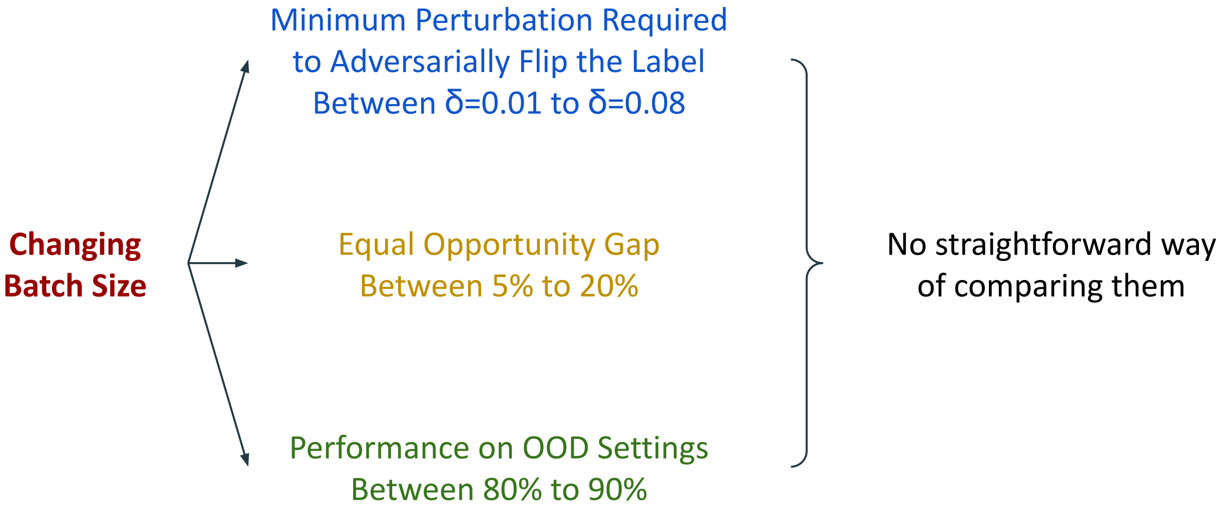
Changing Batch Size

Minimum Perturbation Required to Adversarially Flip the Label Between δ=0.01 to δ=0.08

Equal Opportunity Gap Between 5% to 20%

Performance on OOD Settings Between 80% to 90%

Accuracy Under Intervention



Changing Batch Size

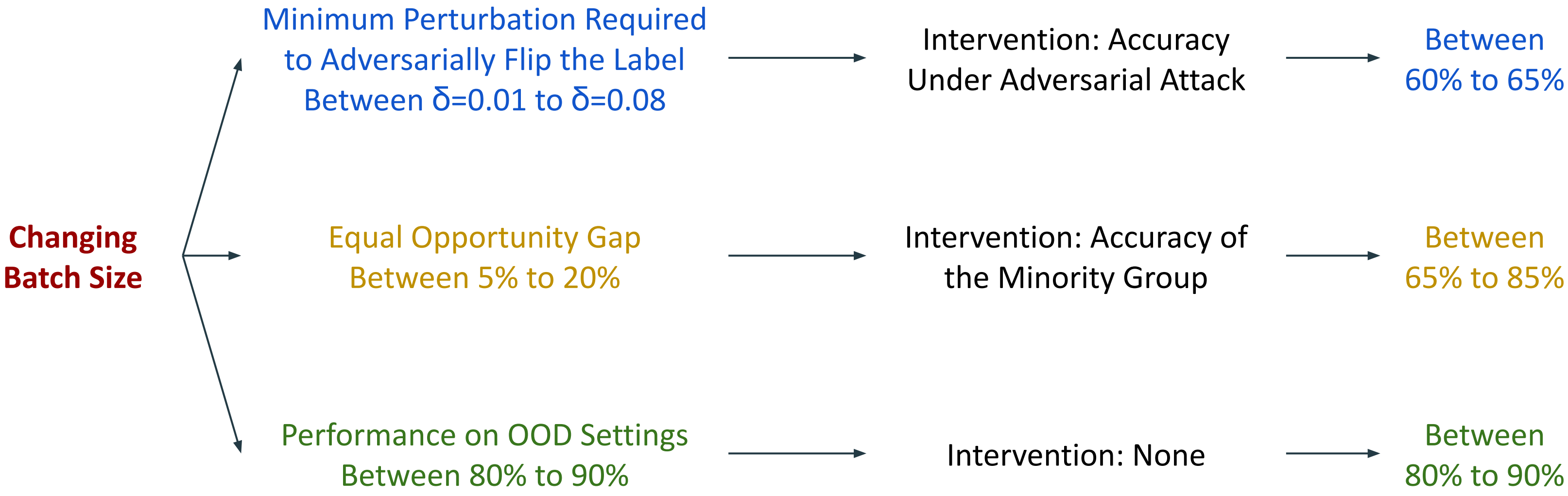
Minimum Perturbation Required to Adversarially Flip the Label Between δ=0.01 to δ=0.08

Equal Opportunity Gap Between 5% to 20%

Performance on OOD Settings Between 80% to 90%

No straightforward way of comparing them

Accuracy Under Intervention



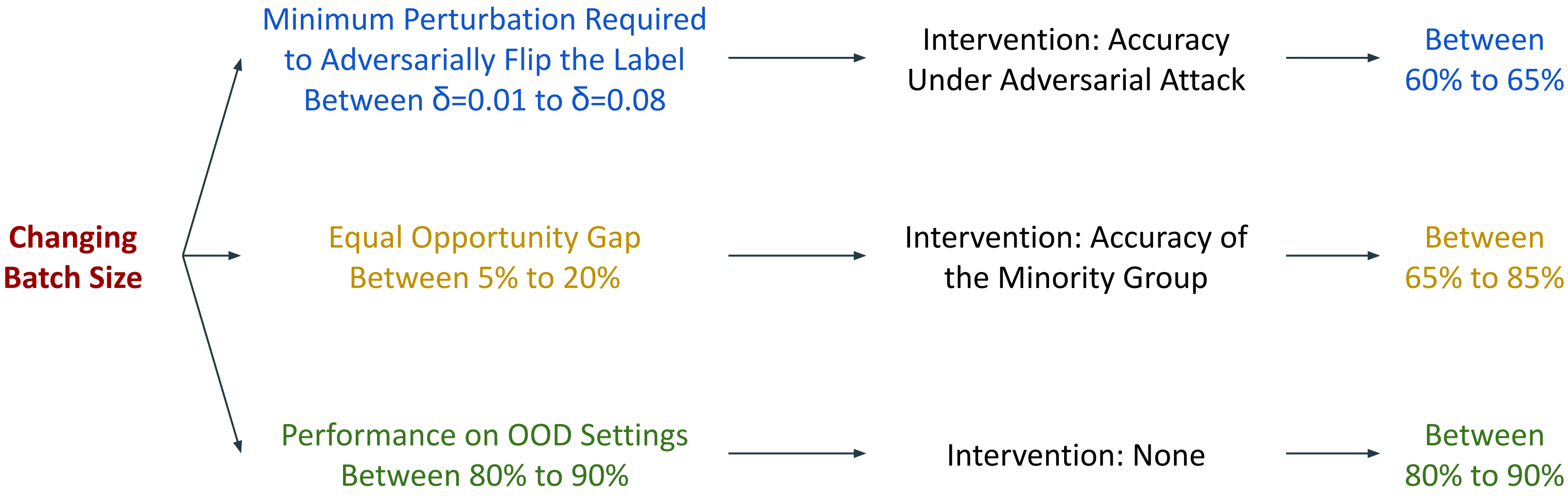
Changing Batch Size

Minimum Perturbation Required to Adversarially Flip the Label Between δ=0.01 to δ=0.08 → Intervention: Accuracy Under Adversarial Attack → Between 60% to 65%

Equal Opportunity Gap Between 5% to 20% → Intervention: Accuracy of the Minority Group → Between 65% to 85%

Performance on OOD Settings Between 80% to 90% → Intervention: None → Between 80% to 90%

Accuracy Under Intervention



Changing Batch Size

Minimum Perturbation Required to Adversarially Flip the Label Between δ=0.01 to δ=0.08 → Intervention: Accuracy Under Adversarial Attack → Between 60% to 65%

Equal Opportunity Gap Between 5% to 20% → Intervention: Accuracy of the Minority Group → Between 65% to 85%

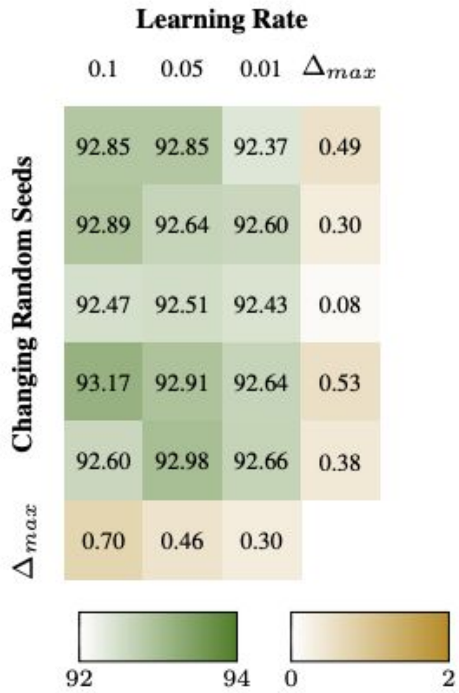
Performance on OOD Settings Between 80% to 90% → Intervention: None → Between 80% to 90%

The impact of changing batch size is a lot more severe on fairness, a little less on OOD robustness, and not that severe on adversarial robustness

Multiplicity Sheets

Multiplicity Sheets

Multiplicity Sheets

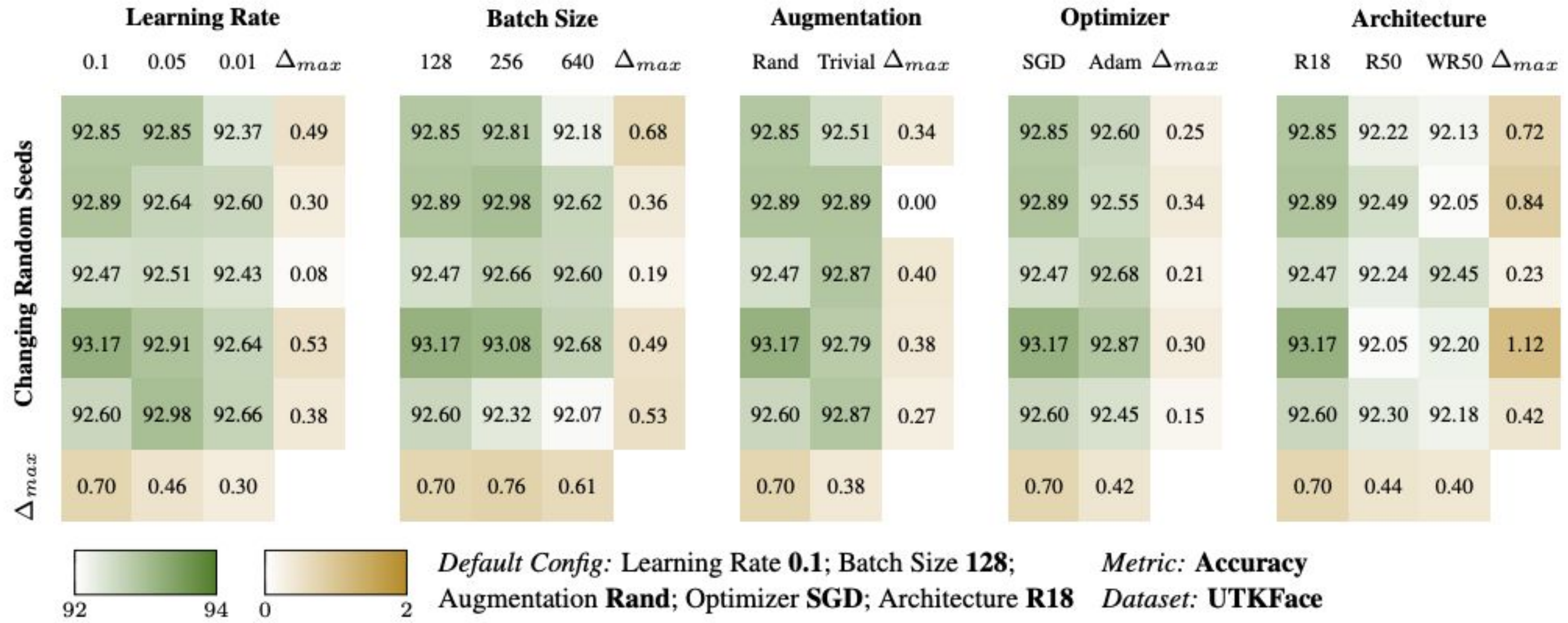


	Learning Rate			
	0.1	0.05	0.01	Δ_{max}
Changing Random Seeds	92.85	92.85	92.37	0.49
	92.89	92.64	92.60	0.30
	92.47	92.51	92.43	0.08
	93.17	92.91	92.64	0.53
	92.60	92.98	92.66	0.38
Δ_{max}	0.70	0.46	0.30	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **Accuracy**
Dataset: **UTKFace**

Multiplicity Sheets

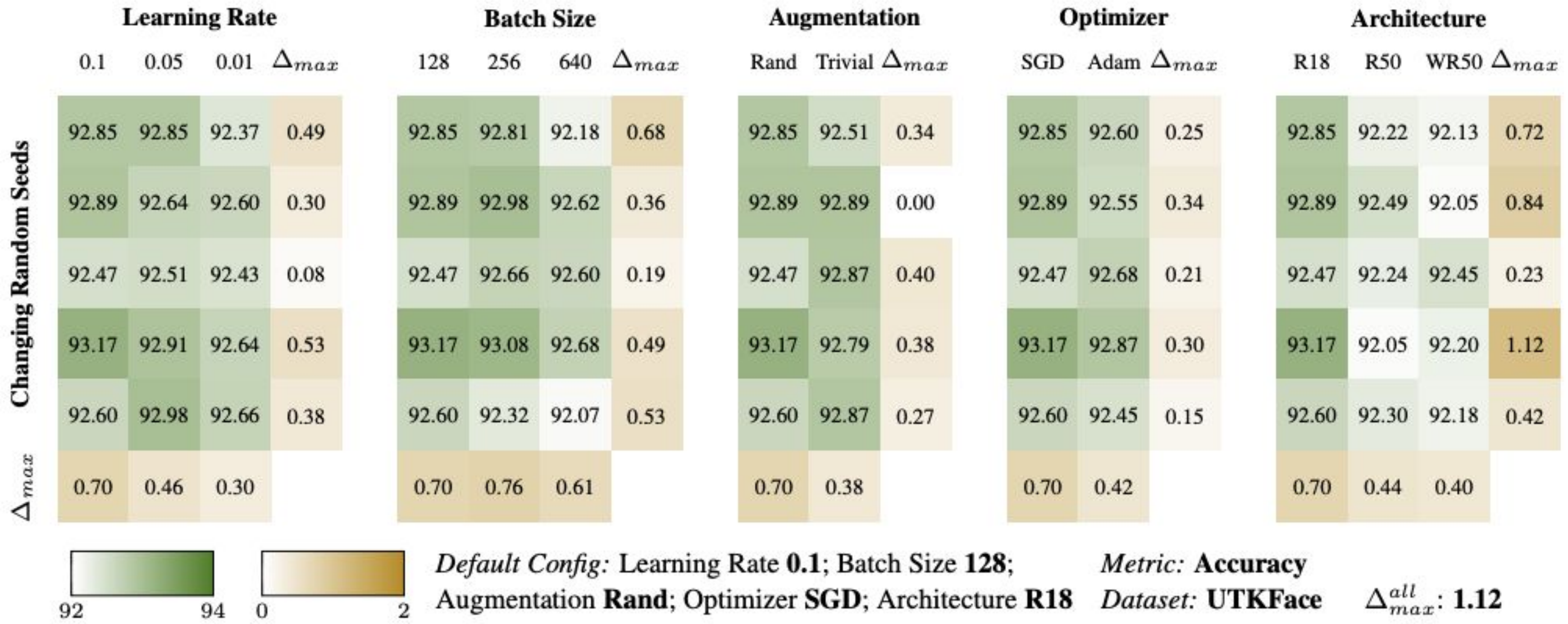


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	92.85	92.85	92.37	0.49	92.85	92.81	92.18	0.68	92.85	92.51	0.34	92.85	92.60	0.25	92.85	92.22	92.13	0.72
	92.89	92.64	92.60	0.30	92.89	92.98	92.62	0.36	92.89	92.89	0.00	92.89	92.55	0.34	92.89	92.49	92.05	0.84
	92.47	92.51	92.43	0.08	92.47	92.66	92.60	0.19	92.47	92.87	0.40	92.47	92.68	0.21	92.47	92.24	92.45	0.23
	93.17	92.91	92.64	0.53	93.17	93.08	92.68	0.49	93.17	92.79	0.38	93.17	92.87	0.30	93.17	92.05	92.20	1.12
	92.60	92.98	92.66	0.38	92.60	92.32	92.07	0.53	92.60	92.87	0.27	92.60	92.45	0.15	92.60	92.30	92.18	0.42
Δ_{max}	0.70	0.46	0.30		0.70	0.76	0.61		0.70	0.38		0.70	0.42		0.70	0.44	0.40	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **Accuracy**
Dataset: **UTKFace**

Multiplicity Sheets



	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	92.85	92.85	92.37	0.49	92.85	92.81	92.18	0.68	92.85	92.51	0.34	92.85	92.60	0.25	92.85	92.22	92.13	0.72
	92.89	92.64	92.60	0.30	92.89	92.98	92.62	0.36	92.89	92.89	0.00	92.89	92.55	0.34	92.89	92.49	92.05	0.84
	92.47	92.51	92.43	0.08	92.47	92.66	92.60	0.19	92.47	92.87	0.40	92.47	92.68	0.21	92.47	92.24	92.45	0.23
	93.17	92.91	92.64	0.53	93.17	93.08	92.68	0.49	93.17	92.79	0.38	93.17	92.87	0.30	93.17	92.05	92.20	1.12
	92.60	92.98	92.66	0.38	92.60	92.32	92.07	0.53	92.60	92.87	0.27	92.60	92.45	0.15	92.60	92.30	92.18	0.42
Δ_{max}	0.70	0.46	0.30		0.70	0.76	0.61		0.70	0.38		0.70	0.42		0.70	0.44	0.40	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

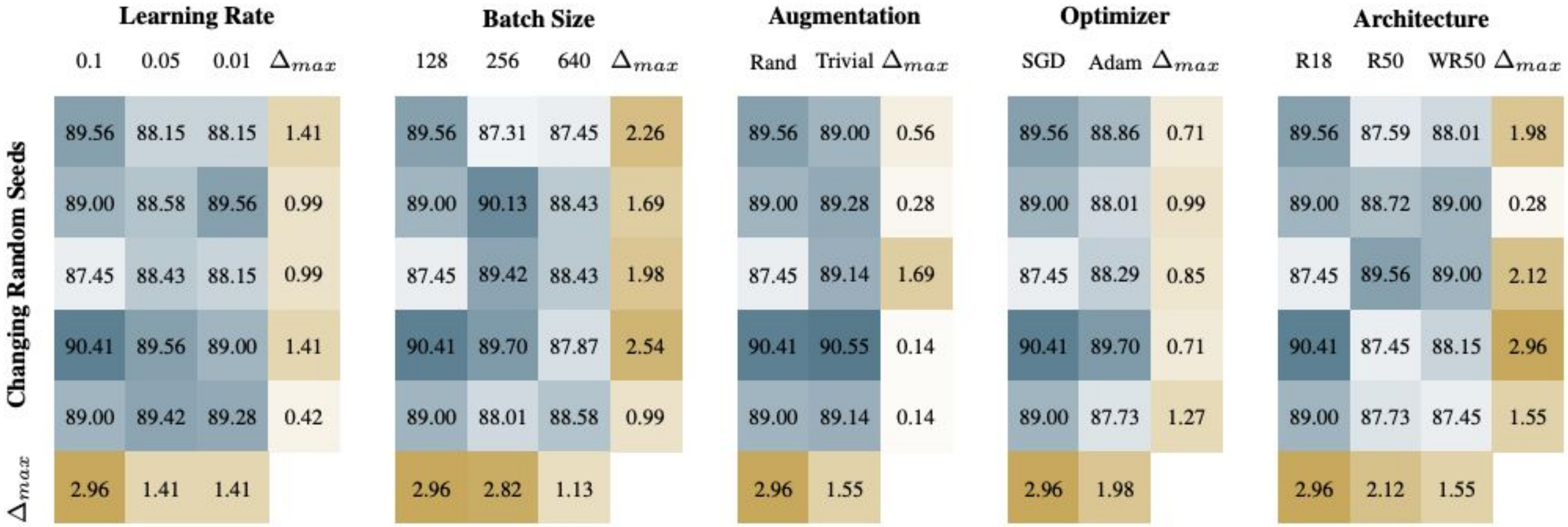
Metric: **Accuracy**
Dataset: **UTKFace**

Δ^{all}_{max}: **1.12**

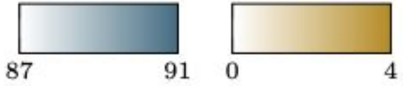
Contributions

- A standardized framework to measure and study model multiplicity.
- **An empirical benchmark study of model multiplicity in image classification.**
- Investigating the impact of model selection on unseen failure cases.

Multiplicity Sheet: Fairness



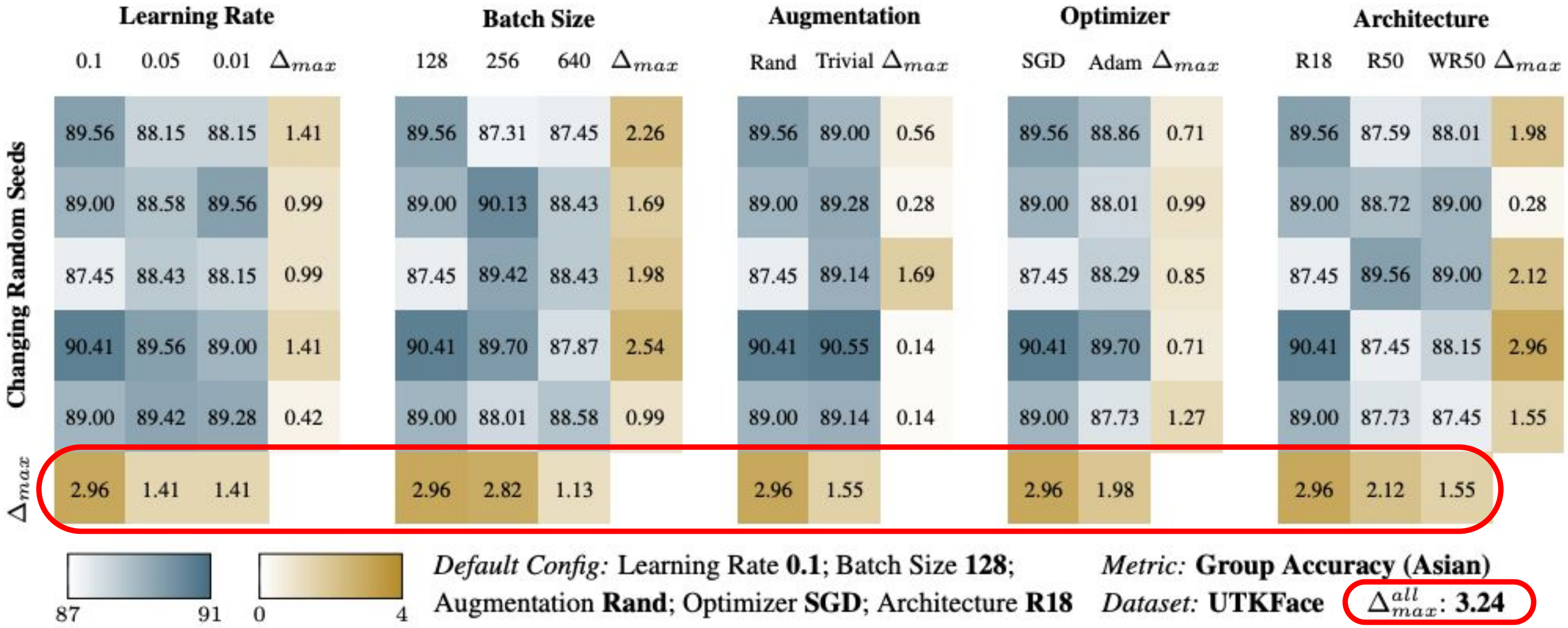
	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	89.56	88.15	88.15	1.41	89.56	87.31	87.45	2.26	89.56	89.00	0.56	89.56	88.86	0.71	89.56	87.59	88.01	1.98
	89.00	88.58	89.56	0.99	89.00	90.13	88.43	1.69	89.00	89.28	0.28	89.00	88.01	0.99	89.00	88.72	89.00	0.28
	87.45	88.43	88.15	0.99	87.45	89.42	88.43	1.98	87.45	89.14	1.69	87.45	88.29	0.85	87.45	89.56	89.00	2.12
	90.41	89.56	89.00	1.41	90.41	89.70	87.87	2.54	90.41	90.55	0.14	90.41	89.70	0.71	90.41	87.45	88.15	2.96
	89.00	89.42	89.28	0.42	89.00	88.01	88.58	0.99	89.00	89.14	0.14	89.00	87.73	1.27	89.00	87.73	87.45	1.55
Δ_{max}	2.96	1.41	1.41		2.96	2.82	1.13		2.96	1.55		2.96	1.98		2.96	2.12	1.55	



Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **Group Accuracy (Asian)**
Dataset: **UTKFace** Δ_{max}^{all}: **3.24**

Multiplicity Sheet: Fairness



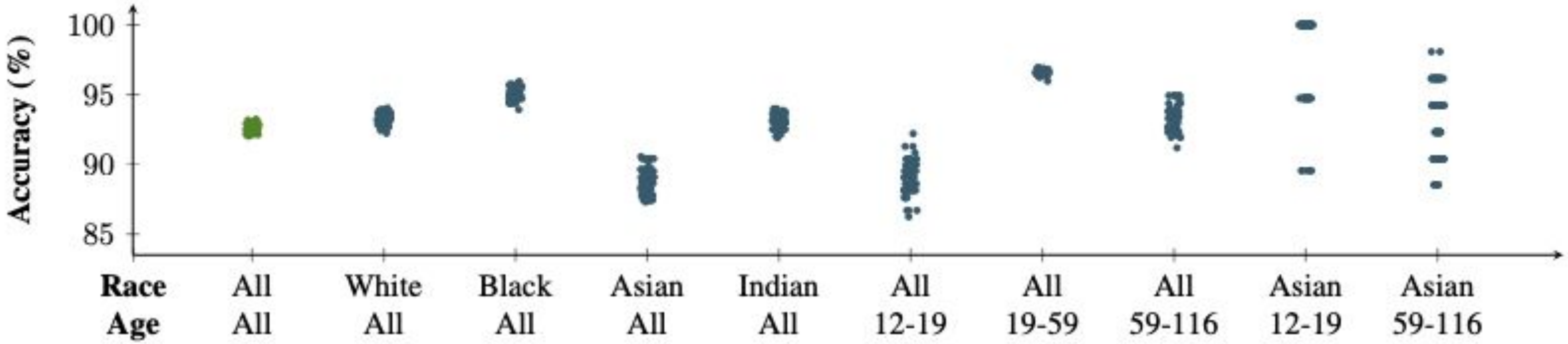
	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	89.56	88.15	88.15	1.41	89.56	87.31	87.45	2.26	89.56	89.00	0.56	89.56	88.86	0.71	89.56	87.59	88.01	1.98
	89.00	88.58	89.56	0.99	89.00	90.13	88.43	1.69	89.00	89.28	0.28	89.00	88.01	0.99	89.00	88.72	89.00	0.28
	87.45	88.43	88.15	0.99	87.45	89.42	88.43	1.98	87.45	89.14	1.69	87.45	88.29	0.85	87.45	89.56	89.00	2.12
	90.41	89.56	89.00	1.41	90.41	89.70	87.87	2.54	90.41	90.55	0.14	90.41	89.70	0.71	90.41	87.45	88.15	2.96
	89.00	89.42	89.28	0.42	89.00	88.01	88.58	0.99	89.00	89.14	0.14	89.00	87.73	1.27	89.00	87.73	87.45	1.55
Δ_{max}	2.96	1.41	1.41		2.96	2.82	1.13		2.96	1.55		2.96	1.98		2.96	2.12	1.55	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

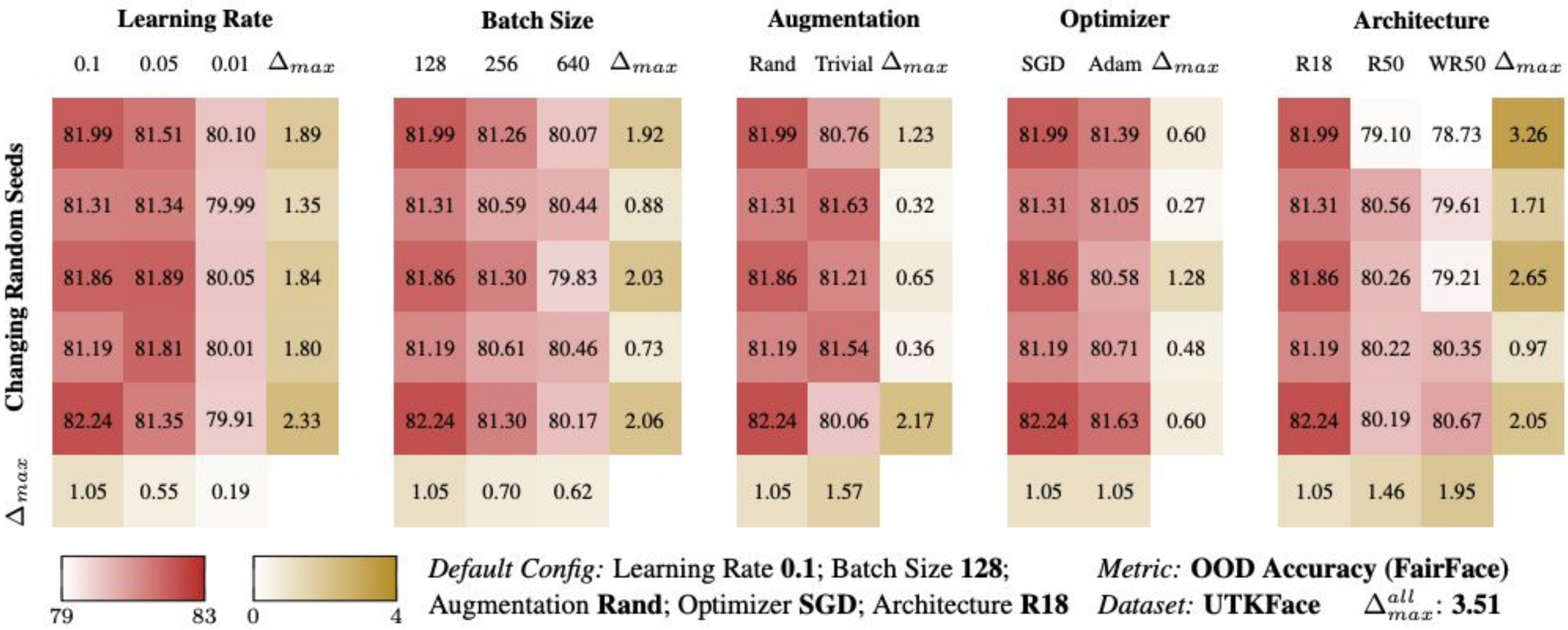
Metric: **Group Accuracy (Asian)**

Dataset: **UTKFace** Δ_{max}^{all}: **3.24**

Intersectionality and Multiplicity



Multiplicity Sheet: Robustness

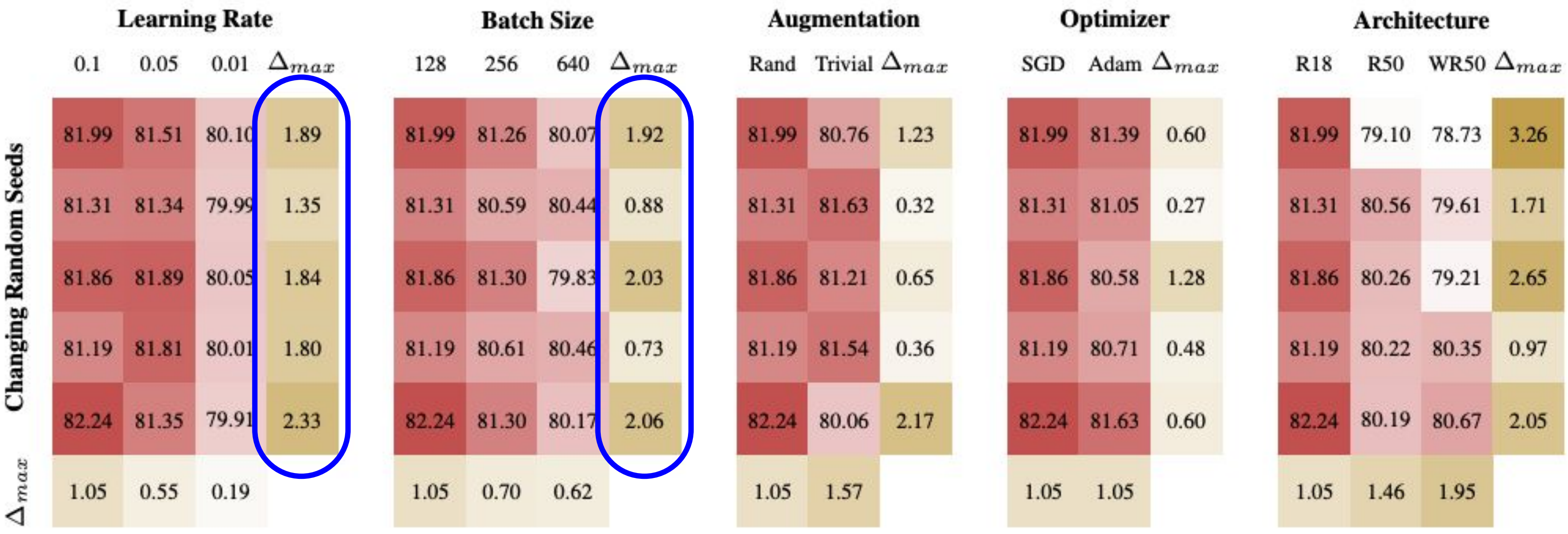


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	81.99	81.51	80.10	1.89	81.99	81.26	80.07	1.92	81.99	80.76	1.23	81.99	81.39	0.60	81.99	79.10	78.73	3.26
	81.31	81.34	79.99	1.35	81.31	80.59	80.44	0.88	81.31	81.63	0.32	81.31	81.05	0.27	81.31	80.56	79.61	1.71
	81.86	81.89	80.05	1.84	81.86	81.30	79.83	2.03	81.86	81.21	0.65	81.86	80.58	1.28	81.86	80.26	79.21	2.65
	81.19	81.81	80.01	1.80	81.19	80.61	80.46	0.73	81.19	81.54	0.36	81.19	80.71	0.48	81.19	80.22	80.35	0.97
	82.24	81.35	79.91	2.33	82.24	81.30	80.17	2.06	82.24	80.06	2.17	82.24	81.63	0.60	82.24	80.19	80.67	2.05
Δ_{max}	1.05	0.55	0.19		1.05	0.70	0.62		1.05	1.57		1.05	1.05		1.05	1.46	1.95	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **OOD Accuracy (FairFace)**
Dataset: **UTKFace** Δ_{max}^{all}: **3.51**

Multiplicity Sheet: Robustness

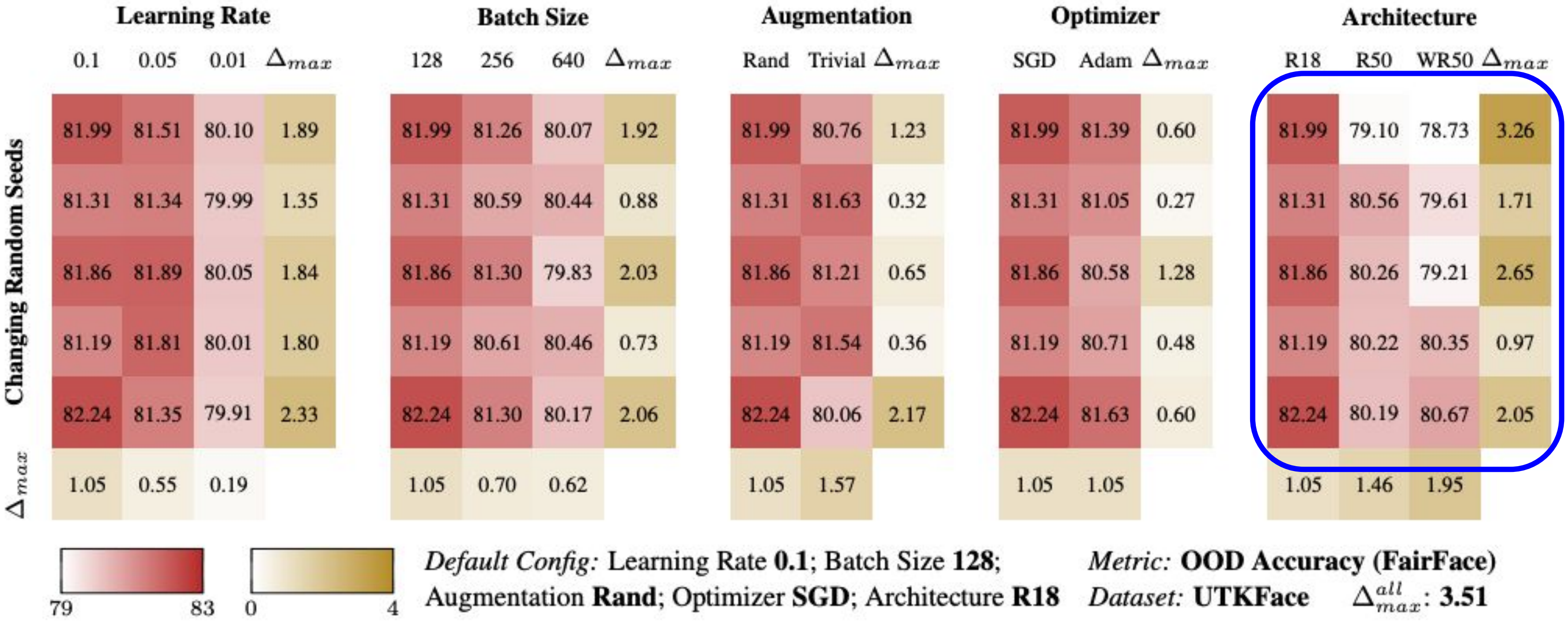


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	81.99	81.51	80.10	1.89	81.99	81.26	80.07	1.92	81.99	80.76	1.23	81.99	81.39	0.60	81.99	79.10	78.73	3.26
	81.31	81.34	79.99	1.35	81.31	80.59	80.44	0.88	81.31	81.63	0.32	81.31	81.05	0.27	81.31	80.56	79.61	1.71
	81.86	81.89	80.05	1.84	81.86	81.30	79.83	2.03	81.86	81.21	0.65	81.86	80.58	1.28	81.86	80.26	79.21	2.65
	81.19	81.81	80.01	1.80	81.19	80.61	80.46	0.73	81.19	81.54	0.36	81.19	80.71	0.48	81.19	80.22	80.35	0.97
	82.24	81.35	79.91	2.33	82.24	81.30	80.17	2.06	82.24	80.06	2.17	82.24	81.63	0.60	82.24	80.19	80.67	2.05
Δ_{max}	1.05	0.55	0.19		1.05	0.70	0.62		1.05	1.57		1.05	1.05		1.05	1.46	1.95	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **OOD Accuracy (FairFace)**
Dataset: **UTKFace** Δ^{all}_{max}: **3.51**

Multiplicity Sheet: Robustness

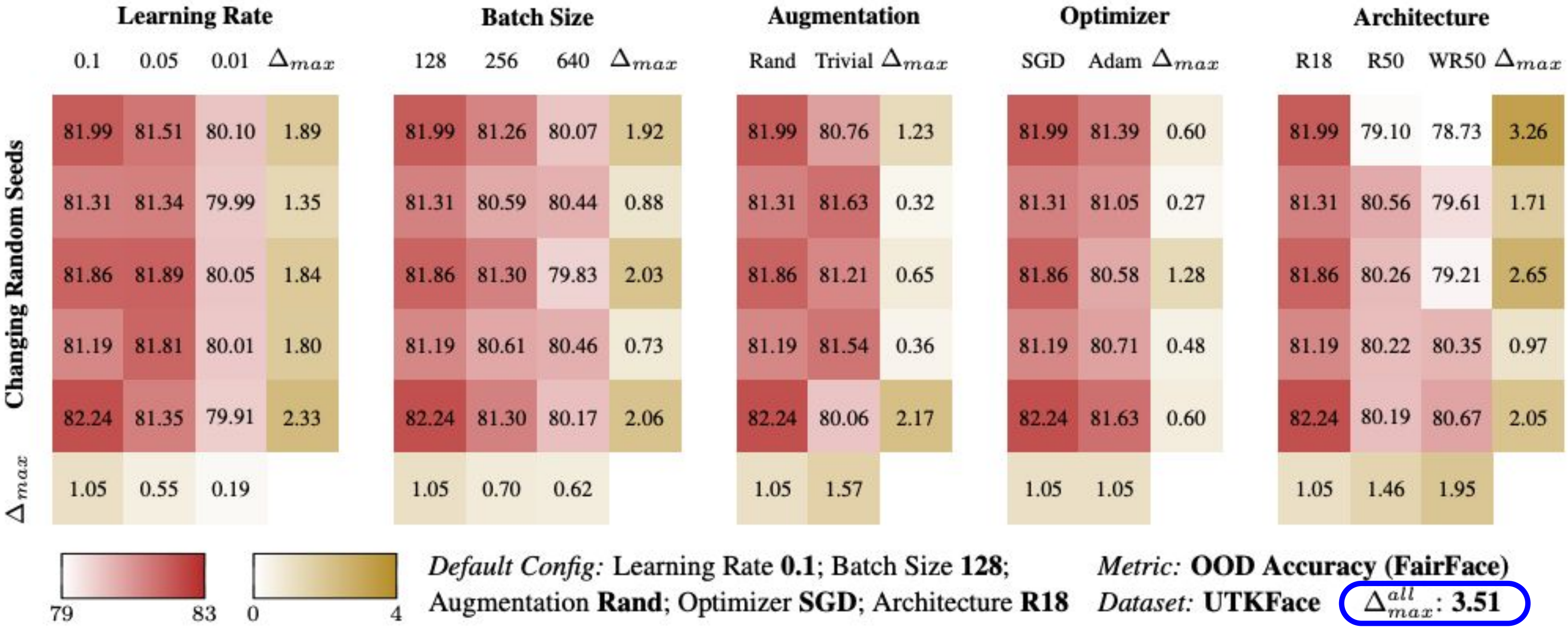


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	81.99	81.51	80.10	1.89	81.99	81.26	80.07	1.92	81.99	80.76	1.23	81.99	81.39	0.60	81.99	79.10	78.73	3.26
	81.31	81.34	79.99	1.35	81.31	80.59	80.44	0.88	81.31	81.63	0.32	81.31	81.05	0.27	81.31	80.56	79.61	1.71
	81.86	81.89	80.05	1.84	81.86	81.30	79.83	2.03	81.86	81.21	0.65	81.86	80.58	1.28	81.86	80.26	79.21	2.65
	81.19	81.81	80.01	1.80	81.19	80.61	80.46	0.73	81.19	81.54	0.36	81.19	80.71	0.48	81.19	80.22	80.35	0.97
	82.24	81.35	79.91	2.33	82.24	81.30	80.17	2.06	82.24	80.06	2.17	82.24	81.63	0.60	82.24	80.19	80.67	2.05
Δ_{max}	1.05	0.55	0.19		1.05	0.70	0.62		1.05	1.57		1.05	1.05		1.05	1.46	1.95	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **OOD Accuracy (FairFace)**
Dataset: **UTKFace** Δ_{max}^{all}: **3.51**

Multiplicity Sheet: Robustness

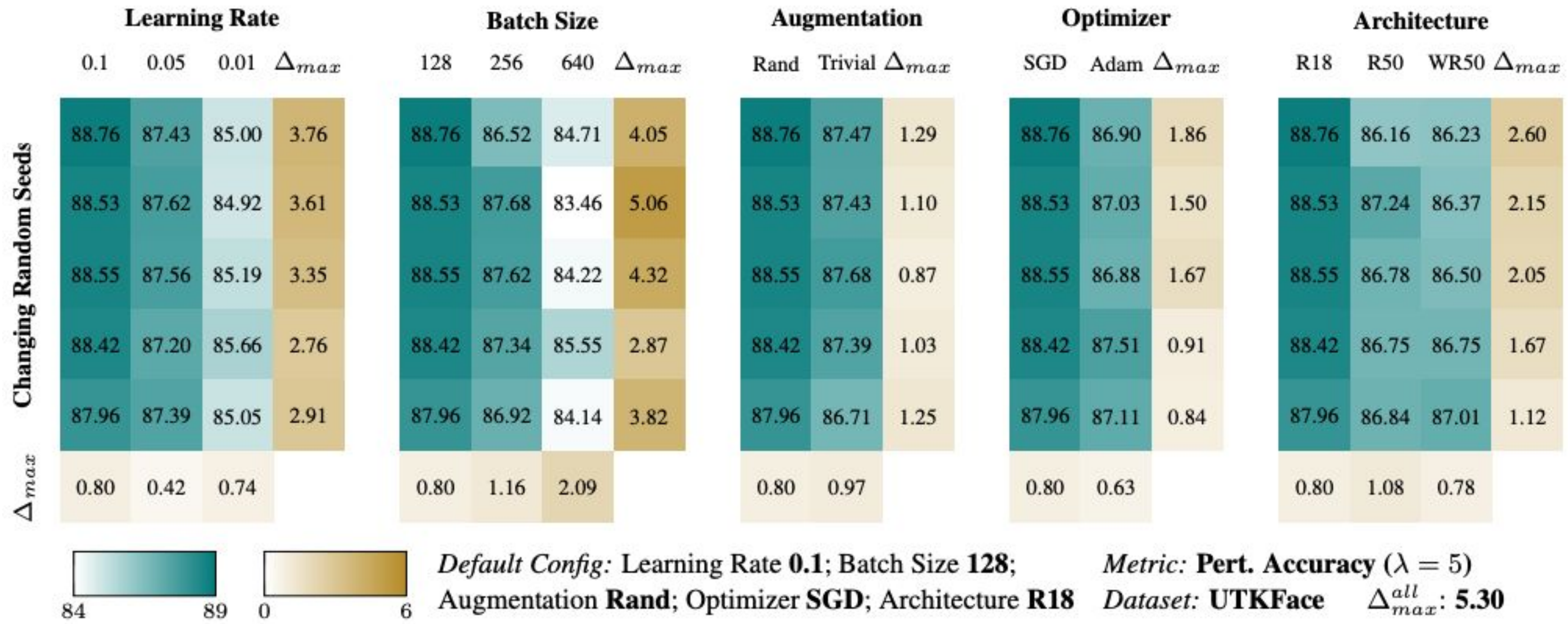


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	81.99	81.51	80.10	1.89	81.99	81.26	80.07	1.92	81.99	80.76	1.23	81.99	81.39	0.60	81.99	79.10	78.73	3.26
	81.31	81.34	79.99	1.35	81.31	80.59	80.44	0.88	81.31	81.63	0.32	81.31	81.05	0.27	81.31	80.56	79.61	1.71
	81.86	81.89	80.05	1.84	81.86	81.30	79.83	2.03	81.86	81.21	0.65	81.86	80.58	1.28	81.86	80.26	79.21	2.65
	81.19	81.81	80.01	1.80	81.19	80.61	80.46	0.73	81.19	81.54	0.36	81.19	80.71	0.48	81.19	80.22	80.35	0.97
	82.24	81.35	79.91	2.33	82.24	81.30	80.17	2.06	82.24	80.06	2.17	82.24	81.63	0.60	82.24	80.19	80.67	2.05
Δ_{max}	1.05	0.55	0.19		1.05	0.70	0.62		1.05	1.57		1.05	1.05		1.05	1.46	1.95	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **OOD Accuracy (FairFace)**
Dataset: **UTKFace** Δ^{all}_{max}: **3.51**

Multiplicity Sheet: Privacy

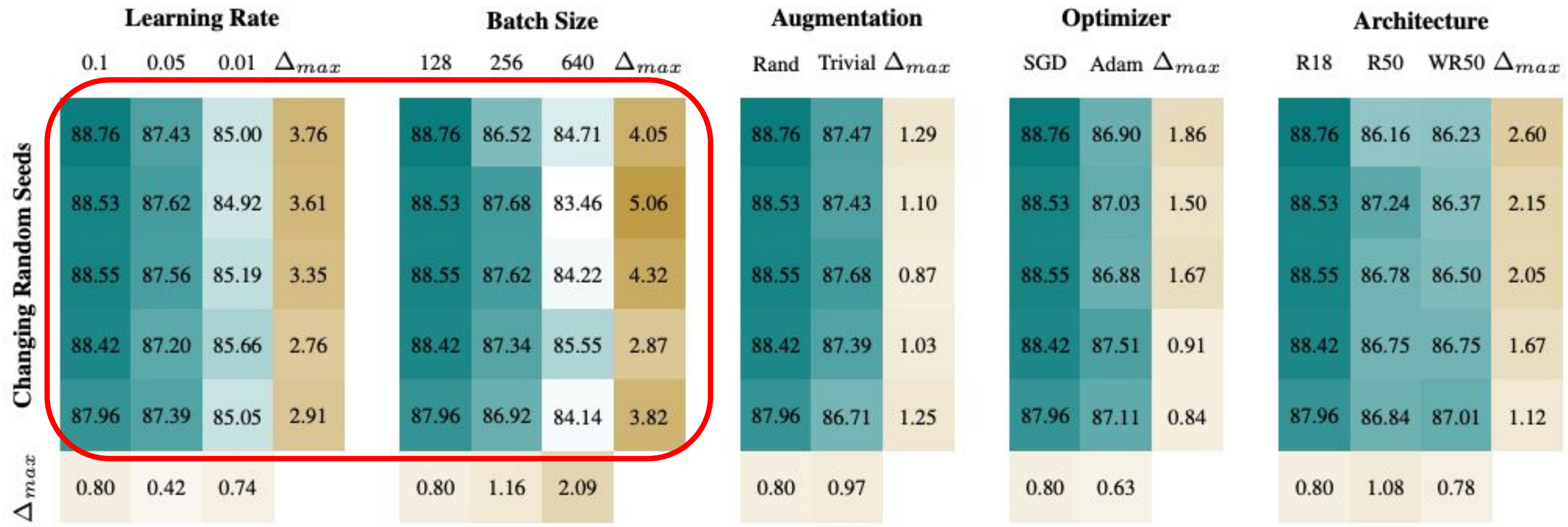


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	88.76	87.43	85.00	3.76	88.76	86.52	84.71	4.05	88.76	87.47	1.29	88.76	86.90	1.86	88.76	86.16	86.23	2.60
	88.53	87.62	84.92	3.61	88.53	87.68	83.46	5.06	88.53	87.43	1.10	88.53	87.03	1.50	88.53	87.24	86.37	2.15
	88.55	87.56	85.19	3.35	88.55	87.62	84.22	4.32	88.55	87.68	0.87	88.55	86.88	1.67	88.55	86.78	86.50	2.05
	88.42	87.20	85.66	2.76	88.42	87.34	85.55	2.87	88.42	87.39	1.03	88.42	87.51	0.91	88.42	86.75	86.75	1.67
	87.96	87.39	85.05	2.91	87.96	86.92	84.14	3.82	87.96	86.71	1.25	87.96	87.11	0.84	87.96	86.84	87.01	1.12
Δ_{max}	0.80	0.42	0.74		0.80	1.16	2.09		0.80	0.97		0.80	0.63		0.80	1.08	0.78	

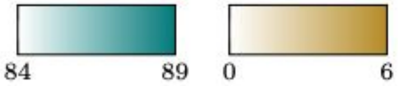
Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **Pert. Accuracy** ($\lambda = 5$)
Dataset: **UTKFace** Δ_{max}^{all}: **5.30**

Multiplicity Sheet: Privacy



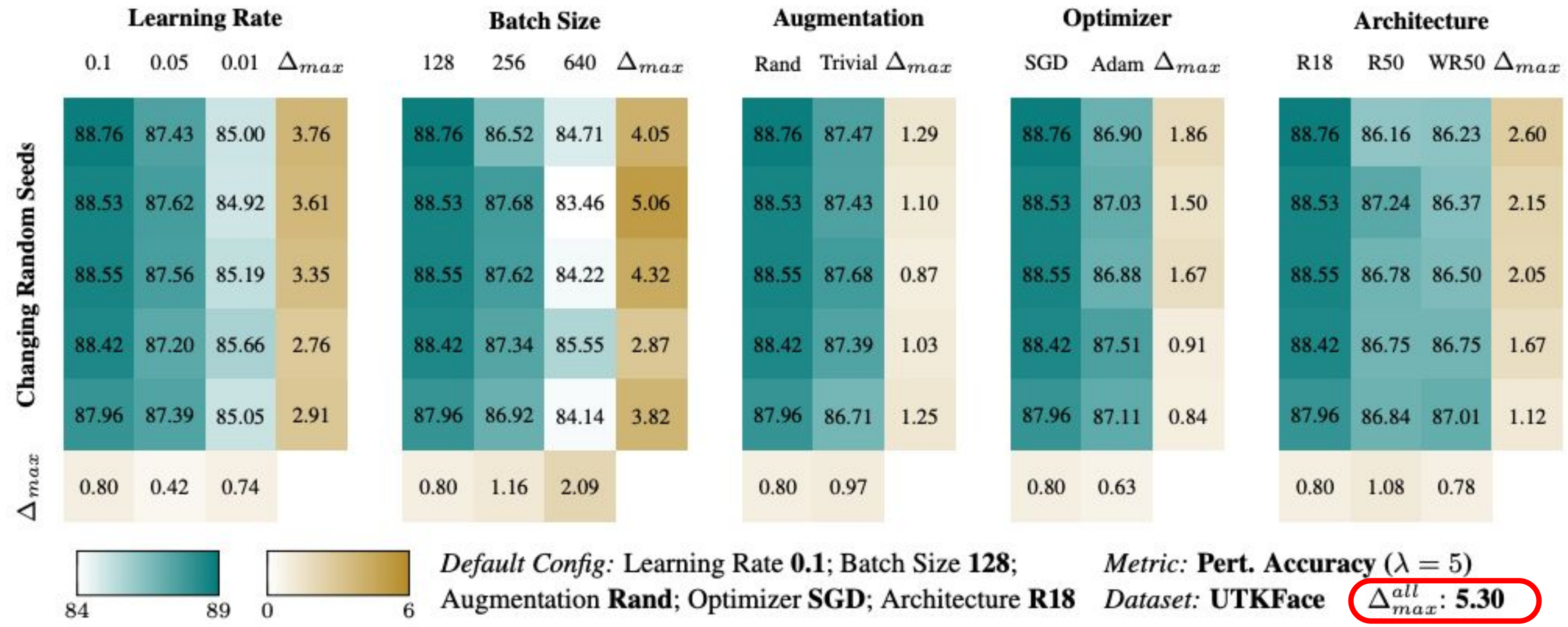
	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	88.76	87.43	85.00	3.76	88.76	86.52	84.71	4.05	88.76	87.47	1.29	88.76	86.90	1.86	88.76	86.16	86.23	2.60
	88.53	87.62	84.92	3.61	88.53	87.68	83.46	5.06	88.53	87.43	1.10	88.53	87.03	1.50	88.53	87.24	86.37	2.15
	88.55	87.56	85.19	3.35	88.55	87.62	84.22	4.32	88.55	87.68	0.87	88.55	86.88	1.67	88.55	86.78	86.50	2.05
	88.42	87.20	85.66	2.76	88.42	87.34	85.55	2.87	88.42	87.39	1.03	88.42	87.51	0.91	88.42	86.75	86.75	1.67
	87.96	87.39	85.05	2.91	87.96	86.92	84.14	3.82	87.96	86.71	1.25	87.96	87.11	0.84	87.96	86.84	87.01	1.12
Δ_{max}	0.80	0.42	0.74		0.80	1.16	2.09		0.80	0.97		0.80	0.63		0.80	1.08	0.78	



Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **Pert. Accuracy** ($\lambda = 5$)
Dataset: **UTKFace** Δ_{max}^{all}: **5.30**

Multiplicity Sheet: Privacy



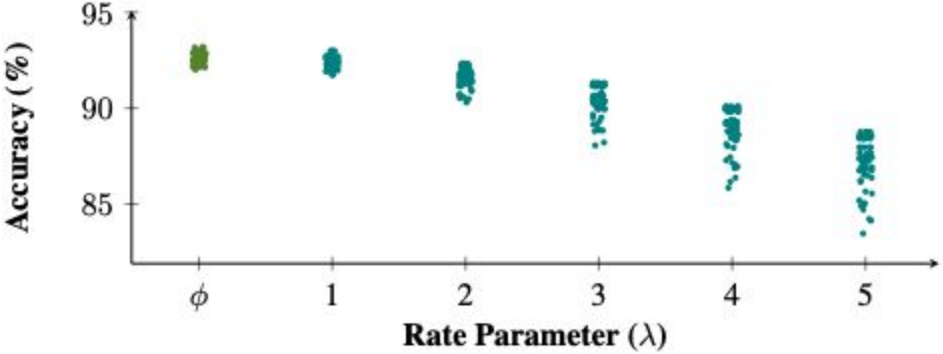
	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	88.76	87.43	85.00	3.76	88.76	86.52	84.71	4.05	88.76	87.47	1.29	88.76	86.90	1.86	88.76	86.16	86.23	2.60
	88.53	87.62	84.92	3.61	88.53	87.68	83.46	5.06	88.53	87.43	1.10	88.53	87.03	1.50	88.53	87.24	86.37	2.15
	88.55	87.56	85.19	3.35	88.55	87.62	84.22	4.32	88.55	87.68	0.87	88.55	86.88	1.67	88.55	86.78	86.50	2.05
	88.42	87.20	85.66	2.76	88.42	87.34	85.55	2.87	88.42	87.39	1.03	88.42	87.51	0.91	88.42	86.75	86.75	1.67
	87.96	87.39	85.05	2.91	87.96	86.92	84.14	3.82	87.96	86.71	1.25	87.96	87.11	0.84	87.96	86.84	87.01	1.12
Δ_{max}	0.80	0.42	0.74		0.80	1.16	2.09		0.80	0.97		0.80	0.63		0.80	1.08	0.78	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

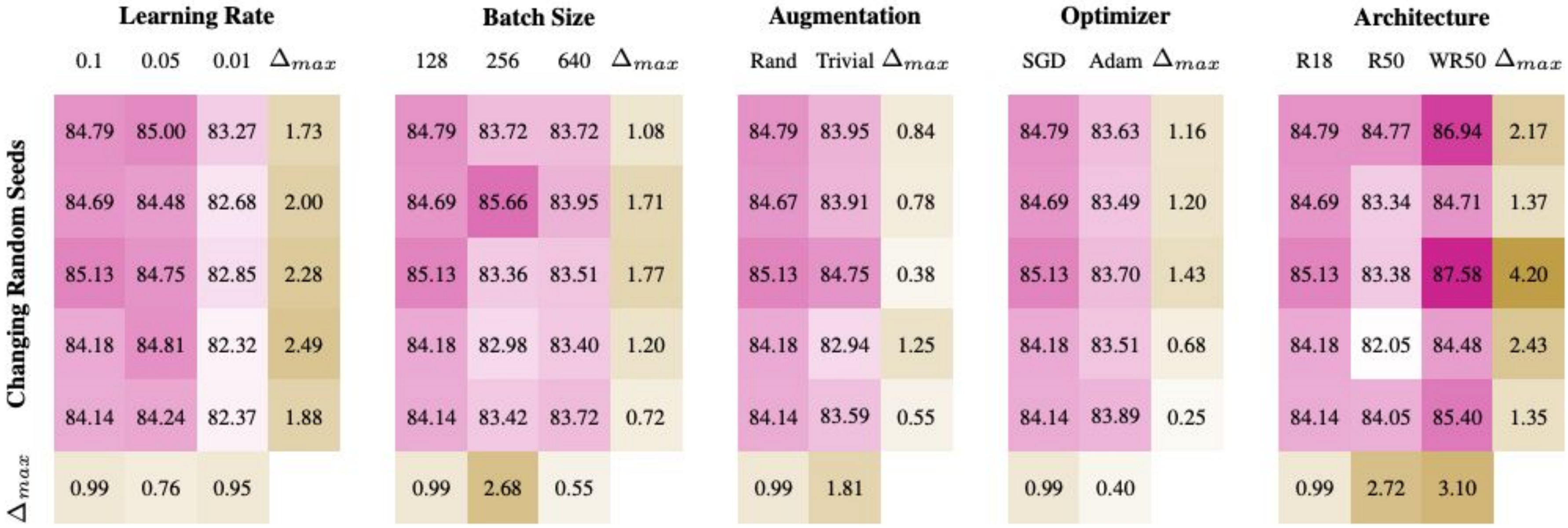
Metric: **Pert. Accuracy** ($\lambda = 5$)

Dataset: **UTKFace** Δ^{all}_{max}: **5.30**

Accuracy Under Intervention: Privacy



Multiplicity Sheet: Security



Learning Rate

	0.1	0.05	0.01	Δ_{max}
Changing Random Seeds	84.79	85.00	83.27	1.73
	84.69	84.48	82.68	2.00
	85.13	84.75	82.85	2.28
	84.18	84.81	82.32	2.49
	84.14	84.24	82.37	1.88
Δ_{max}	0.99	0.76	0.95	

Batch Size

	128	256	640	Δ_{max}
Changing Random Seeds	84.79	83.72	83.72	1.08
	84.69	85.66	83.95	1.71
	85.13	83.36	83.51	1.77
	84.18	82.98	83.40	1.20
	84.14	83.42	83.72	0.72
Δ_{max}	0.99	2.68	0.55	

Augmentation

	Rand	Trivial	Δ_{max}
Changing Random Seeds	84.79	83.95	0.84
	84.67	83.91	0.78
	85.13	84.75	0.38
	84.18	82.94	1.25
	84.14	83.59	0.55
Δ_{max}	0.99	1.81	

Optimizer

	SGD	Adam	Δ_{max}
Changing Random Seeds	84.79	83.63	1.16
	84.69	83.49	1.20
	85.13	83.70	1.43
	84.18	83.51	0.68
	84.14	83.89	0.25
Δ_{max}	0.99	0.40	

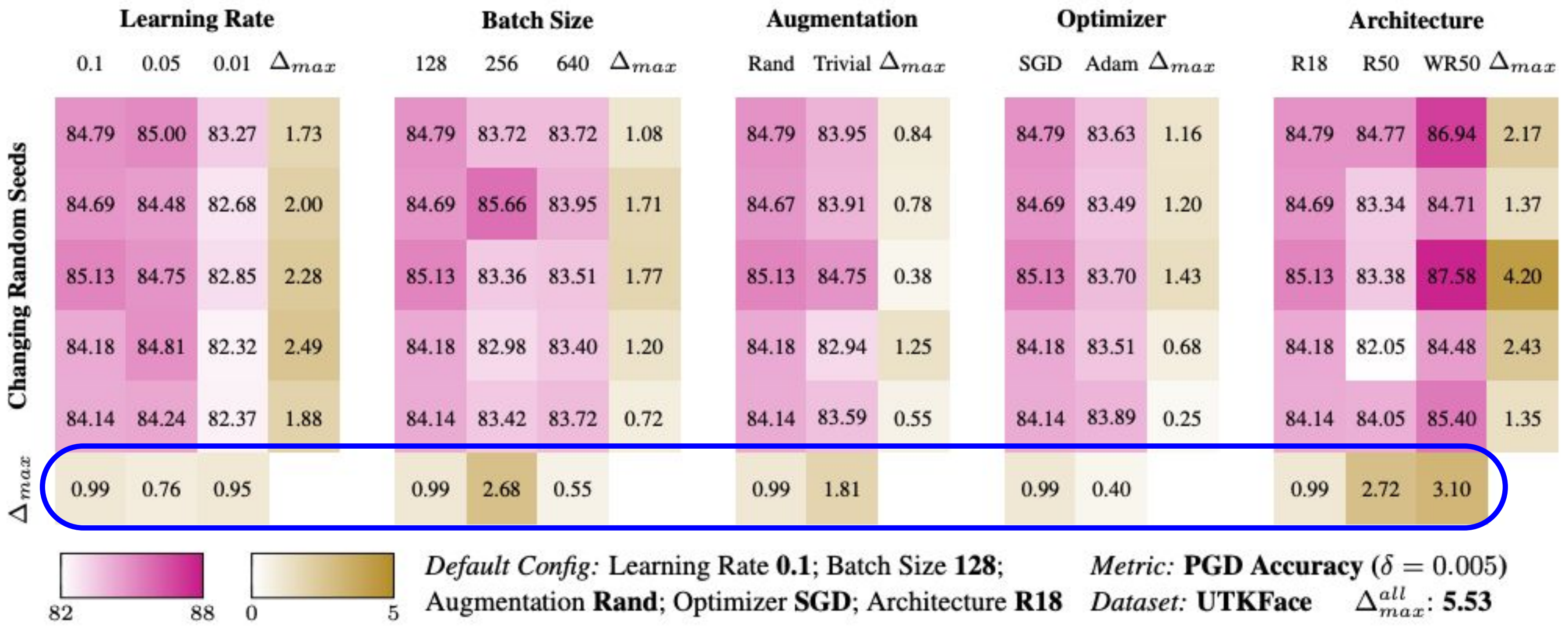
Architecture

	R18	R50	WR50	Δ_{max}
Changing Random Seeds	84.79	84.77	86.94	2.17
	84.69	83.34	84.71	1.37
	85.13	83.38	87.58	4.20
	84.18	82.05	84.48	2.43
	84.14	84.05	85.40	1.35
Δ_{max}	0.99	2.72	3.10	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **PGD Accuracy** ($\delta = 0.005$)
Dataset: **UTKFace** Δ_{max}^{all}: **5.53**

Multiplicity Sheet: Security

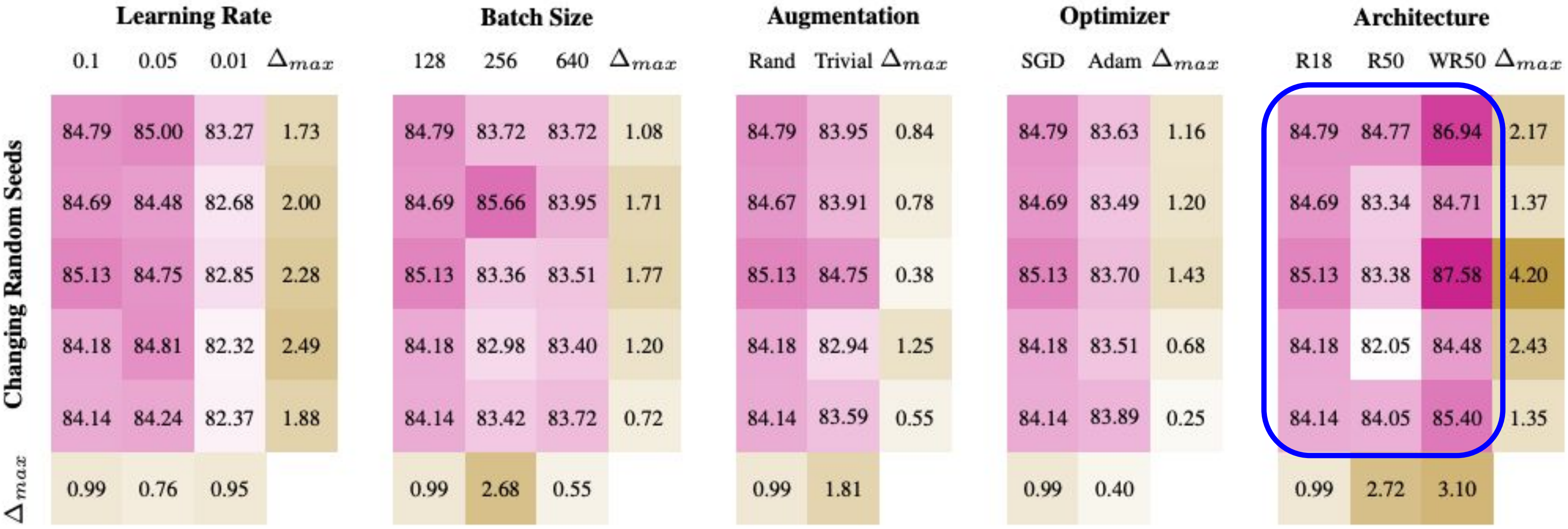


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	84.79	85.00	83.27	1.73	84.79	83.72	83.72	1.08	84.79	83.95	0.84	84.79	83.63	1.16	84.79	84.77	86.94	2.17
	84.69	84.48	82.68	2.00	84.69	85.66	83.95	1.71	84.67	83.91	0.78	84.69	83.49	1.20	84.69	83.34	84.71	1.37
	85.13	84.75	82.85	2.28	85.13	83.36	83.51	1.77	85.13	84.75	0.38	85.13	83.70	1.43	85.13	83.38	87.58	4.20
	84.18	84.81	82.32	2.49	84.18	82.98	83.40	1.20	84.18	82.94	1.25	84.18	83.51	0.68	84.18	82.05	84.48	2.43
	84.14	84.24	82.37	1.88	84.14	83.42	83.72	0.72	84.14	83.59	0.55	84.14	83.89	0.25	84.14	84.05	85.40	1.35
Δ_{max}	0.99	0.76	0.95		0.99	2.68	0.55		0.99	1.81		0.99	0.40		0.99	2.72	3.10	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **PGD Accuracy** ($\delta = 0.005$)
Dataset: **UTKFace** Δ^{all}_{max}: **5.53**

Multiplicity Sheet: Security

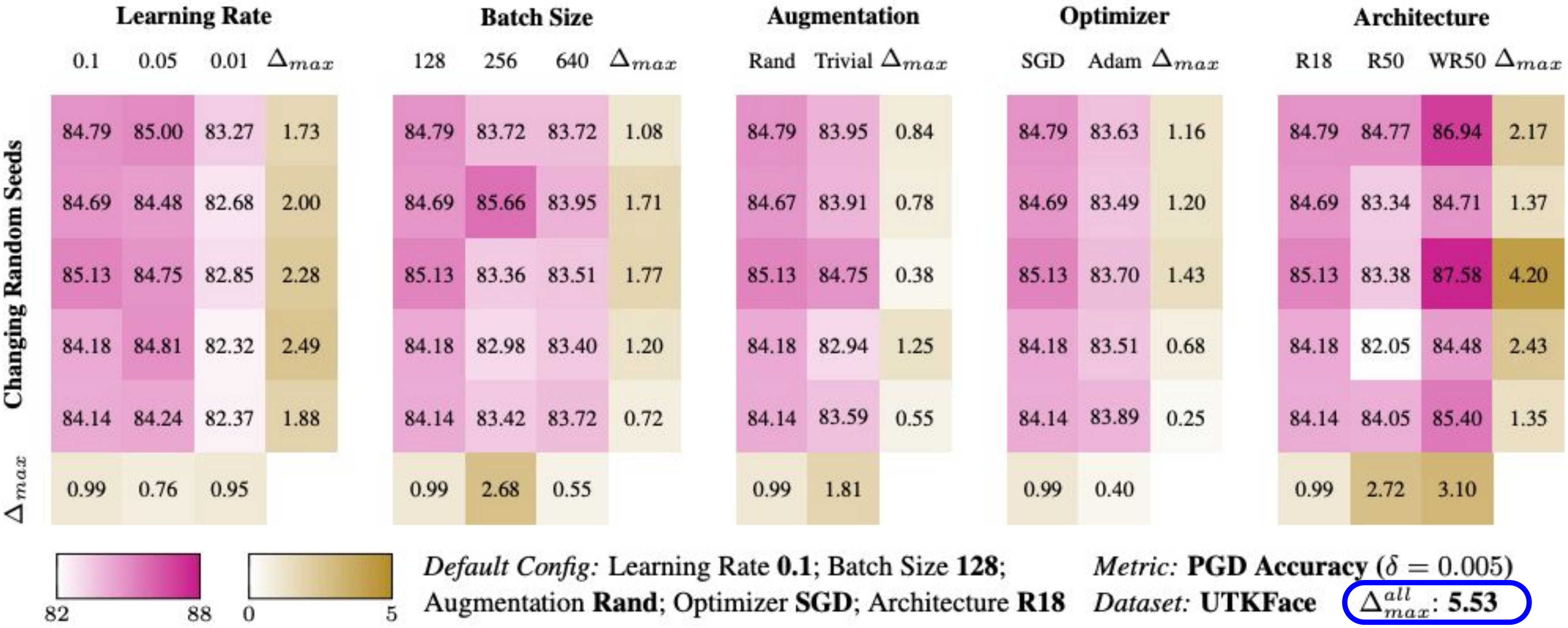


	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	84.79	85.00	83.27	1.73	84.79	83.72	83.72	1.08	84.79	83.95	0.84	84.79	83.63	1.16	84.79	84.77	86.94	2.17
	84.69	84.48	82.68	2.00	84.69	85.66	83.95	1.71	84.67	83.91	0.78	84.69	83.49	1.20	84.69	83.34	84.71	1.37
	85.13	84.75	82.85	2.28	85.13	83.36	83.51	1.77	85.13	84.75	0.38	85.13	83.70	1.43	85.13	83.38	87.58	4.20
	84.18	84.81	82.32	2.49	84.18	82.98	83.40	1.20	84.18	82.94	1.25	84.18	83.51	0.68	84.18	82.05	84.48	2.43
	84.14	84.24	82.37	1.88	84.14	83.42	83.72	0.72	84.14	83.59	0.55	84.14	83.89	0.25	84.14	84.05	85.40	1.35
Δ_{max}	0.99	0.76	0.95		0.99	2.68	0.55		0.99	1.81		0.99	0.40		0.99	2.72	3.10	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **PGD Accuracy** ($\delta = 0.005$)
Dataset: **UTKFace** Δ^{all}_{max}: **5.53**

Multiplicity Sheet: Security



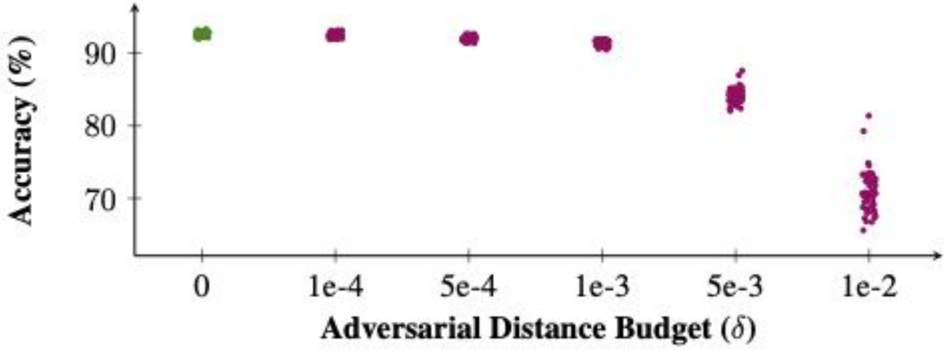
	Learning Rate				Batch Size				Augmentation			Optimizer			Architecture			
	0.1	0.05	0.01	Δ_{max}	128	256	640	Δ_{max}	Rand	Trivial	Δ_{max}	SGD	Adam	Δ_{max}	R18	R50	WR50	Δ_{max}
Changing Random Seeds	84.79	85.00	83.27	1.73	84.79	83.72	83.72	1.08	84.79	83.95	0.84	84.79	83.63	1.16	84.79	84.77	86.94	2.17
	84.69	84.48	82.68	2.00	84.69	85.66	83.95	1.71	84.67	83.91	0.78	84.69	83.49	1.20	84.69	83.34	84.71	1.37
	85.13	84.75	82.85	2.28	85.13	83.36	83.51	1.77	85.13	84.75	0.38	85.13	83.70	1.43	85.13	83.38	87.58	4.20
	84.18	84.81	82.32	2.49	84.18	82.98	83.40	1.20	84.18	82.94	1.25	84.18	83.51	0.68	84.18	82.05	84.48	2.43
	84.14	84.24	82.37	1.88	84.14	83.42	83.72	0.72	84.14	83.59	0.55	84.14	83.89	0.25	84.14	84.05	85.40	1.35
Δ_{max}	0.99	0.76	0.95		0.99	2.68	0.55		0.99	1.81		0.99	0.40		0.99	2.72	3.10	

Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**; Architecture **R18**

Metric: **PGD Accuracy** ($\delta = 0.005$)

Dataset: **UTKFace** Δ^{all}_{max}: **5.53**

Accuracy Under Intervention: Security



Contributions

- A standardized framework to measure and study model multiplicity.
- An empirical benchmark study of model multiplicity in image classification.
- **Investigating the impact of model selection on unseen failure cases.**

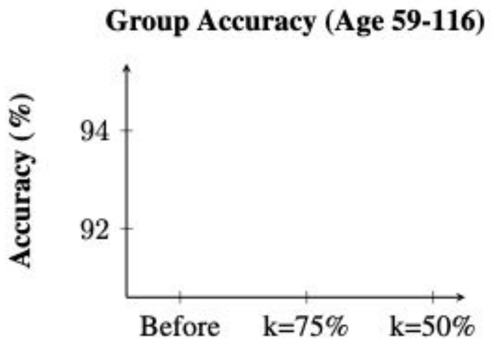
Model Selection to Counter Multiplicity

- We collect all 45 models in previous multiplicity sheets.
- We only select models which rank in the top k% of each metric, i.e.,
  - Accuracy for racial group ‘Asian’
  - Accuracy on OOD dataset ‘FairFace’
  - Accuracy under Output Perturbations for Privacy
  - Accuracy under PGD Adversarial Attacks δ=0.005

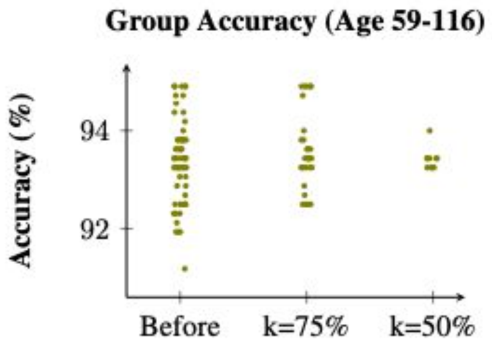
Model Selection to Counter Multiplicity: Unseen Metrics

- To ‘simulate’ unseen failure cases
  - Accuracy for **age group ‘59-116’**
  - Accuracy on OOD dataset **‘CelebA’**
  - Accuracy under **Input** Perturbations for Privacy
  - Accuracy under PGD Adversarial Attacks **δ=0.01**

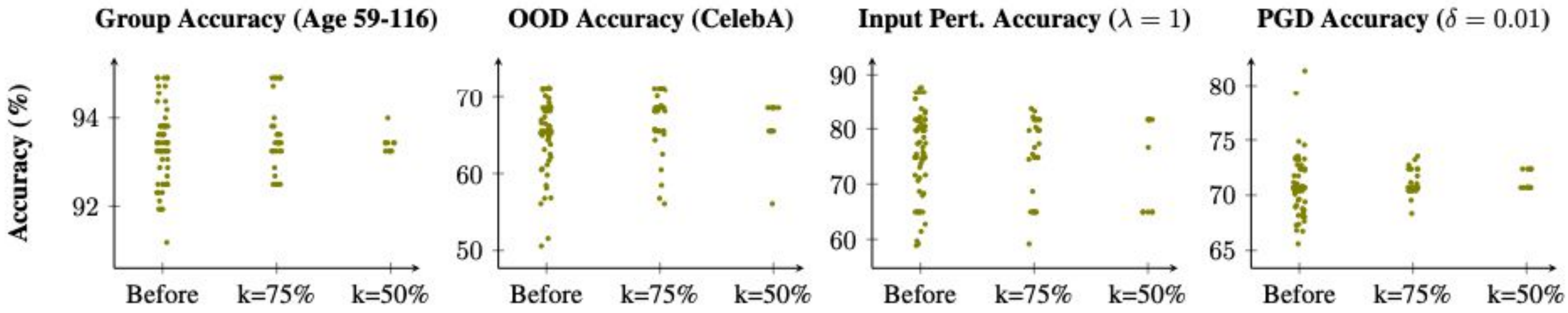
Model Selection to Counter Multiplicity



Model Selection to Counter Multiplicity



Model Selection to Counter Multiplicity



Summary

Summary

- We created an empirical language to talk about multiplicity (accuracy under intervention and multiplicity sheets)

Summary

- We created an empirical language to talk about multiplicity (accuracy under intervention and multiplicity sheets)
- We performed a detailed case study and benchmarked model multiplicity of various trustworthy ML metrics for image classification.

Summary

- We created an empirical language to talk about multiplicity (accuracy under intervention and multiplicity sheets)
- We performed a detailed case study and benchmarked model multiplicity of various trustworthy ML metrics for image classification.
- We showed empirically that the concerns of model multiplicity persist even beyond model selection.