
ACML
Asian Conference on Machine Learning

Re-assessing Accuracy Degradation: A Framework for Understanding DNN Behavior on Similar-but-non-identical Test Datasets

Esla Timothy Anzaku, Hoahan Wang, Ajiboye Babalola, Arnout Van Messem, Wesley De Neve

GHENT UNIVERSITY

GHENT UNIVERSITY GLOBAL CAMPUS

LIÈGE université Mathematics

MASON KOREA George Mason University

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Introduction (1/3)

DNN Evaluation Trends

- **Established**: Use well-known benchmarks
- **Emerging**: Create replicate test datasets to assess generalization
- **Challenge**: Unexpected top-1 accuracy gap on similar test datasets

Introduction (2/3)

Test Dataset Examples

Established Benchmarks

ImageNet-1k Val. Set

50,000 Images | 1,000 Classes

Published in 2009

CIFAR-10 Test Set

10,000 Images | 10 Classes

Published in 2009

Replicate Test Datasets

ImageNetV2

10,000 Images

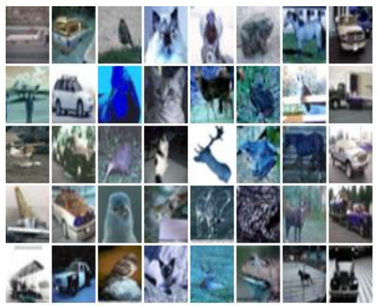
Published in 2019

CIFAR-10.1

2,000 Images

Published in 2019

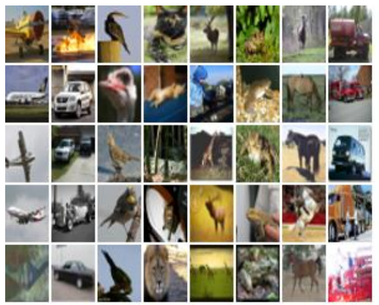
CIFAR-10.2



10,000 Images

Published in 2020

CINIC

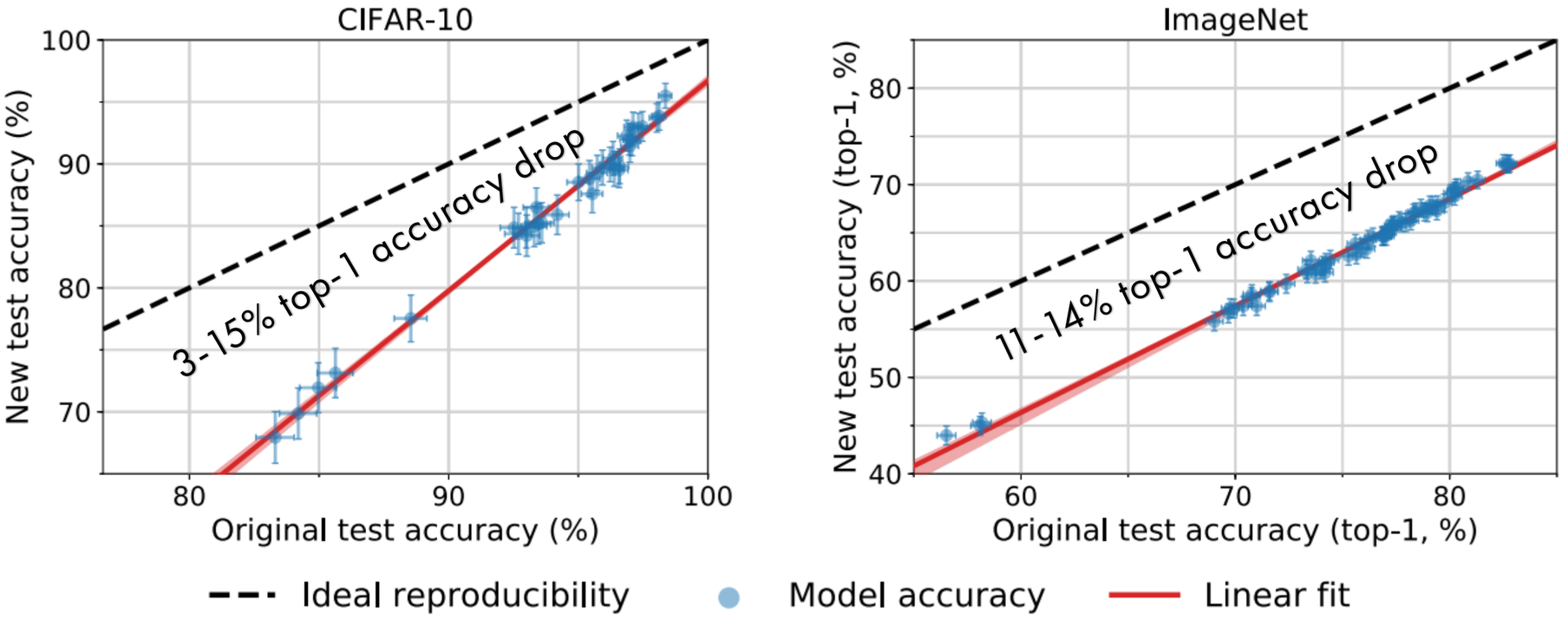


90,000 Images

Published in 2018

Introduction (3/3)

Recht et al: Do ImageNet Classifiers Generalize to ImageNet?

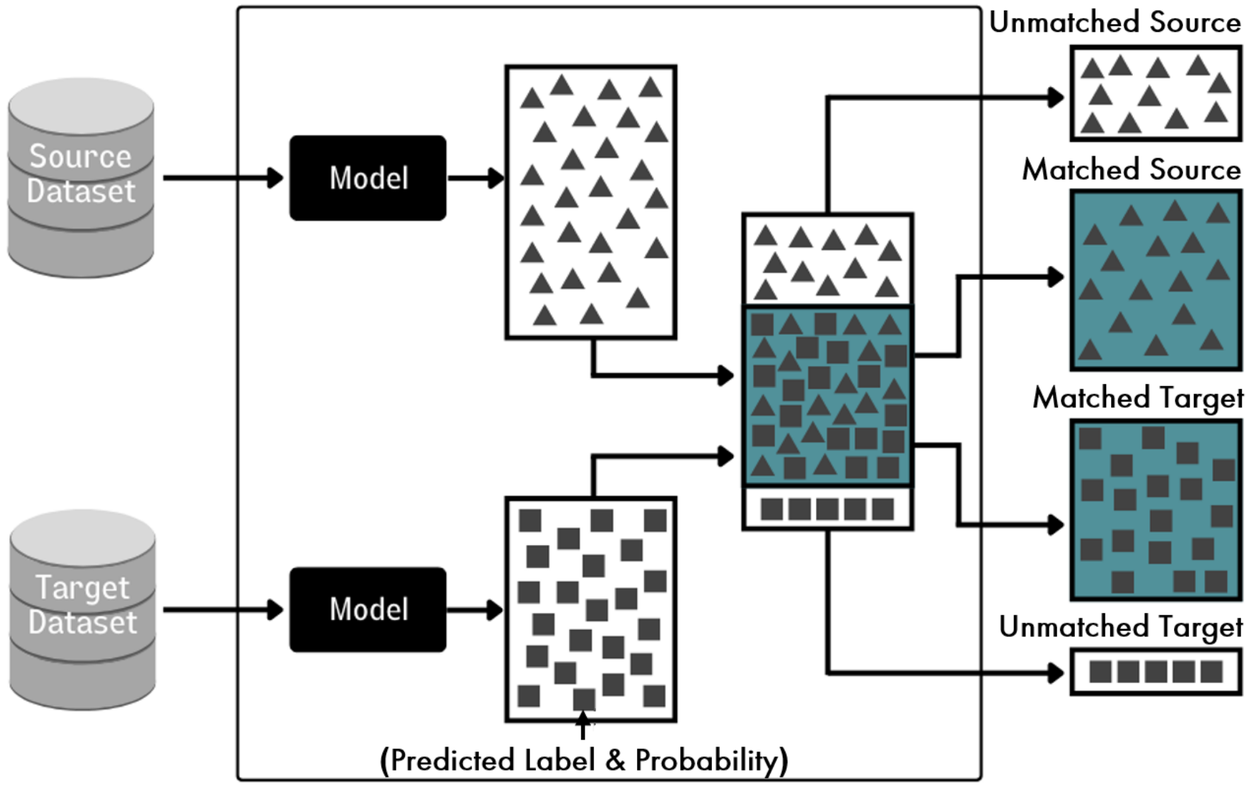


The Challenge:

- Similar dataset curation protocol
- Yet, unexpected and unexplained top-1 accuracy degradation

Proposed Framework (1/2)

Leverage DNN Uncertainty in Model Assessment



1. Get model predictions

2. Match predictions & generate test subsets

3. Assess the test subsets

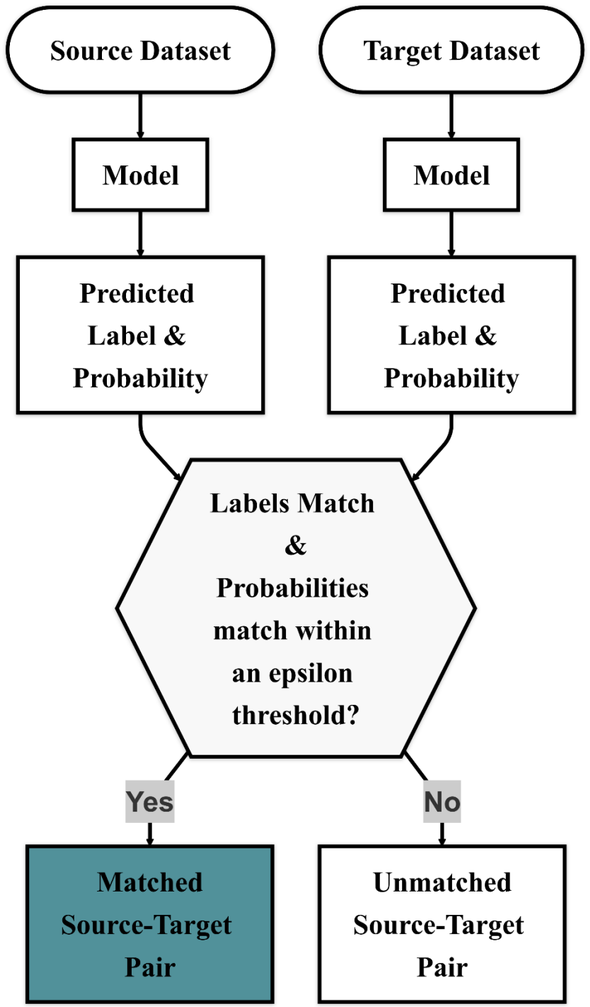
Model Behavior is similar on Source and Target Datasets

IF

- ❑ Accuracy gap on matched subsets is substantially smaller
- ❑ All subsets have similar accuracy versus uncertainty relationship

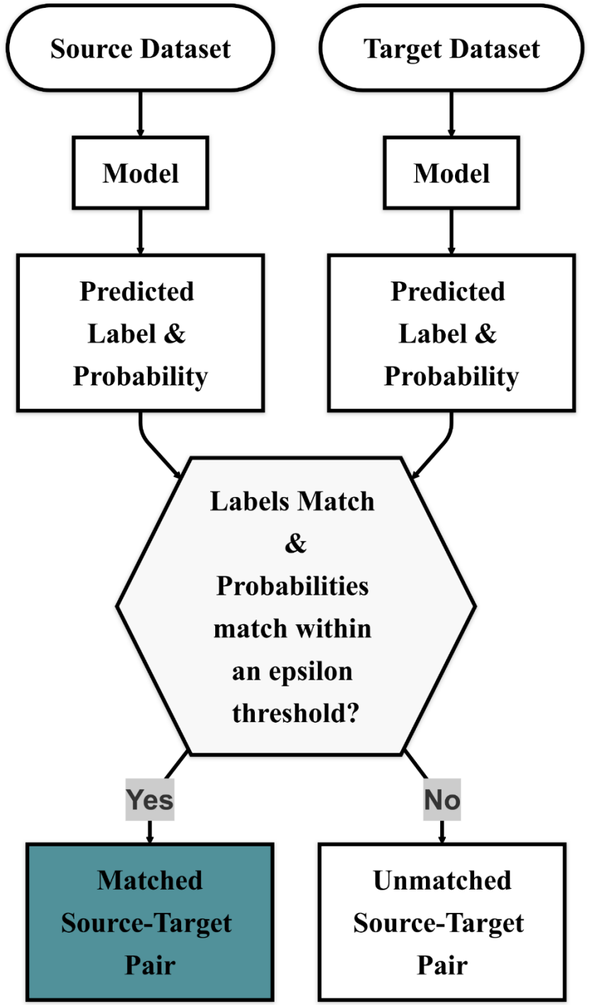
Proposed Framework (2/2)

Leverage DNN Uncertainty in Model Assessment



Proposed Framework (2/2)

Leverage DNN Uncertainty in Model Assessment



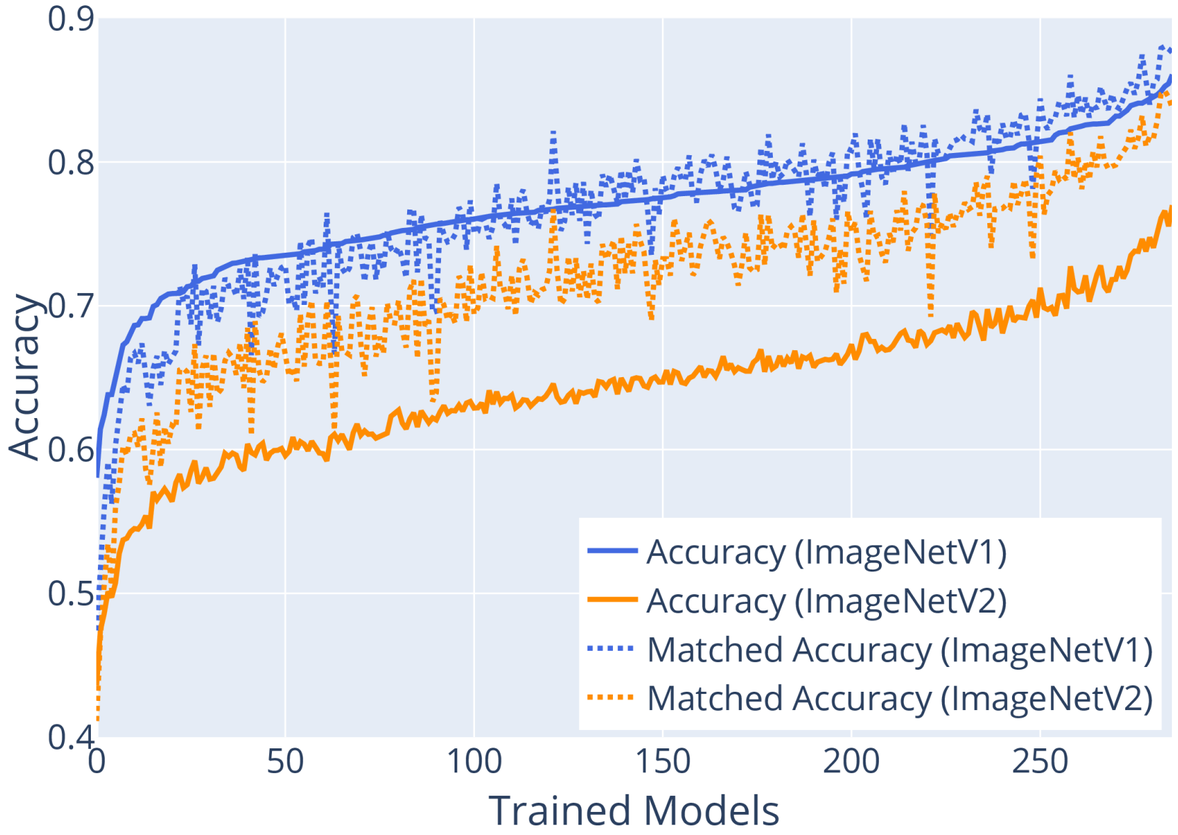
Conventional top-1 Accuracy Assessment		Proposed Evaluation Framework
➢ Uses all datapoints		➢ Matches similar predictions
➢ Treats all predictions equally	VS	➢ Creates fair comparison subsets
➢ Ignores model uncertainty		➢ Leverages model uncertainty
➢ Assumes dataset characteristics are the same		➢ Accounts for differences in dataset characteristics

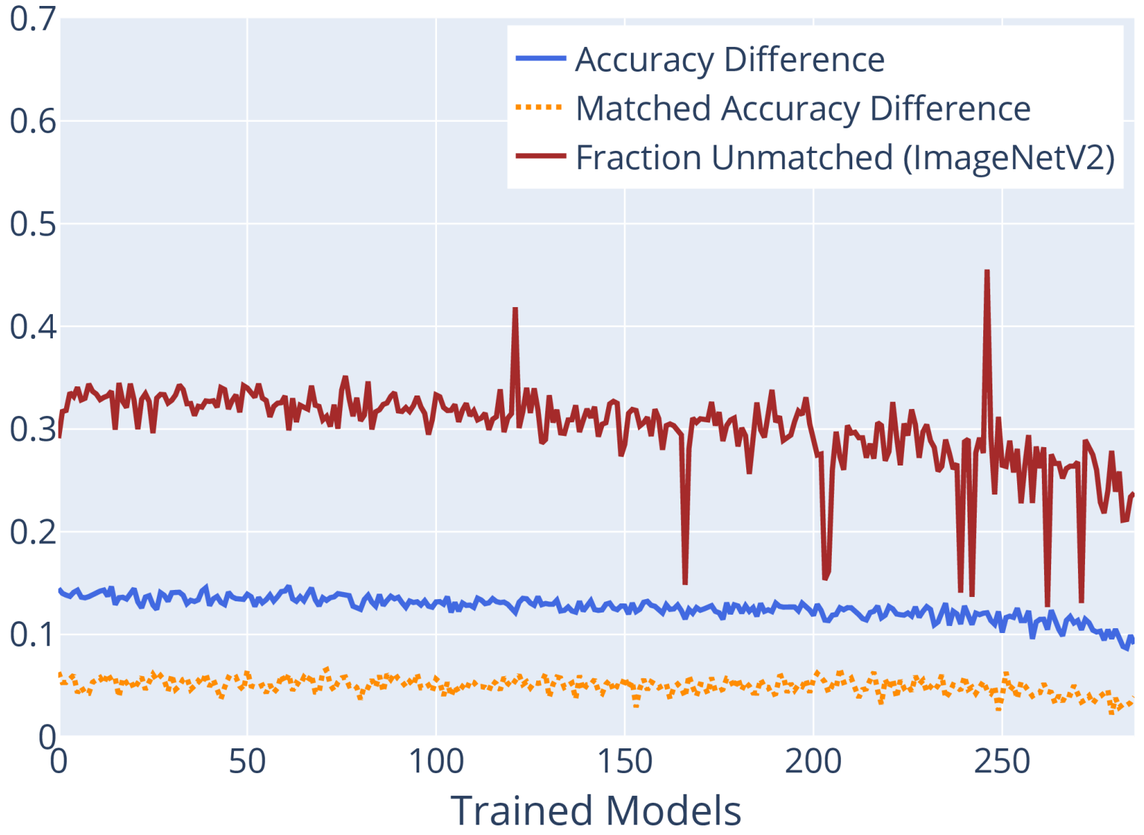
Experimental Setup

Extensive Assessment

- ❑Test dataset pairs: ImageNetV1/V2;CIFAR-10/10.1, CIFAR-10/10.2, CINIC/CIFAR-10 (278 models)
- ❑286 pre-trained ImageNet model
  - Diverse Architectures: ResNet, EfficientNet, MobileNet, ConvNeXt V2, ViTs, etc.
  - Baseline training dataset is ImageNet-1K
    - Some models were pre-trained on larger datasets like ImageNet-22k and JFT-300M

Results (1/2)

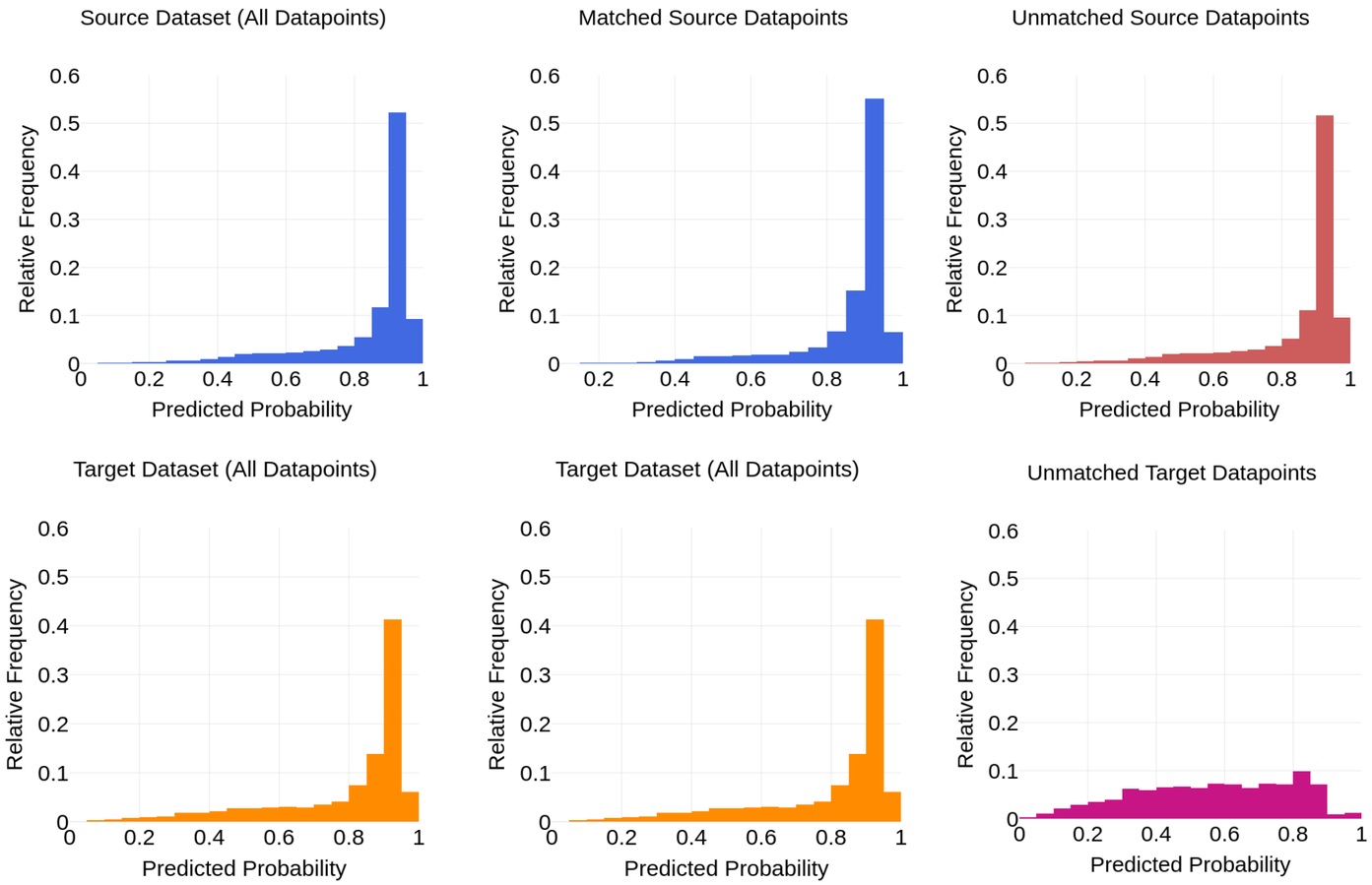


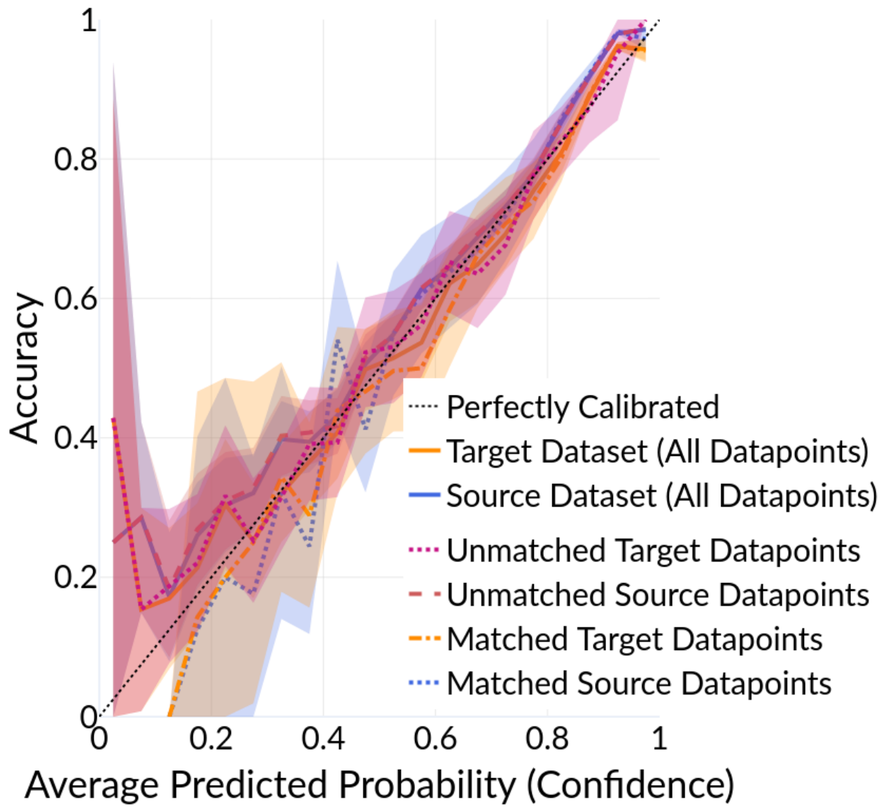


Take-away Message:

Leveraging uncertainty leads to significantly lower top-1 accuracy gap

Results (2/2)





Take-away Messages:

- Different test subsets with different accuracies and uncertainty distributions
- Yet, similar accuracy-uncertainty relationship

Conclusions

- Top-1 accuracy gaps are substantially lower than earlier reported
- Accuracy-uncertainty profiles are consistent across matched and unmatched subsets
- DNNs demonstrated better robustness on replicate test datasets than earlier reported
- Test and replicate datasets differ in subtle ways that warrants further investigation

Thank You!

GHENT
UNIVERSITY