Friedrich-Alexander-Universität
Technische Fakultät

Machine Learning
Data Analytics

FAU

Improving Robustness against Real-World and Worst-Case Distribution Shifts through Decision Region Quantification

Leo Schwinn, Leon Bungert, An Nguyen, René Raab, Falk Pulsmeyer, Doina Precup, Björn Eskofier, Dario Zanca

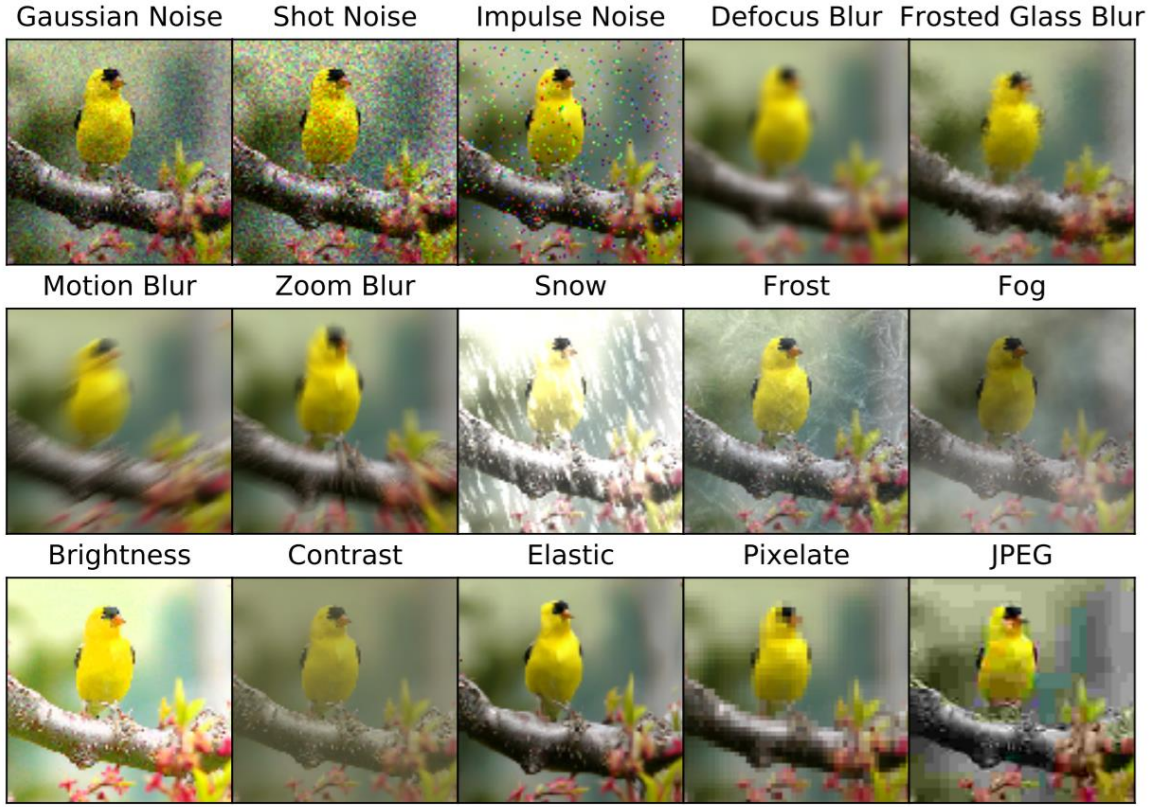
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Presented at ICML 2022

Vulnerable to real-world distribution shifts

ImageNet-R

@Hendrycks et al. “The Many Faces of Robustness” In ICCV, 2021

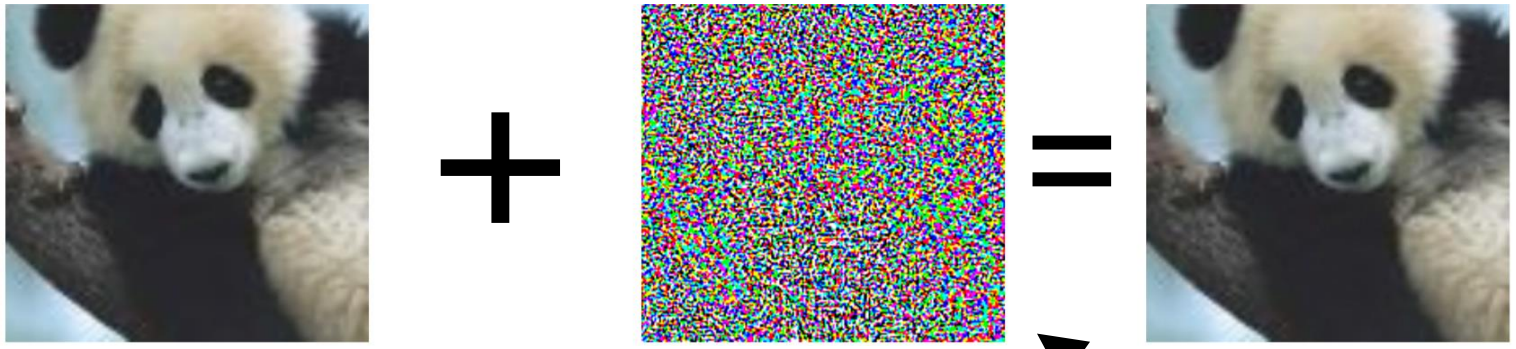
ImageNet-C



@Hendrycks et al. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations” In ICLR, 2019

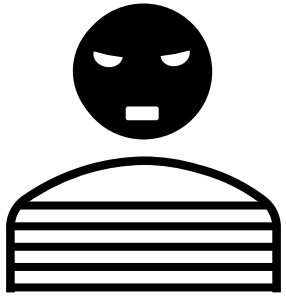
Vulnerable to worst-case distribution shifts (adversarial examples)

Panda ✓ + **Perturbation** = **Gibbon** ✗

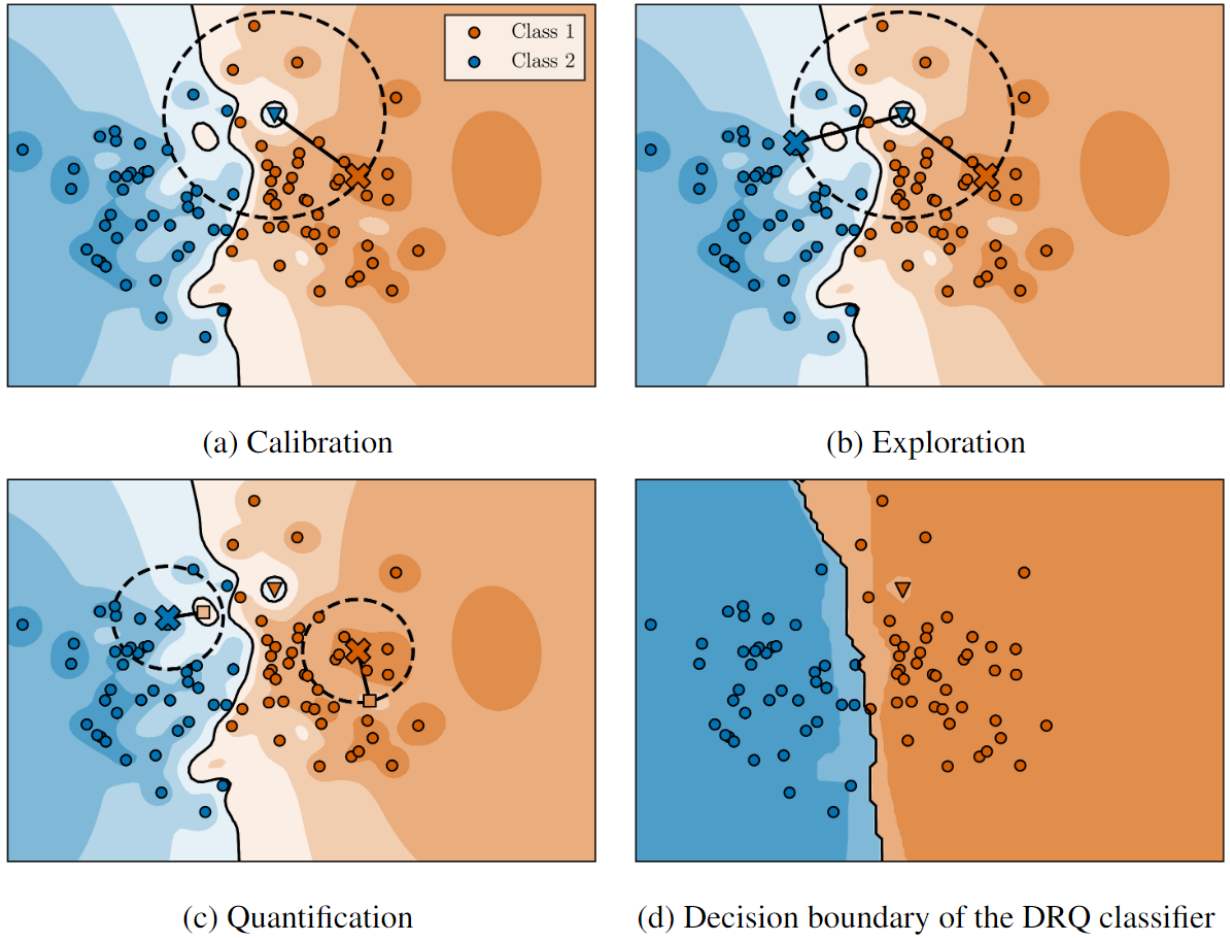


@Goodfellow et al. “Explaining and Harnessing Adversarial Examples” In ICLR, 2015

Malicious Actor



Quantify Robustness of Decision to improve predictions



(a) Calibration

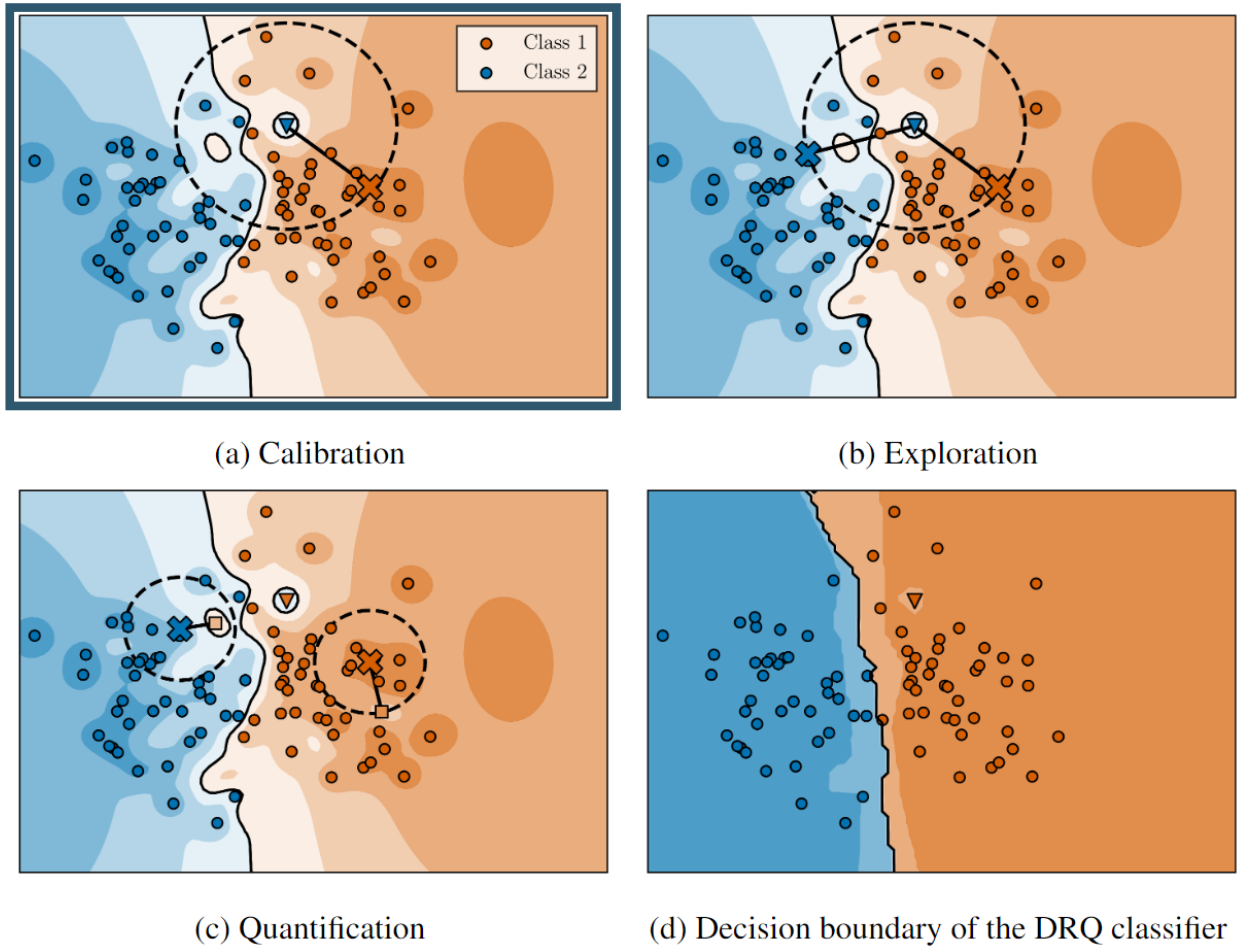
(b) Exploration

(c) Quantification

(d) Decision boundary of the DRQ classifier

Quantify Robustness of Decision to improve predictions

(a) Calibrate search radius



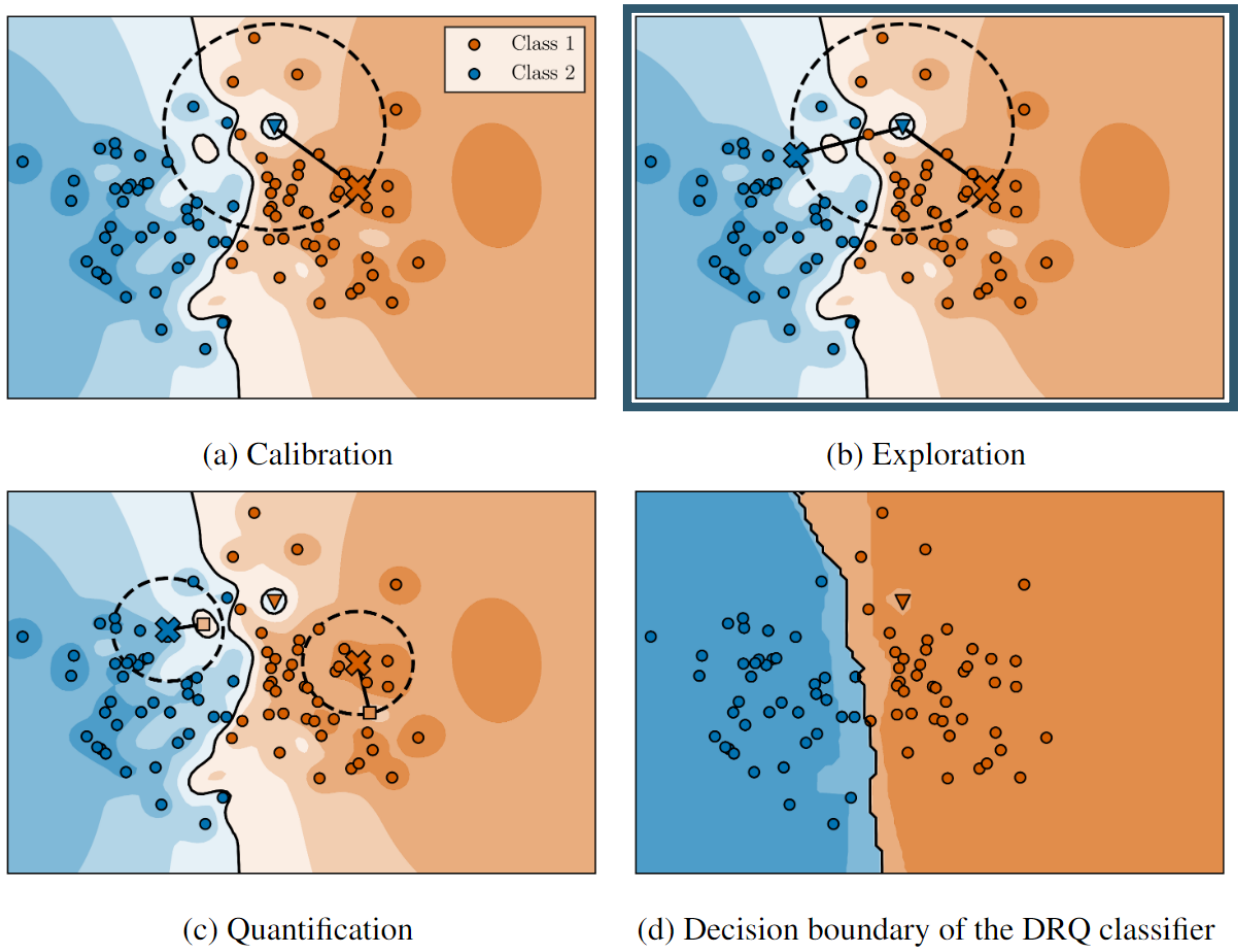
(a) Calibration (b) Exploration

(c) Quantification (d) Decision boundary of the DRQ classifier

Quantify Robustness of Decision to improve predictions

(a) Calibrate search radius

(b) Explore candidate predictions



(a) Calibration (b) Exploration

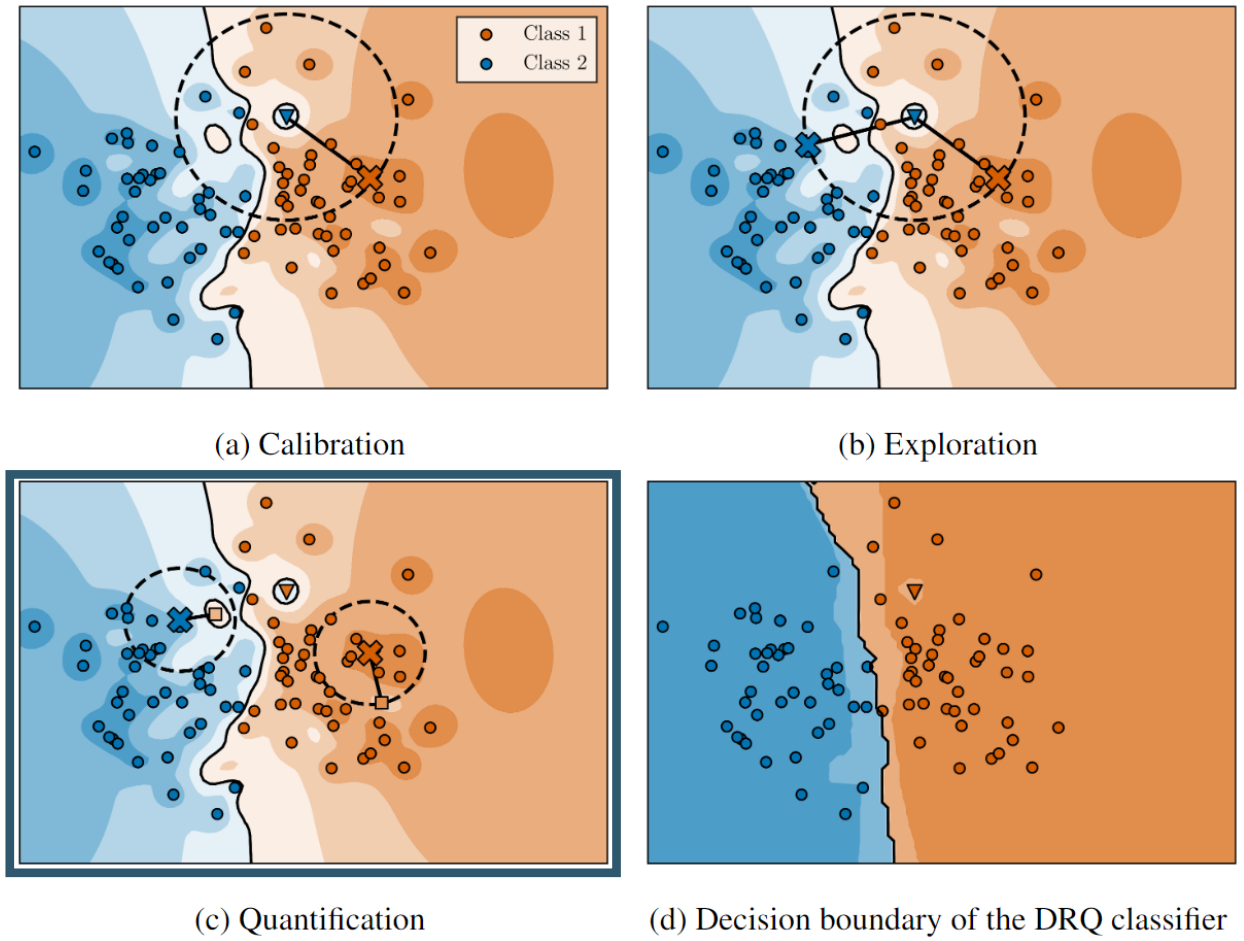
(c) Quantification (d) Decision boundary of the DRQ classifier

Quantify Robustness of Decision to improve predictions

(a) Calibrate search radius

(b) Explore candidate predictions

(c) Quantify robustness of candidates



(a) Calibration (b) Exploration

(c) Quantification (d) Decision boundary of the DRQ classifier

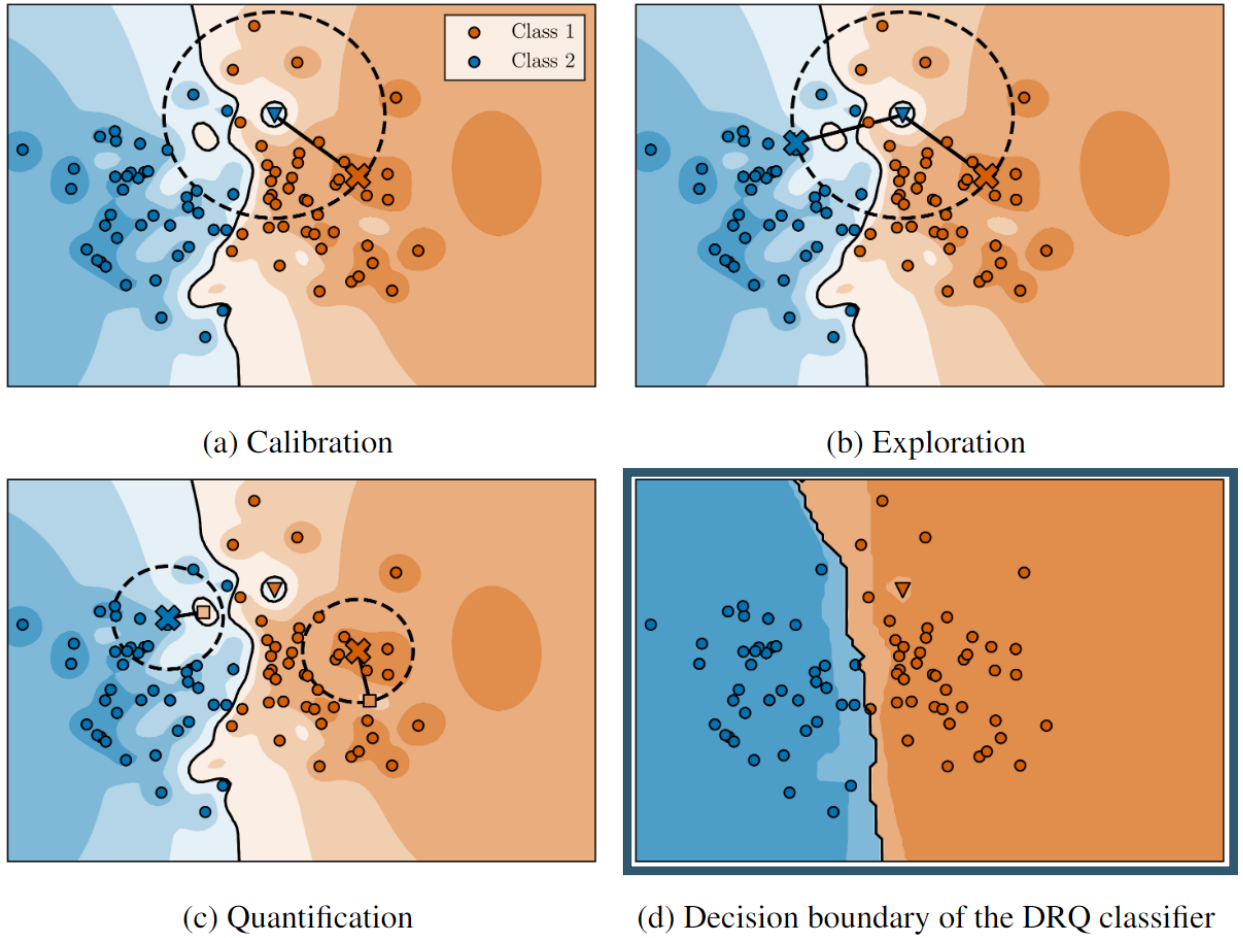
Quantify Robustness of Decision to improve predictions

(a) Calibrate search radius

(b) Explore candidate predictions

(c) Quantify robustness of candidates

(d) Decision boundary after DRQ



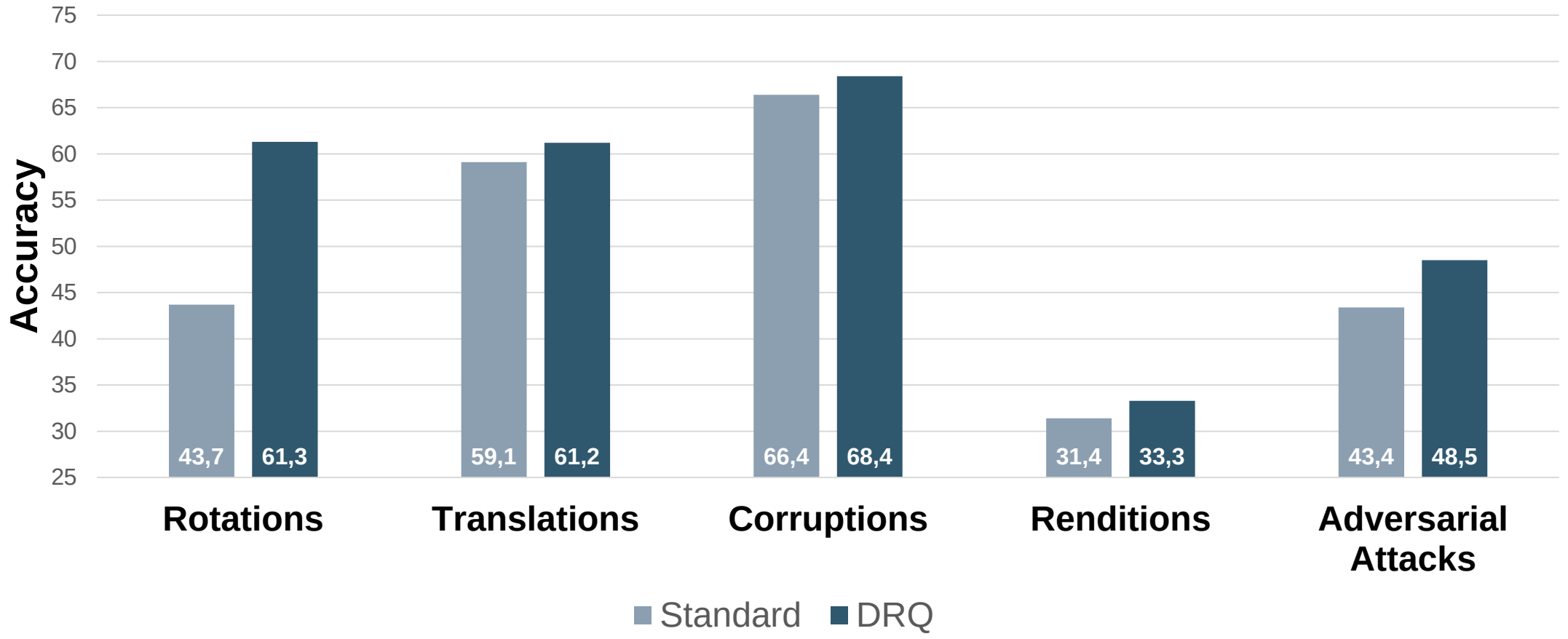
(a) Calibration (b) Exploration

(c) Quantification (d) Decision boundary of the DRQ classifier

Decision Region Quantification (DRQ)

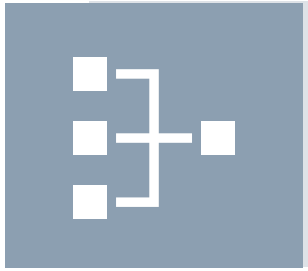
Results

Effectivness of DRQ on various benchmark datasets



Summary

DRQ can increase the robustness for various distribution shift types at the same time



DRQ can be used with any pre-trained differentiable model

Summary

DRQ can increase the robustness for various distribution shift types at the same time

DRQ can be used with any pre-trained differentiable model

Outlook

Can it stand the test of time?

→ Stronger attacks within the algorithm increase the robustness

Towards Responsable AI