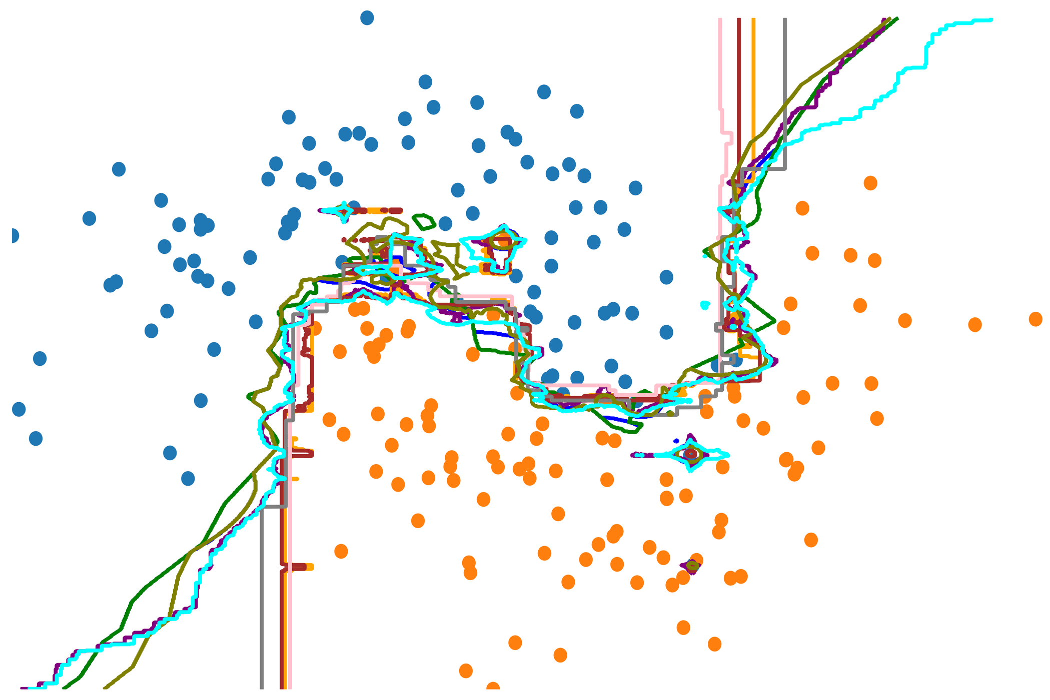
Explaining Local Discrepancies between Image Classification Models

T. Laugel[1,2] X. Renard[1,2] M. Detyniecki[1,2,3]

[1]AXA Group, Paris, France [2]TRAIL, Sorbonne Université, CNRS, LIP6, Paris, France
[3]Polish Academy of Science, Warsaw, Poland

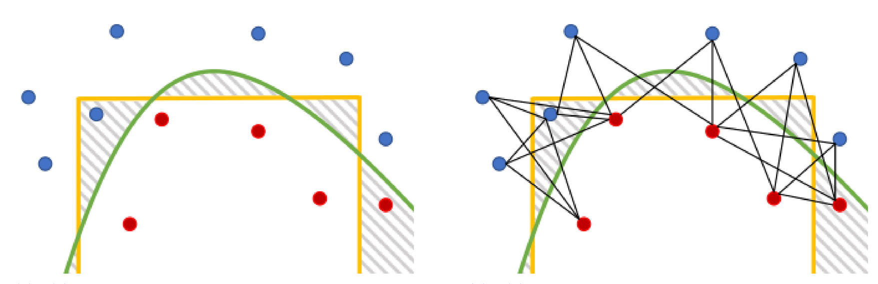
Model discrepancy: context and issues raised

- **Rashomon Effect**: numerous models achieve identical / similar predictive performance despite having very diverse behaviors
- ML practitioners often solely focus on optimizing performance metrics, leaving this **discrepancy** between models **unobserved**
- This phenomenon results in an arbitrary selection of one model over its competitors
- Multiple hazardous consequences can arise: unfair treatment, loss of opportunity...

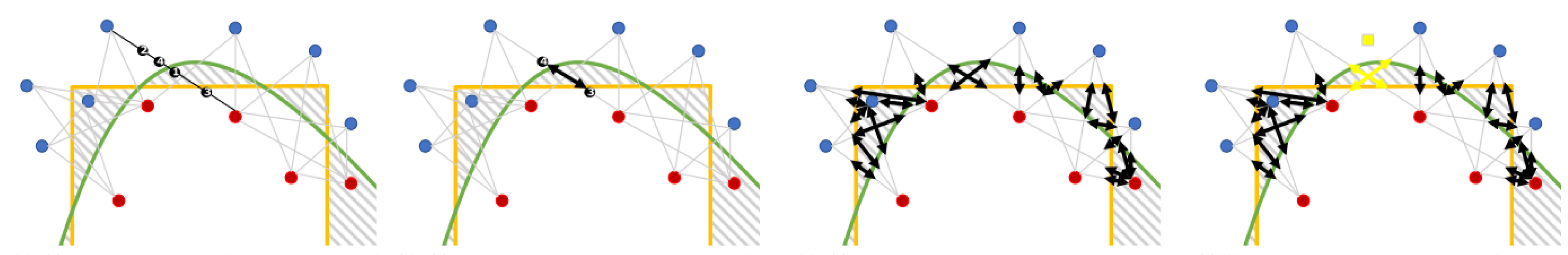


Our proposition: explaining local differences with DIG-CV

- **DIG** (Discrepancy Interval Generation) aims at **detecting** and **explaining** differences between models trained on the same data
- DIG is **model-agnostic** and was designed for **tabular data**



((a)) Initial state: classifiers are trained on the training set.
((b)) Step 1.1: instantiation of the graph ($k_{init} = 3$)



((c)) Step 1.2: refinement of one edge ($n_{epochs} = 4$)
((d)) Step 1.3: extraction of one discrepancy interval
((e)) Step 1.3: all the discrepancy intervals
((f)) Step 2: retrieval of the relevant discrepancy interval for a new instance

Figure: Illustration of the principle of DIG on a toy dataset (red and blue points, colored depending on their true label) and a pool of 2 classifiers (yellow/green lines). Discrepancy regions to be detected by DIG are represented by hatched areas.

- Several shortcomings make it impossible to apply DIG to images:
  - The proposed sampling strategy (convex paths leads to **unrealistic examples**
  - Pixel-by-pixel discrepancy description hurts the **actionnability** of the explanations
- We address these shortcomings by adding a **model-agnostic feature learning step**
- We fit a β-VAE to the training data and apply DIG to the learned latent space

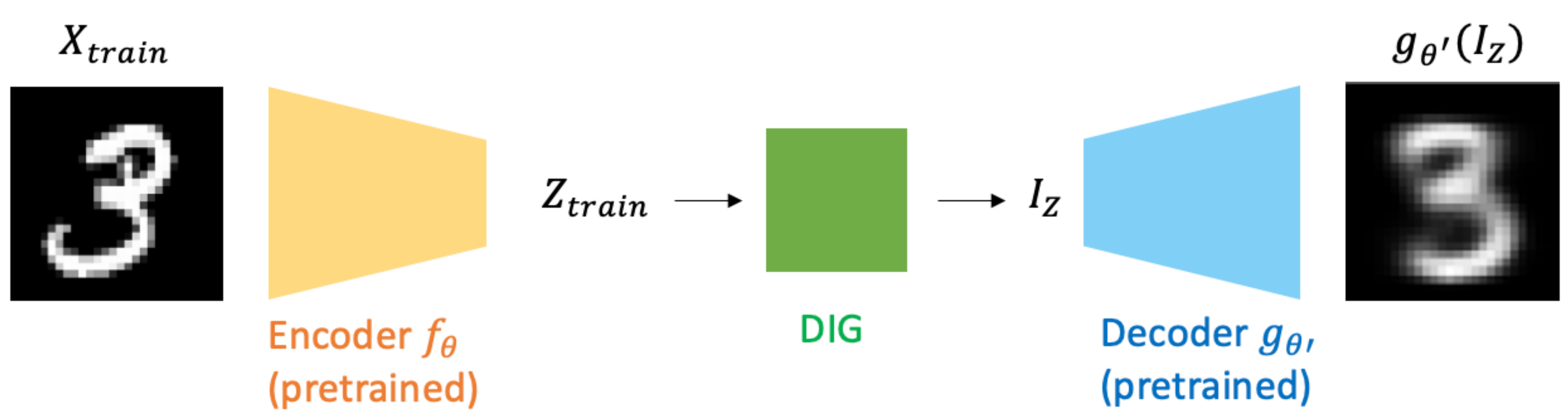


Figure: Proposed framework for DIG-CV

Experimental Results on MNIST and F-MNIST

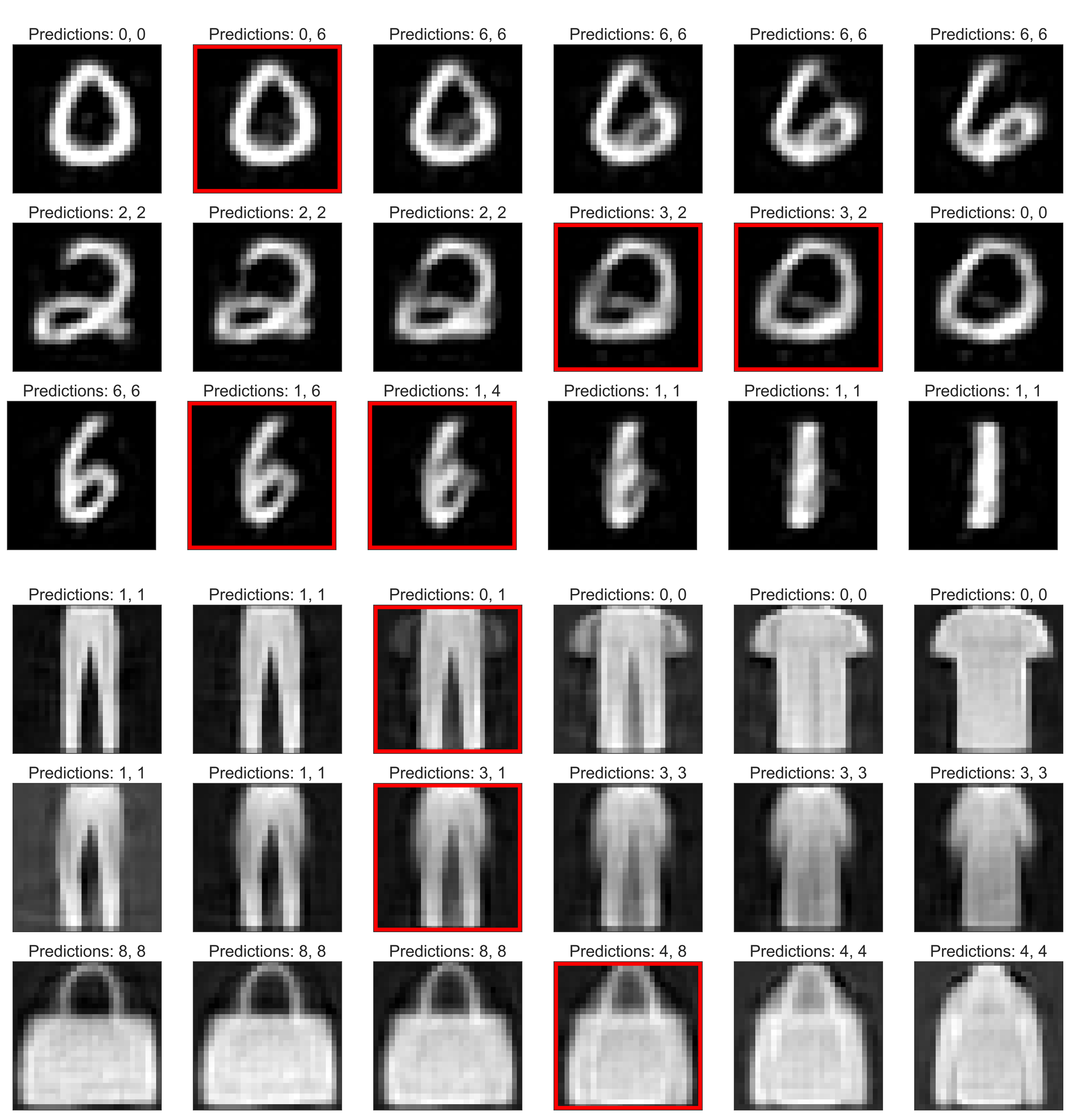


Figure: Results obtained for three instances of MNIST (top) and F-MNIST (bottom). Highlighted images are the ones over which models are disagreeing.

Conclusion and Perspectives

- Image classification models suffer from discrepancy
- We help addressing the issue by generating local explanations of discrepancy areas
- Future works include:
  - Empirical study of the approach's efficiency
  - User experiments to assess the benefits of the approach

Contact Information and Further Reading

- Initial paper: https://arxiv.org/abs/2104.05467
- Contact authors: thibault.laugel@axa.com xavier.renard@axa.com
- Code available at: https://github.com/axa-rev-research/discrepancies-in-machine-learning