Harnessing Explainability to Build Resilient Ensembles

Abraham Chan,

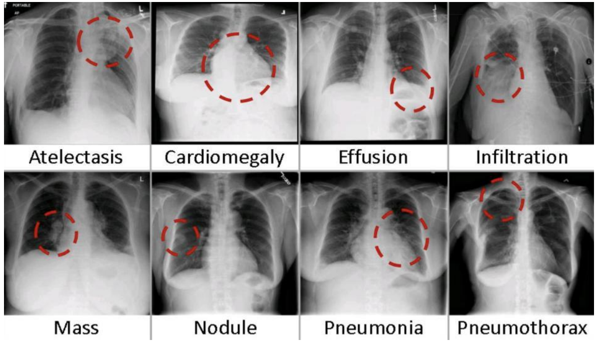
Arpan Gujarati, Karthik Pattabiraman, Sathish Gopalakrishnan

Training Data Faults

70% of Lyft dataset missing, mislabelled [Kang et al, 2022]

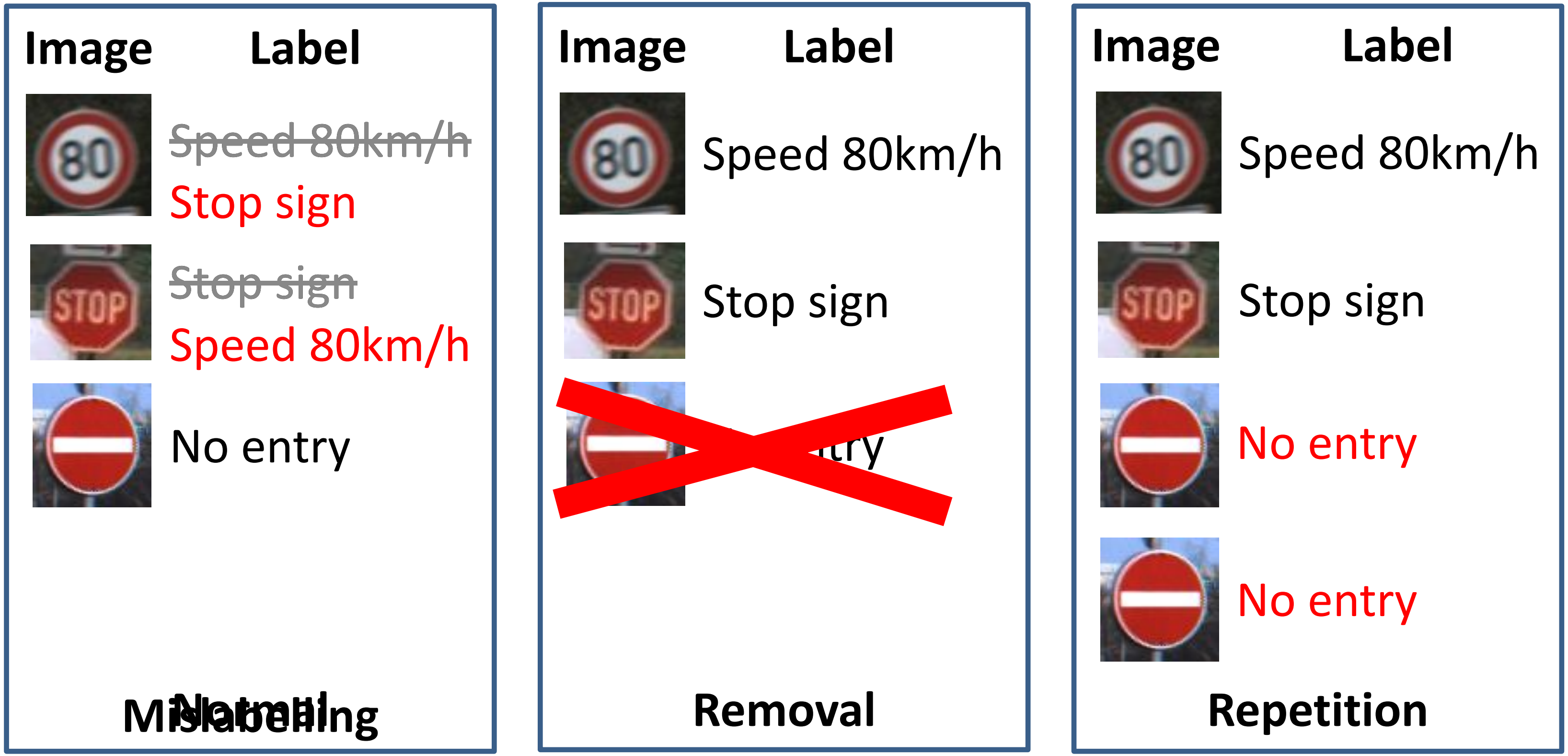
Autonomous Vehicles

20% of ChestX-ray mislabelled [Tang et al, 2021]

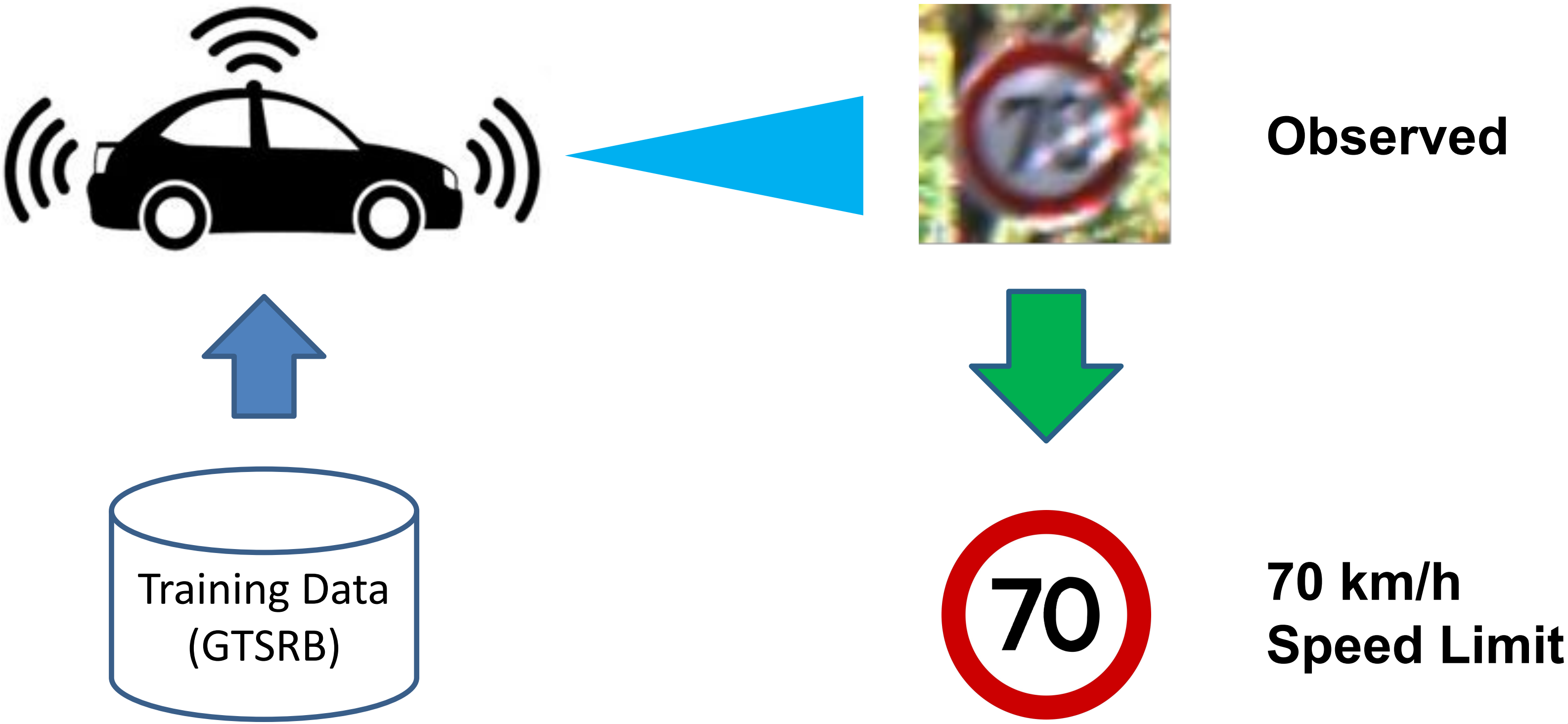


Healthcare

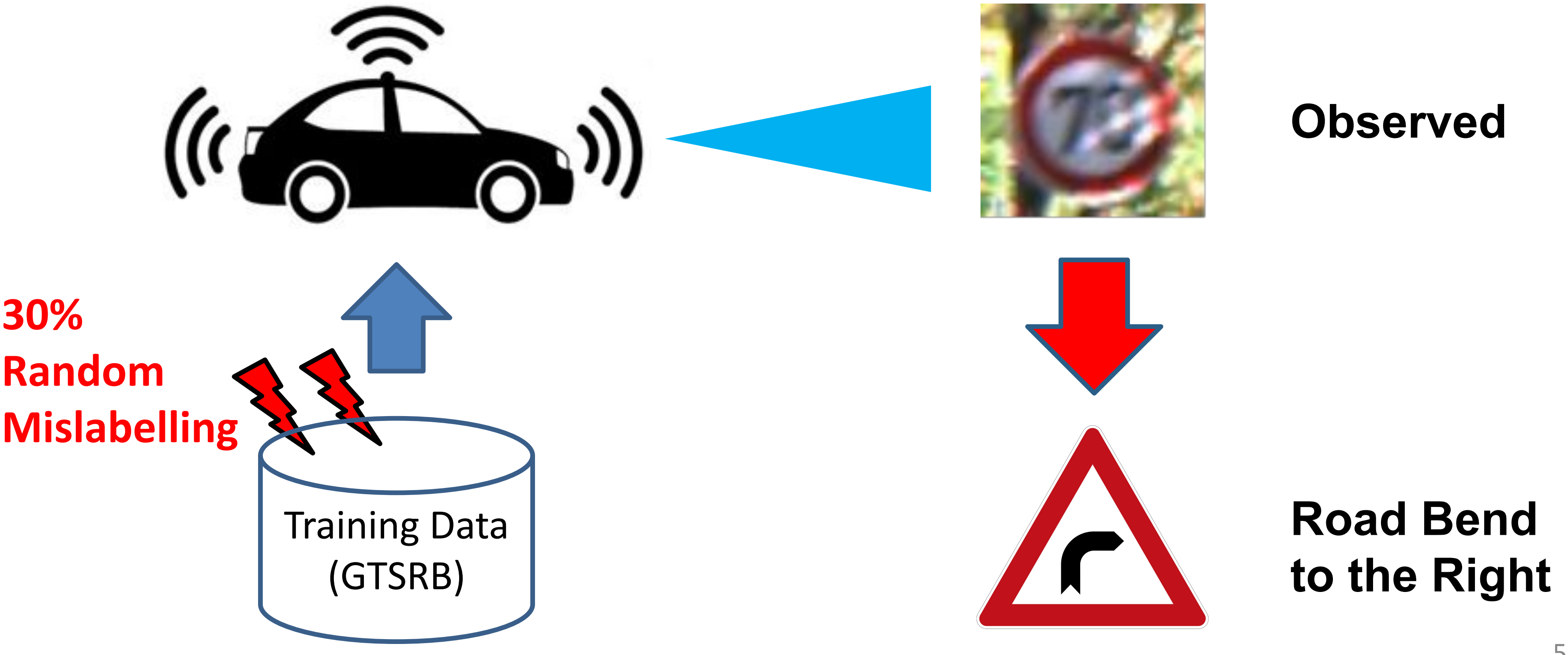
Training Data Faults



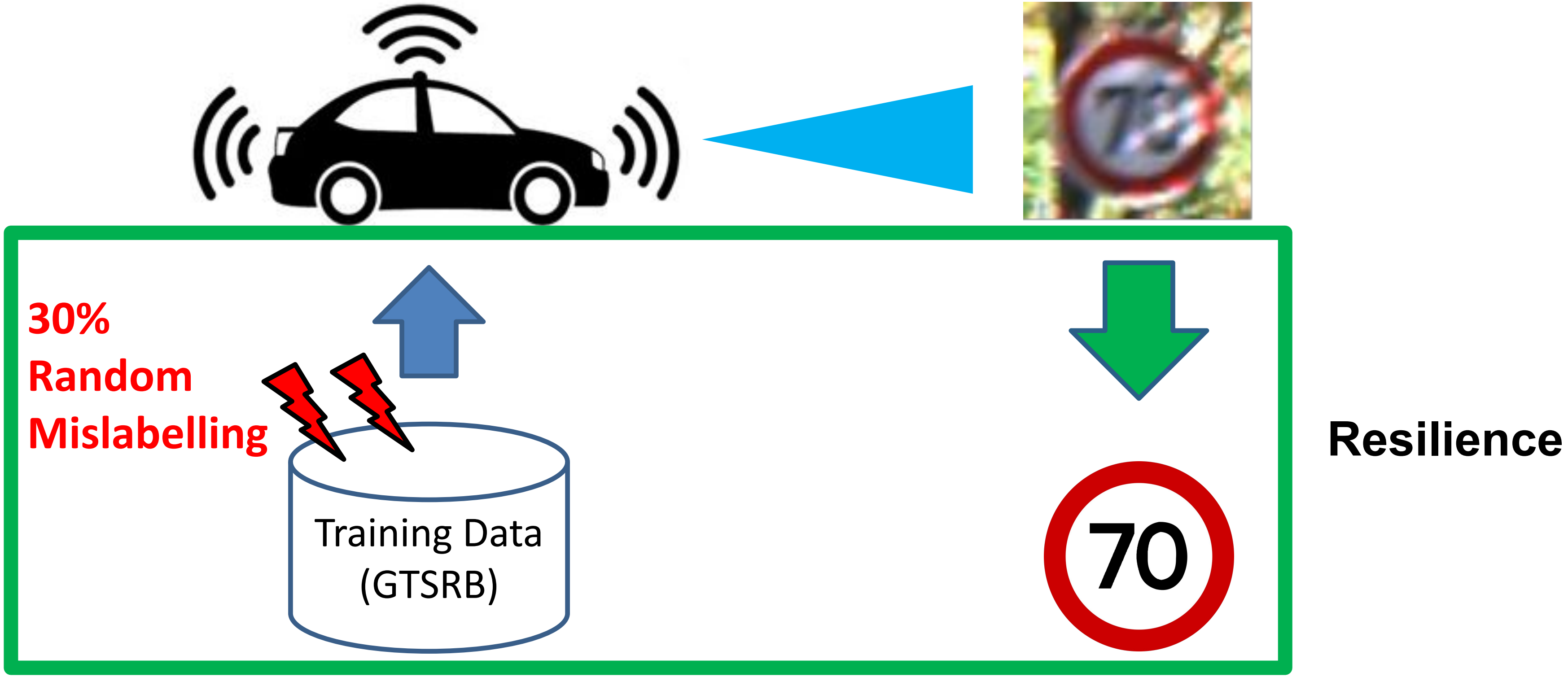
Autonomous Vehicle Example



Random Mislabelling



Resilience against Faulty Training Data

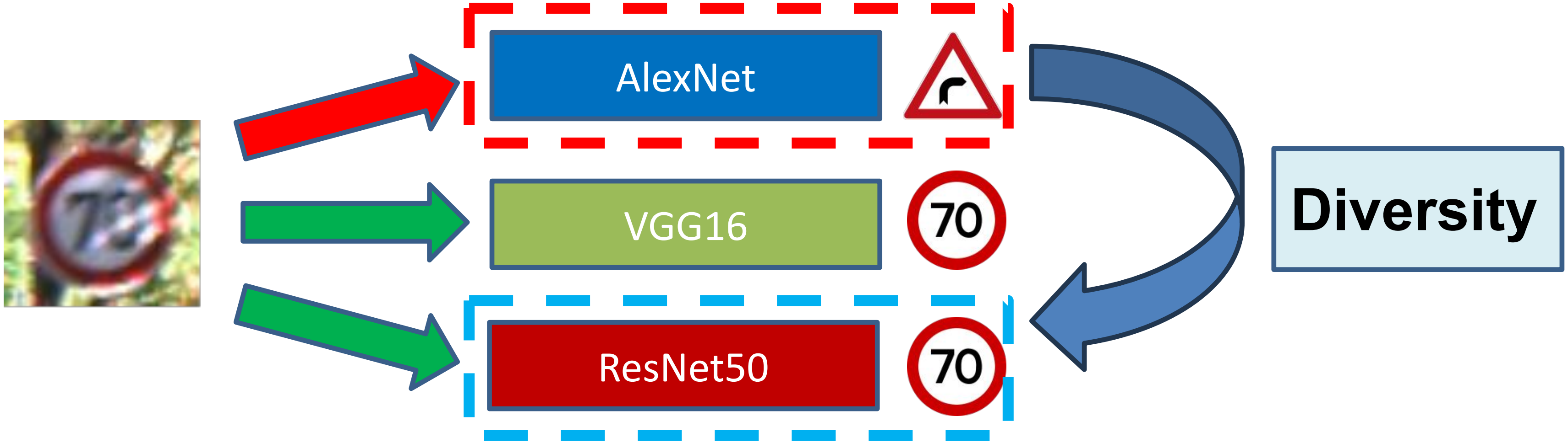


How to mitigate training data faults with minimal human effort?

Our Solution: Build Resilient Ensembles

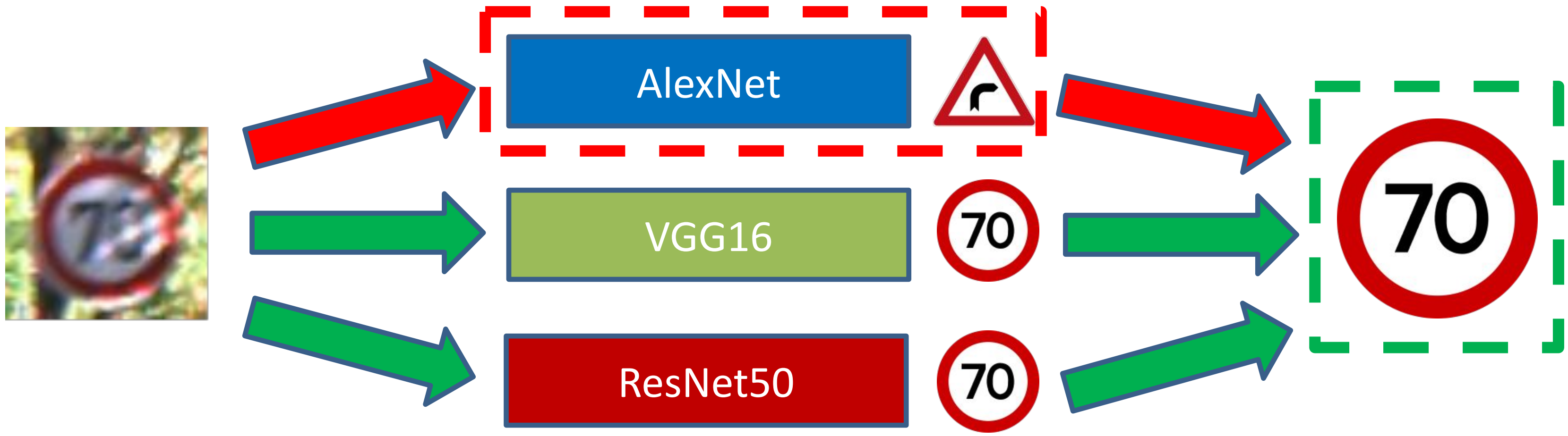
Our Work: The Fault in Our Data Stars: Studying Mitigation Techniques against Faulty Training Data in ML Applications **[DSN'22]**

Resilient Ensembles (NVP)



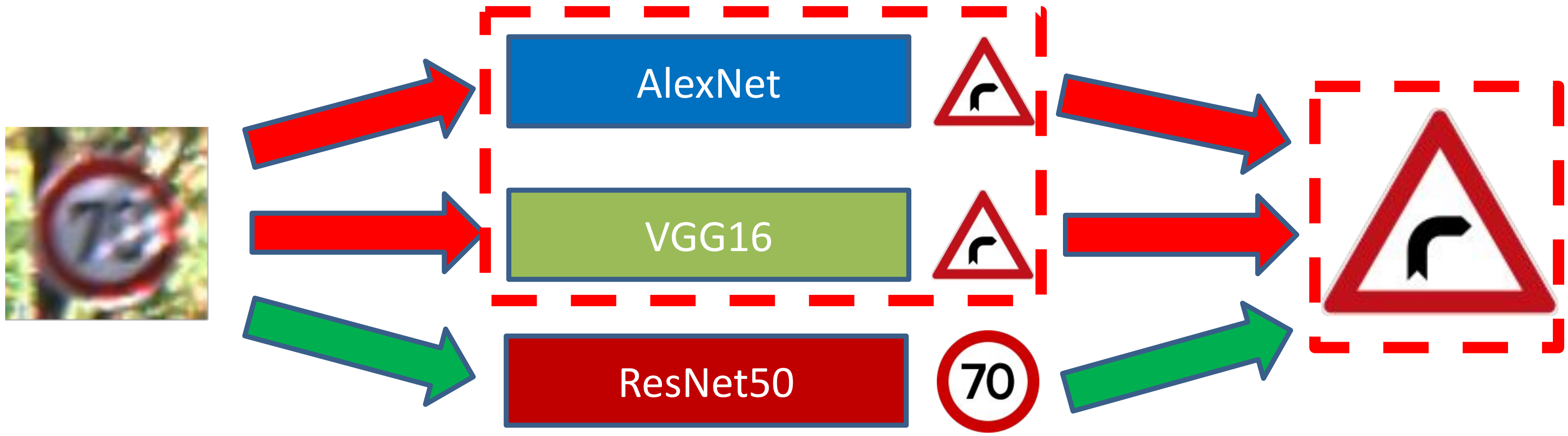
Our Work: The Fault in Our Data Stars: Studying Mitigation Techniques against Faulty Training Data in ML Applications **[DSN'22]**

Resilient Ensembles

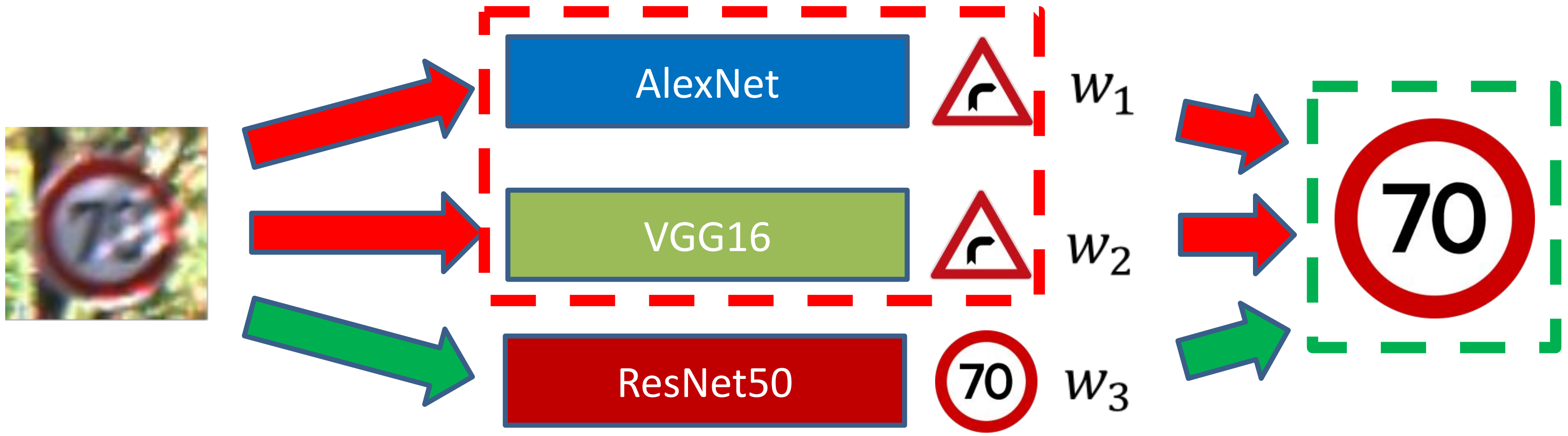


Our Work: Understanding the Resilience of Neural Network Ensembles against Faulty Training Data **[QRS'21]**

When Ensembles Misclassify?



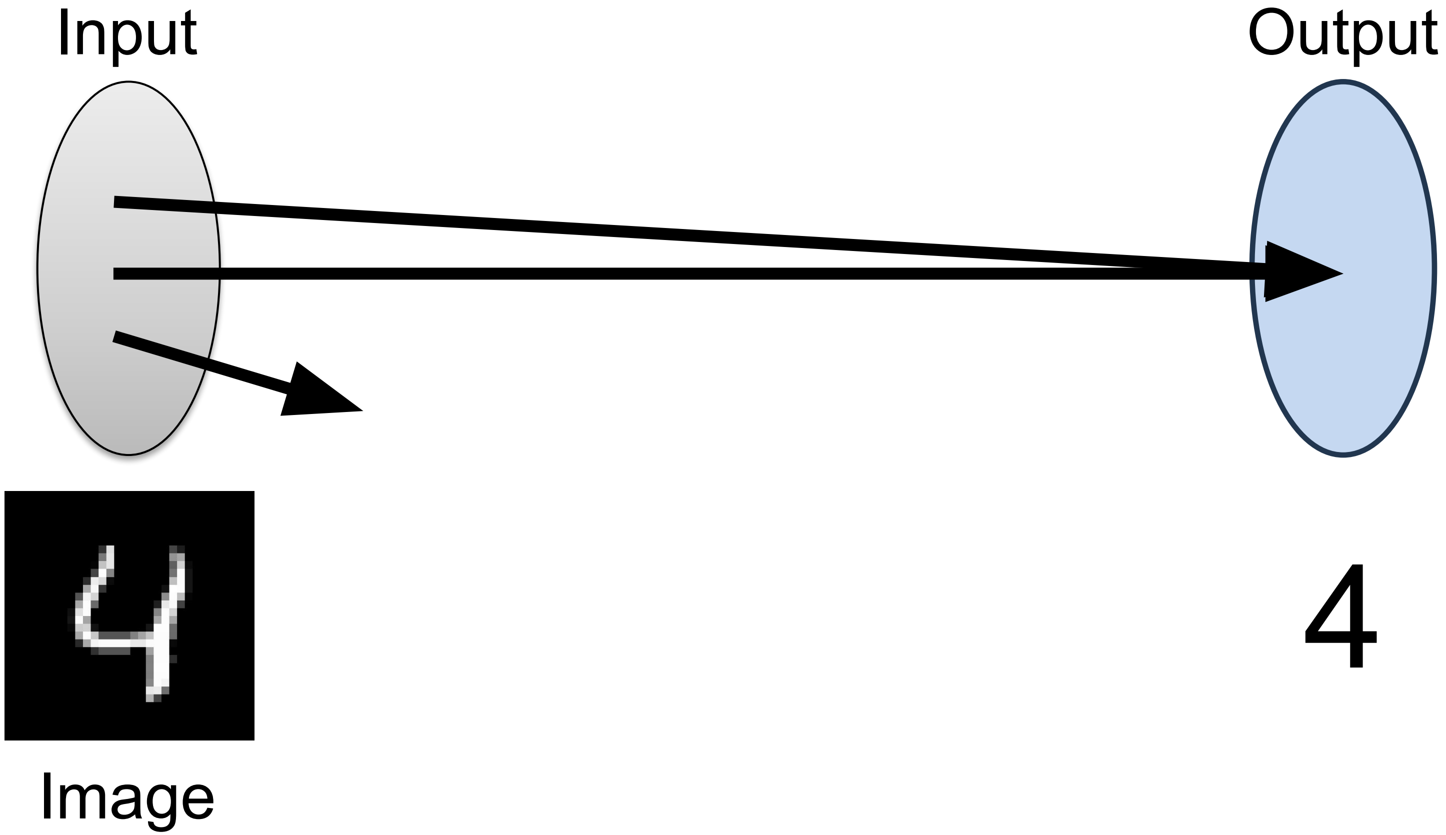
Dynamically Weighted Ensembles



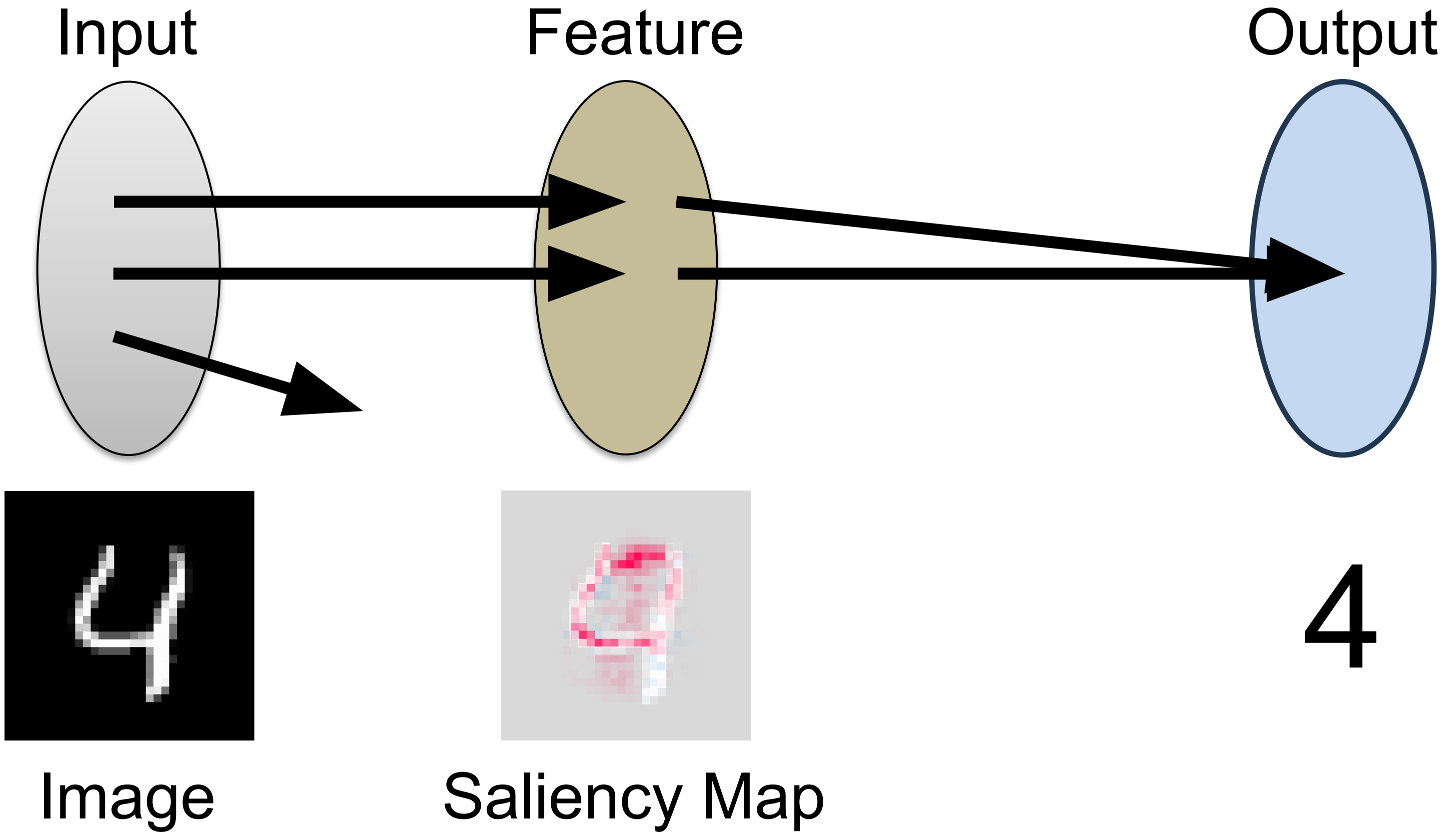
How to determine weights?
Feature Space Diversity?

$$w_1, w_2 < w_3$$

Input - Output Space

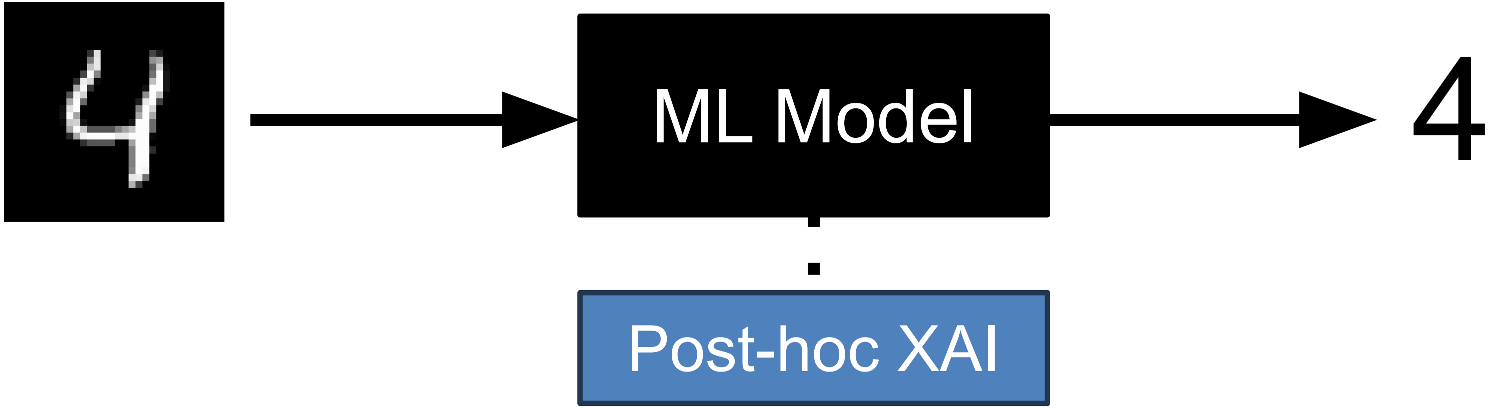


Feature Space



Explainable AI (XAI)

- Local post-hoc techniques (black box ML):
  - SHAP
  - Counterfactual Explanations
  - Integrated Gradients

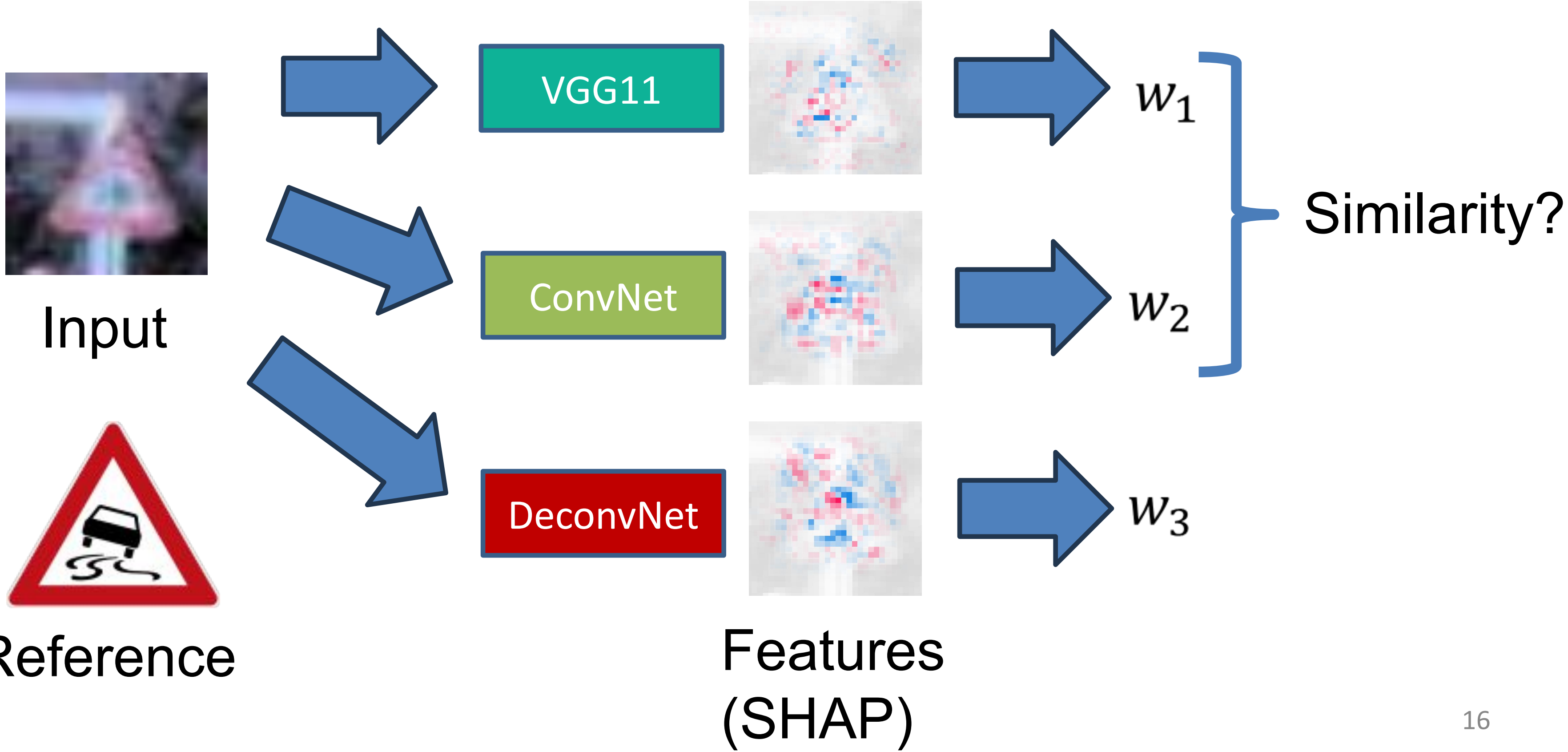


SHAP

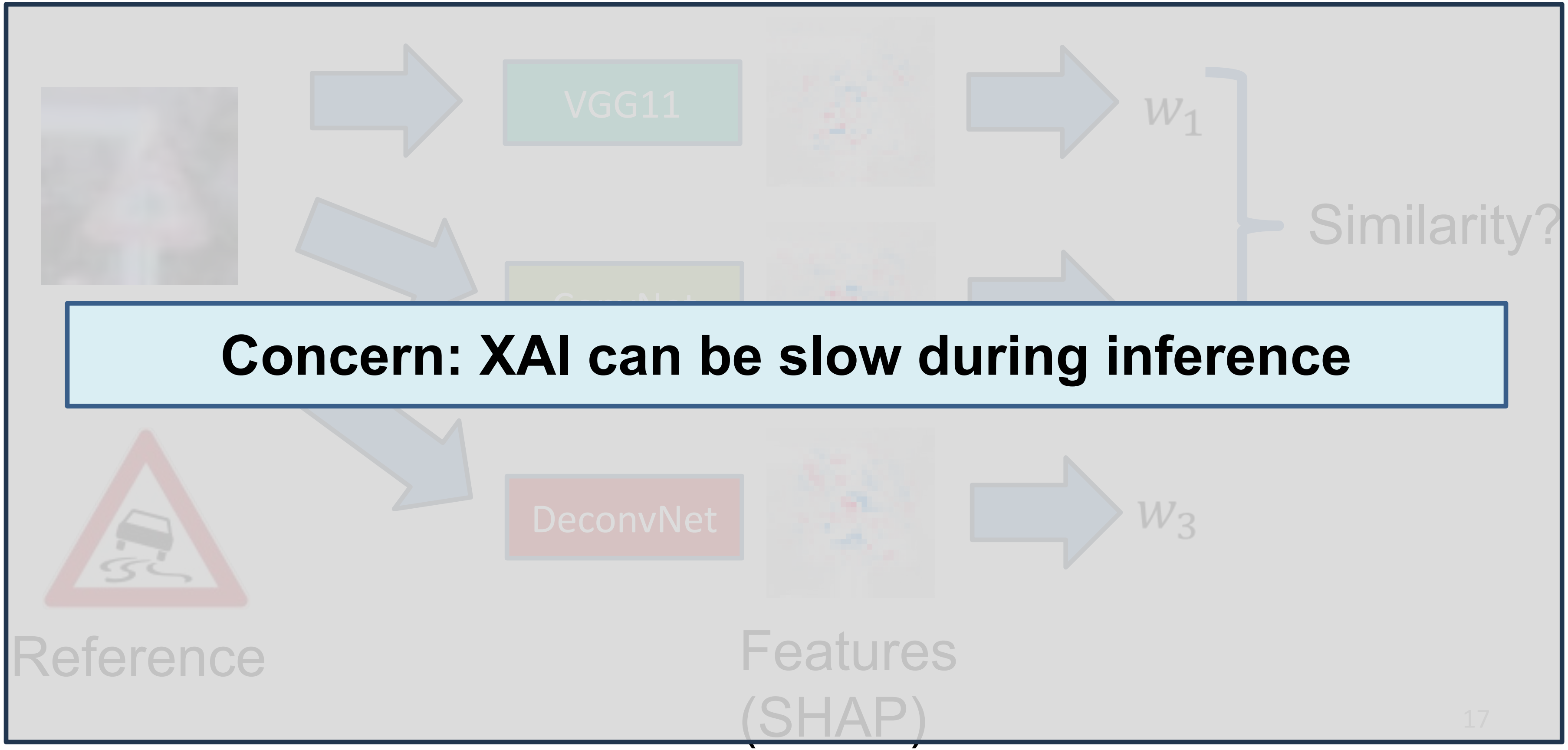
- Which pixels contribute most to decision?

Input　　　　Features (Saliency Map)

Example: Feature Diversity using SHAP

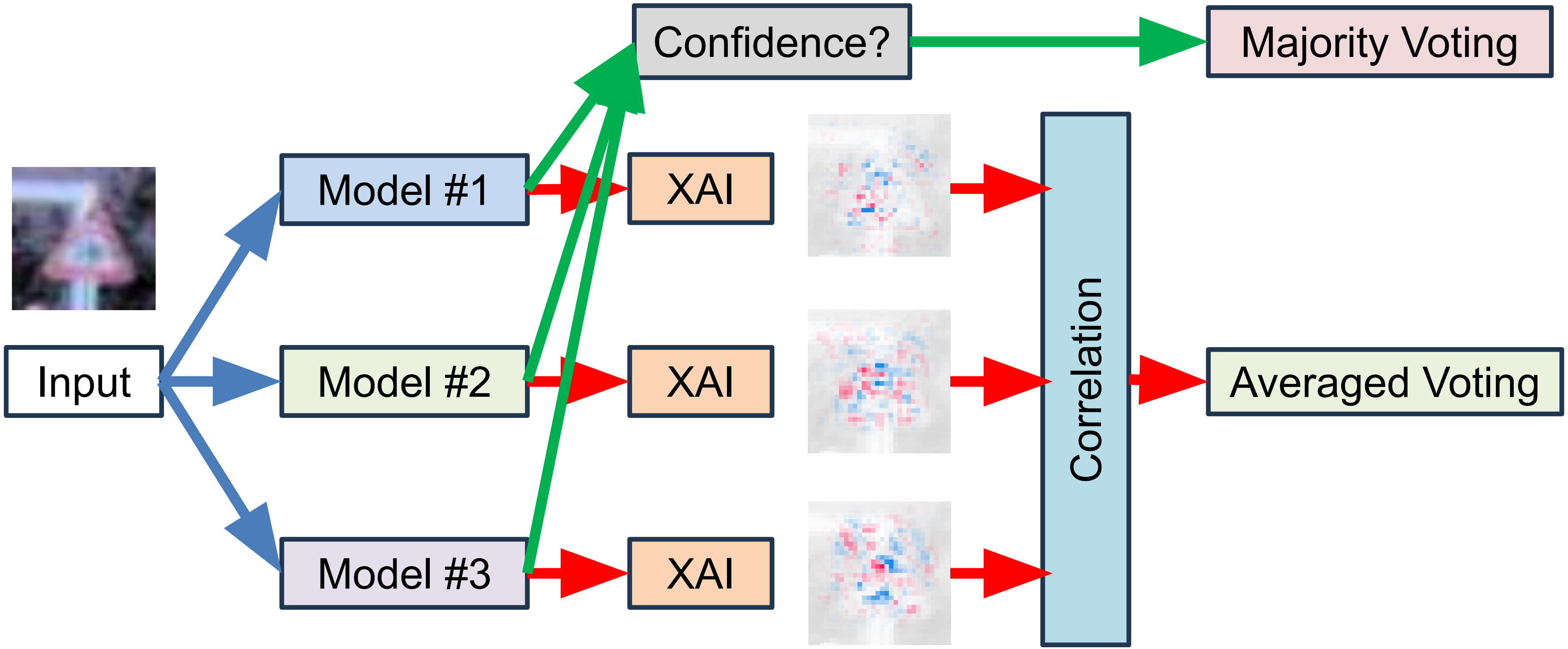


Example: Feature Diversity using SHAP



Concern: XAI can be slow during inference

Optimized Workflow



Intuition: SHAP Correlations

- Similarity Metric: R^2 Correlation
- Benchmark: GTSRB with 30% Mislabelling

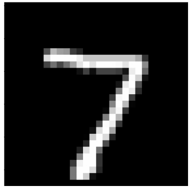
	ConvNet	DeconvNet	VGG11
ConvNet	1	$\boxed{0.72}$	$\boxed{0.53}$
DeconvNet	X	1	$\boxed{0.36}$
VGG11	X	X	1

From Similarity to Weights

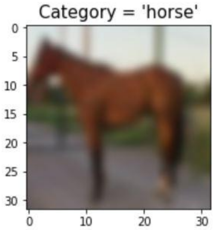
- Saliency maps -> matrices (M_1, M_2)

- Distance between M_1 and M_2
  - R^2
  - Cosine similarity
  - Wasserstein Distance (Earth Mover's Distance)

Exp Setup: Evaluation Datasets



MNIST
Handwritten Digits



CIFAR-10
Object Identification

GTSRB
Self-Driving Cars

Pneumonia
Medical Diagnosis

Safety-Critical Applications

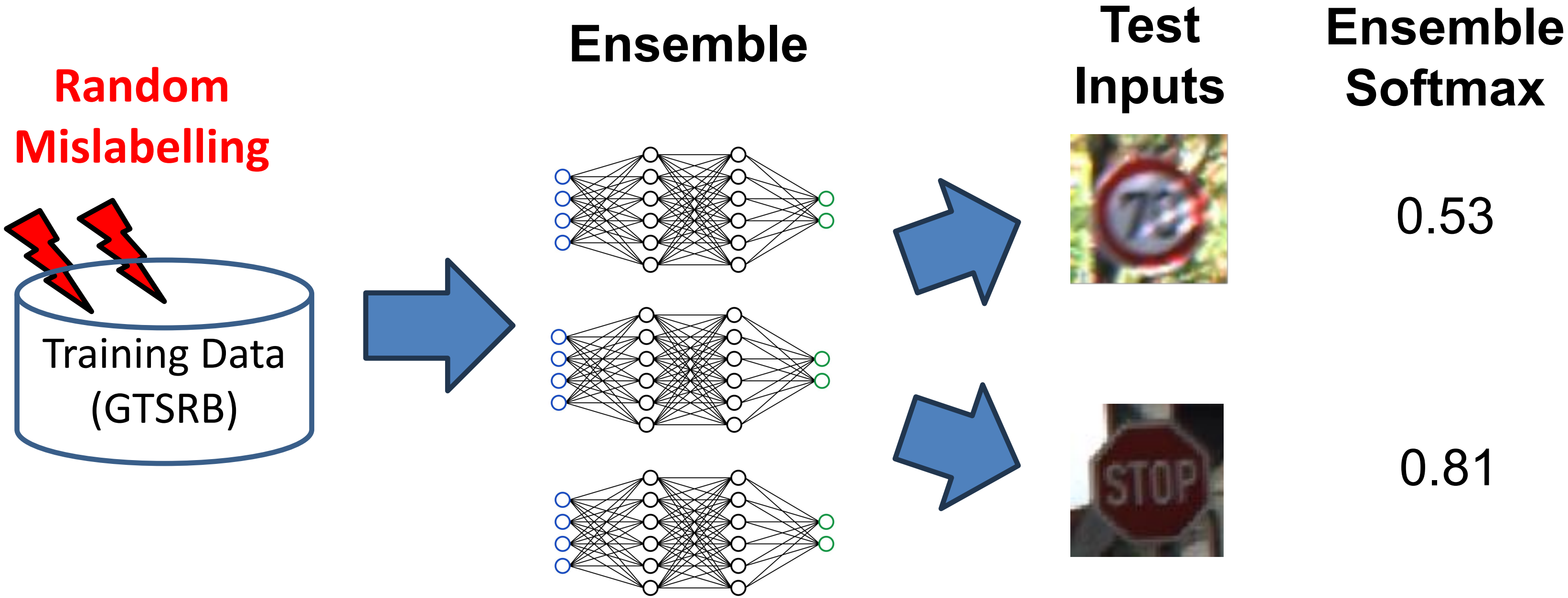
Exp Setup: Deep Neural Networks

ML Model Name	Depth (# of Layers)
ConvNet	Shallow
DeconvNet	Shallow
MobileNet	Deep
ResNet18	Deep
ResNet50	Deep
VGG11	Deep
VGG16	Deep

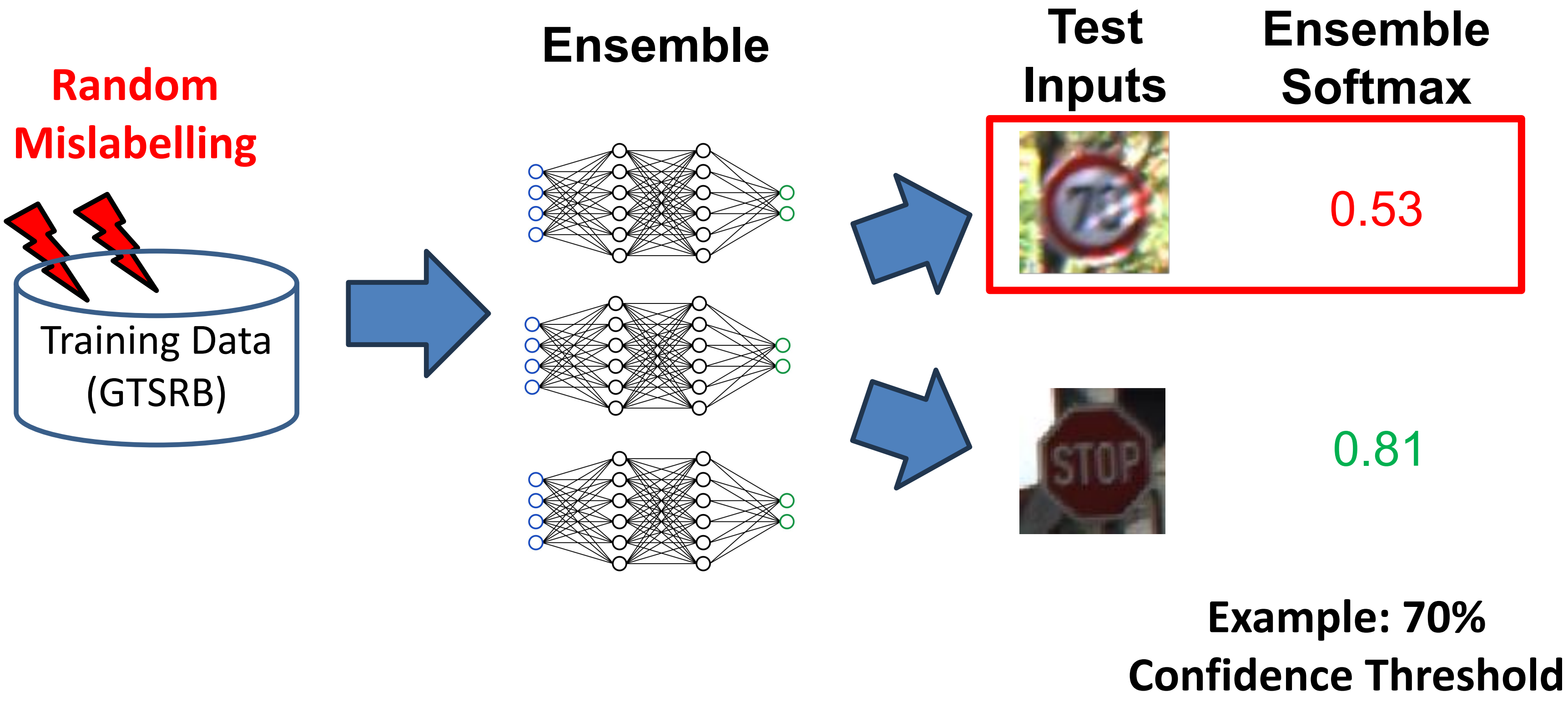
Research Questions

Feasibility	1. How many ensemble predictions are low confidence? 2. How diverse are ensembles in the feature space, compared to prediction confidence?
Runtime	3. When to use XAI?
Design	4. Which XAI technique? 5. How to determine dynamic weights?

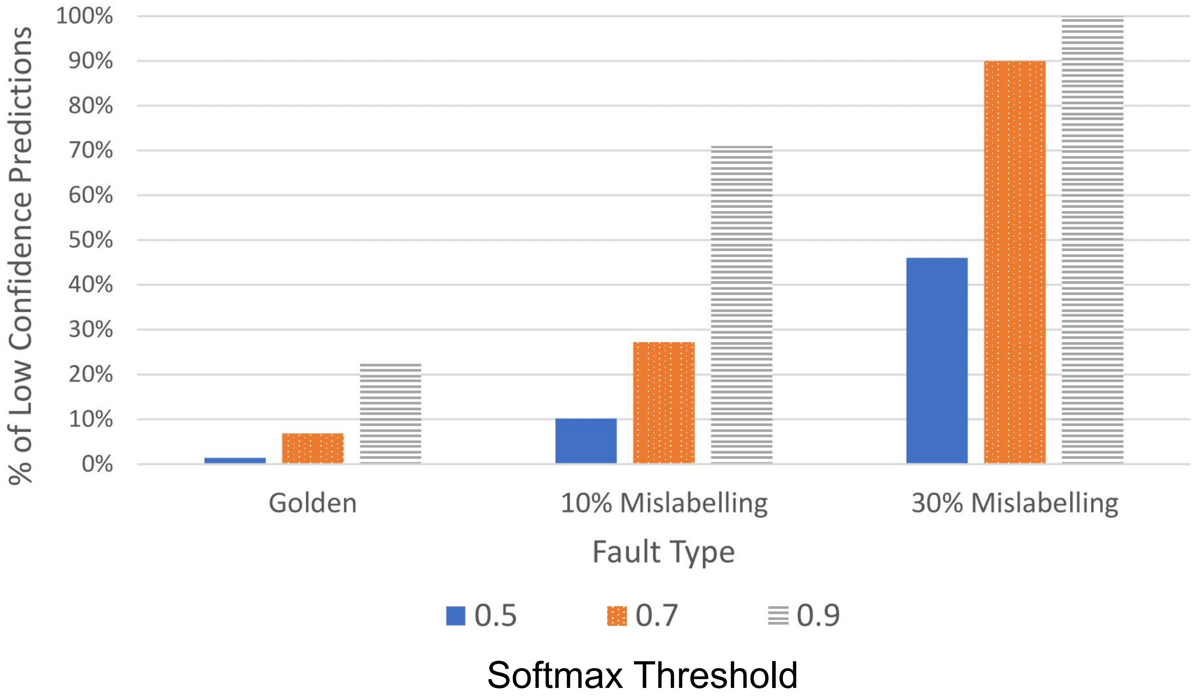
RQ1: Confidence under Training Faults



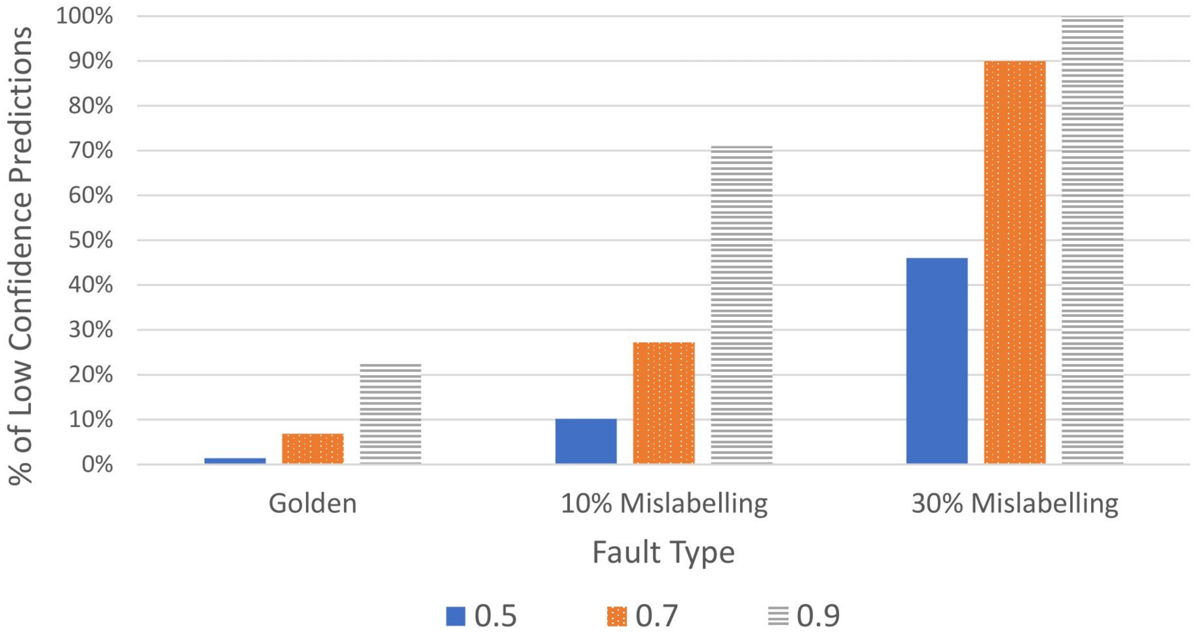
RQ1: Confidence under Training Faults



RQ1: Confidence under Training Faults



RQ1: Confidence under Training Faults



Softmax Threshold

Prediction confidence drops as faulty training data increases

RQ2: Confidence - Feature Correlation

- How diverse are ensembles in the feature space compared to their prediction confidence?

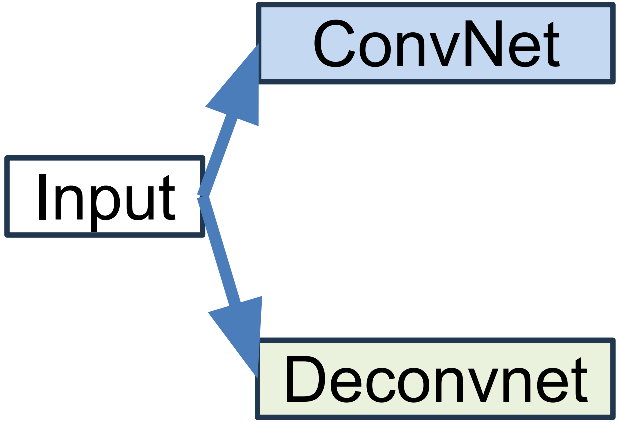
RQ2: Confidence - Feature Correlation

Goal:

We could use SHAP with confidence to determine dynamic weights

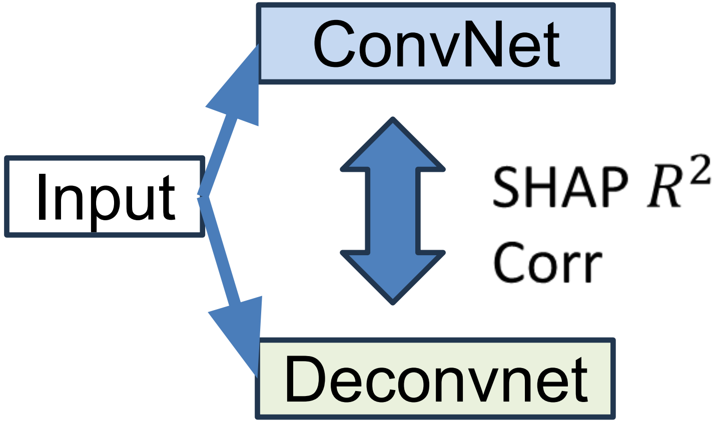
RQ2: Confidence - Feature Correlation

GTSRB, 30% mislabelling



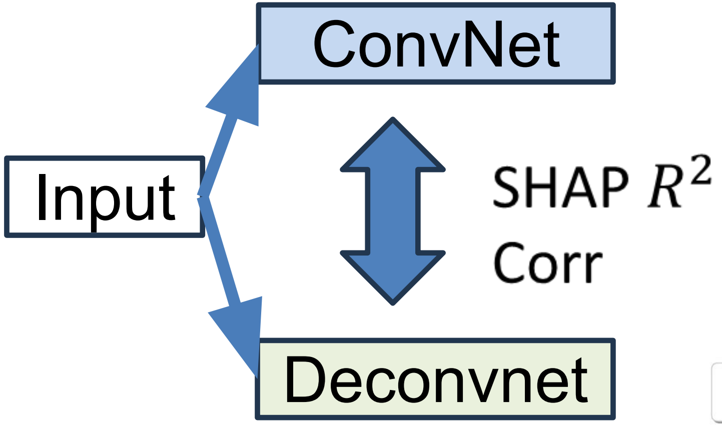
RQ2: Confidence - Feature Correlation

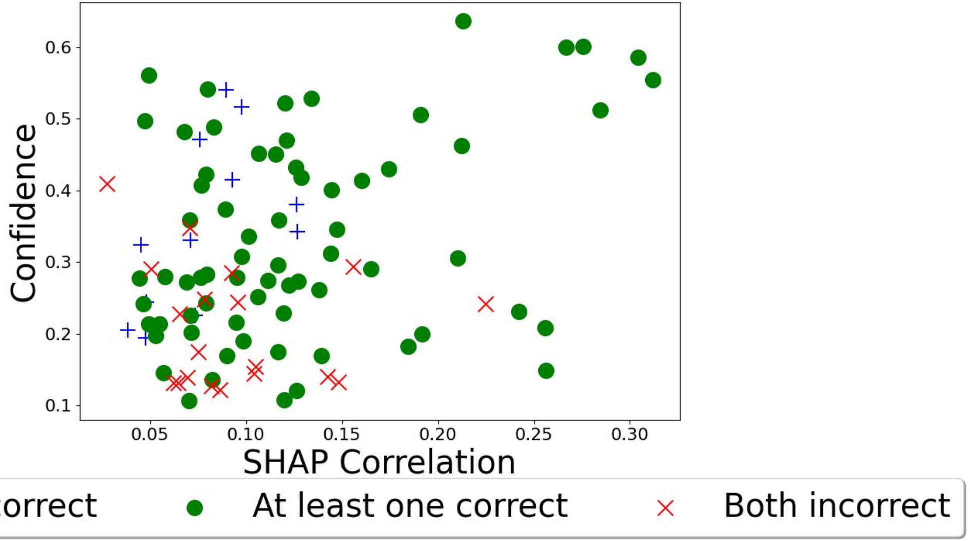
GTSRB, 30% mislabelling



RQ2: Confidence - Feature Correlation

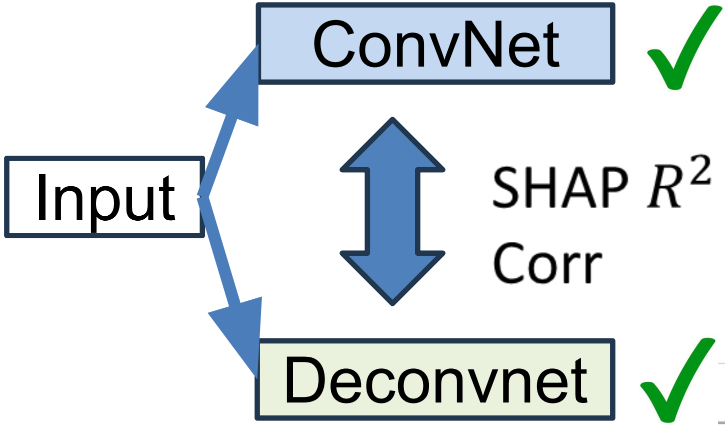
GTSRB, 30% mislabelling

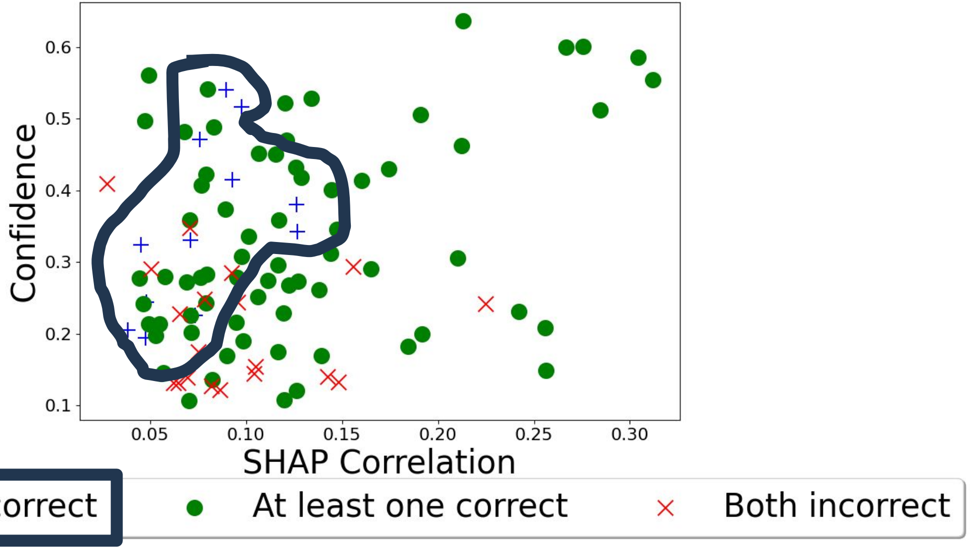




RQ2: Confidence - Feature Correlation

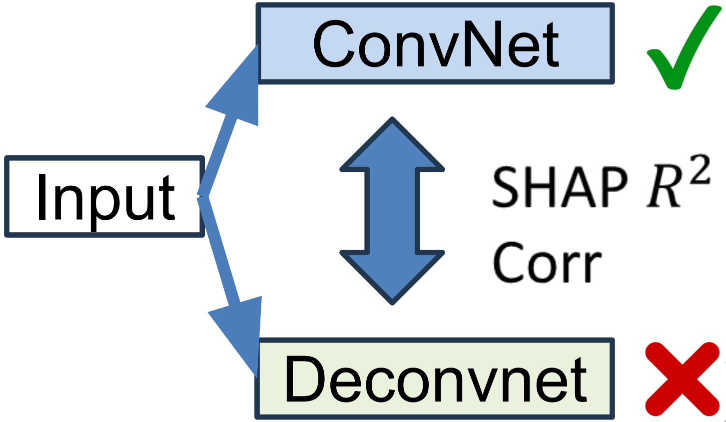
GTSRB, 30% mislabelling

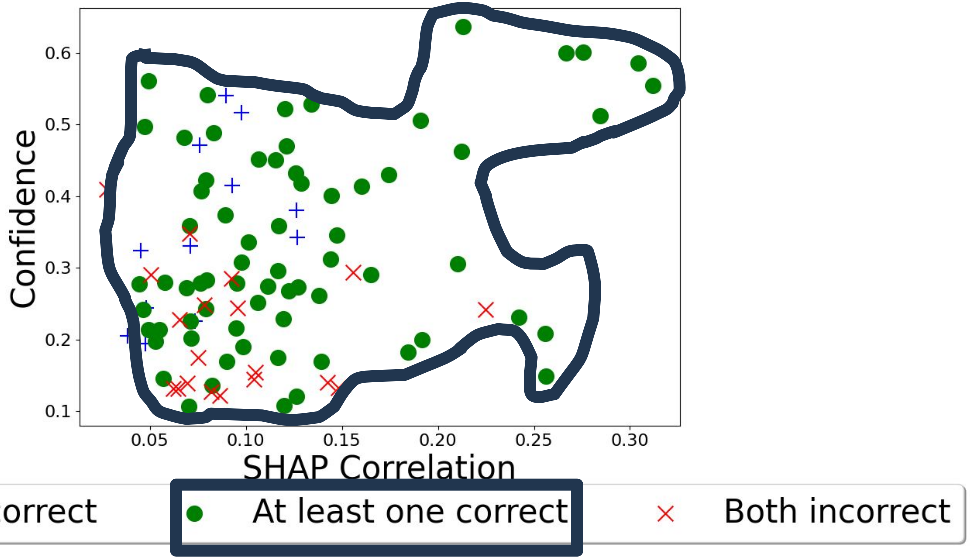




RQ2: Confidence - Feature Correlation

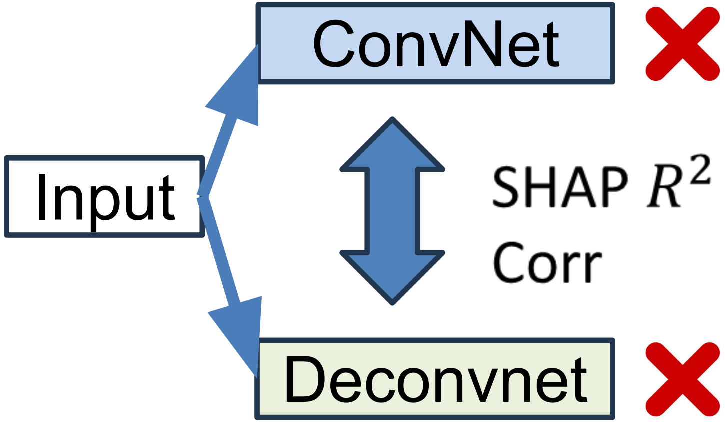
GTSRB, 30% mislabelling

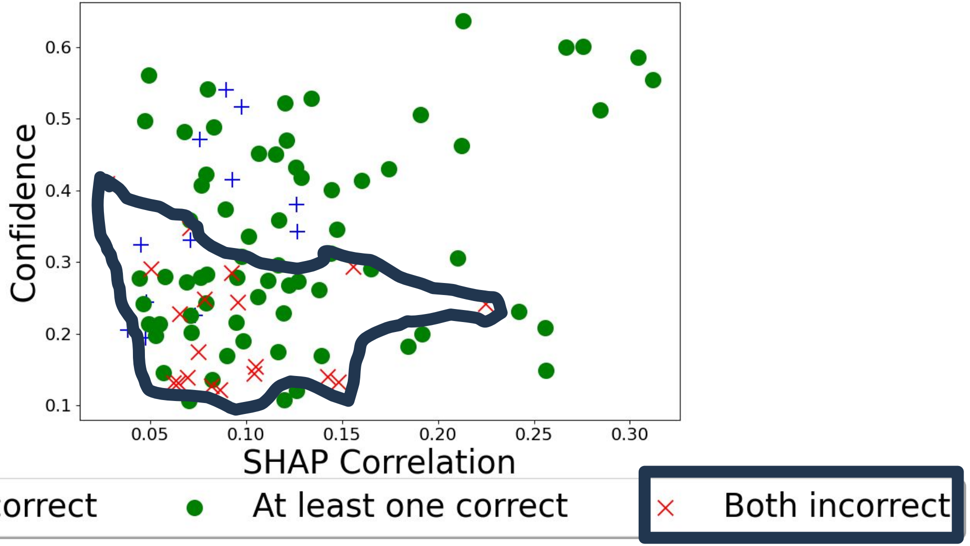




RQ2: Confidence - Feature Correlation

GTSRB, 30% mislabelling



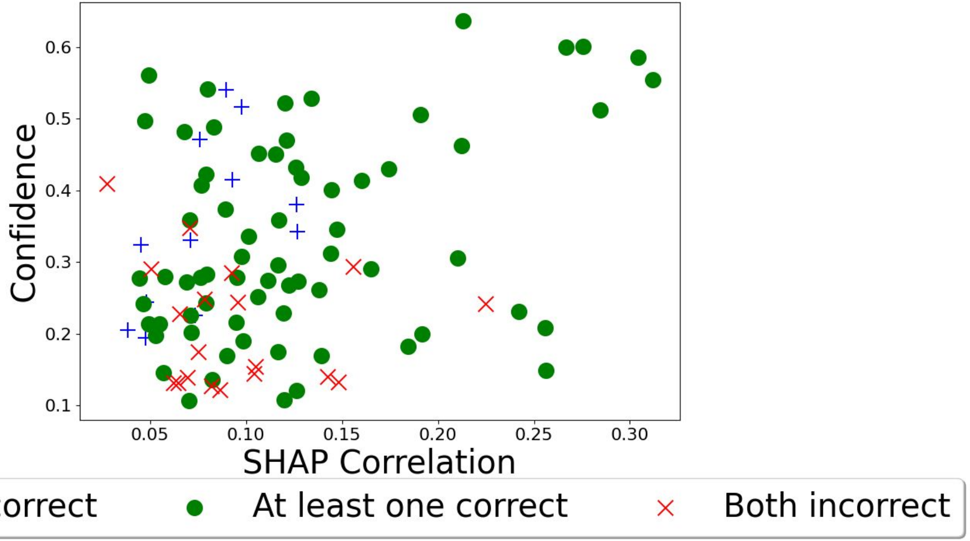


RQ2: Confidence - Feature Correlation

Observations:

- SHAP correlations are low for both correct and incorrect cases

- Not necessarily true that high confidence have higher SHAP correlation

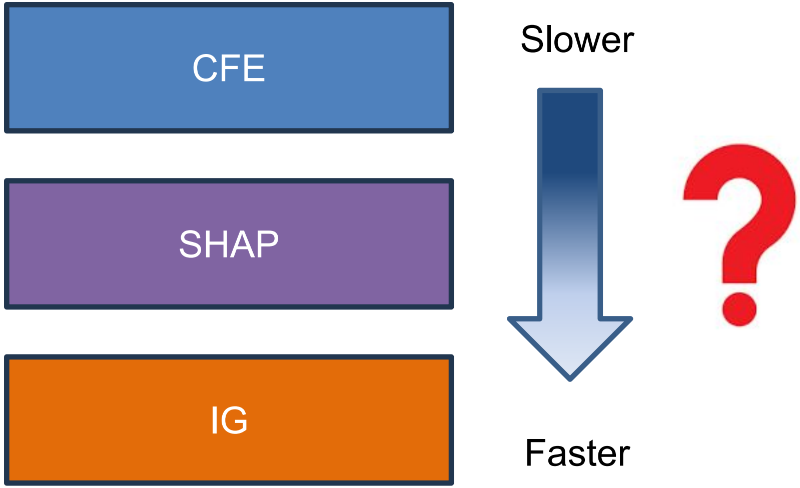


RQ3: Runtime Overhead

- **Intuition:** Only low confidence inputs needs to be checked by XAI

- 3.3x faster if XAI only applied on <u>low confidence</u> rather than every input

RQ4: Which XAI Technique?



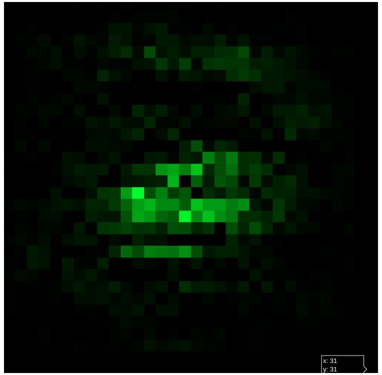
Criteria:

- Consistency
- Contrastivity
- Runtime

RQ5: How to determine weights?

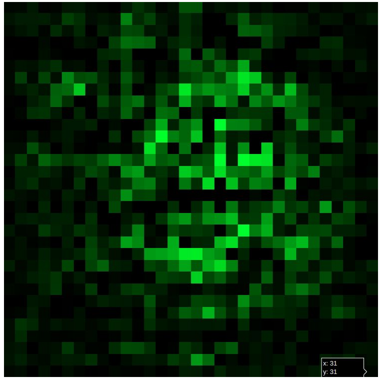
Image

ConvNet



Focused

VGG11



Dispersed

- All models, trained with GTSRB with 30% Mislabelling

Summary

1. Ensembles are resilient, but need dynamic weights

2. Use XAI to determine ensemble weights

3. Combining XAI with prediction confidence

Email: abrahamc@ece.ubc.ca