ReMlX: Resilience for ML Ensembles using XAI at Inference against Faulty Training Data

Abraham Chan,

Arpan Gujarati, Karthik Pattabiraman, Sathish Gopalakrishnan

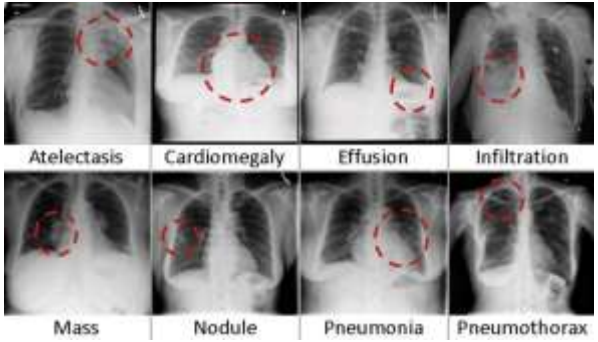
THE UNIVERSITY OF BRITISH COLUMBIA

Training Data Faults in Practice

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

Image **Label**

~~Speed 80km/h~~
Stop sign

~~Stop sign~~
Speed 80km/h


No entry

Normal
Mislabelling

Image **Label**

Speed 80km/h

Stop sign


~~No entry~~

Removal

Image **Label**

Speed 80km/h

Stop sign


No entry


No entry

Repetition

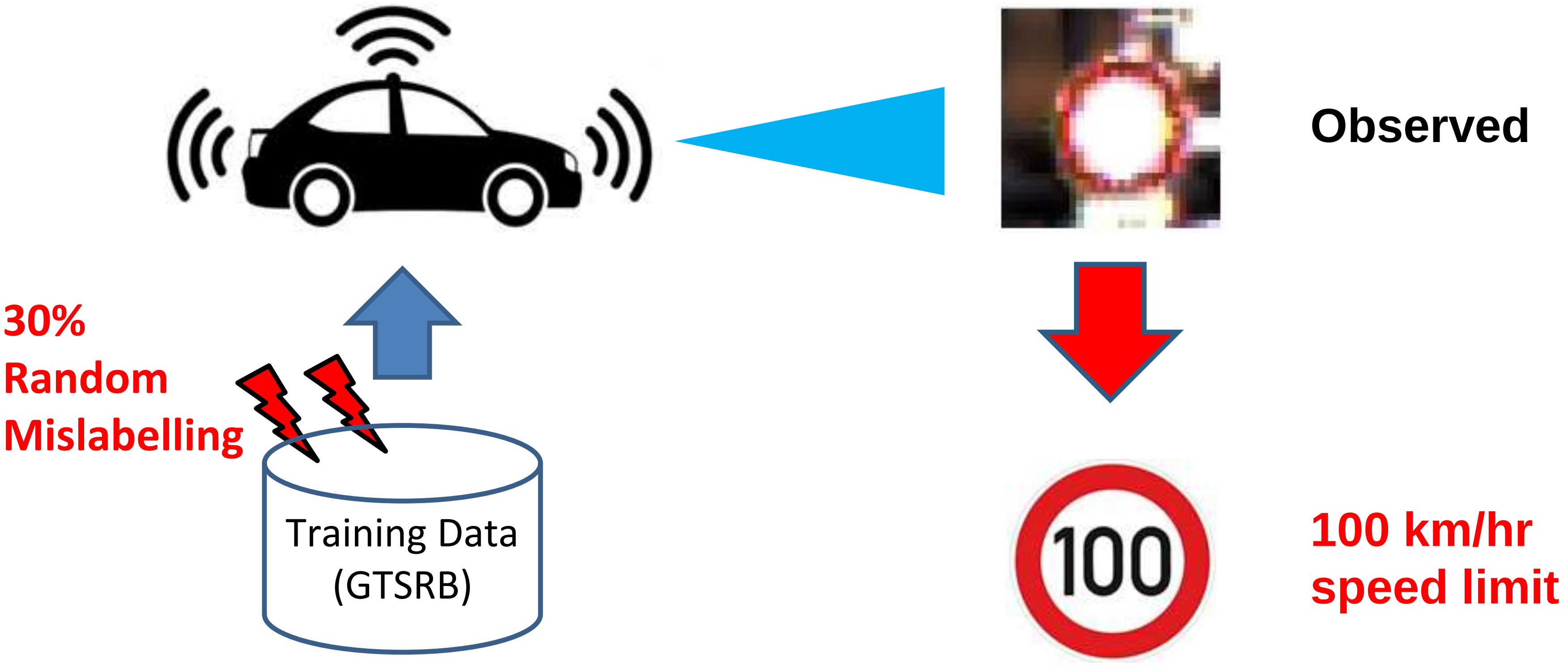
Autonomous Vehicle Example

Observed

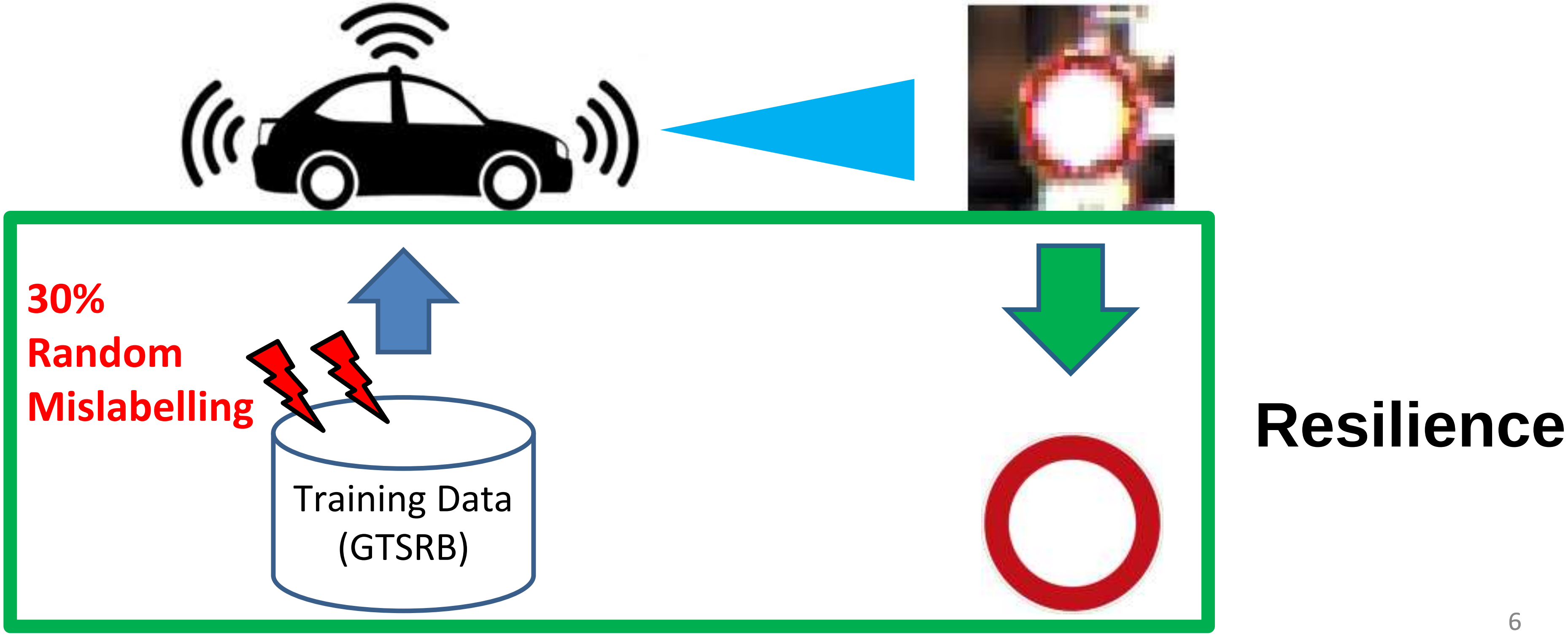
Training Data
(GTSRB)

No vehicles permitted

Random Mislabelling



Resilience against Faulty Training Data



How to mitigate training data faults with minimal human effort?

1. Label Correction
2. Knowledge Distillation
3. Robust Loss
4. Label Smoothing
5. Ensembles

More Practitioner Effort

Less Practitioner Effort

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

How to mitigate training data faults with minimal human effort?

1. Label Correction
2. Knowledge Distillation
3. Robust Loss
4. Label Smoothing
5. **Ensembles**

More Practitioner Effort

Less Practitioner Effort

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

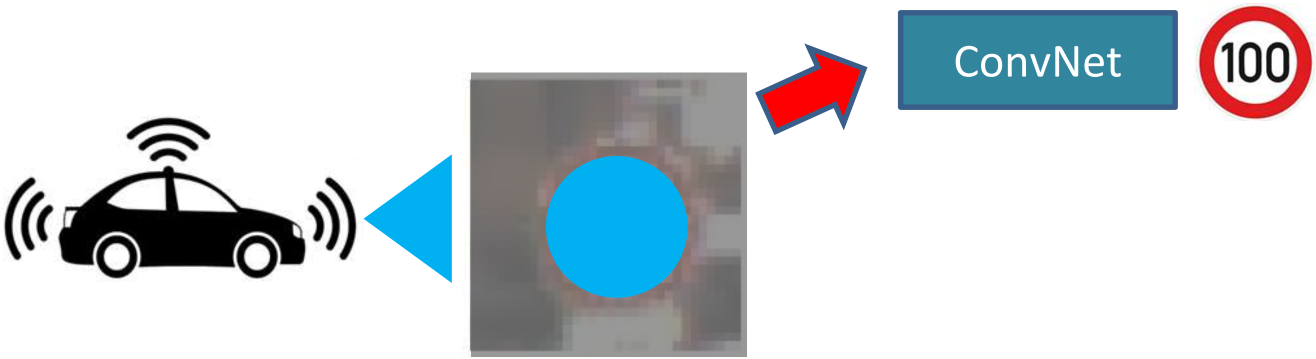
How to mitigate training data faults with minimal human effort?

1. Label Correction
2. Knowledge Distillation
4. Label Smoothing
5. Ensembles

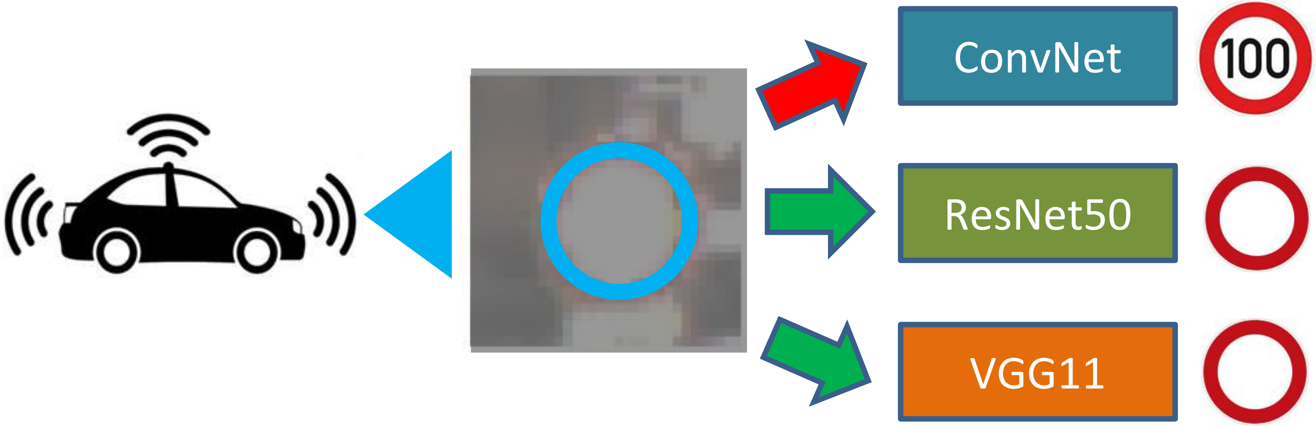
Our Solution: Building Resilient Ensembles

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

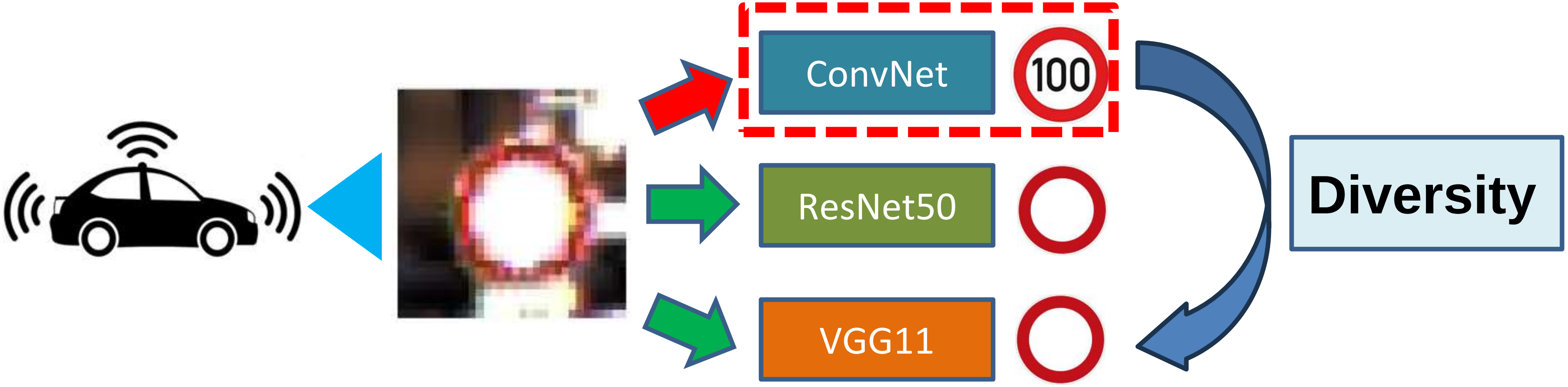
Resilient Ensembles



Resilient Ensembles

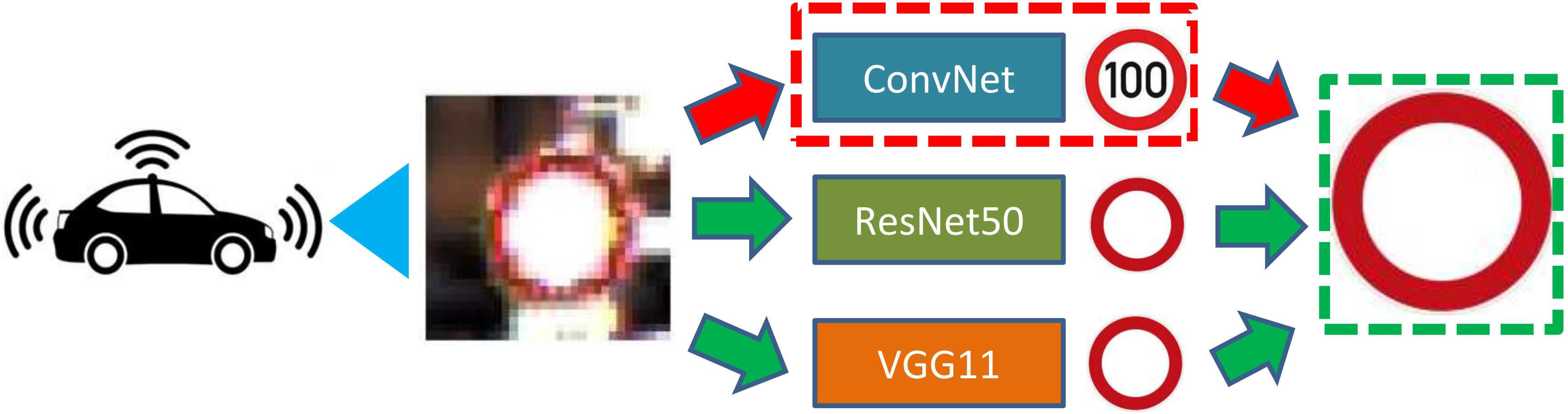


Resilient Ensembles



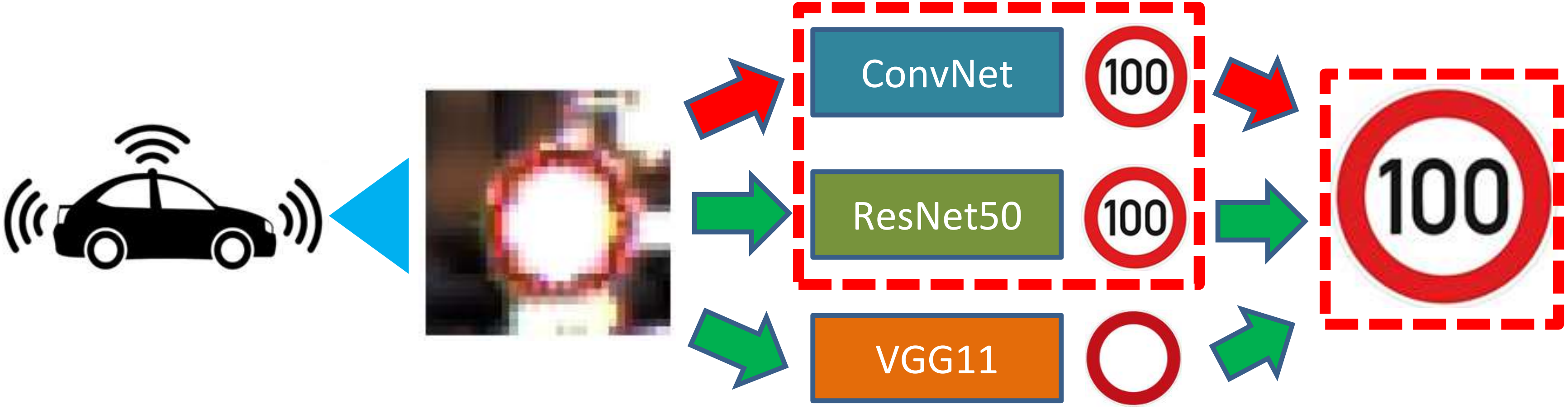
Our Prior Work: Understanding the Resilience of Neural Network Ensembles against Faulty Training Data **[Chan, QRS'21]**

Resilient Ensembles



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

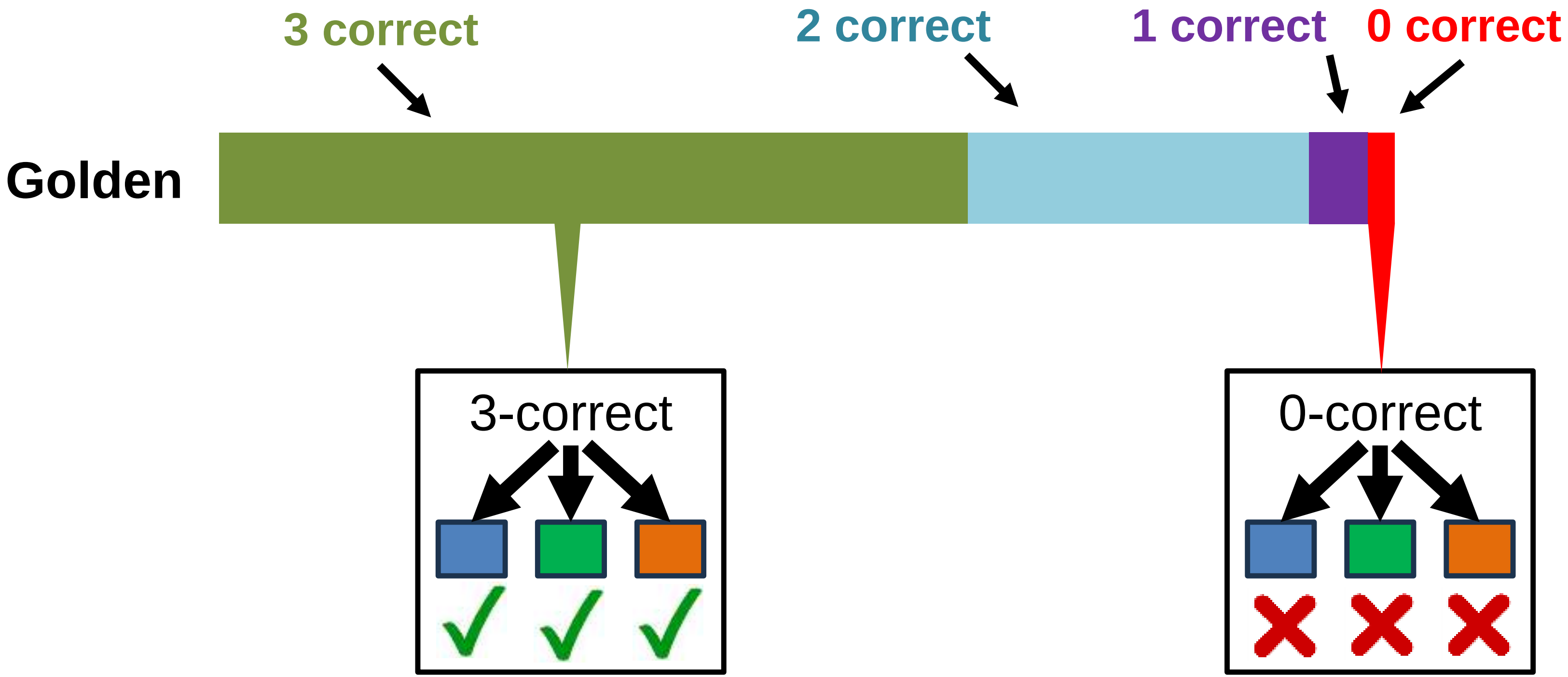
When Ensembles Misclassify?



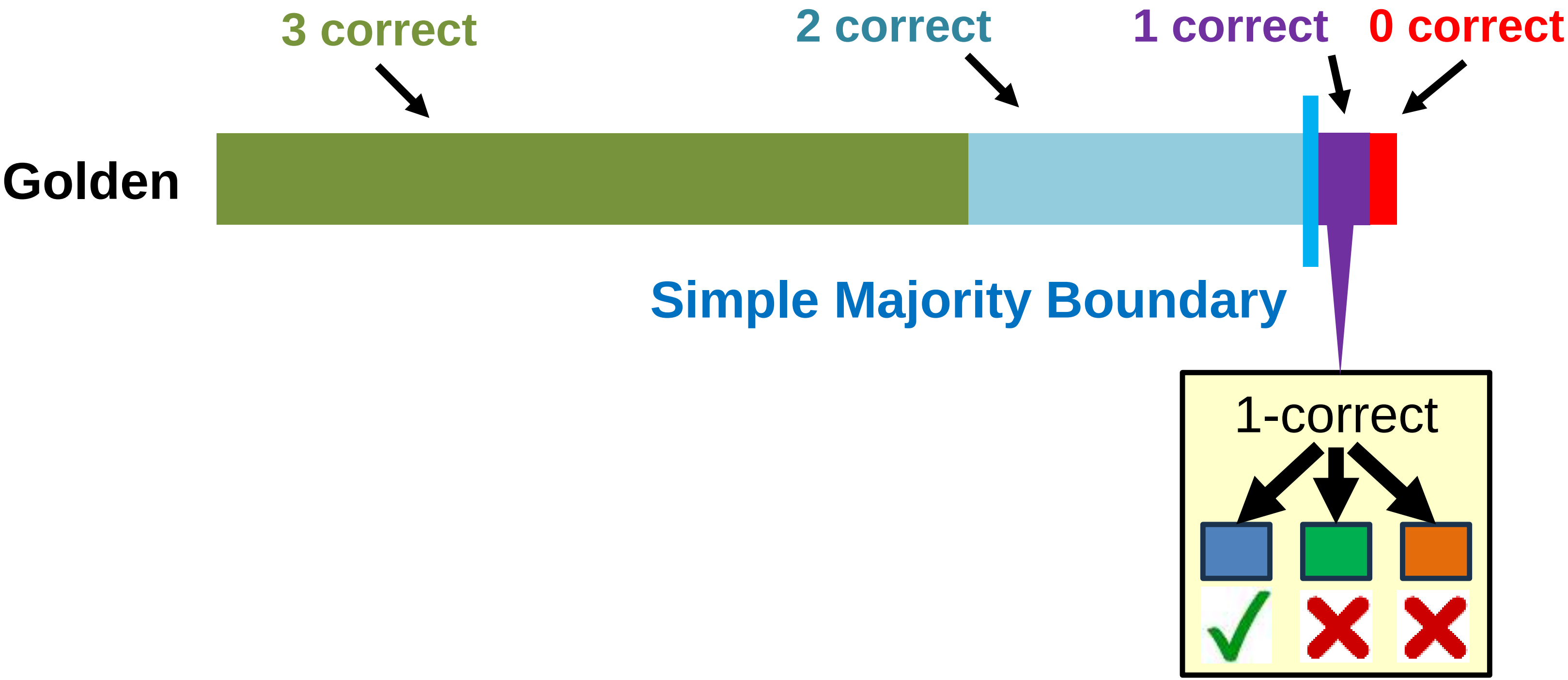
This paper's contribution:

ReMIX reduces ensemble misclassifications

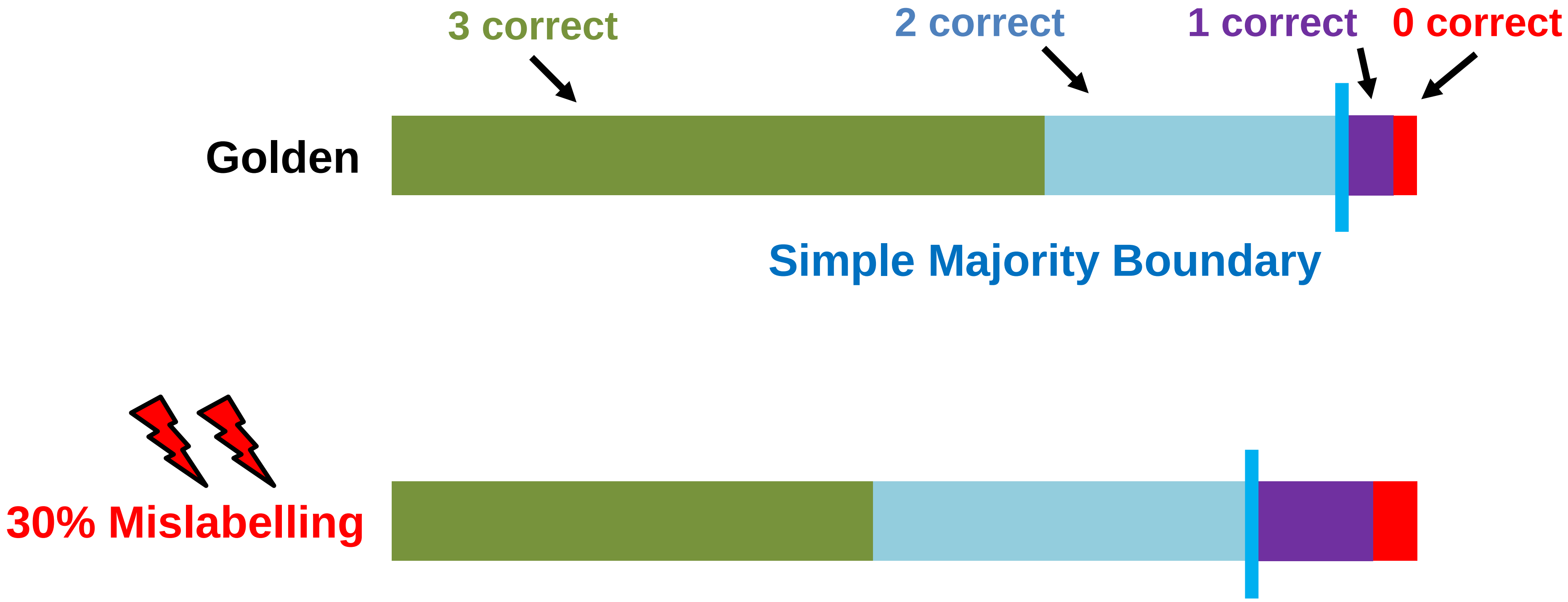
Correct Models in Ensemble



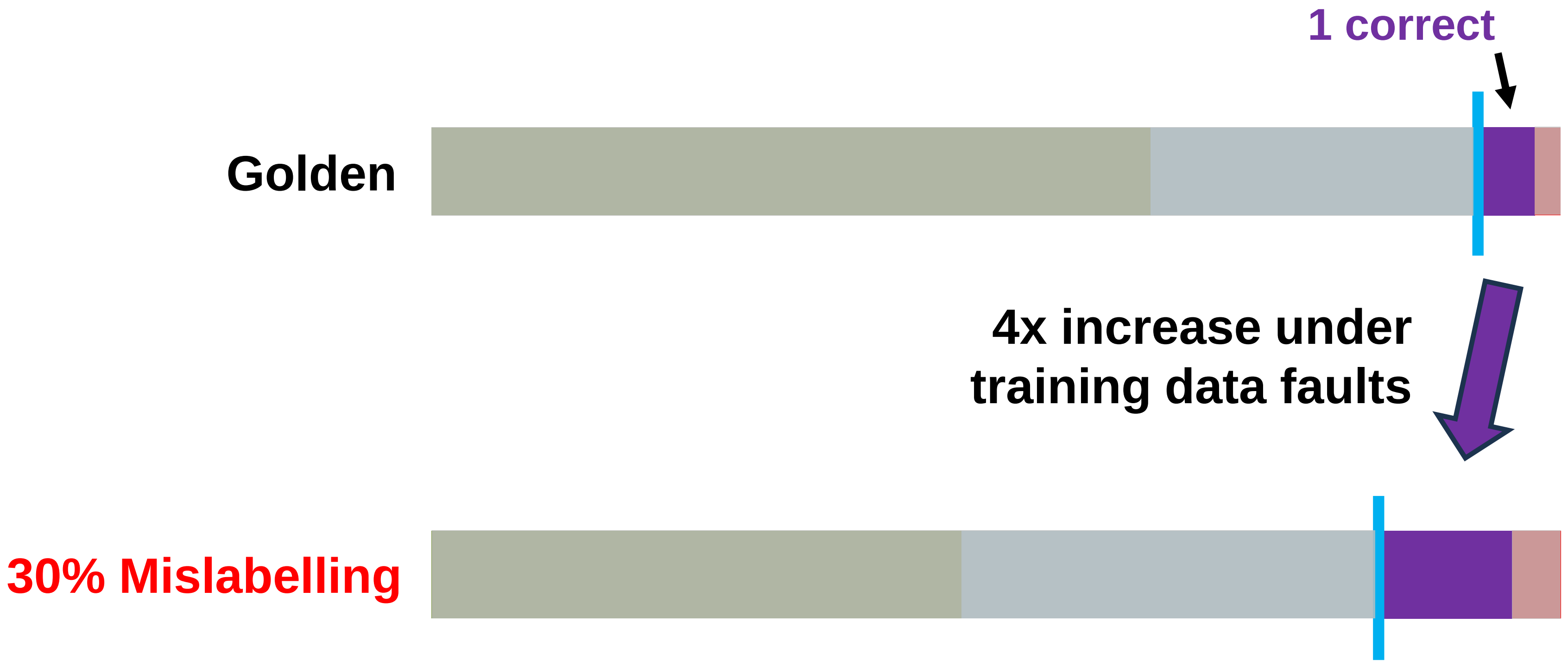
Correct Models in Ensemble



Correct Models in Ensemble

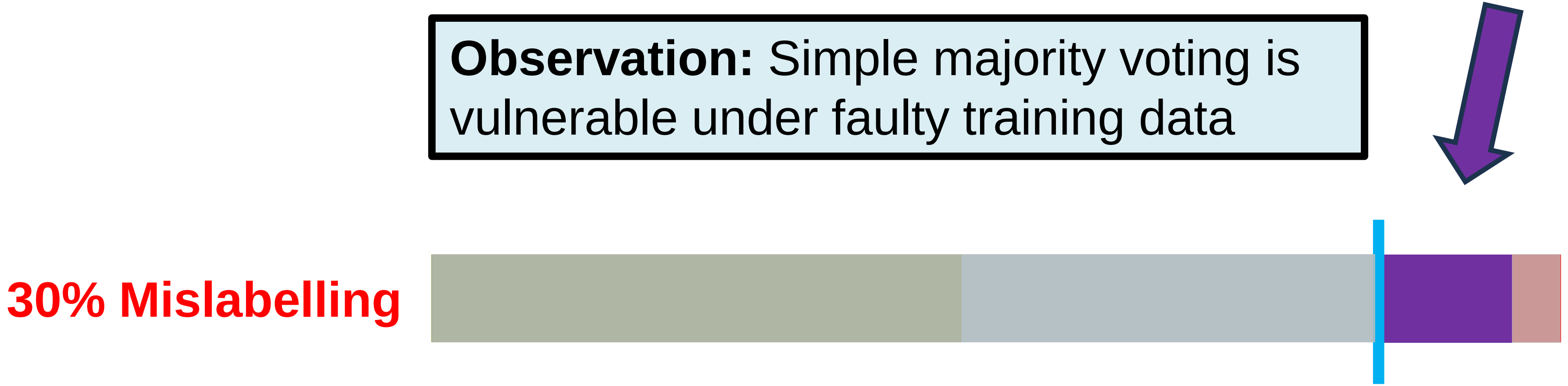


Correct Models in Ensemble

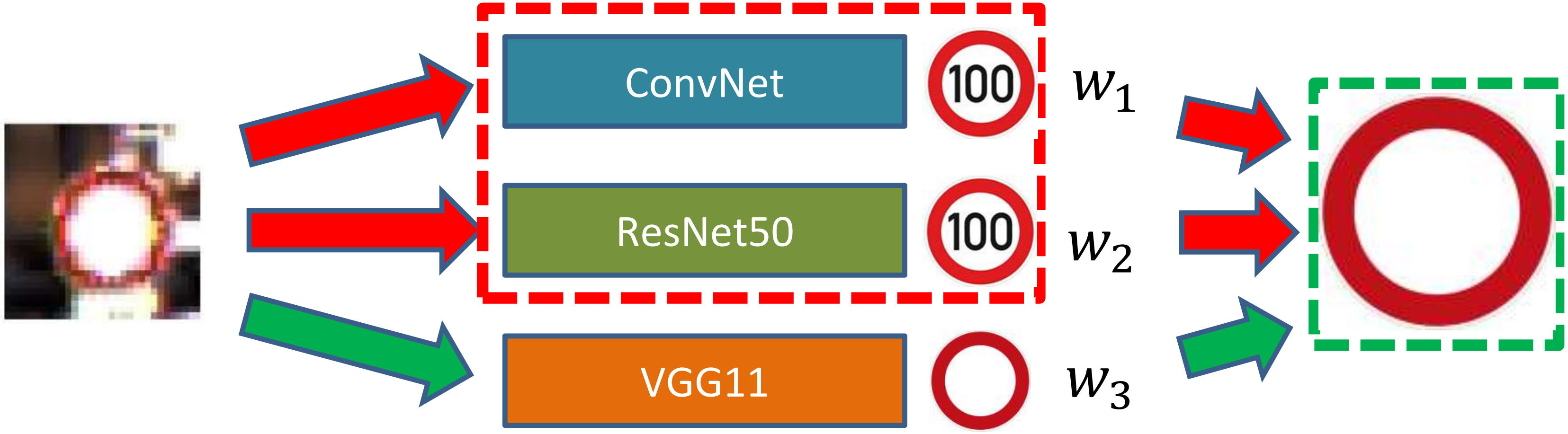


Correct Models in Ensemble

Observation: Simple majority voting is vulnerable under faulty training data



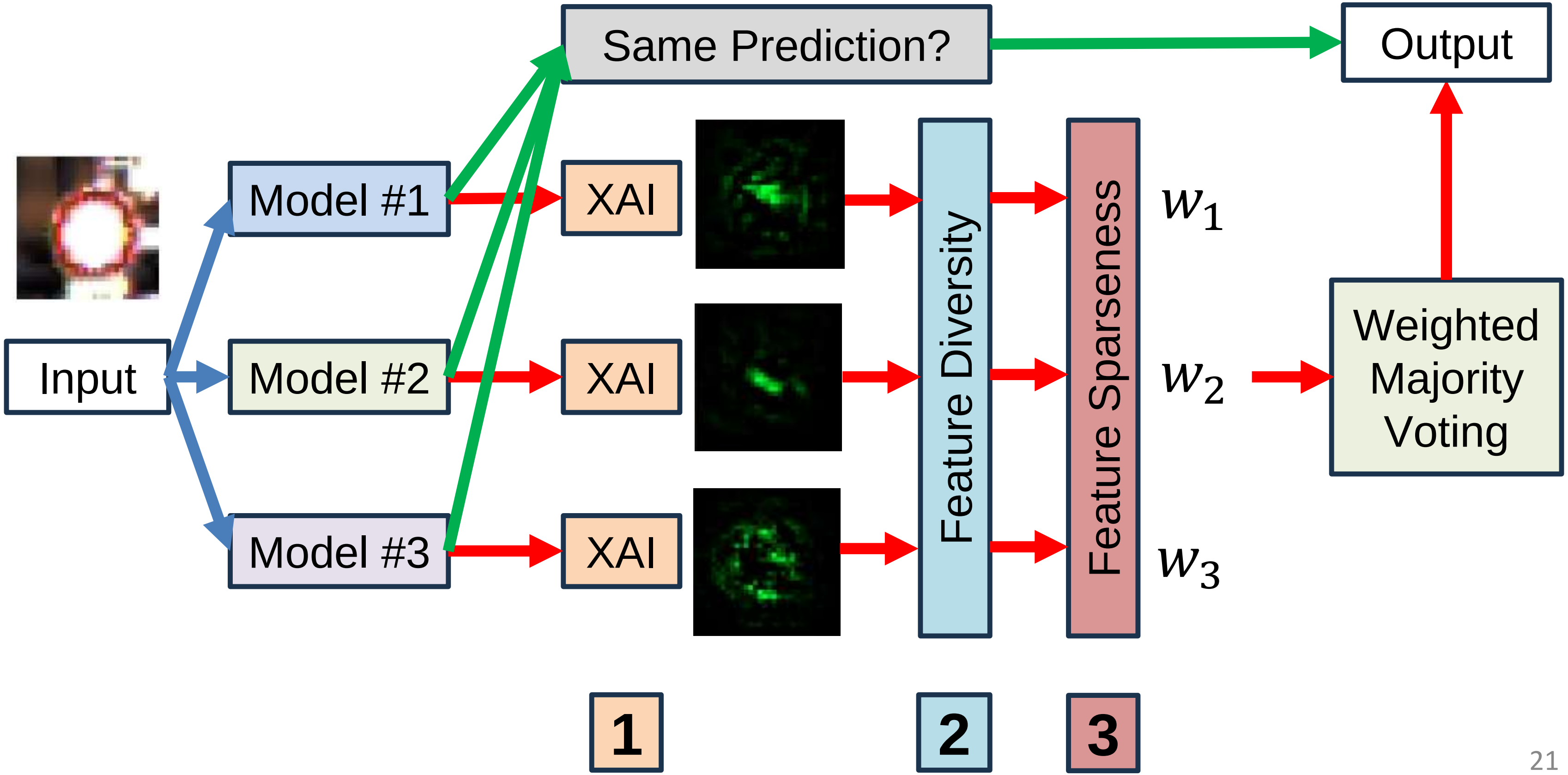
Dynamically Weighted Ensembles



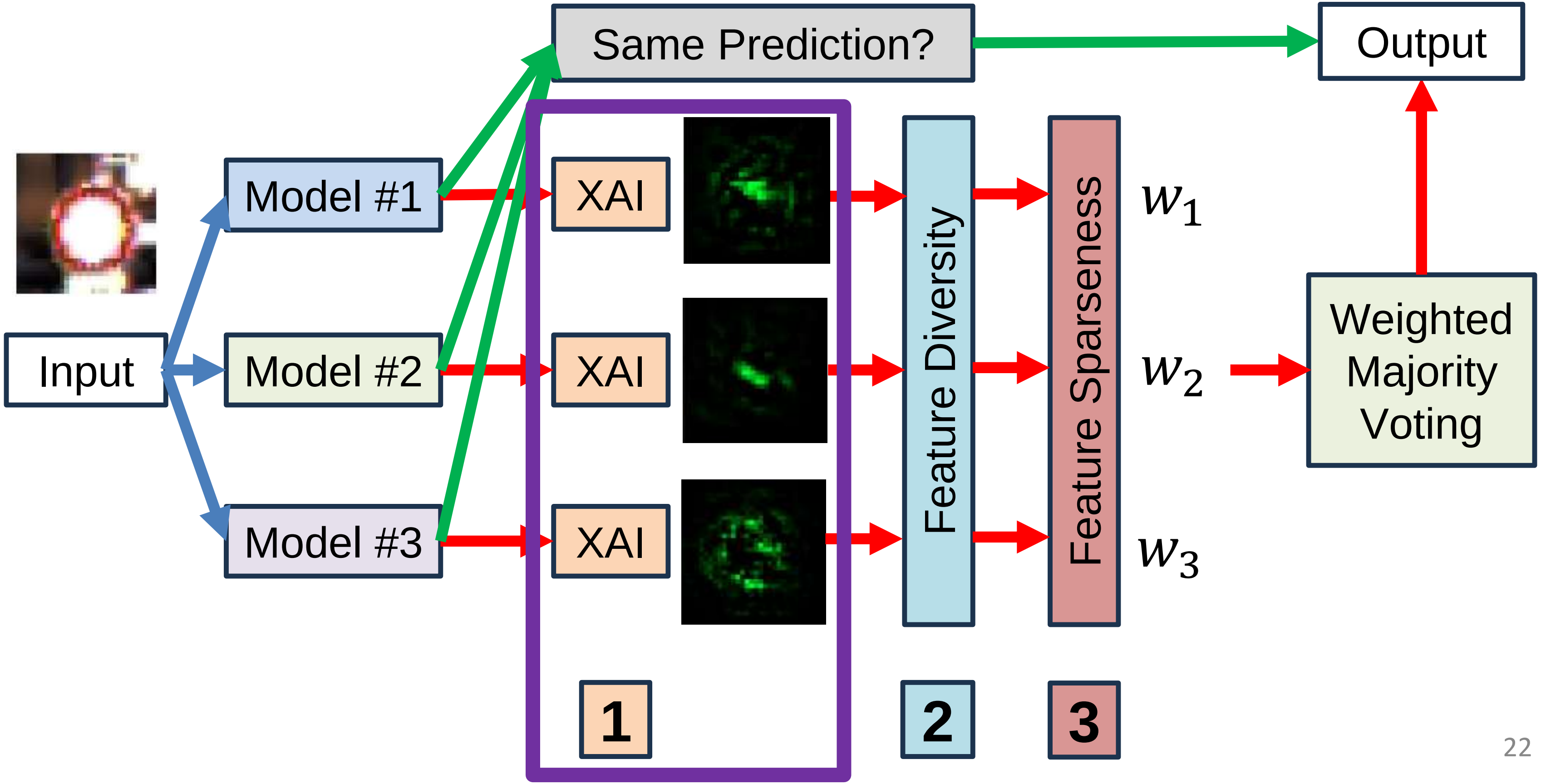
How to determine weights?
ReMIX uses Feature Space Diversity!

$$w_1, w_2 < w_3$$

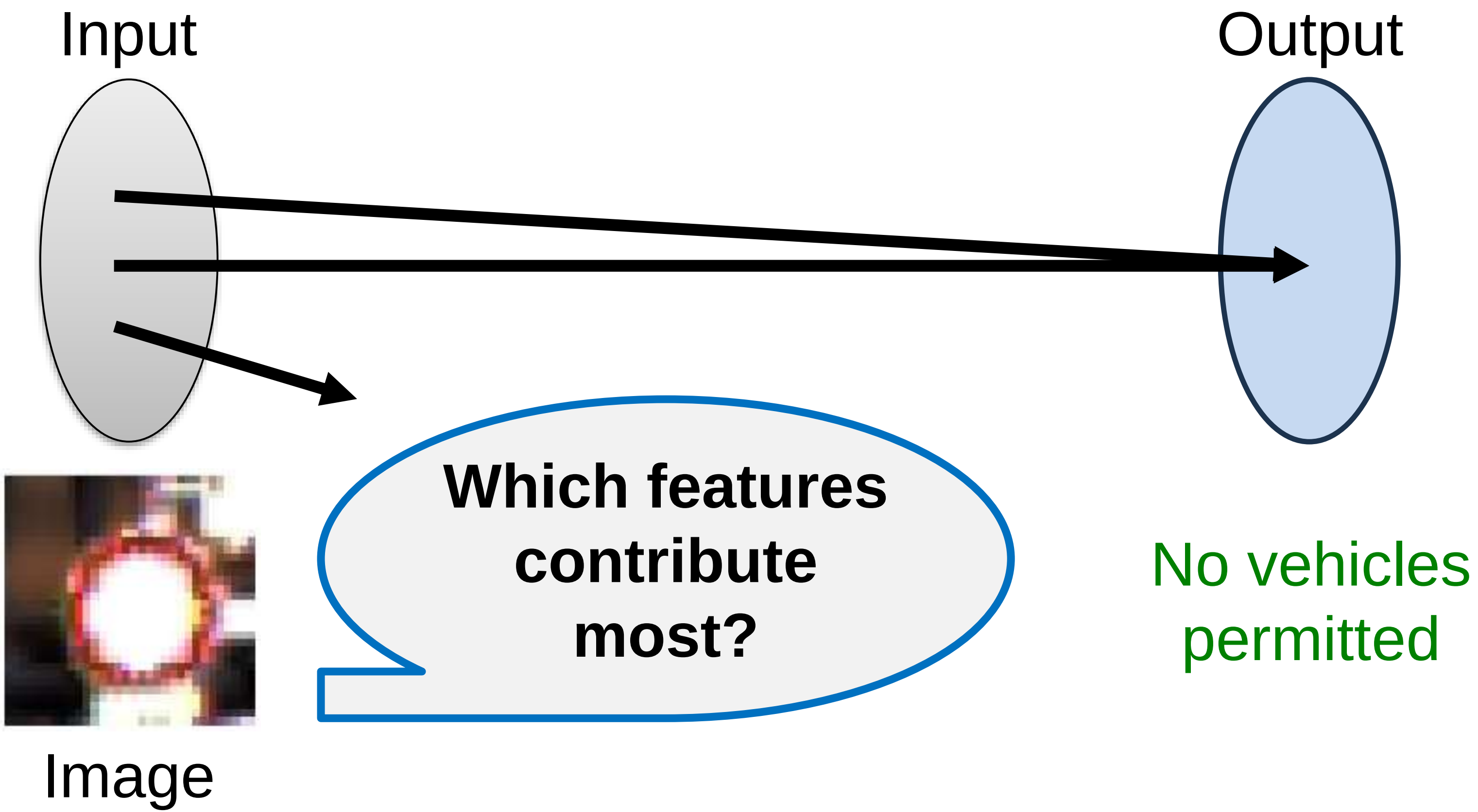
ReMIX Workflow



Step 1 – Explainable AI (XAI)

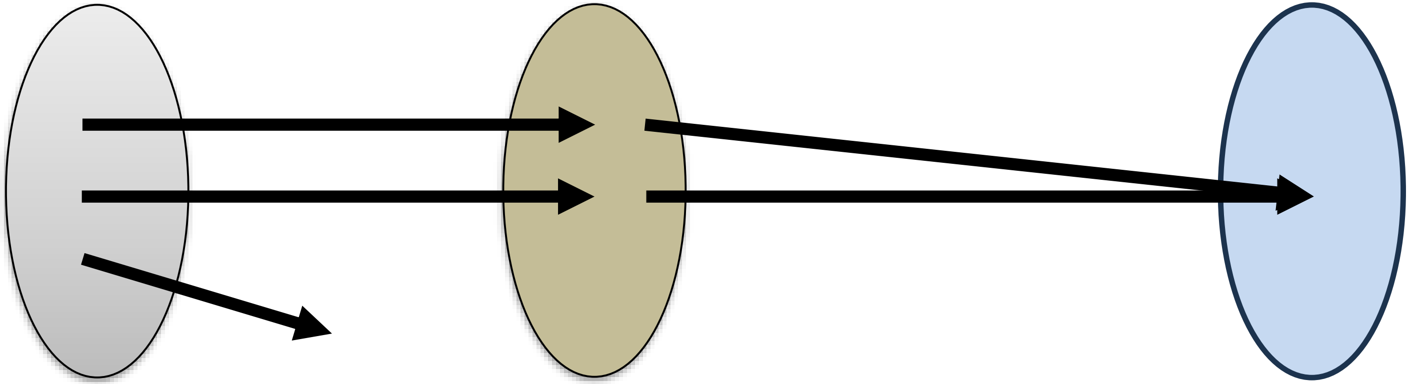


Input - Output Space

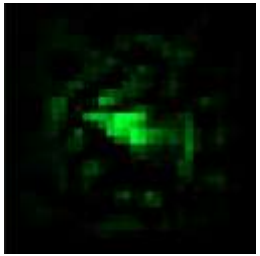


Feature Space

Input Feature Output



Image



Feature Map

No vehicles permitted

Post-Hoc Local Explainable AI (XAI)

Post-Hoc XAI

ML Model

No vehicles permitted

Post-Hoc Local Explainable AI (XAI)

1. Smooth Gradients (SG)
2. Integrated Gradients (IG)
3. SHAP
4. LIME
5. Counterfactual Explanations

Faithful?

Robust?

Efficient?

ML Model

No vehicles permitted

Post-Hoc Local Explainable AI (XAI)

1. **Smooth Gradients (SG)**
2. Integrated Gradients (IG)
3. SHAP
4. LIME
5. Counterfactual Explanations

Faithful?

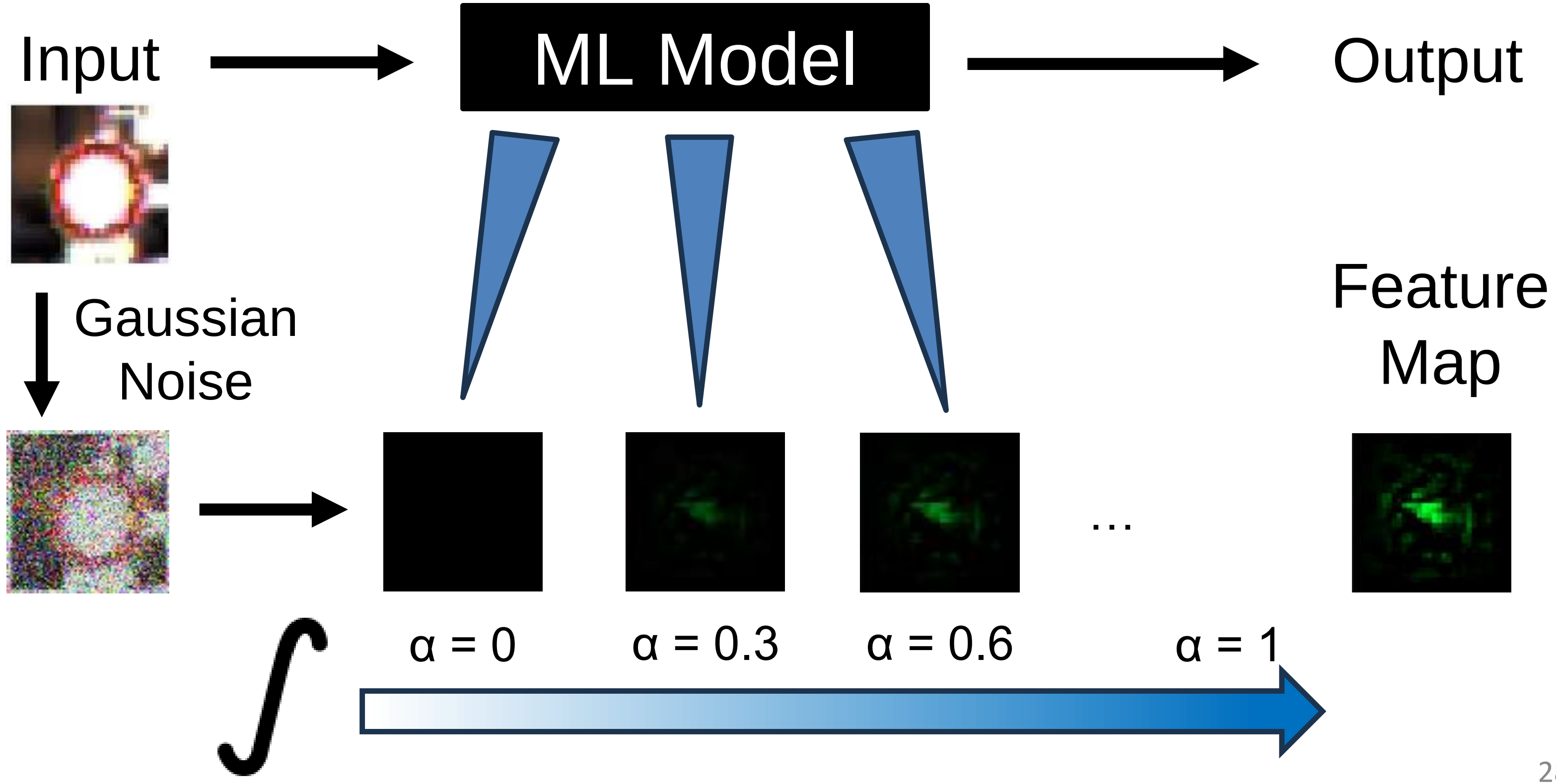
Robust?

Efficient?

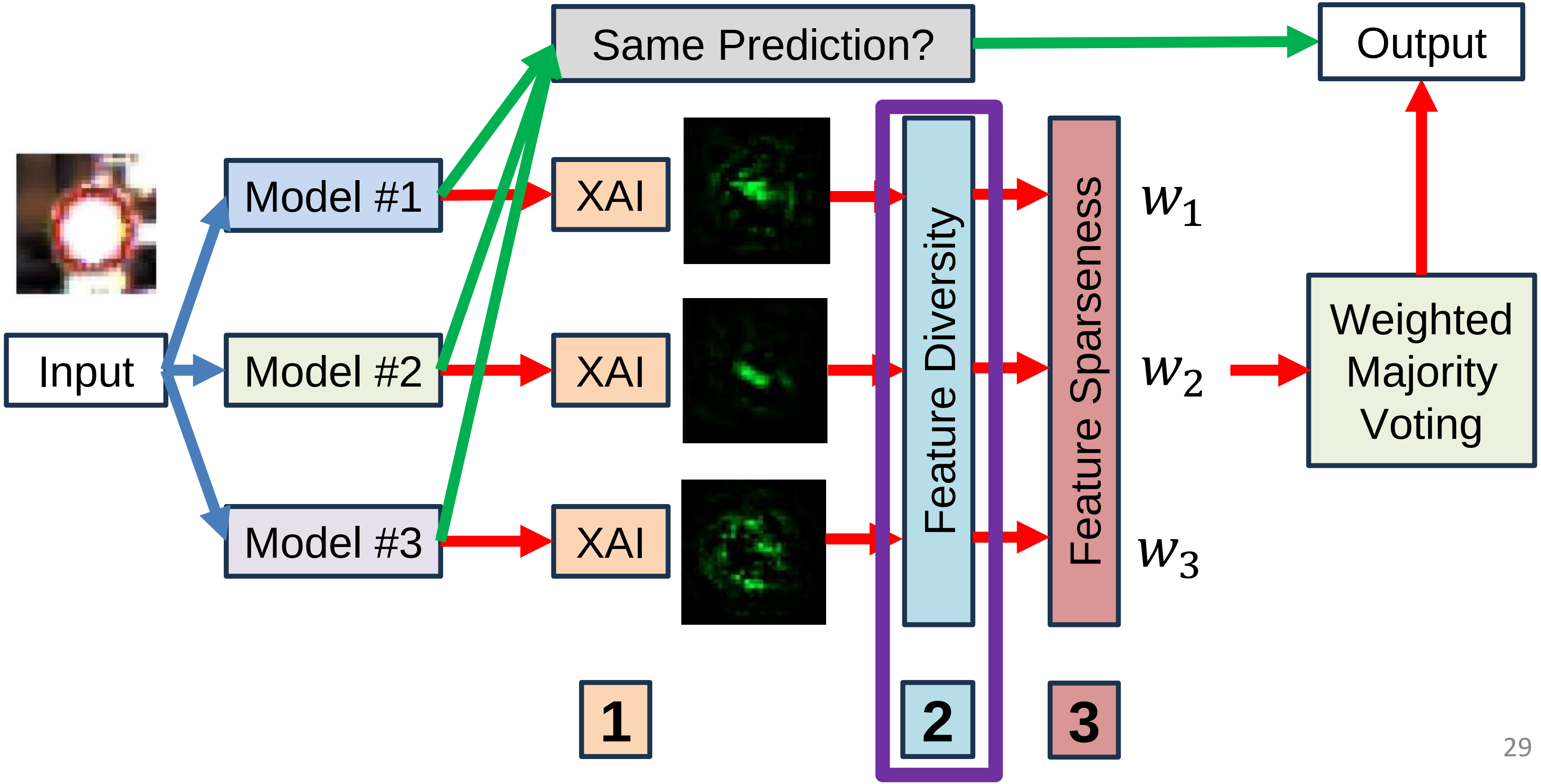
ML Model

No vehicles permitted

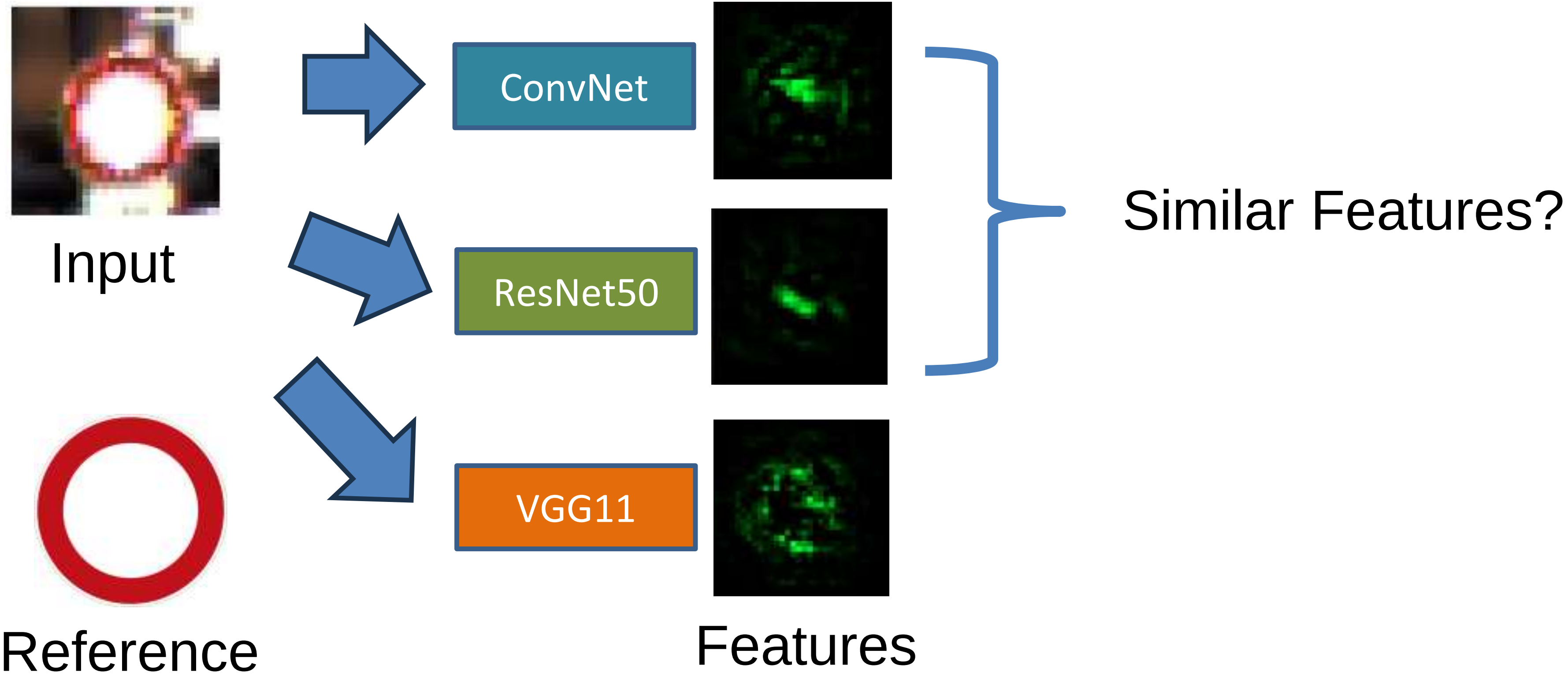
Smooth Gradients (SG)



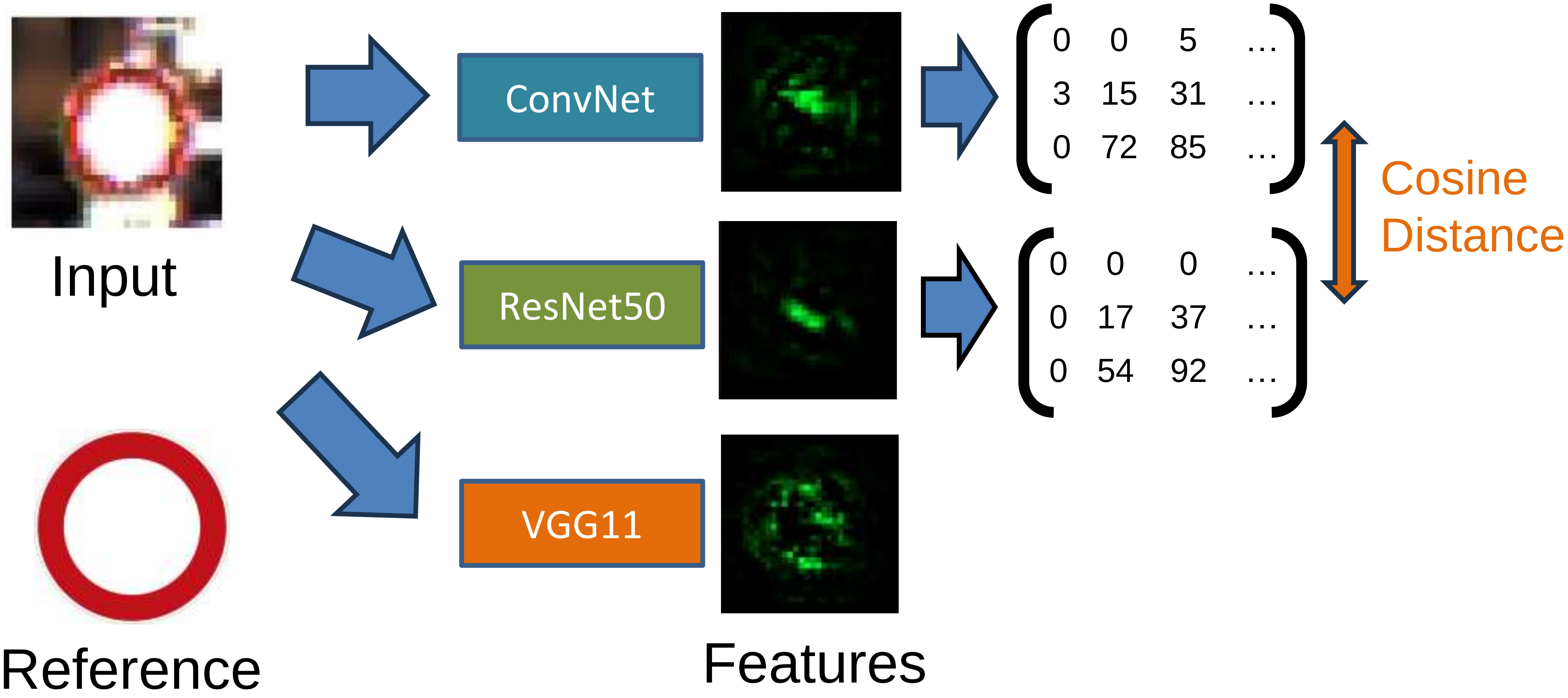
Step 2 – Feature Diversity



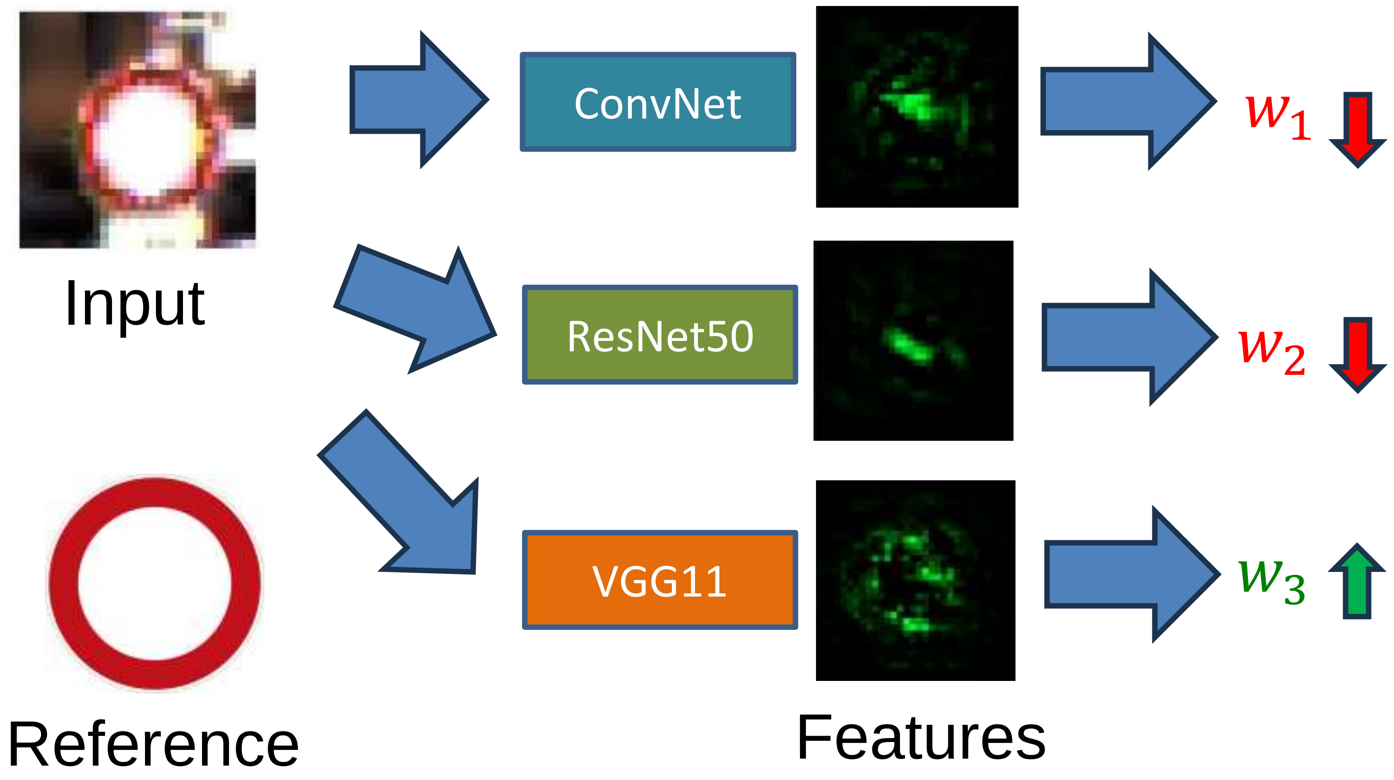
Feature Diversity using SG



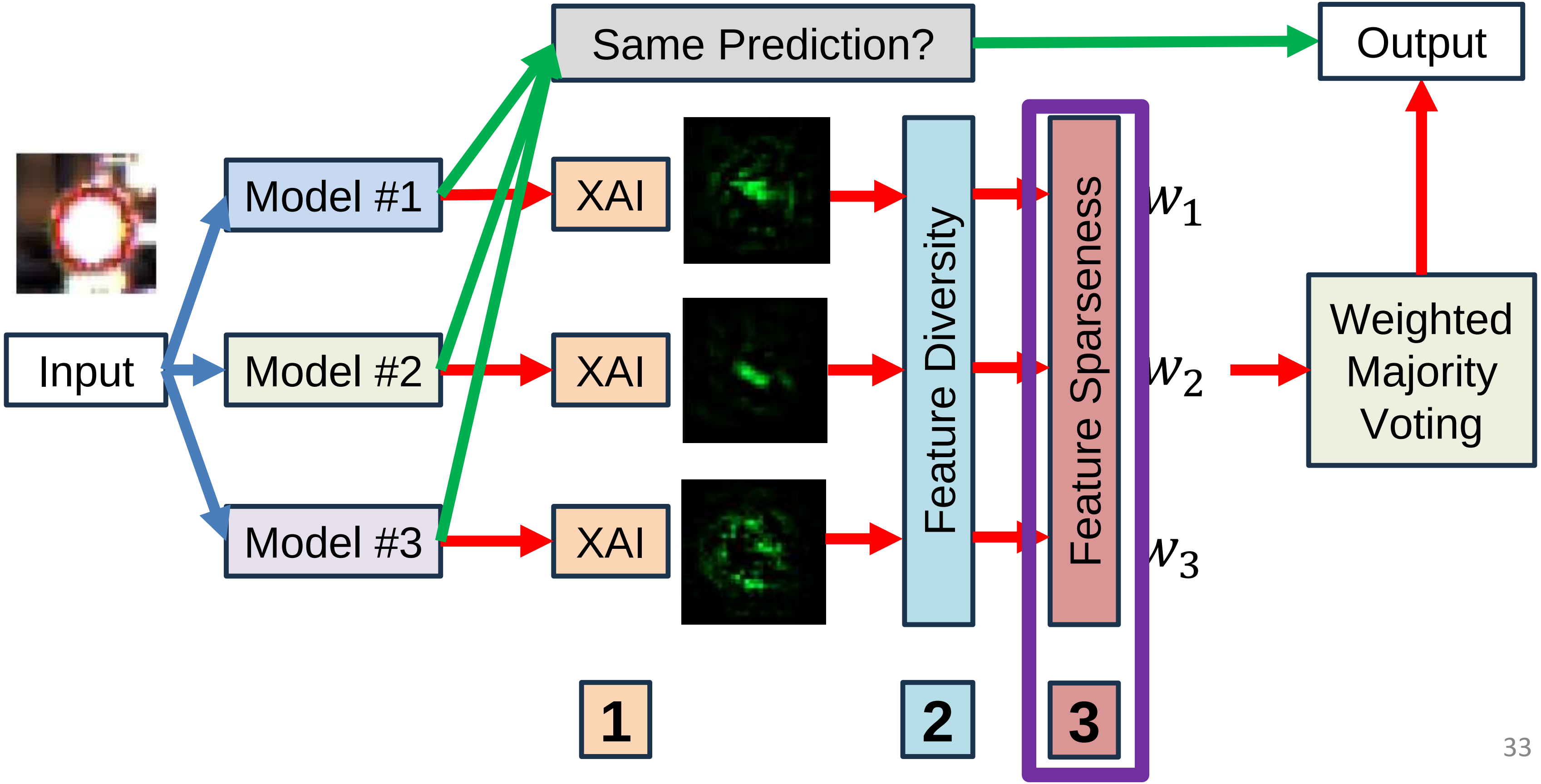
Feature-Space Diversity



Dynamic Weights using Feature Diversity

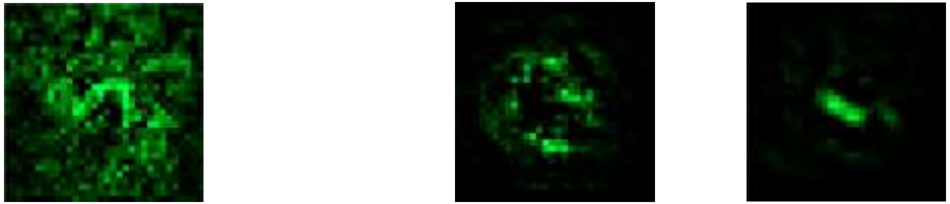


Step 3 – Feature Sparseness



Feature Sparseness

Input



Low

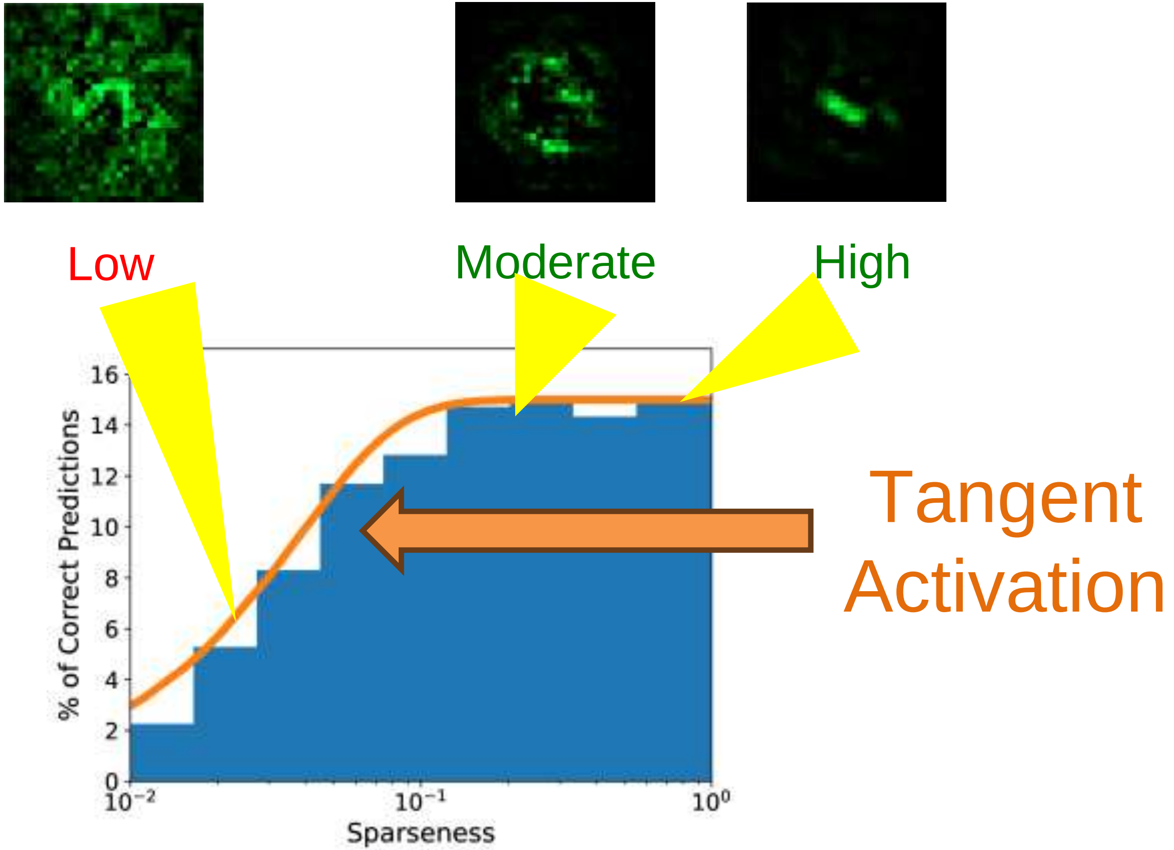
Moderate

High

?Useful Diversity?

Feature Sparseness

Input



Evaluation Datasets

CIFAR-10
Object Classification

GTSRB
Self-Driving Cars

Pneumonia
Medical Diagnosis

$\underbrace{\text{Self-Driving Cars} \quad \text{Medical Diagnosis}}_{\text{Safety-Critical Applications}}$

Neural Networks

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

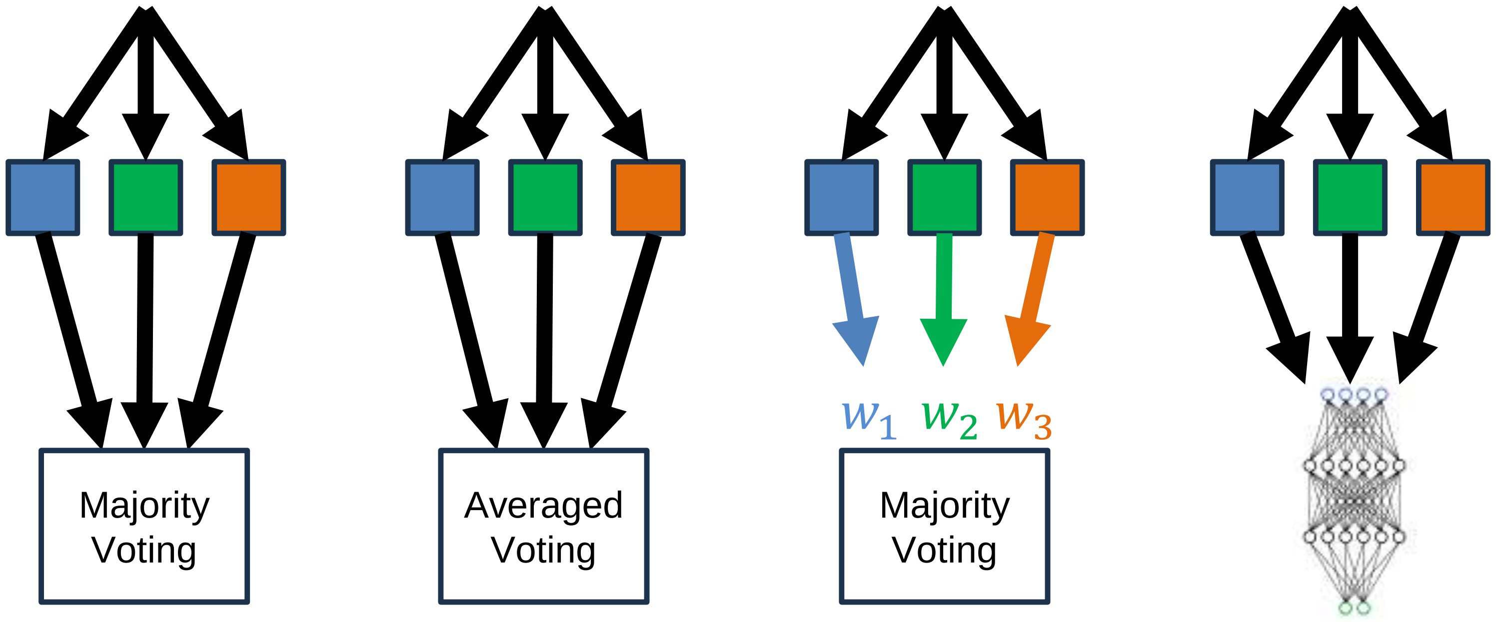
Resilience Metrics

- Balanced Accuracy
  - Compatible with imbalanced datasets
- F1 score
  - Focus on false positives/negatives than true negatives (e.g. Pneumonia [focus case] vs Benign)

1 = Most Resilient

0 = Least Resilient

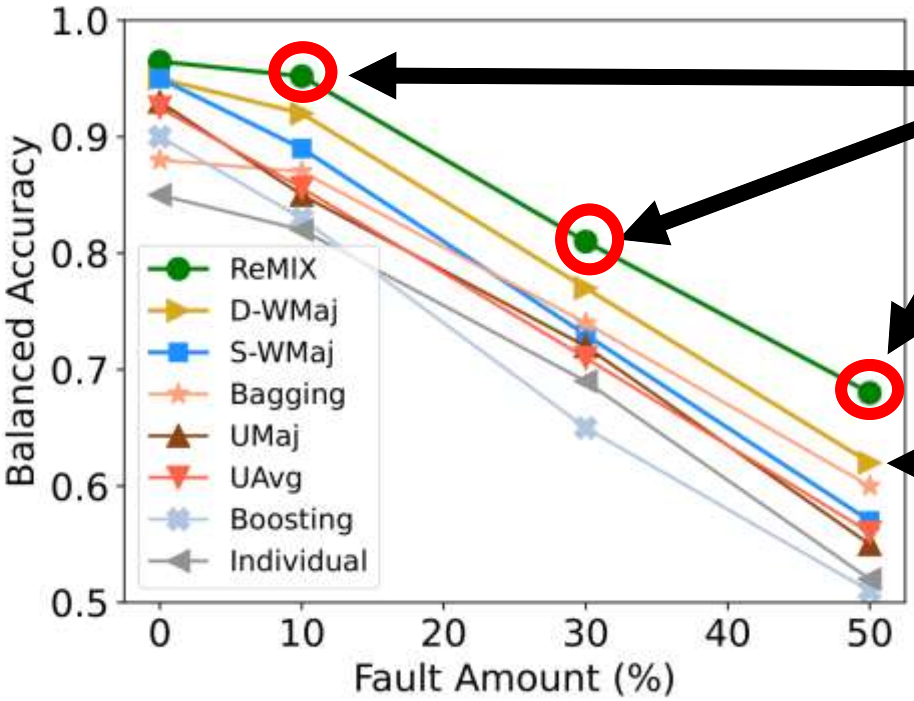
Ensembling Baselines



UMaj **UAvg** **S-WMaj** **D-WMaj (Stacking)**

RQ1: Resilience of ReMIX vs Baselines

Dataset: GTSRB



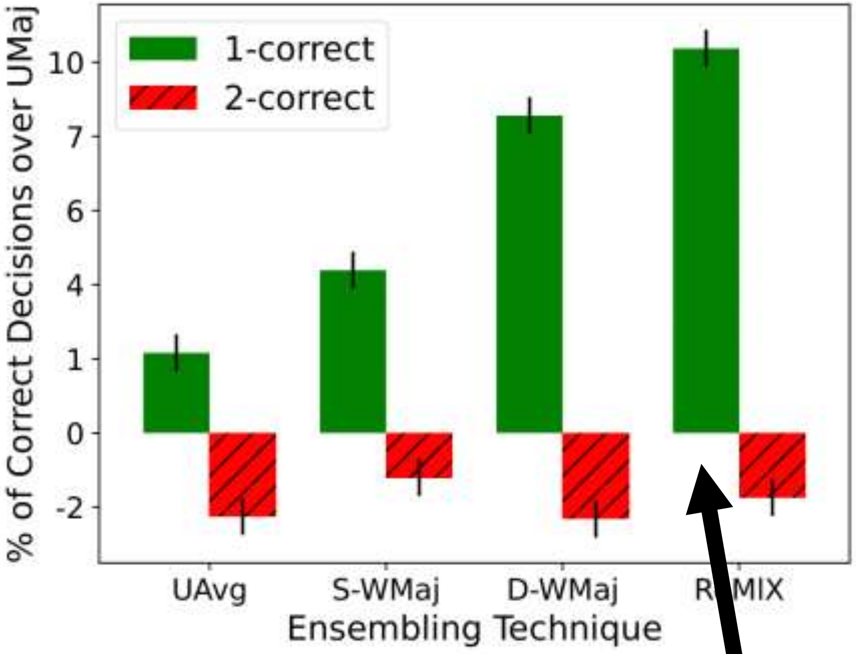
ReMIX

12% more resilient on average

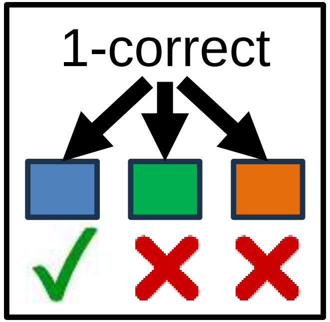
D-WMaj

RQ1: Resilience of ReMIX vs Baselines

Dataset: GTSRB



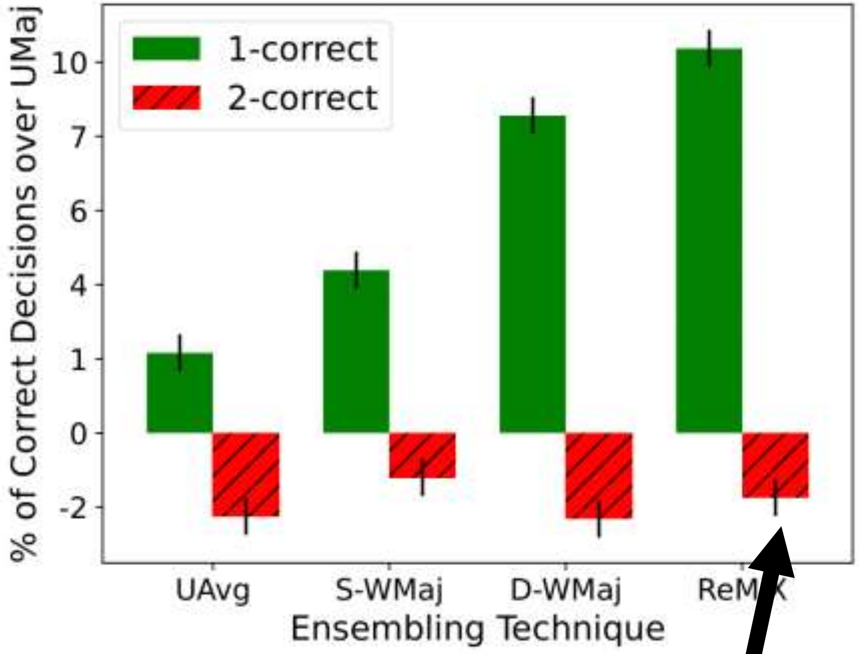
ReMIX



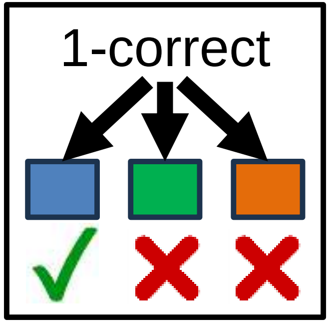
ReMIX maximizes predictions of 1-correct cases

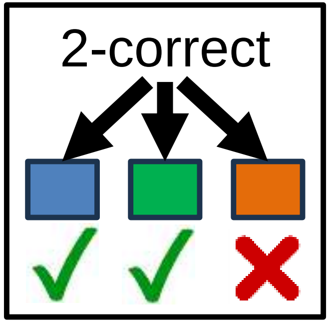
RQ1: Resilience of ReMIX vs Baselines

Dataset:
GTSRB

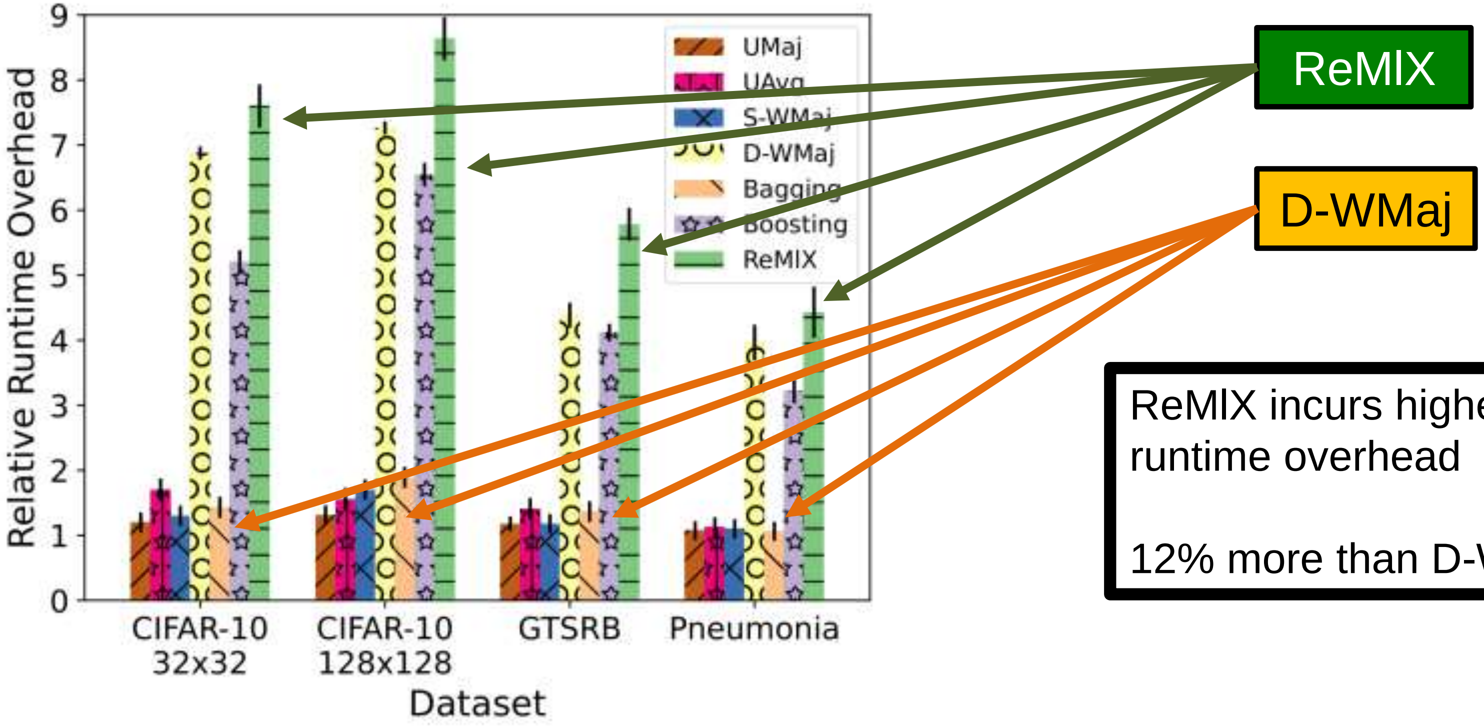


ReMIX





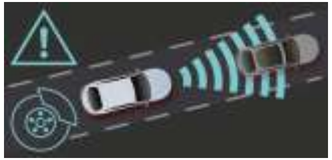
RQ2: Runtime Overhead



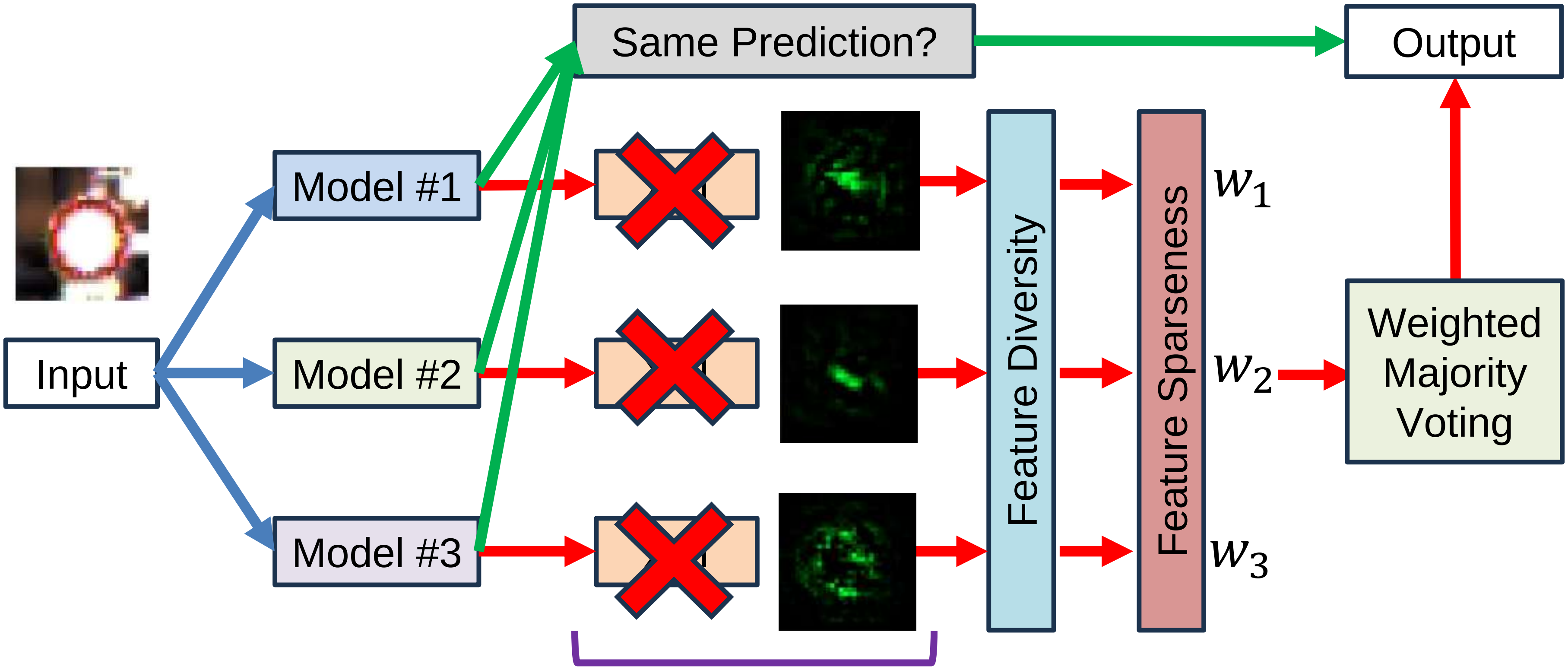
ReMIX incurs highest runtime overhead

12% more than D-WMaj

RQ2: Runtime Overhead

Application	ReMIX Average	Industry Maximum	Associated Risk	
 AVs - GTSRB	0.07s	**0.83s**	Safe braking	✓
 VR Telesurgery - Pneumonia	0.31s	**0.50s**	VR sickness	✓

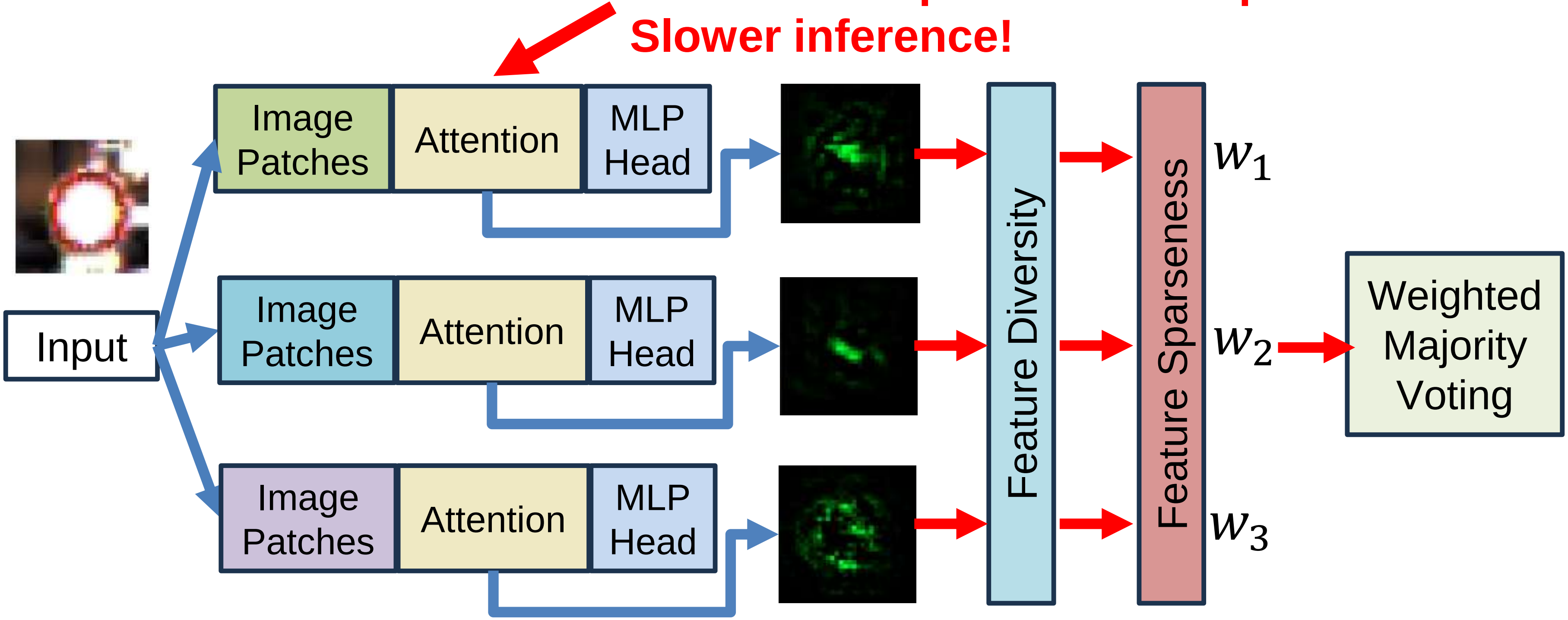
Optimization using Ante-Hoc XAI



67% Runtime Overhead

Optimization using Ante-Hoc XAI

More trained parameters required!
Slower inference!

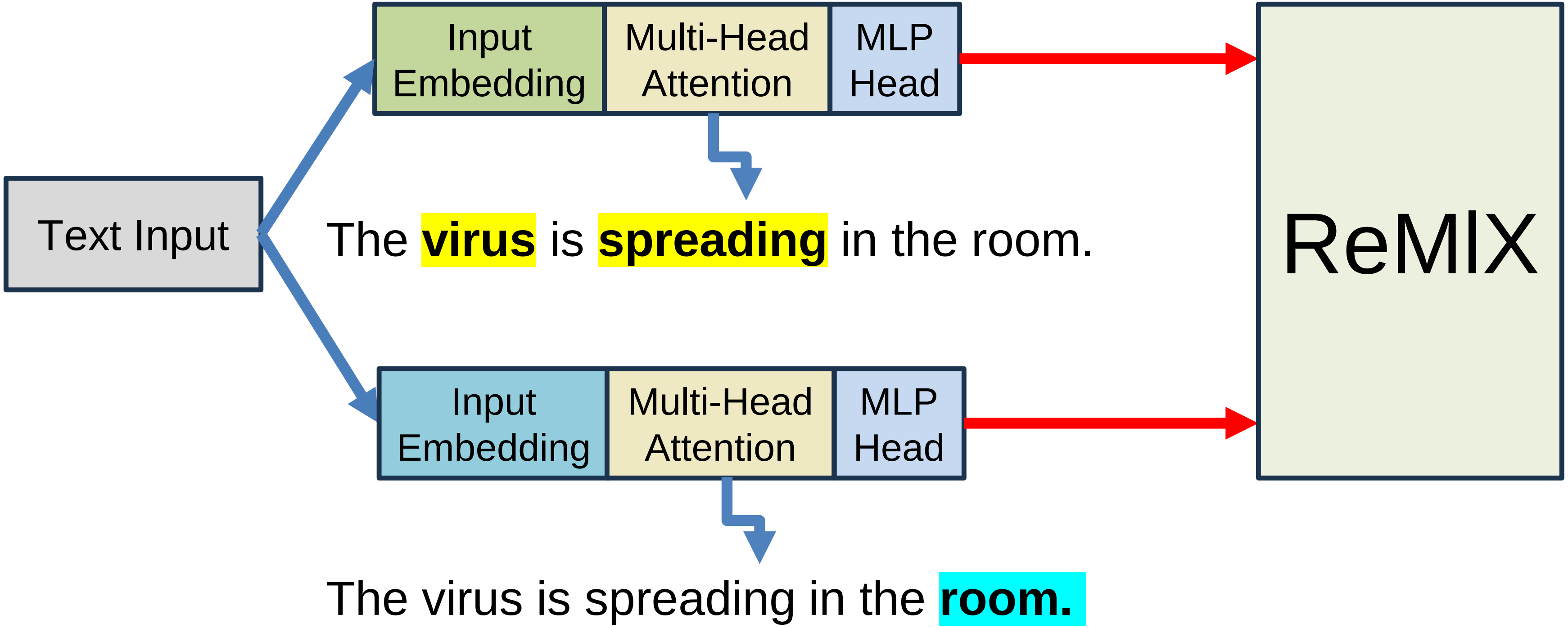


Future – ReMIX for Sentiment Analysis

Text Input

The virus is spreading in the room.

Future – ReMIX for Sentiment Analysis



Summary

1. **Problem:** Reducing misclassifications by ensembles

2. **Approach:** (**ReMlX**) **Re**silience of **ML** Ensembles using **X**AI

3. **Results:** **ReMlX** improves **resilience by 12%** but with **15% overhead** over D-WMaj / Stacking (best baseline)

Email: abrahamc@ece.ubc.ca

Website: https://people.ece.ubc.ca/abrahamc/

Paper

Code

Code Available

Code Reviewed

Code Reproducible