The Fault in Our Data Stars: **Studying Mitigation Techniques** against **Faulty Training Data** in Machine Learning Applications

Abraham Chan, Arpan Gujarati, Karthik Pattabiraman, Sathish Gopalakrishnan

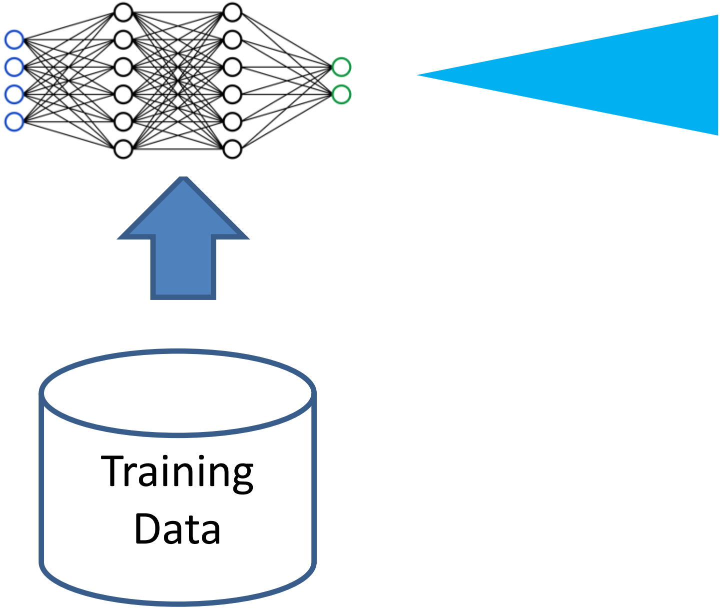
The University of British Columbia

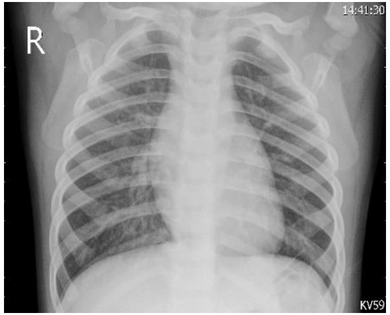
Training Data

Image	**Label**
	Normal
	Pneumonia

Modern ML Applications

Dataset: Pneumonia

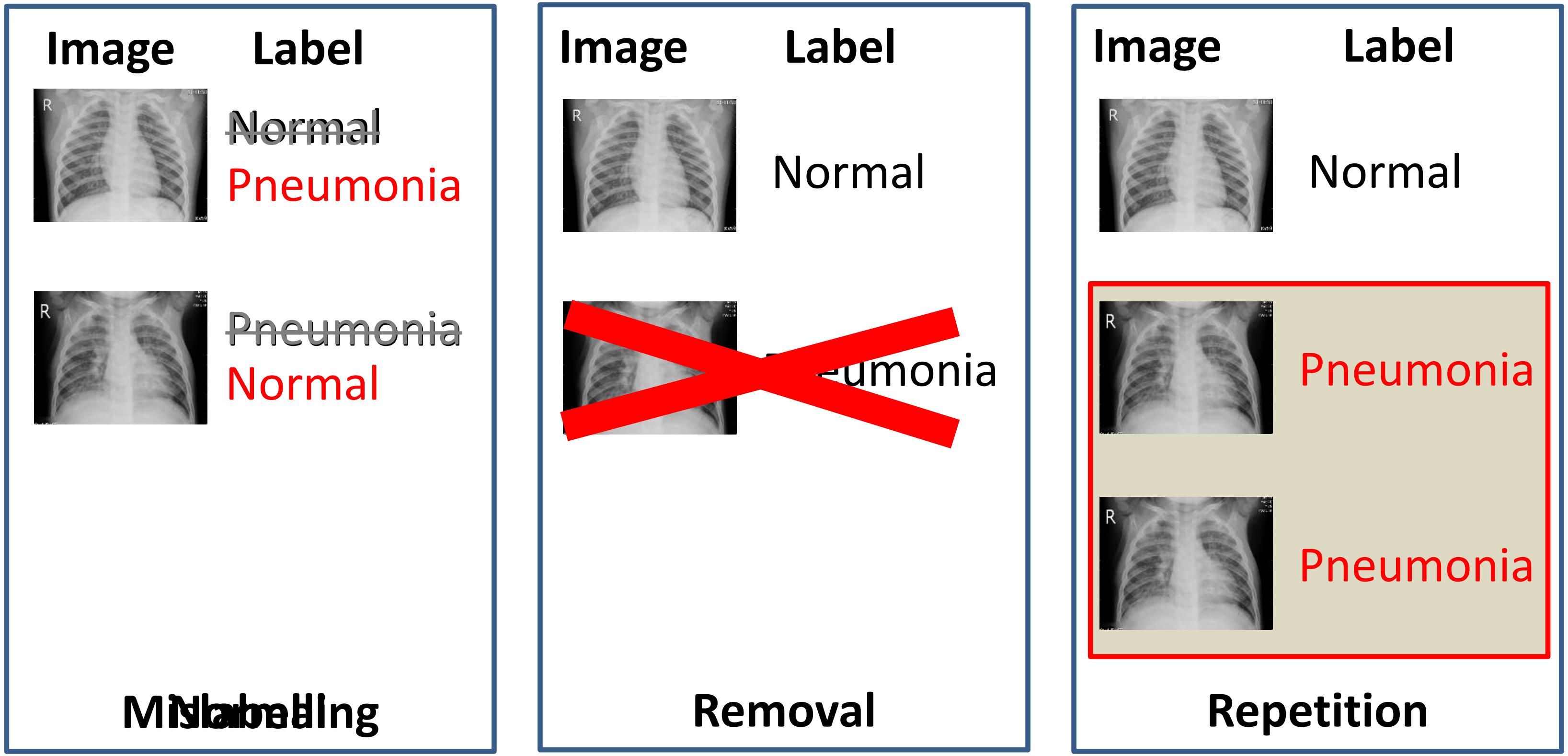




Prediction: **Normal**

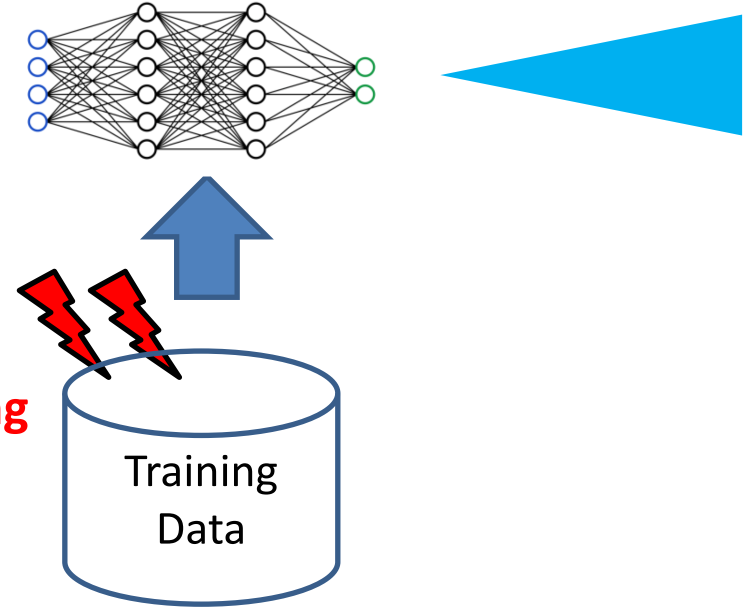
Model Accuracy: **90%**

Training Data Faults

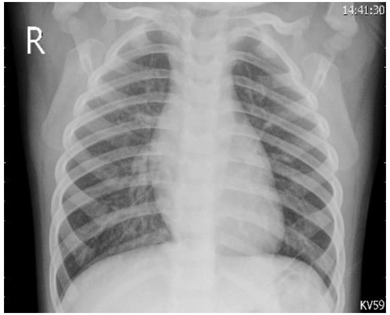


Random Mislabelling

10% Random Mislabelling



Dataset: Pneumonia



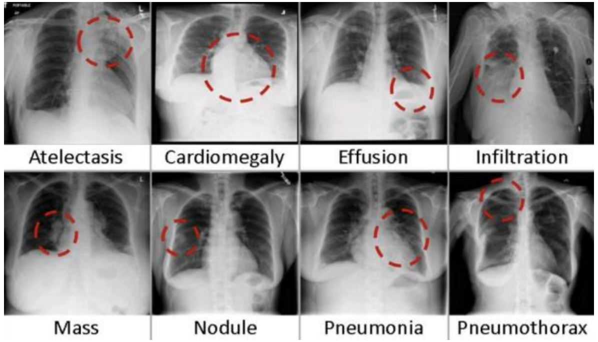
Actual: **Normal**
Prediction: **Pneumonia**

Model Accuracy: **55%**

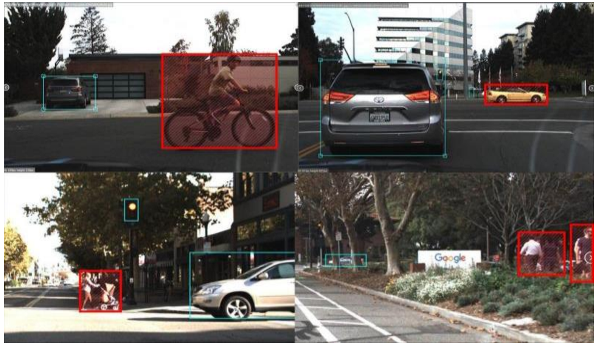
Original Accuracy: 90%

Training Data Faults in Practice

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



33% of the popular Udacity Dataset2 mislabelled or missing labels [Dwyer, 2020]



How can we mitigate training data faults?

(From the P.O.V of a practitioner)

Selection Criteria

- Code Available?
- Architecture agnostic?
- Artificial noise?
- Not pre-trained?
- Standalone?

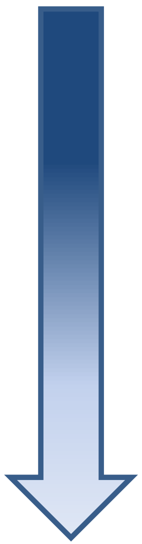
Techniques against Mislabelled Data

1. Loss Correction (LC)
2. Knowledge Distillation (KD)
3. Robust Loss (RL)
4. Label Smoothing (LS)
5. Ensemble Learning (Ens)

Our Contribution:
How do we choose a technique?

Techniques against Mislabelled Data

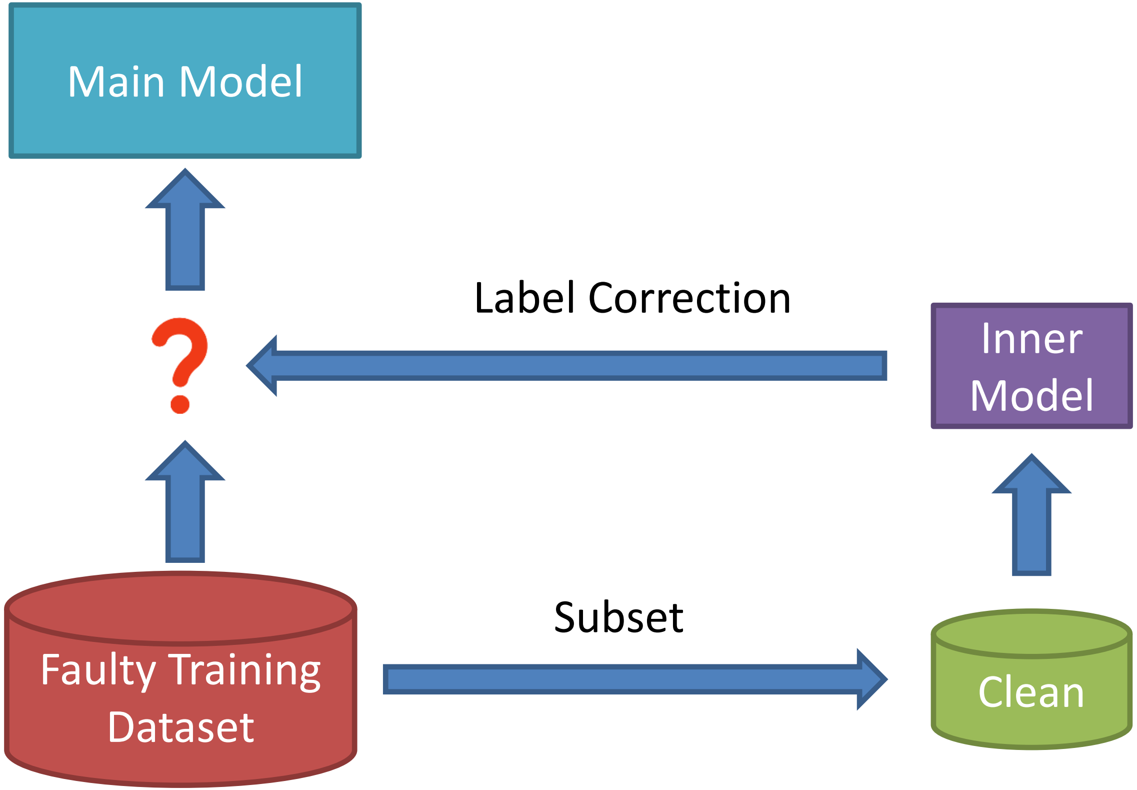
More Practitioner Effort



Less Practitioner Effort

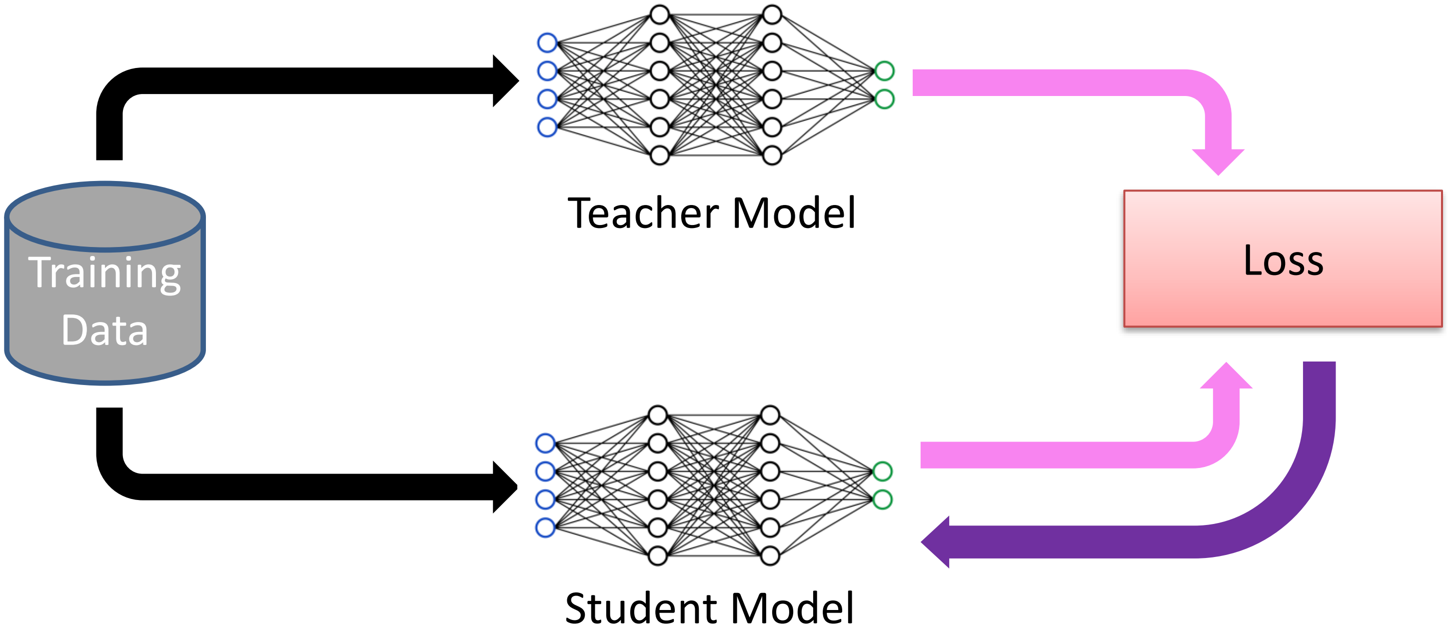
1. Loss Correction (LC)
2. Knowledge Distillation (KD)
3. Robust Loss (RL)
4. Label Smoothing (LS)
5. Ensemble Learning (Ens)

Loss Correction (LC)

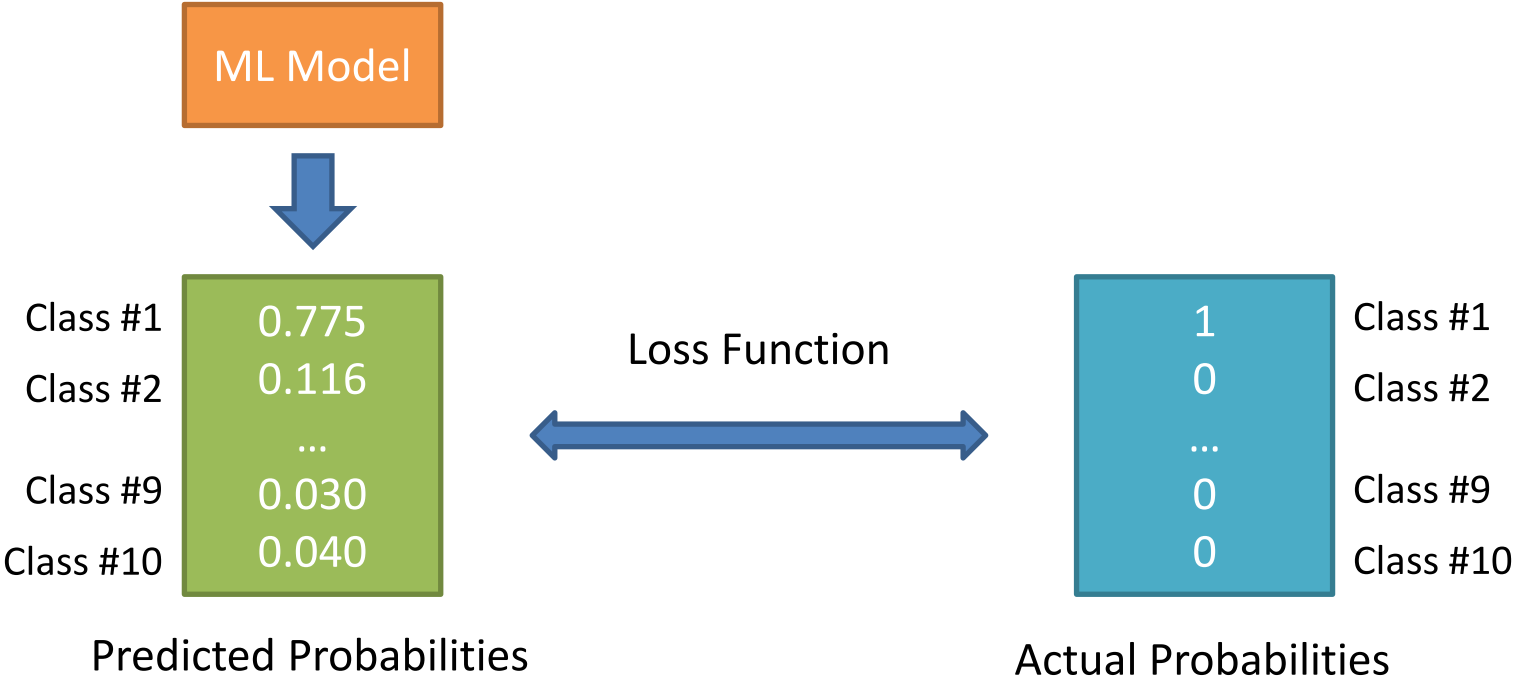


Self Knowledge Distillation (KD)

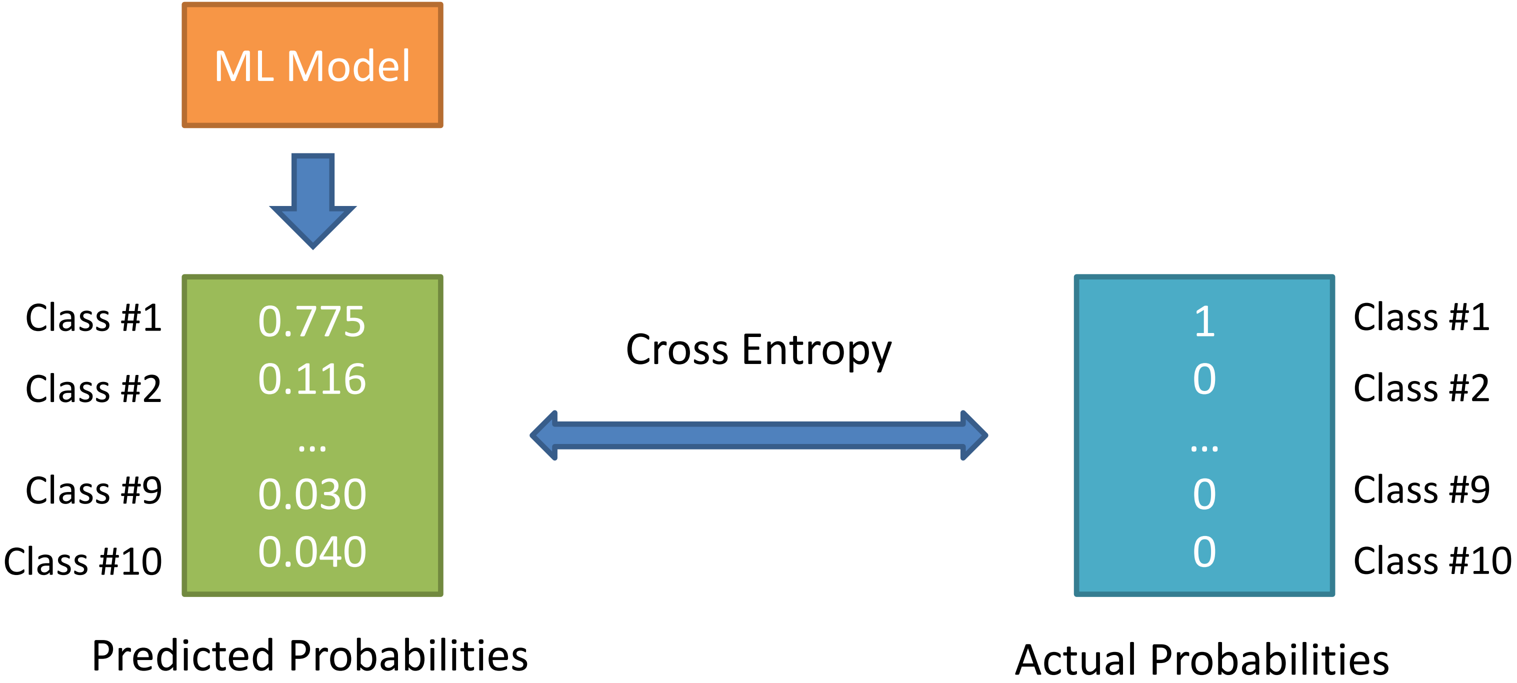
- Teacher = Student = (i.e., ResNet50)



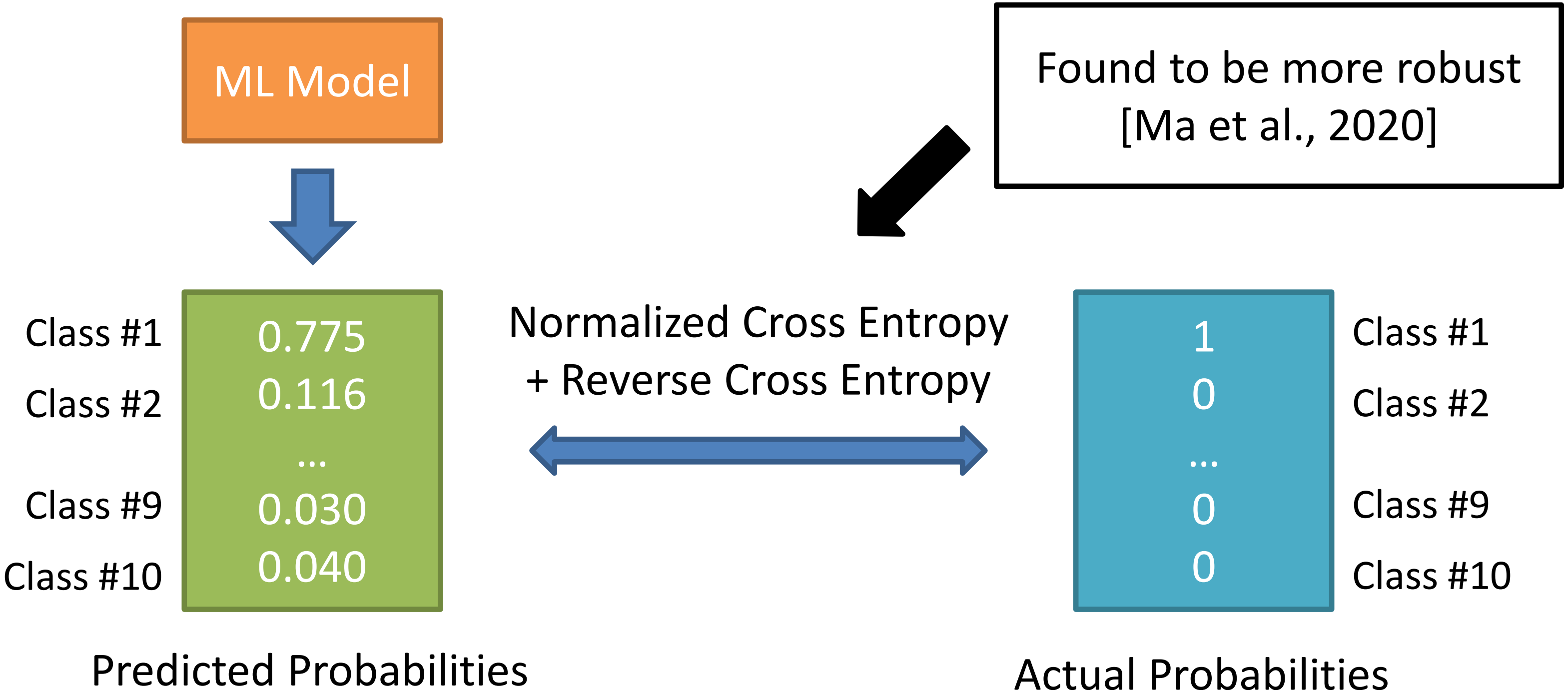
Robust Loss (RL)



Robust Loss (RL)

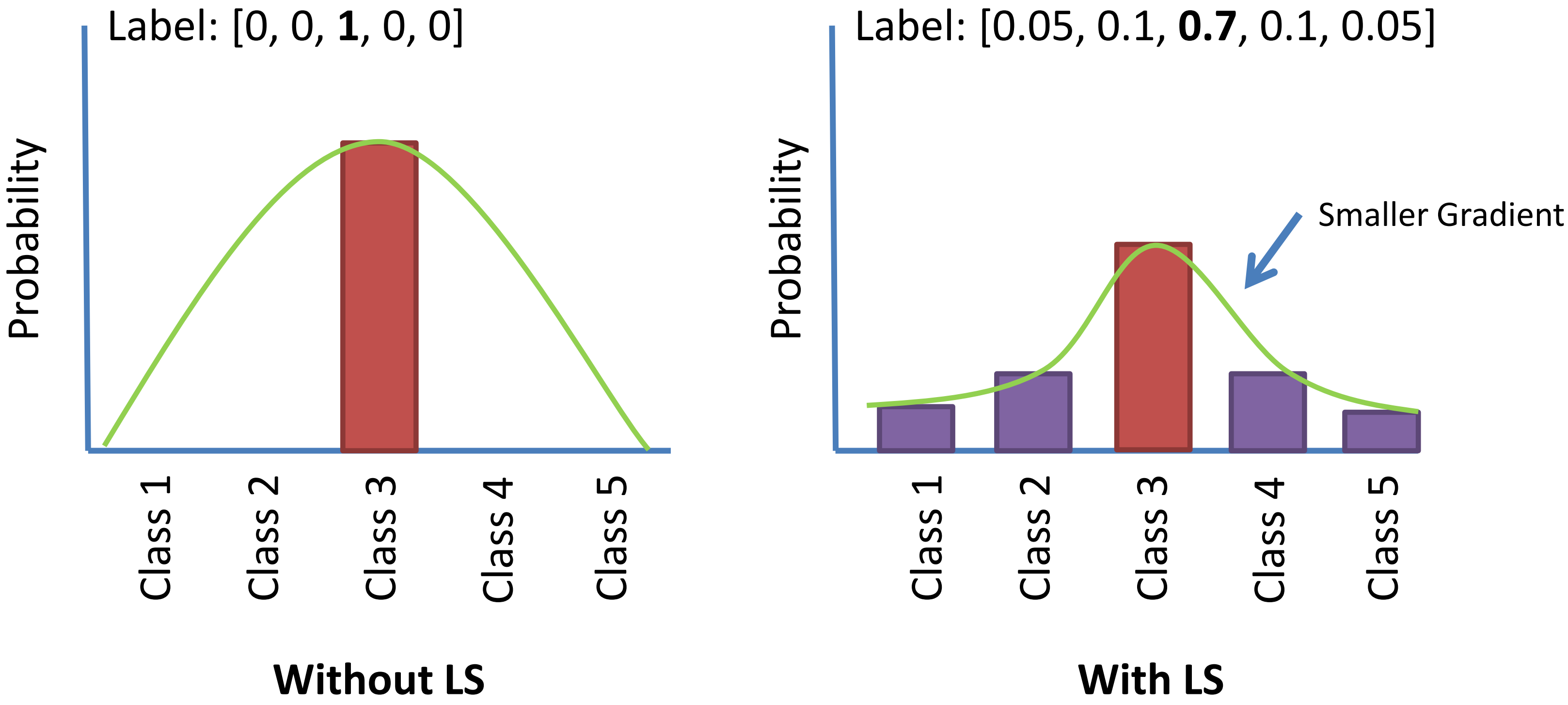


Robust Loss (RL)

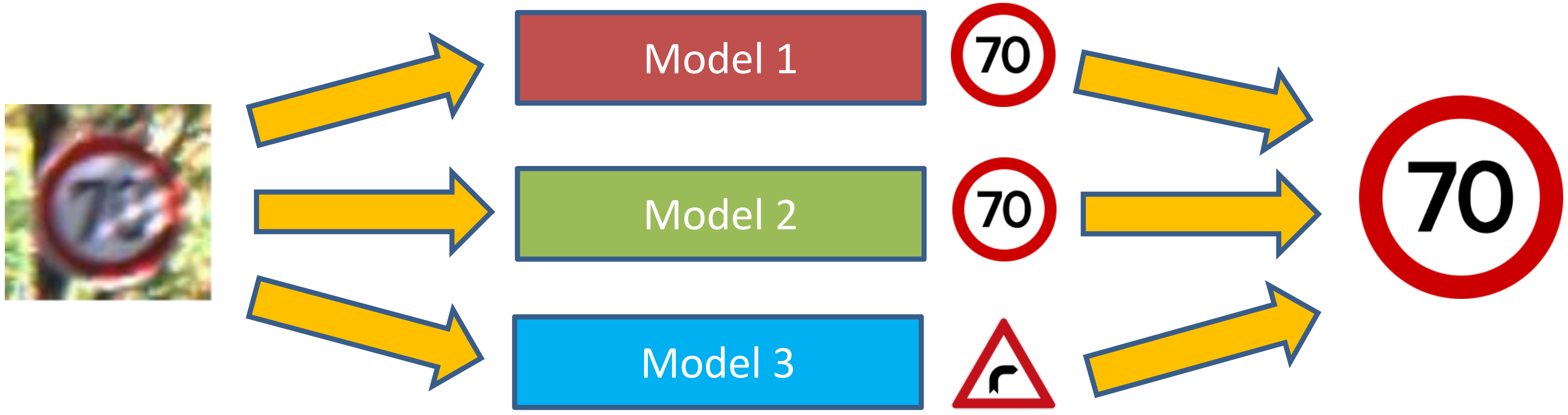


Techniques: LC KD RL LS Ens

Label Smoothing (LS)



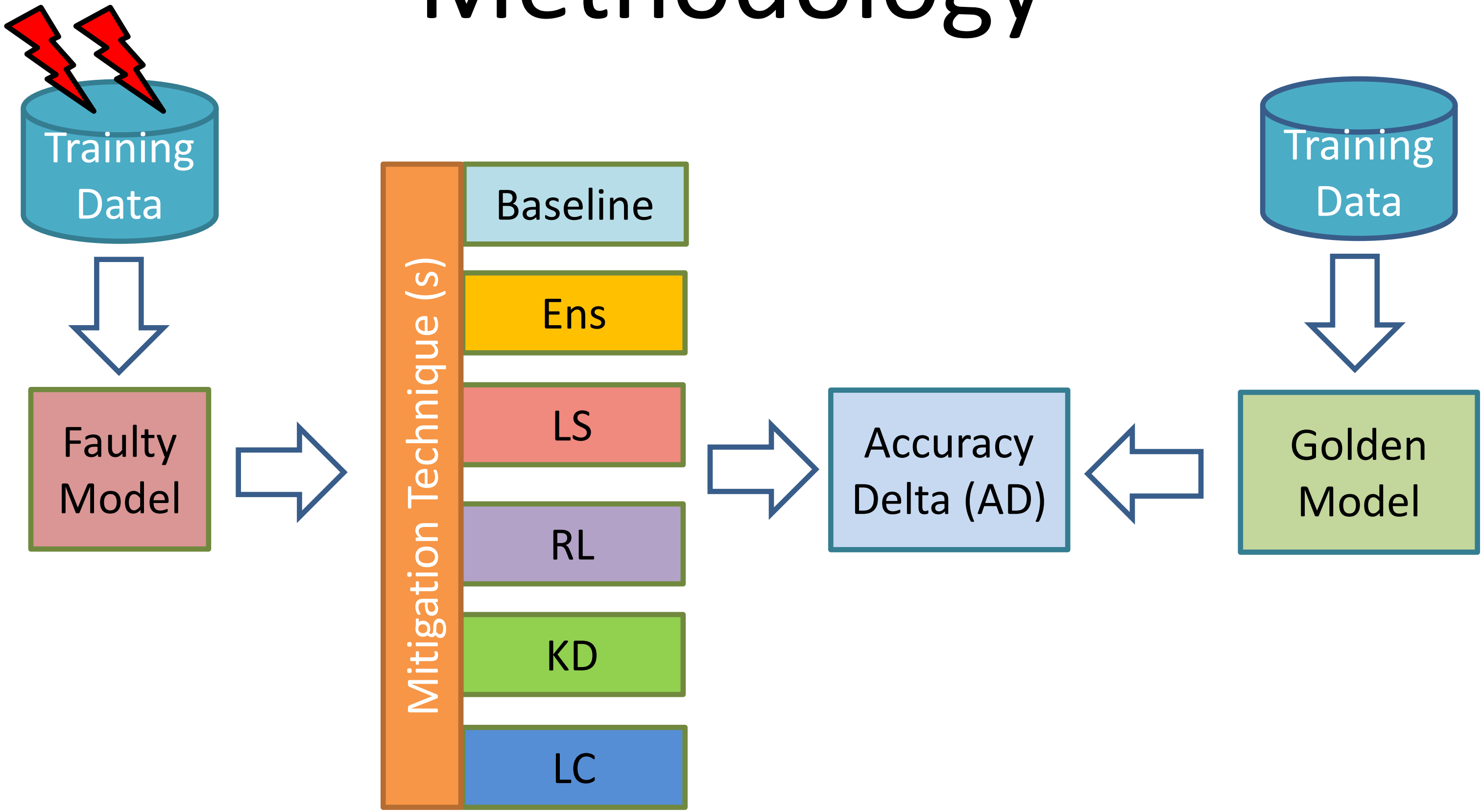
Ensemble Learning (Ens)



Understanding the Resilience of Neural Network Ensembles against Faulty Training Data

Our Prior Work: [QRS'21]

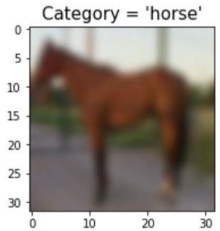
Methodology



Neural Networks

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

Evaluation Datasets



CIFAR-10
Object Detection

GTSRB
Self-Driving Cars

Pneumonia
Medical Diagnosis

Safety-Critical Applications

Reliability Metric: Accuracy Delta (AD)

Model trained with golden data

- Test Image 1 ✓
- Test Image 2 ✓
- Test Image 3 ✓
- Test Image 4 ✗

Model trained with faulty data

- Test Image 1 ✓
- Test Image 2 ✗
- Test Image 3 ✗

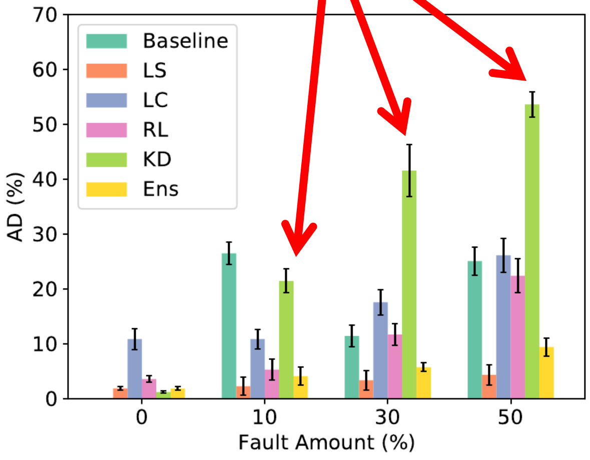
Accuracy Delta (AD) = 2 / 3 = 67% in this case

AD Across: Models | Fault Types | Datasets

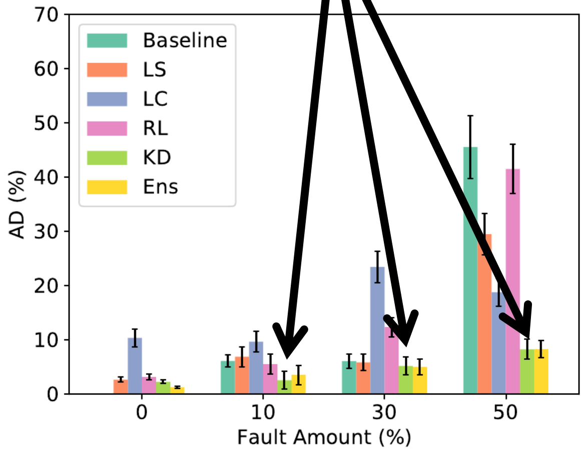
KD is not effective here

KD is effective here

Higher AD is worse



GTSRB, ResNet50, Mislabelling



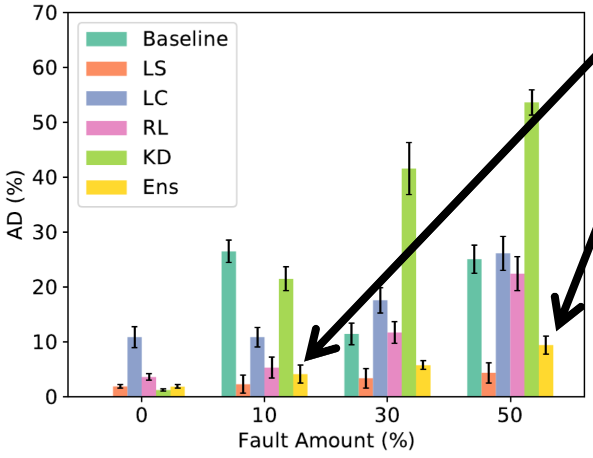
GTSRB, VGG16, Mislabelling

Higher AD is worse

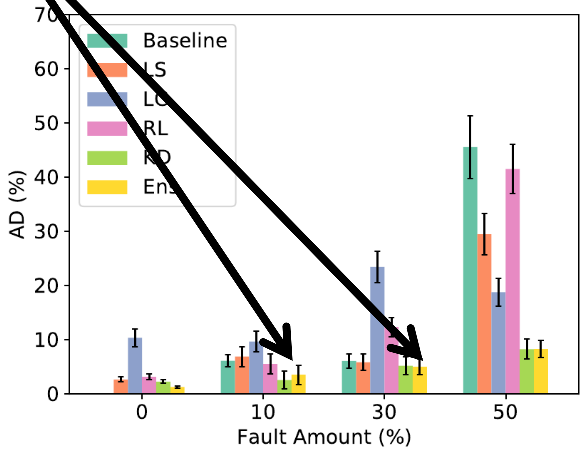
AD Across: Models | Fault Types | Datasets

Ensembles are effective across models

Higher AD is worse



GTSRB, ResNet50, Mislabelling



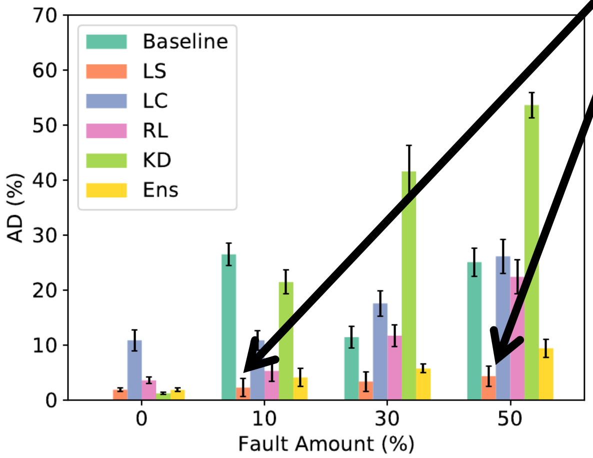
GTSRB, VGG16, Mislabelling

Higher AD is worse

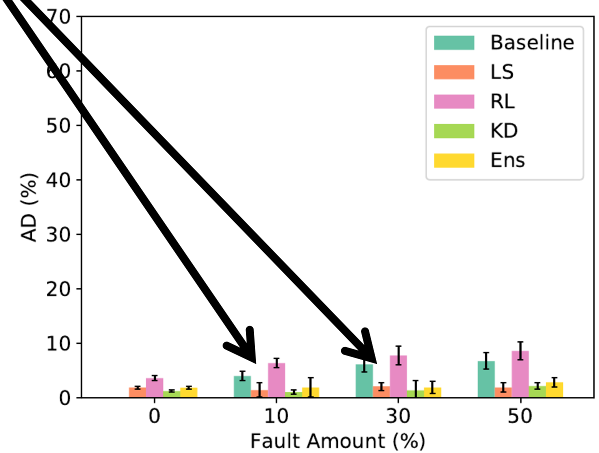
AD Across: Models | Fault Types | Datasets

LS is also effective across fault types

Higher AD is worse



GTSRB, ResNet50, Mislabelling



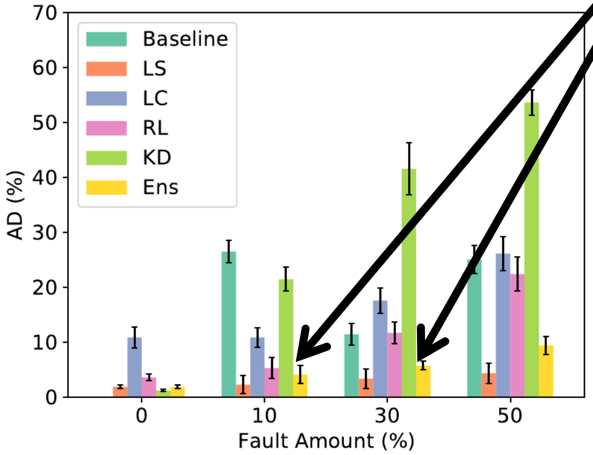
GTSRB, ResNet50, Removal

Higher AD is worse

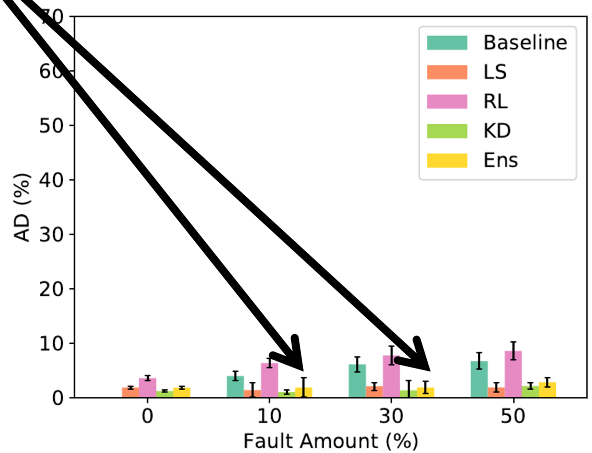
AD Across: Models | Fault Types | Datasets

Ensembles are also effective across fault types

Higher AD is worse



GTSRB, ResNet50, Mislabelling



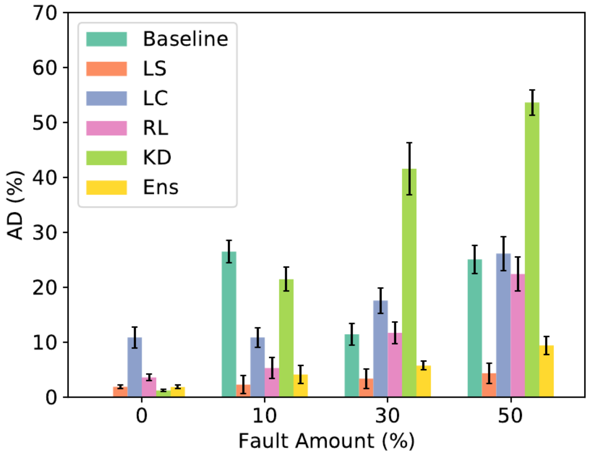
GTSRB, ResNet50, Removal

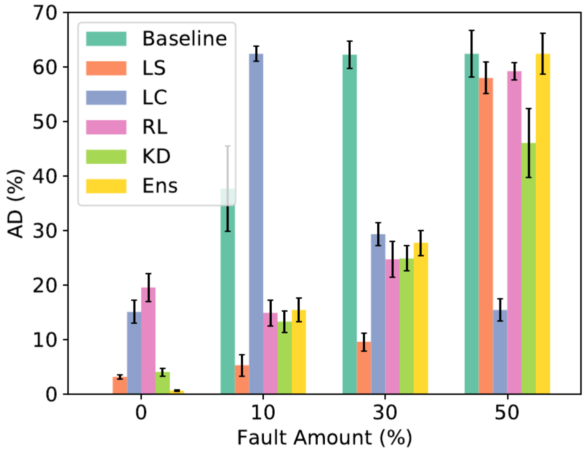
Higher AD is worse

AD Across: Models | Fault Types | Datasets

Finding: Ensemble is generally effective, followed by LS

Higher AD is worse





Higher AD is worse

GTSRB
Self-Driving Cars

Pneumonia
Medical Diagnosis

Takeaways

- **Ensembles** performed best overall but **Label Smoothing** surprisingly effective (second place)
- Dataset size did not have an impact on **Loss Correction** (but works well for datasets with fewer classes)
- **Knowledge Distillation** and **Robust Loss** performed well only at low fault amounts

Summary

- **Problem:** Choose a mitigation technique against faulty training data
- **Approach:** Evaluate techniques on 7 models across 3 datasets
- **Results:**
  - **Ensembles** effective across all configurations
  - **Label smoothing** is second in effectiveness, with less overhead

Email: abrahamc@ece.ubc.ca

More Info: https://github.com/DependableSystemsLab/TDFM-Techniques