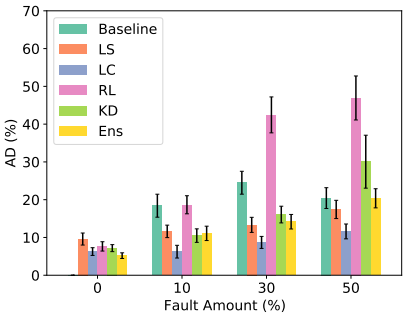
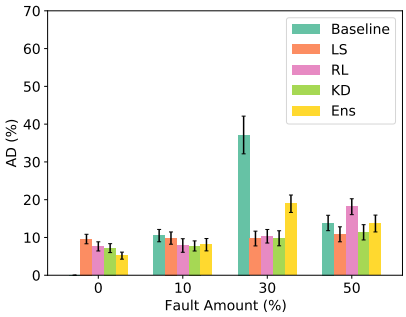
Accuracy Delta by Model, Fault Type, Dataset

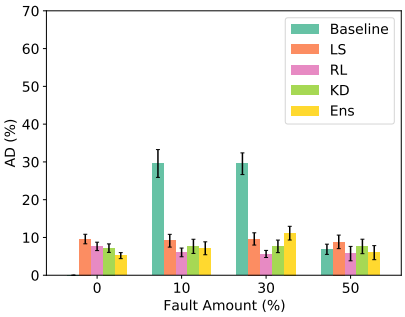
I. CIFAR-10



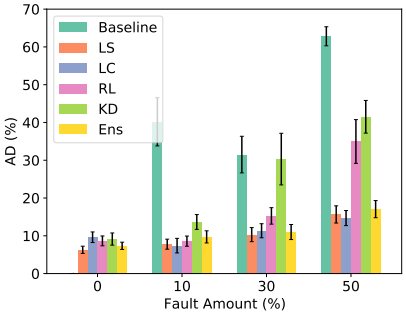
(a) CIFAR-10, ConvNet, Mislabelling



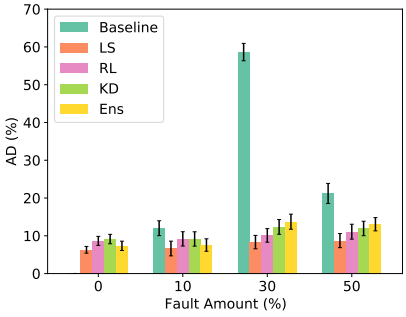
(b) CIFAR-10, ConvNet, Removal



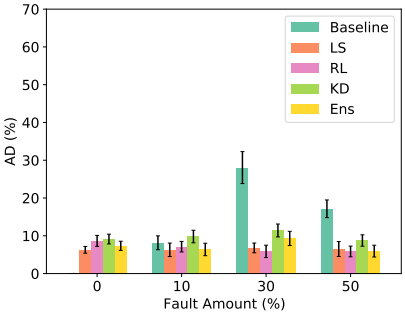
(c) CIFAR-10, ConvNet, Repetition



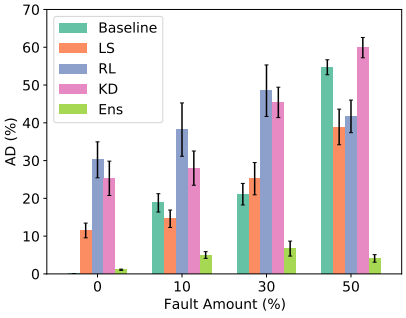
(d) CIFAR-10, DeconvNet, Mislabelling



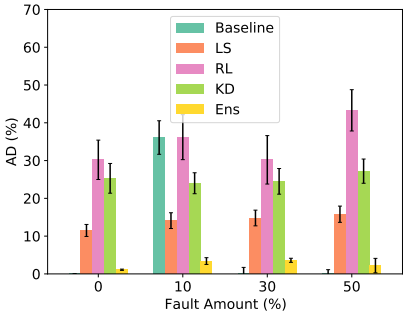
(e) CIFAR-10, DeconvNet, Removal



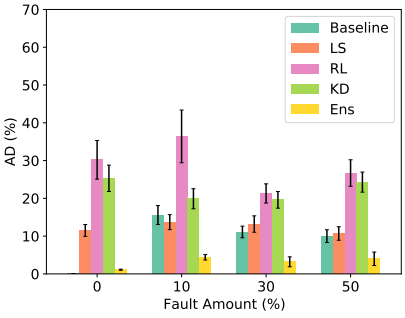
(f) CIFAR-10, DeconvNet, Repetition



(g) CIFAR-10, MobileNet, Mislabelling

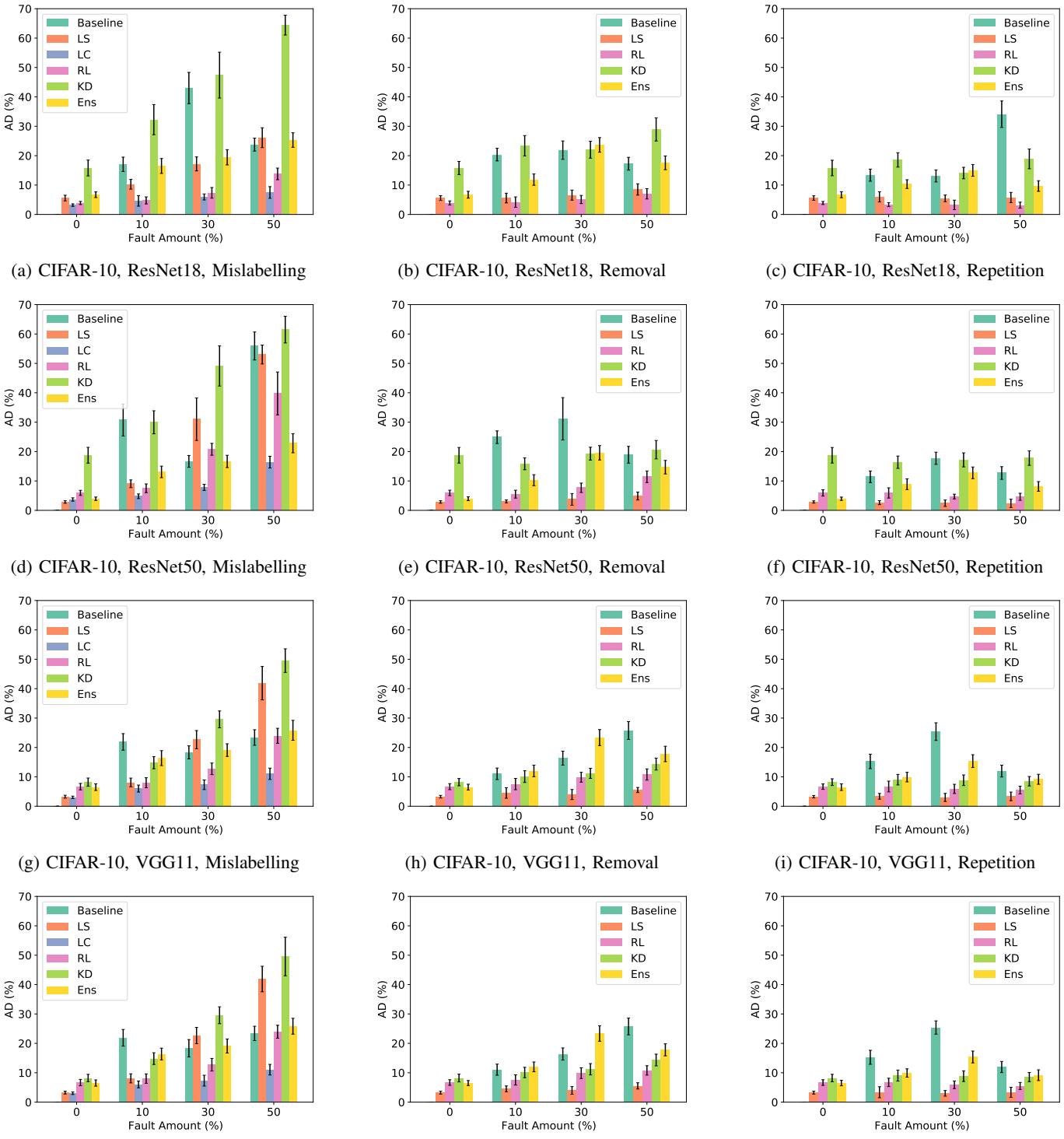


(h) CIFAR-10, MobileNet, Removal



(i) CIFAR-10, MobileNet, Repetition

Fig. 1: AD of individual models, compared with models protected with TDFM techniques when trained with faulty CIFAR-10 datasets. The error bars in the results indicate the 95% confidence intervals. Lower values are better.



(a) CIFAR-10, ResNet18, Mislabelling (b) CIFAR-10, ResNet18, Removal (c) CIFAR-10, ResNet18, Repetition

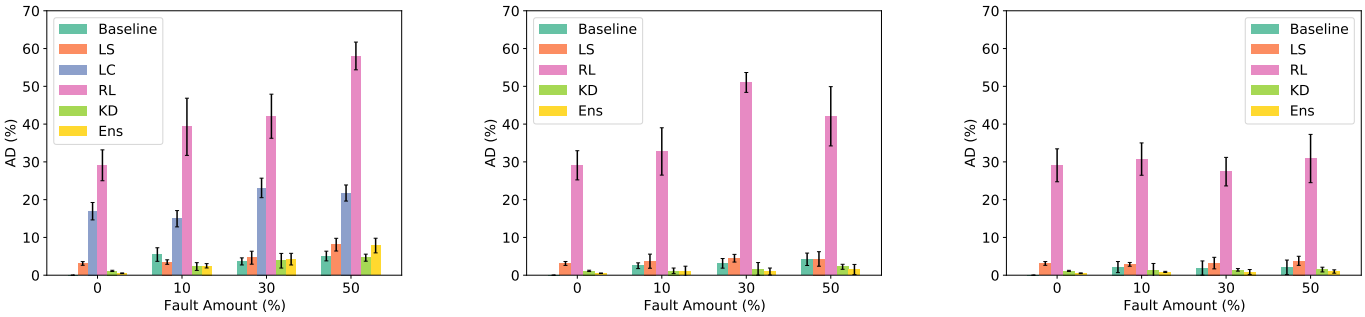
(d) CIFAR-10, ResNet50, Mislabelling (e) CIFAR-10, ResNet50, Removal (f) CIFAR-10, ResNet50, Repetition

(g) CIFAR-10, VGG11, Mislabelling (h) CIFAR-10, VGG11, Removal (i) CIFAR-10, VGG11, Repetition

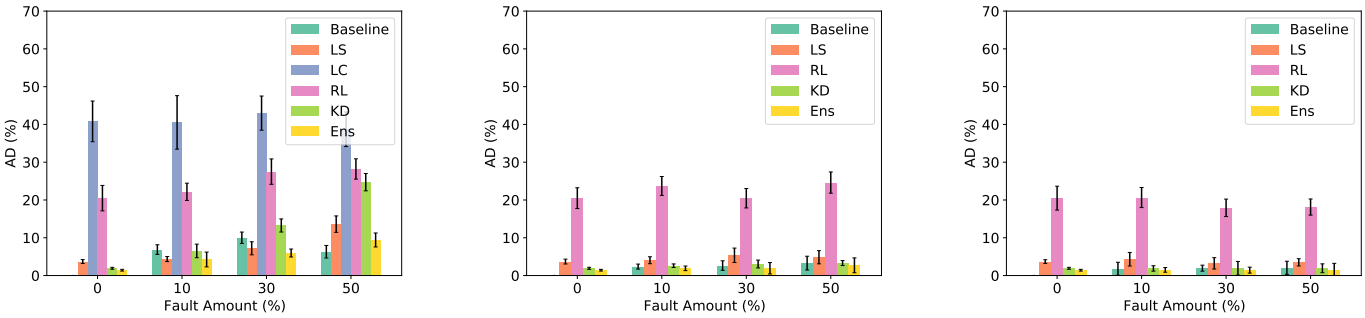
(j) CIFAR-10, VGG16, Mislabelling (k) CIFAR-10, VGG16, Removal (l) CIFAR-10, VGG16, Repetition

Fig. 2: AD of individual models, compared with models protected with TDFM techniques when trained with faulty CIFAR-10 datasets. The error bars in the results indicate the 95% confidence intervals. Lower values are better.

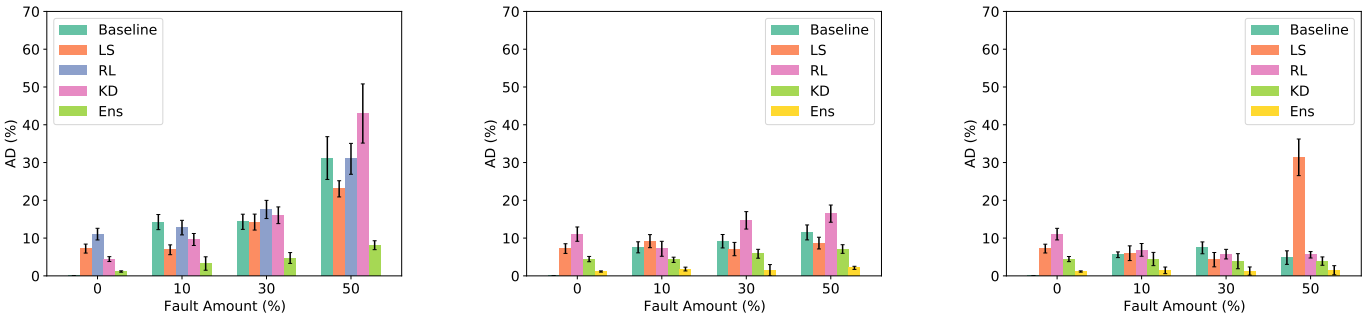
II. GTSRB



(a) GTSRB, ConvNet, Mislabelling (b) GTSRB, ConvNet, Removal (c) GTSRB, ConvNet, Repetition



(d) GTSRB, DeconvNet, Mislabelling (e) GTSRB, DeconvNet, Removal (f) GTSRB, DeconvNet, Repetition



(g) GTSRB, MobileNet, Mislabelling (h) GTSRB, MobileNet, Removal (i) GTSRB, MobileNet, Repetition

Fig. 3: AD of individual models, compared with models protected with TDFM techniques when trained with faulty GTSRB datasets. The error bars in the results indicate the 95% confidence intervals. Lower values are better.

(a) GTSRB, ResNet18, Mislabelling (b) GTSRB, ResNet18, Removal (c) GTSRB, ResNet18, Repetition

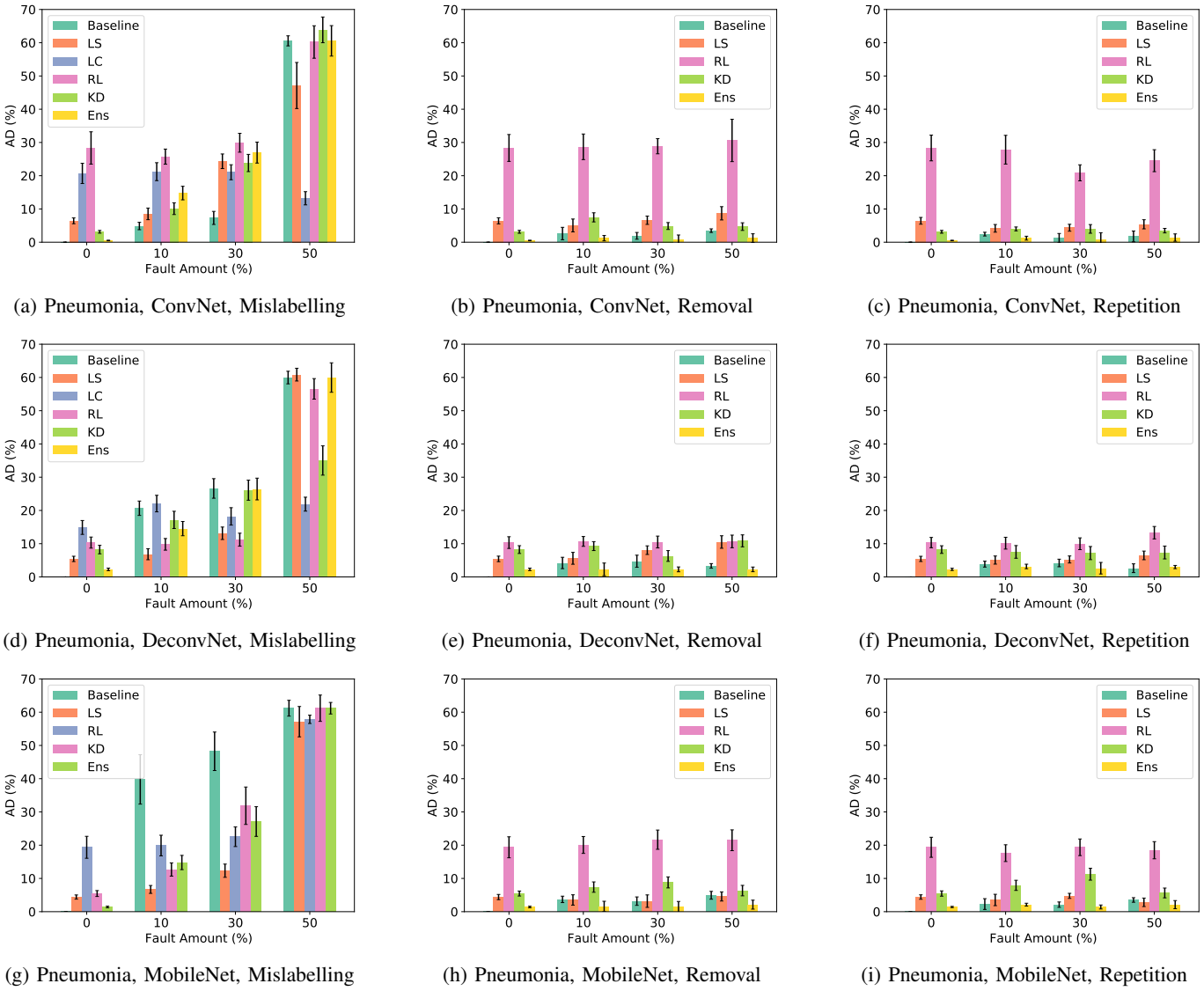
(d) GTSRB, ResNet50, Mislabelling (e) GTSRB, ResNet50, Removal (f) GTSRB, ResNet50, Repetition

(g) GTSRB, VGG11, Mislabelling (h) GTSRB, VGG11, Removal (i) GTSRB, VGG11, Repetition

(j) GTSRB, VGG16, Mislabelling (k) GTSRB, VGG16, Removal (l) GTSRB, VGG16, Repetition

Fig. 4: AD of individual models, compared with models protected with TDFM techniques when trained with faulty GTSRB datasets. The error bars in the results indicate the 95% confidence intervals. Lower values are better.

III. PNEUMONIA



(a) Pneumonia, ConvNet, Mislabelling

(b) Pneumonia, ConvNet, Removal

(c) Pneumonia, ConvNet, Repetition

(d) Pneumonia, DeconvNet, Mislabelling

(e) Pneumonia, DeconvNet, Removal

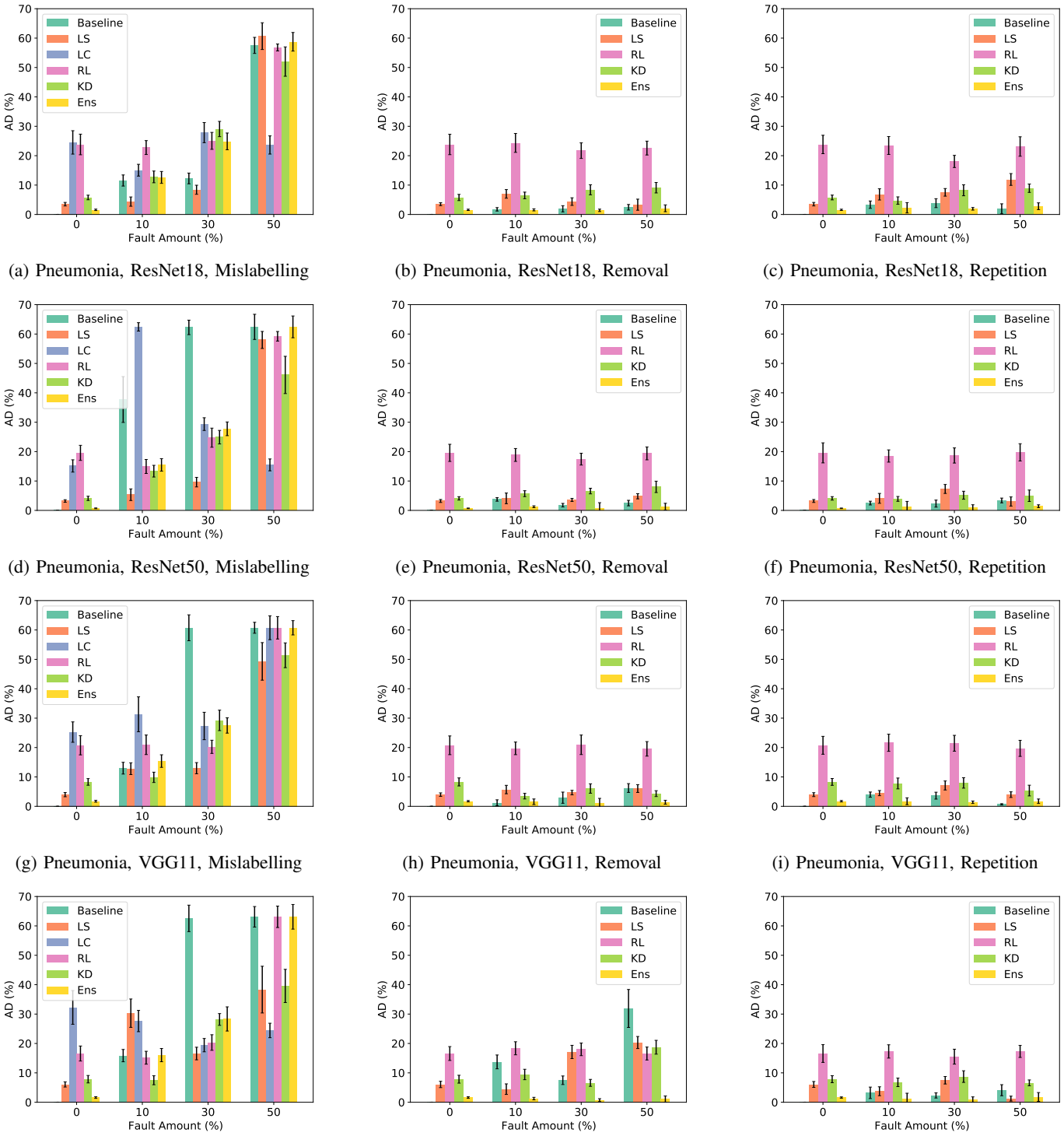
(f) Pneumonia, DeconvNet, Repetition

(g) Pneumonia, MobileNet, Mislabelling

(h) Pneumonia, MobileNet, Removal

(i) Pneumonia, MobileNet, Repetition

Fig. 5: AD of individual models, compared with models protected with TDFM techniques when trained with faulty Pneumonia datasets. The error bars in the results indicate the 95% confidence intervals. Lower values are better.



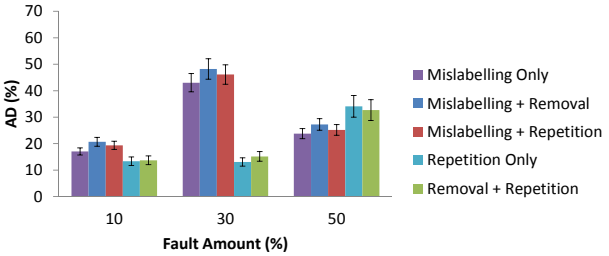
(a) Pneumonia, ResNet18, Mislabelling (b) Pneumonia, ResNet18, Removal (c) Pneumonia, ResNet18, Repetition

(d) Pneumonia, ResNet50, Mislabelling (e) Pneumonia, ResNet50, Removal (f) Pneumonia, ResNet50, Repetition

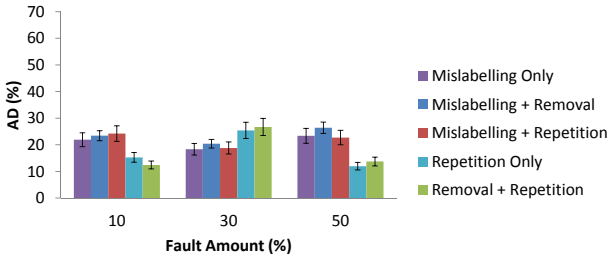
(g) Pneumonia, VGG11, Mislabelling (h) Pneumonia, VGG11, Removal (i) Pneumonia, VGG11, Repetition

(j) Pneumonia, VGG16, Mislabelling (k) Pneumonia, VGG16, Removal (l) Pneumonia, VGG16, Repetition

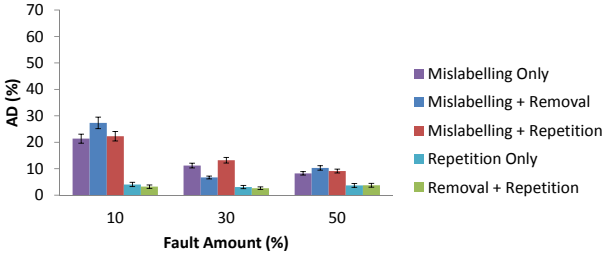
Fig. 6: AD of individual models, compared with models protected with TDFM techniques when trained with faulty Pneumonia datasets. The error bars in the results indicate the 95% confidence intervals. Lower values are better.



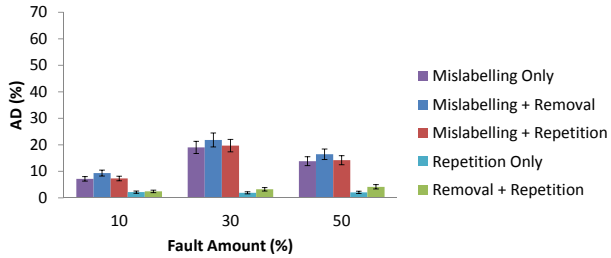
(a) CIFAR-10, ResNet18



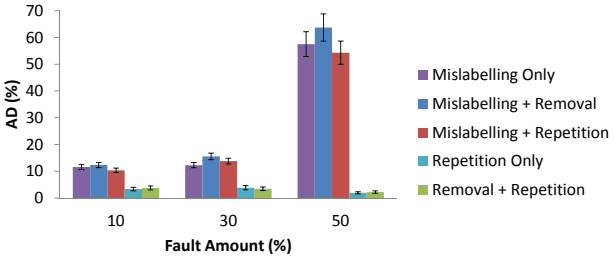
(b) CIFAR-10, VGG11



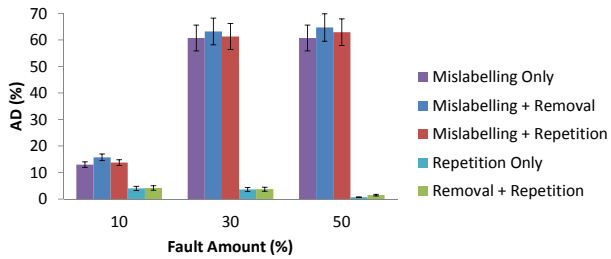
(c) Pneumonia, ResNet18



(d) Pneumonia, VGG11



(e) Pneumonia, ResNet18



(f) Pneumonia, VGG11

Fig. 7: AD of individual models, trained on datasets injected with multiple fault types. The error bars in the results indicate the 95% confidence intervals. Lower values are better.

V. Runtime Cost Analysis

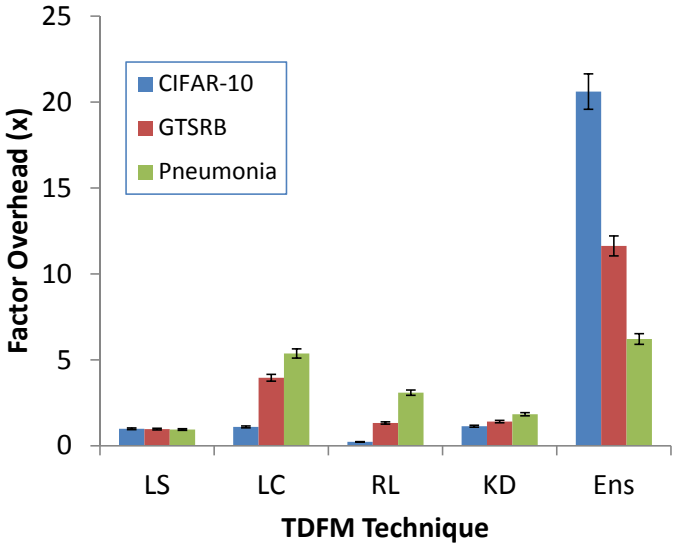


Fig. 8: Average training time overheads across TDFM techniques and datasets. Lower values are better.