



Figure 1: Experiments from inserting malignant labels into the MNIST dataset. For example, inserting two malignant labels per training sample would add two additional "1" values to the one-hot representation of the training label of the training sample, causing the true label to be one among 3 indistinguishable labels. The test set was not injected with malignant labels.