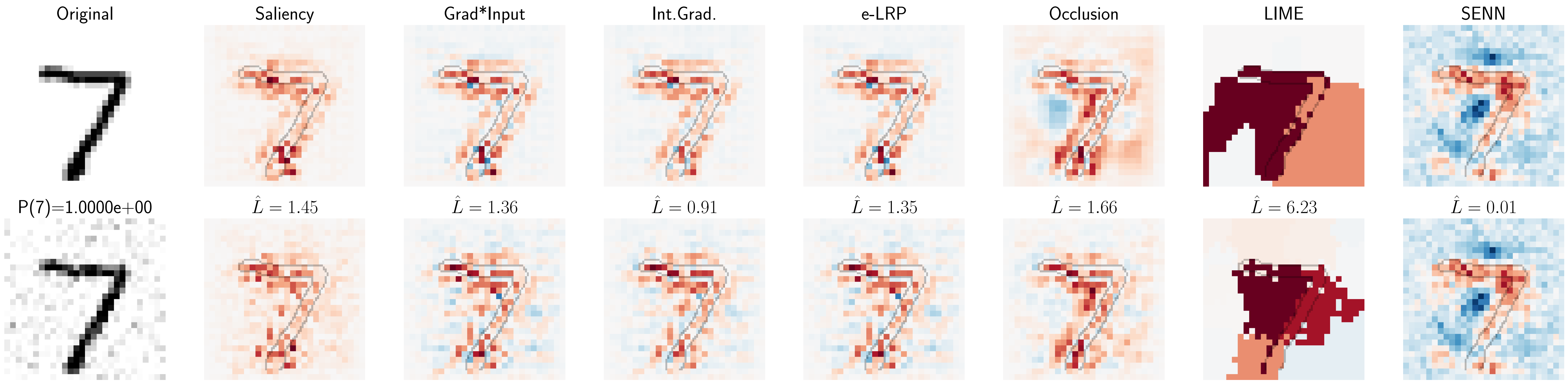
SENN
Robustness

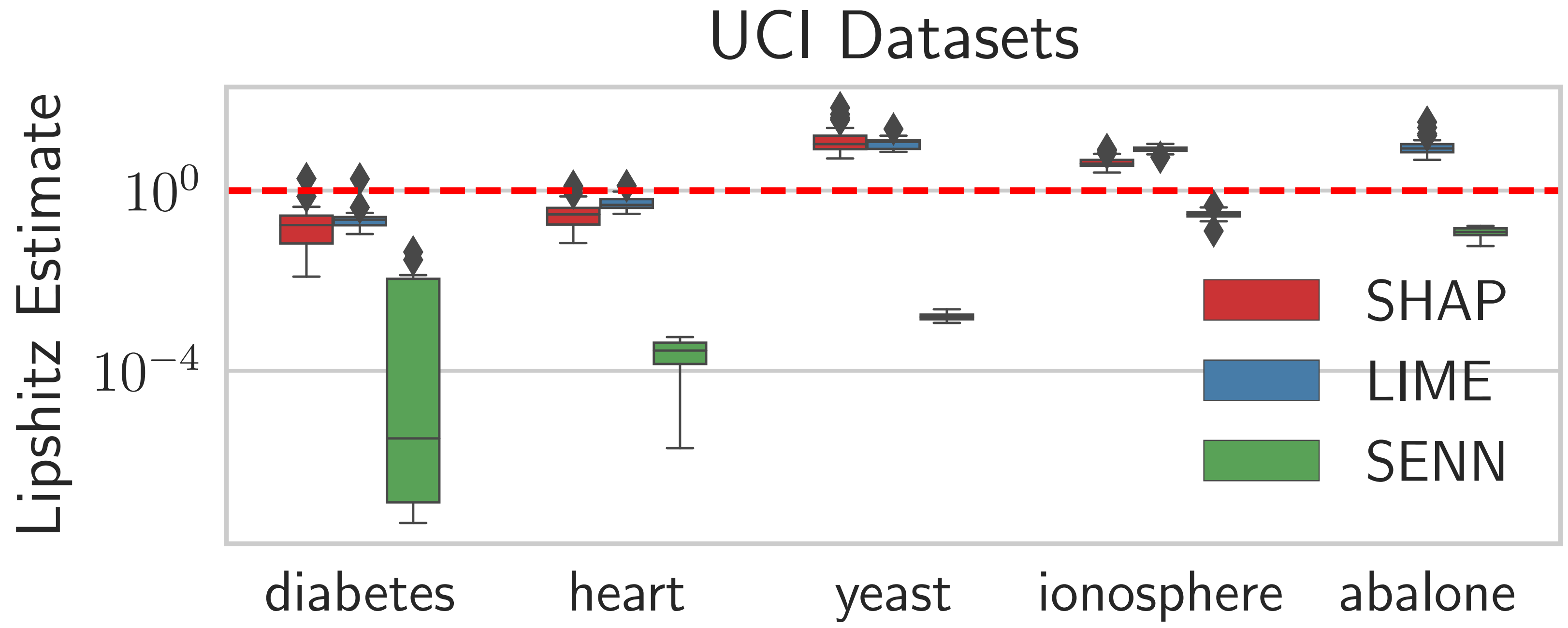


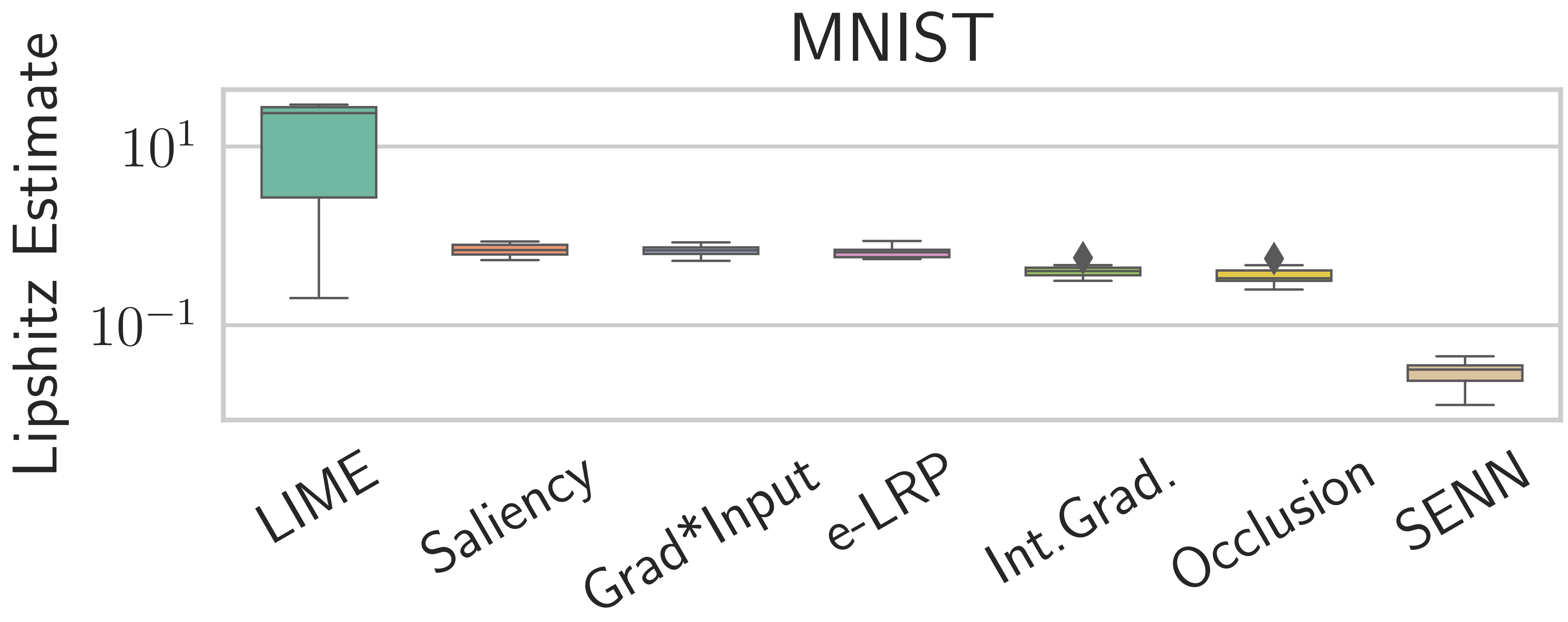
SENN
Robustness

Adversarial robustness estimation:

$$\hat{L}(x) = \arg \max_{\hat{x}\in B_\epsilon(x)} \|f_{expl}(\hat{x}) - f_{expl}(x)\|_2 / \|h(\hat{x}) - h(x)\|$$

Results aggregated over full dataset:


UCI Datasets
Lipshitz Estimate
10^0
10^{-4}
diabetes
heart
yeast
ionosphere
abalone
🟥 SHAP
🟦 LIME
🟩 SENN


MNIST
Lipshitz Estimate
10^1
10^{-1}
LIME
Saliency
Grad*Input
e-LRP
Int.Grad.
Occlusion
SENN