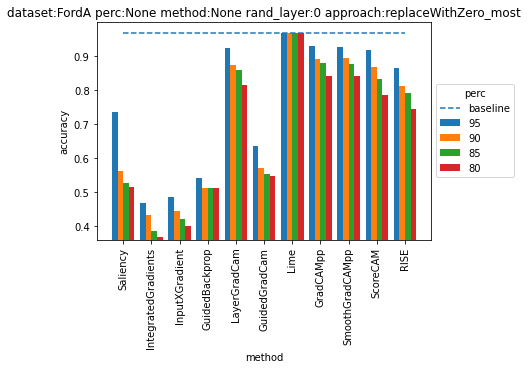
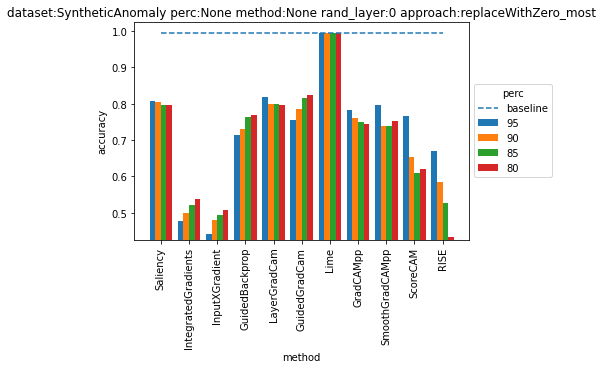
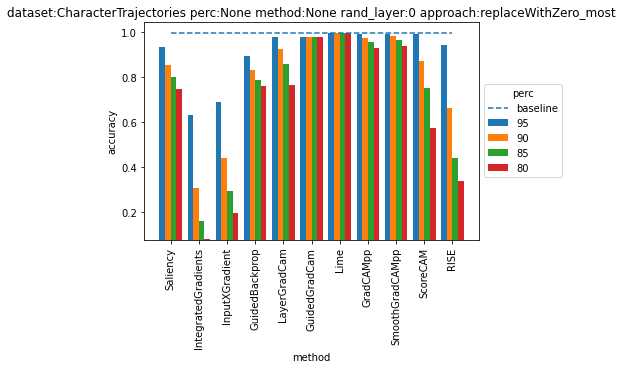
1. Effect of attribution methods:

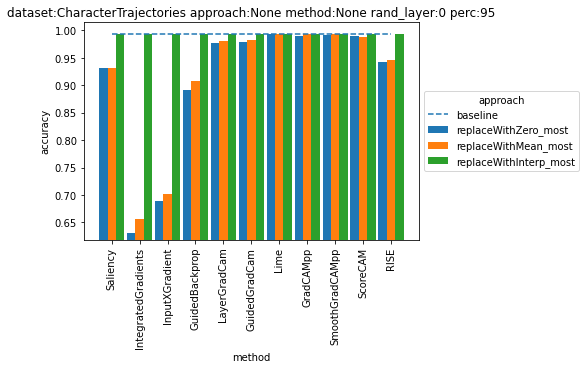
IntegratedGradients is the most effective, followed by InputXGradients

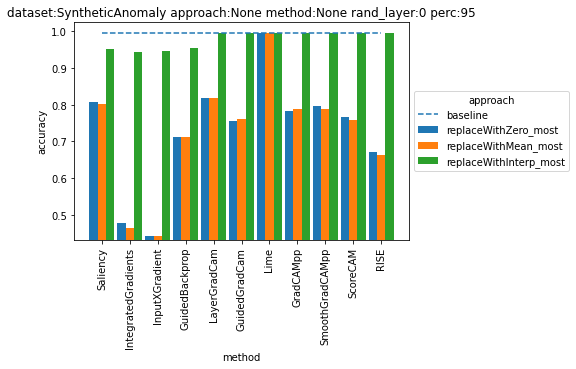
As we can see from the plots that removing points based on these attribution methods leads to a sharp drop in the accuracy.

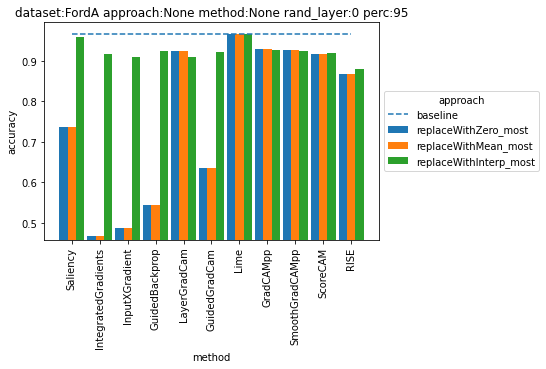
1. Effect of replacement strategy:

Replacing most important points with interpolated values has the minimum effect.

Replacing with zero is slightly more effective than replacing with channel mean.







1. Effect of layer randomization:

The GuidedBackprop method is not an effective attribution method since its attribution map after randomizing its layers is strongly correlated with the attribution map without randomization. Similarly methods like InputXGradient and IntegratedGradients are very effective.

