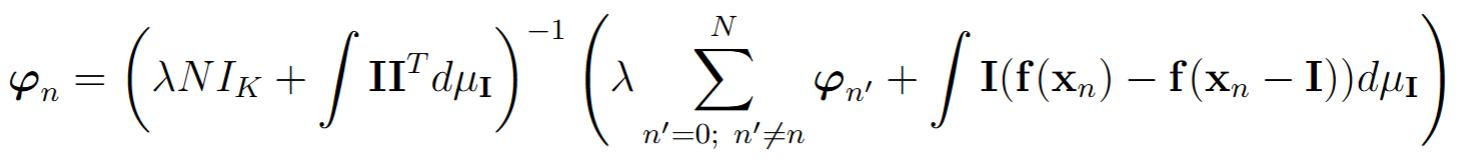
CS2822R Final Project Checkpoint 3

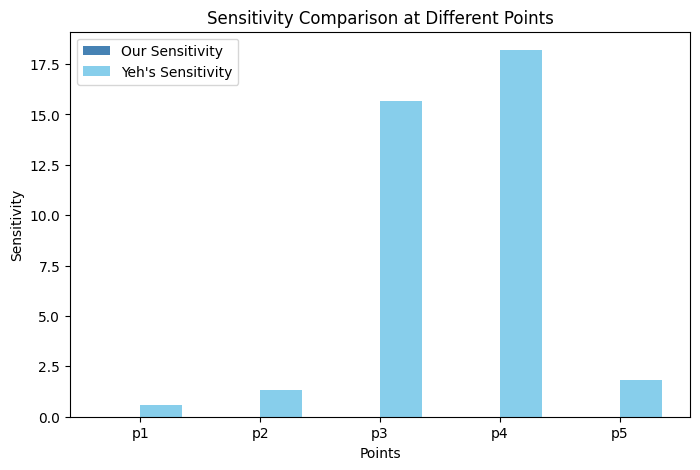
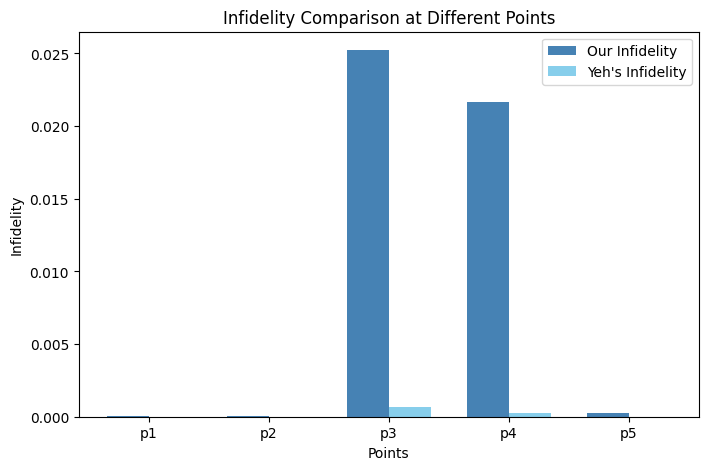
Group 21: Dennis Du, Sezim Yertanatov, Jared Ni

(a) **Research Question**: How would we generalize metrics optimization for explanation methods where there are two or more target metrics to optimize?

Our research focuses on optimizing explanation methods by balancing multiple target metrics, specifically infidelity and sensitivity. In Checkpoint 2, we discussed limitations in existing approaches, particularly the process of optimizing for infidelity and sensitivity separately. This method, while very effective, includes trade-offs: minimizing infidelity can increase sensitivity, and conversely, reducing sensitivity may compromise infidelity. Our approach aims to address these trade-offs by developing a multi-objective optimization method that simultaneously minimizes both metrics through a combined objective function: 

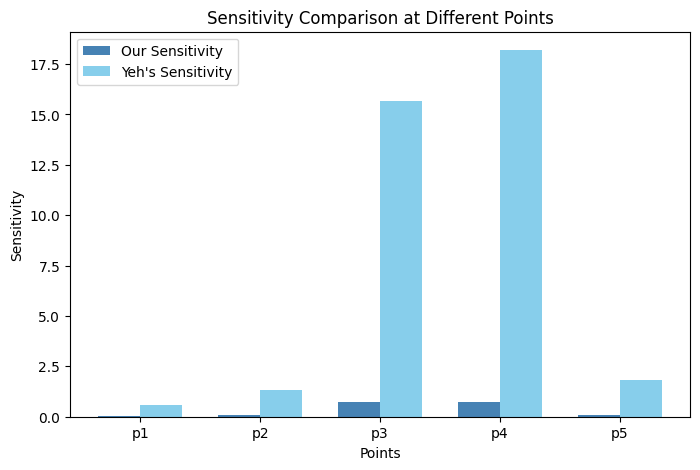
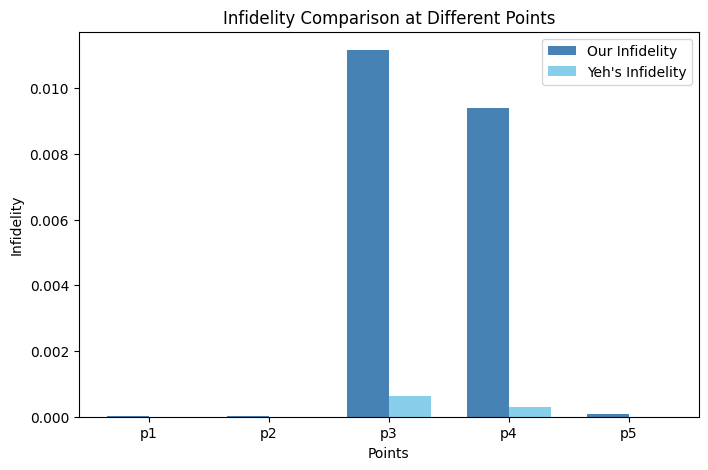
We ran some experiments on a CNN model trained to classify the handwritten digits in the MNIST dataset. The goal was to measure how the probability of the most likely class changes as we perturb the input image, and use this to create an explanation for the model’s classification of that image. For lambda = 0.02, we get the following results, in the table and bar plot below. It appears this causes sensitivity to decrease by orders of magnitude, but at the cost of increasing infidelity by orders of magnitude.

| **Input image** | **Our**  **infidelity** | **Yeh et al.’s infidelity** | **Our sensitivity** | **Yeh et al.’s sensitivity** |
| --- | --- | --- | --- | --- |
| A | 5.1594419e-05 | 2.1025200e-07 | 8.094177e-05 | 0.57809184938 |
| B | 5.9437465e-05 | 2.9644426e-07 | 0.0002139108 | 1.30852101063 |
| C | 0.02523050040 | 0.00064388191 | 0.0020639885 | 15.6712735859 |
| D | 0.02168123601 | 0.00028564136 | 0.0019590314 | 18.1957866131 |
| E | 0.00022400882 | 2.2588810e-07 | 0.0002691616 | 1.82162021248 |

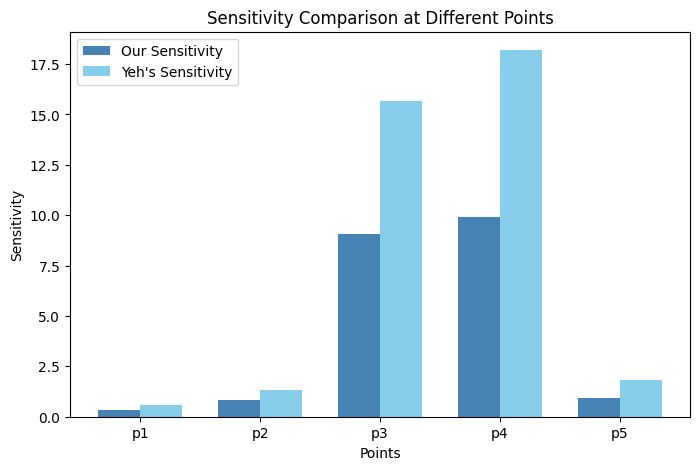
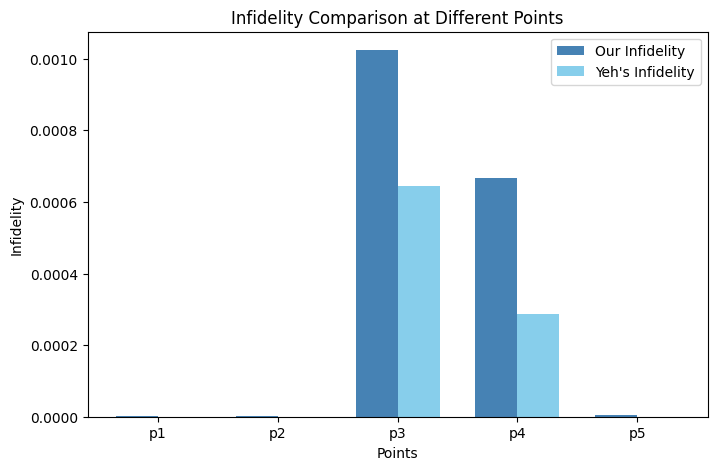


To address this, we adjust lambda to be smaller, in order to put more weight on infidelity and less on sensitivity. We see the following:

lambda = 5e-4:



lambda = 1e-5:



This final chart shows a reasonable tradeoff between infidelity and sensitivity: we can get lower sensitivity by allowing infidelity to increase by a similar amount.