For the dataset I ended up choosing Cracks-and-Potholes due to the inclusion of segmentation masks. As part of the safety algorithm that approximates the relative safety of a pothole detected on the road, having a segmentation mask is much more valuable due to the exact shape and size available rather than a rectangular bounding box which can both under and over approximate these data values. I researched a way to convert bounding boxes into segmentation masks however that process is more time consuming and less reliable. In the future should my model need more data, the possibility of converting the bounding boxes from the RDD2022 dataset to segmentation masks and incorporating them into the training dataset. When it came to cleaning the data, there were several outliers that I looked to trim out. Primarily very small potholes that are clustered together or microcracks that when driven over are essentially harmless to both the vehicle and passengers.

As seen above, there are a bunch of small potholes that show up on the segmentation mask which may create false positives when fed through the safety algorithm. In such cases it makes sense to remove. While reviewing the dataset, the ratio of images that have sizable potholes to images without any potholes were quite skewed. As such, I trimmed down the number of pothole-less samples such that the ratio was around 1:2. Due to the dataset being a university published dataset there were no missing values I had to account for. However I did have to generate the individual pixel coordinate data from the segmentation masks as an input for training.