Part 3 : AI system design

To efficiently identify the most promising 1000 images from an extensive dataset of 1 million, with the goal of enhancing our algorithm's performance,
we can deploy an automated approach based on Active Learning Methods: **entropy-based sampling**, while also recognizing the presence of alternative active learning techniques.

This method starts by running our pretrained algorithm on the entire dataset, which gives confidence scores for each image. Then, we calculate the entropy for each image using these confidence scores. Higher entropy means the model is less sure about its predictions. Our main goal is to pick the top 1000 images with the highest entropy scores because these are the cases where the model is most uncertain and can benefit the most from additional annotation.
When choosing these images, it's important to make sure they represent a wide range of scenarios, objects, or variations related to our algorithm. This diversity enhances the dataset by covering different aspects of the problem. After making this selection, we test the updated model on a separate test dataset to see how it performs.

If it’s necessary, this process can be iteratively repeated with new images to continue the algorithm's enhancement. The iterative nature of this approach allows for continuous learning and adaptation.