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 approach commences by subjecting the entire dataset to our pre-trained algorithm, which yields confidence scores for each image. Subsequently, we compute the entropy for each image based on these confidence scores. Higher entropy values signify greater uncertainty in the model's predictions. The primary objective is to select the top 1000 images with the highest entropy scores, as these are the instances where the model exhibits the most uncertainty and is, therefore, likely to gain the most from further annotation. While curating this selection, it is essential to consider dataset diversity. We should ensure that the chosen images span a wide spectrum of scenarios, objects, or variations pertinent to our algorithm. This diversity enriches the dataset by encompassing various facets of the problem space. Following the selection, we evaluate the updated model's performance on a distinct test dataset.

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.