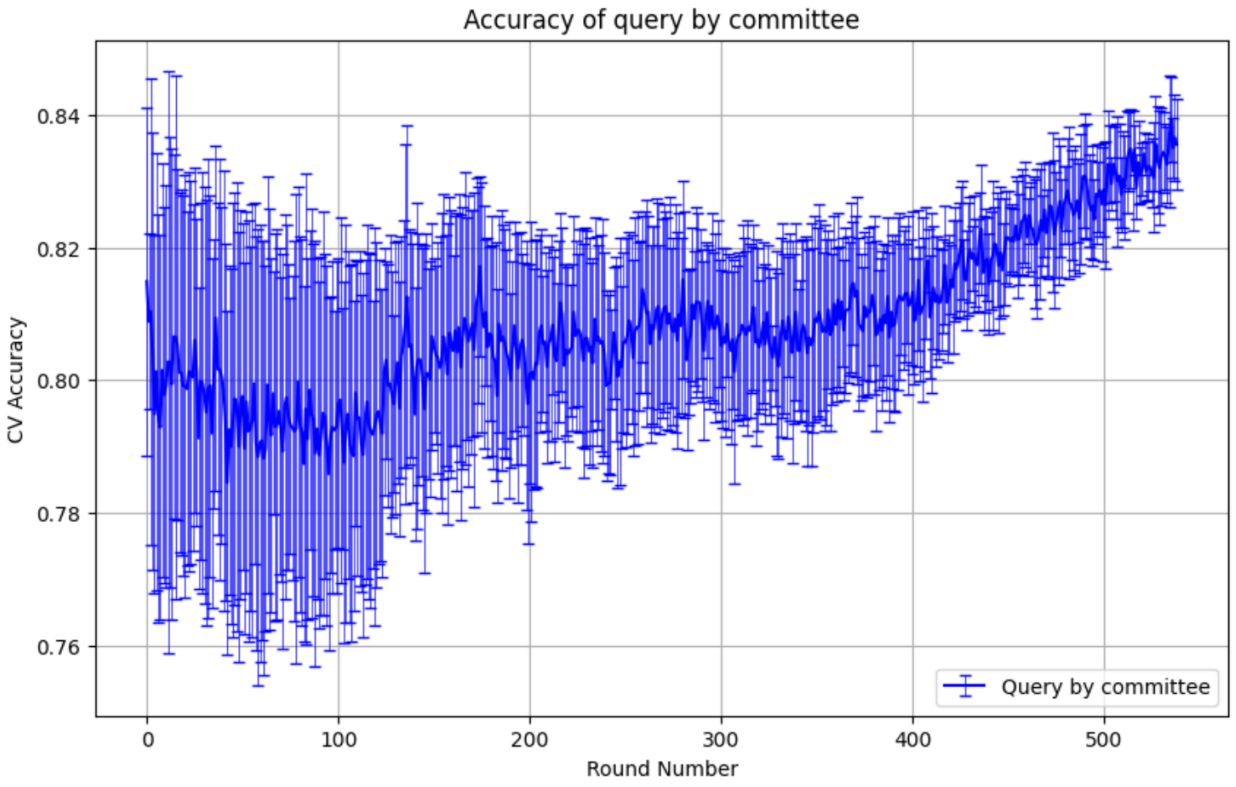
Automation HW2 Chang Liu (cliu8)

Exercise 1.

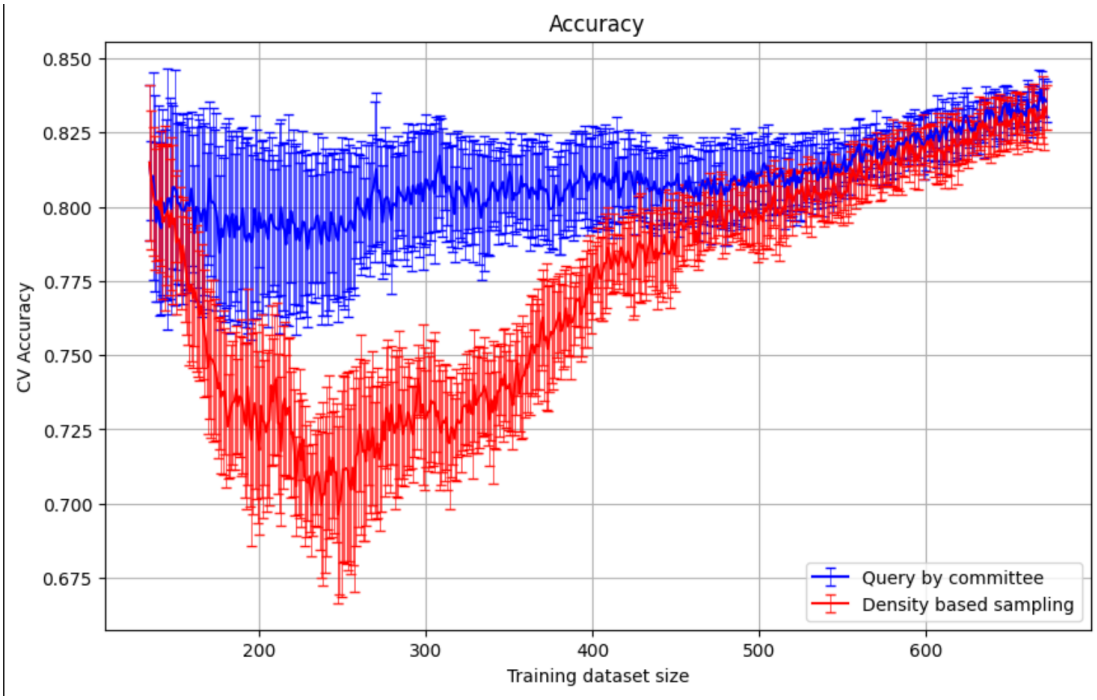
a) The dataset has a shape of (673, 9), with 673 data records, 8 columns as features and the last column as the true label. The data is classified in either 0 or 1, and there are 429 samples categorized as 0, 244 samples categorized as 1.

I used random forest for this binary classification as base learner, the corresponding default loss function applied by the package sklearn is the gini impurity function, $G = 1 - \sum_{i=1}^{C} p_i^2$, where p_i is the proportion of class i at the node.

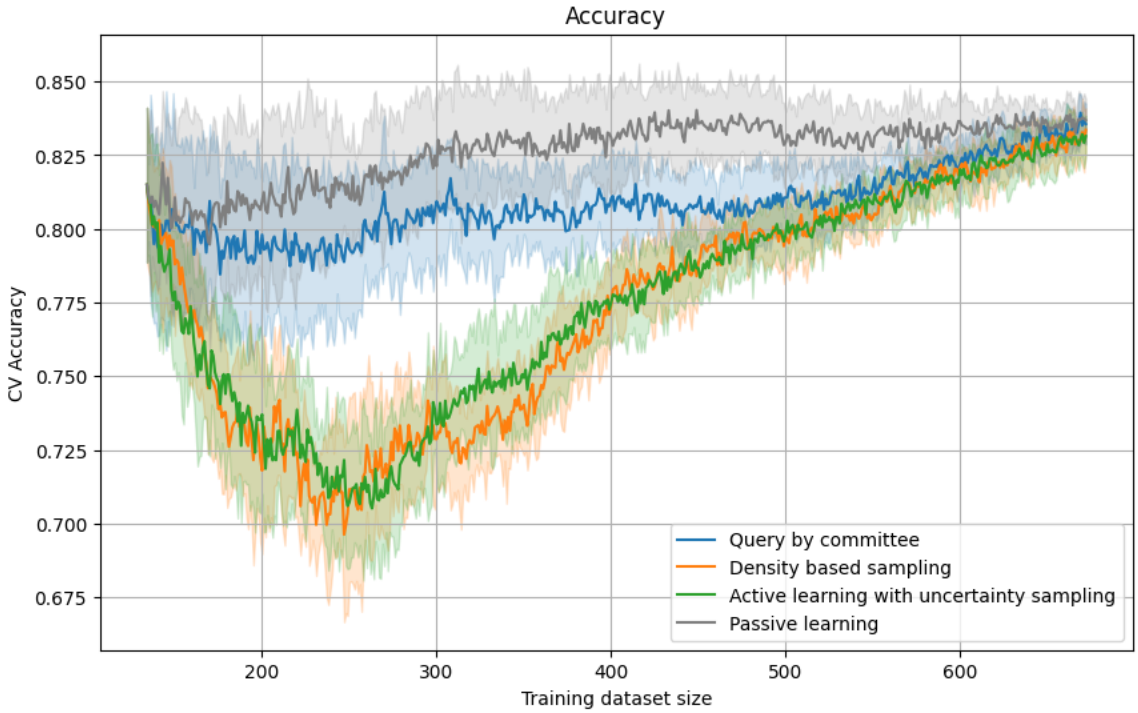
b) Implementation of query by committee:
Random forest is naturally suitable for constructing a committee. I built a random forest classifier with 50 trees and select 5 trees as the committee. Those 5 trees will decide the next input sample based on the KL divergence, and the selected sample will be input to retrain random forest classifier.



c) I computed Euclidean distance between each point pairs as dissimilarity measurement, which is $\text{dist}(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x_i')^2}$. The similarity is then expressed as $e^{-\text{dist}}$, together with a beta parameter equal to 0.5. The query strategy is uncertainty sampling here with entropy-based uncertainty.



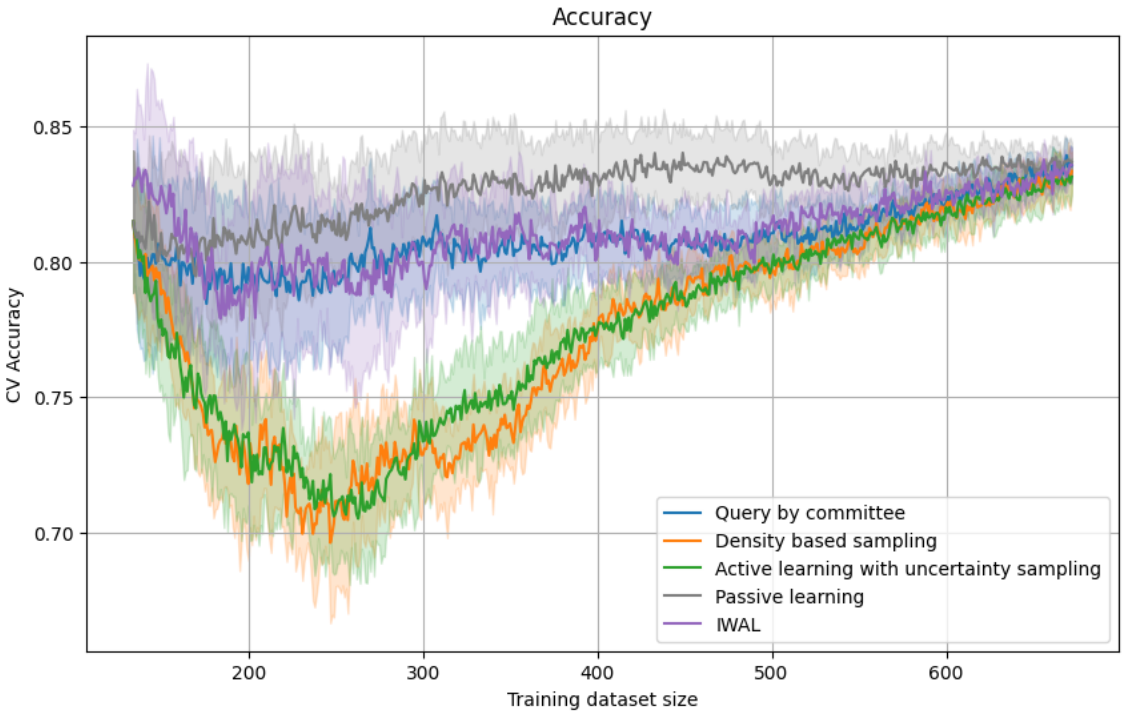
d) Here the active learning computes entropy-based uncertainty and do uncertainty sampling, and passive learning randomly pick one sample for next input.



e) Those curves follow my expectation. We can find the passive learning simulation has the most stable and high performance throughout the whole simulation, while the one using query by committee has a relatively stable performance but has some slight drop on accuracy when it includes the samples that are most disagreed within the small committee in the beginning, which indicates the difficulty on correctly identifying them.
This difficulty of being correctively predicted of specific samples are more obvious in active learning with entropy-based uncertainty sampling and density-based sampling, with these two simulations share very similar performance curve and similar sharp decrease on accuracy. This can be explained by the relatively small beta parameter (weight = 0.5) I chose for the density-based sampling, so it has less impact due to its local density (similarity). They all reach similar performance finally as they receive the whole data, showing by the convergence of the standard deviation.

Exercise 2.

a) For this experiment, I used a p_min = 0.1, k = 5 (committee number).

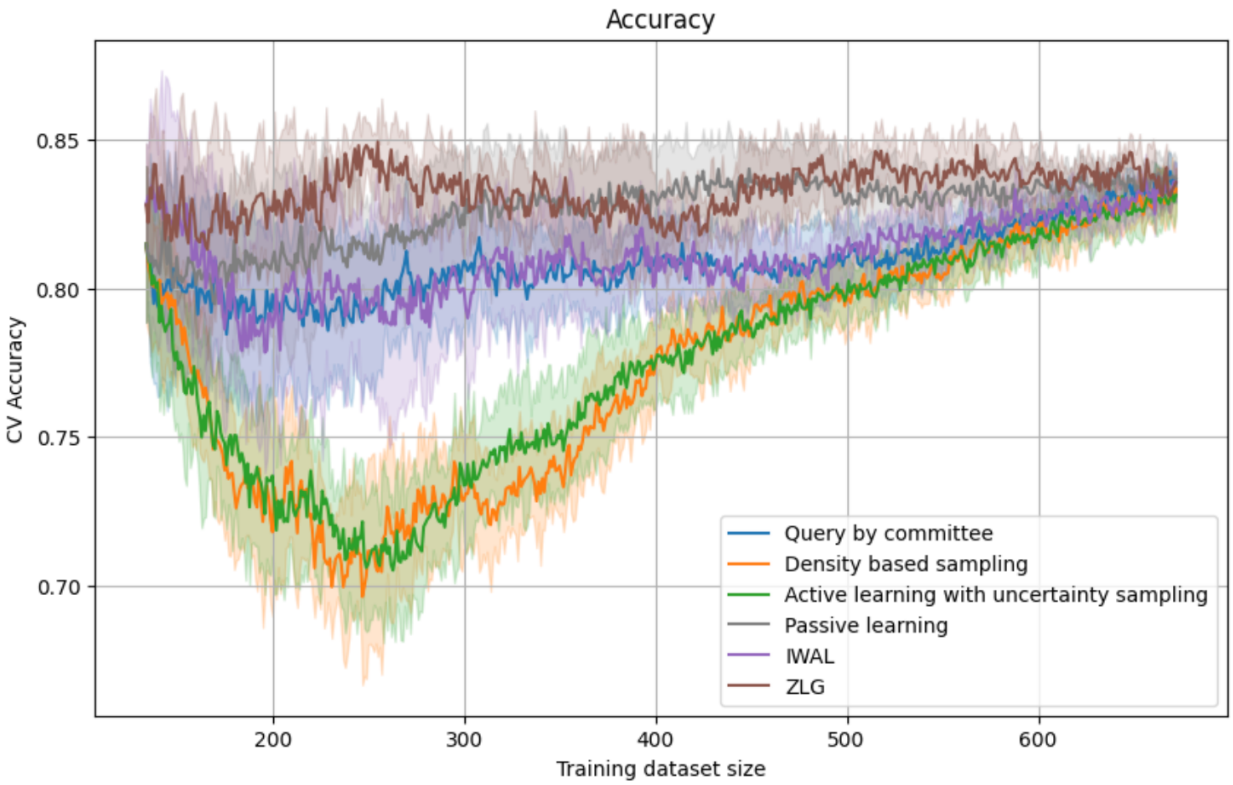


b) Yes. As for IWAL we only ran for 5 different seeds, so it has a relatively different starting point compared to other curves. We can find the IWAL algorithm has a similar trend on its prediction accuracy, which decreases in the beginning and soon pick up its accuracy with more training data included. This is because IWAL

algorithm is also a type 1 algorithm that does hypothesis elimination, so it displays a similar curve to other active learning method with the same strategy.

Exercise 3.

a) I implemented ZLG algorithm, with a t = 0.1.



b) Yes, the IWAL algorithm and ZLG algorithm have the same starting point as they all ran 5 different seeds. As ZLG does not assume a typical hypothesis elimination strategy, but rather a type 2 algorithm, it doesn't include the hardest to predict sample in the beginning, thus it doesn't witness a steep decrease on accuracy in the beginning. Instead, it selects the next sample by minimizing the future risk, which is a measurement based on similarity graph. It might choose some samples that can reinforce the graph structure that leads to the slight increase on accuracy and shows slight decrease for the addition of hard to predict samples.