Analyzing different active learning strategies with scikit-activeml

Selected topics in AI- 2

Name	ID
Abdallah Mamdouh	20200308
Abdelhamed adel	20200275
Sohaila Ahmed Sayed	20200783
Rawan Khaled	20201072

Dataset Brief:

Here is a brief of the dataset used for the experiment:

Digits Dataset:

- The digits dataset is a collection of handwritten digits commonly used in machine learning for classification tasks. It consists of images of handwritten digits from 0 to 9, each digit represented as a grayscale image with a resolution of 8 pixels by 8 pixels. In total, there are 10 classes corresponding to the 10 possible digits.
- It is frequently used for training and testing machine learning algorithms, particularly for tasks like digit recognition. Each image in the dataset is represented as a vector of 64 numerical values,

Iris Dataset:

- The Iris dataset is a collection of data about different species of iris flowers.
- It includes measurements of the length and width of both the sepals (the green leaf-like structures) and the petals of the flowers.
- There are three species of iris in the dataset: setosa, versicolor, and virginica.
- The objective of the dataset is typically to build a model that can classify iris flowers into one of these three species based on their measurements.

Breast Cancer Wisconsin Dataset:

- The Breast Cancer Wisconsin dataset contains information about breast cancer tumors.
- The dataset aims to help in diagnosing whether a tumor is malignant (cancerous) or benign (non-cancerous).
- It contains 569 instances, with each instance having 30 numeric features. These features are computed from images of cell nuclei and include attributes such as radius, texture, perimeter, area,

smoothness, compactness, concavity, symmetry, fractal dimension, etc.

Stroke Dataset:

- The Stroke Dataset comprises medical records of individuals, including demographics, habits, and clinical details, to investigate stroke occurrences.
- Attributes such as age, gender, hypertension, and smoking status are recorded.
- This dataset aids in analyzing risk factors and developing predictive models for stroke occurrence, contributing to preventive healthcare strategies.

Discussing the query strategies used next...

Query strategies:

Here's an overview of the sampling techniques used:

1. Query by Committee (QBC): This technique involves asking a group (committee) of models or experts for their opinions on which data points are most informative for learning. By considering multiple perspectives, the aim is to select the instances about which there is the most disagreement among the committee members, as these are likely to be the most beneficial for improving the model's performance.

2. Margin Sampling: Margin sampling focuses on selecting data points that lie close to the decision boundary of a classification model. These points are often the most challenging to classify accurately, as they

are situated in regions where the model is uncertain about the correct label. By focusing on these instances, margin sampling aims to improve the model's performance by addressing its weaknesses in classification.

3. Uncertainty Sampling: This technique involves selecting instances for which the model is the most uncertain about their correct label. By prioritising these ambiguous cases, uncertainty sampling aims to reduce the model's uncertainty and improve its overall performance. This is typically achieved by measuring the model's confidence or certainty in its predictions and selecting instances where this confidence is low.

Moving on to analysing the data...

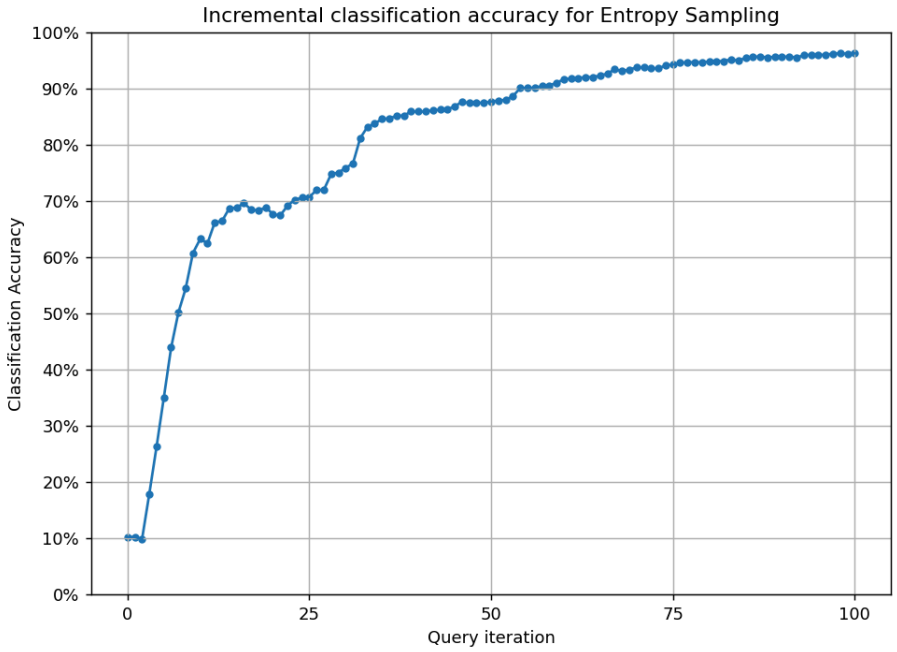
Digits Dataset Analysis:

Afterward, we utilised a pool-based sampling technique for training and applied the ParzenWindowClassifier for making predictions. The model attained an accuracy of 0.10 without active learning. Here are the results following its application.

Using uncertainty sampling strategy

```
Query 1: The accuracy score is 0.10239287701725097.
Query 2: The accuracy score is 0.09849749582637729.
Query 3: The accuracy score is 0.1791875347801892.
Query 4: The accuracy score is 0.2632164718976071.
Query 5: The accuracy score is 0.35058430717863104.
Query 6: The accuracy score is 0.43906510851419034.
Query 7: The accuracy score is 0.5013912075681691.
Query 8: The accuracy score is 0.5442404006677797.
Query 9: The accuracy score is 0.6071229827490262.
```

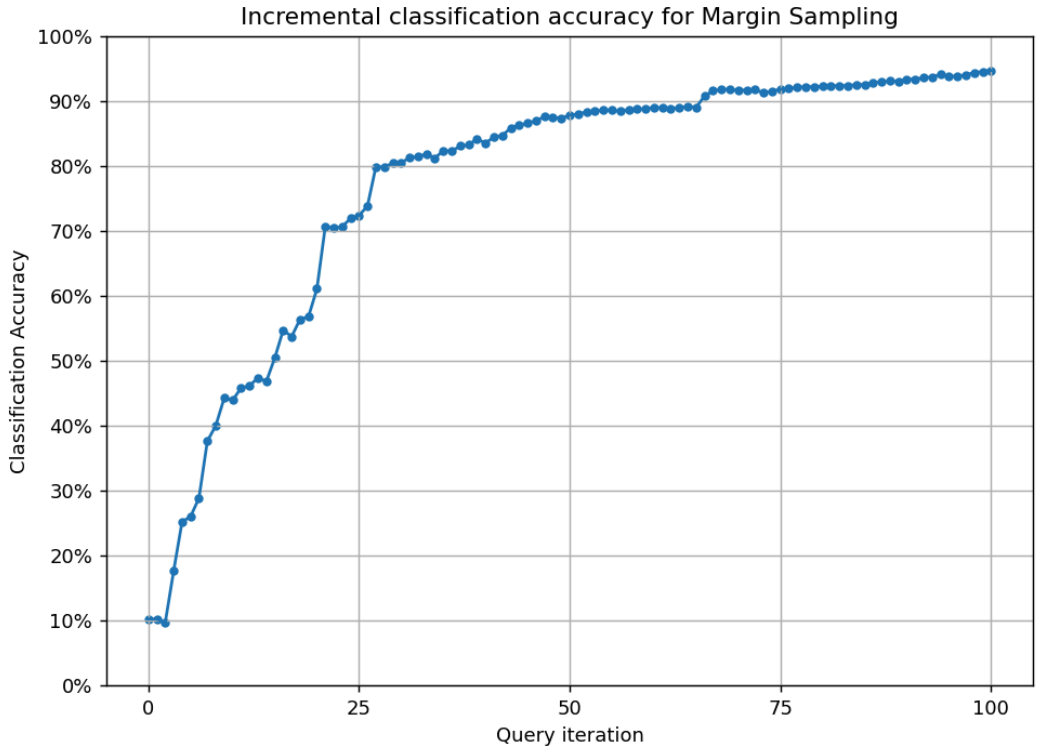
```
Query 91: The accuracy score is 0.9565943238731218.
Query 92: The accuracy score is 0.9543683917640512.
Query 93: The accuracy score is 0.9588202559821926.
Query 94: The accuracy score is 0.9593767390094602.
Query 95: The accuracy score is 0.9593767390094602.
Query 96: The accuracy score is 0.9593767390094602.
Query 97: The accuracy score is 0.9616026711185309.
Query 98: The accuracy score is 0.9621591541457986.
Query 99: The accuracy score is 0.9616026711185309.
Query 100: The accuracy score is 0.9621591541457986.
```



using margin sampling strategy:

```
Query 1: The accuracy score is 0.10239287701725097.
Query 2: The accuracy score is 0.09682804674457429.
Query 3: The accuracy score is 0.17640511964385086.
Query 4: The accuracy score is 0.25208681135225375.
Query 5: The accuracy score is 0.2598775737340011.
Query 6: The accuracy score is 0.2882582081246522.
Query 7: The accuracy score is 0.3761825264329438.
Query 8: The accuracy score is 0.4001112966054535.
Query 9: The accuracy score is 0.44240400667779634.
```

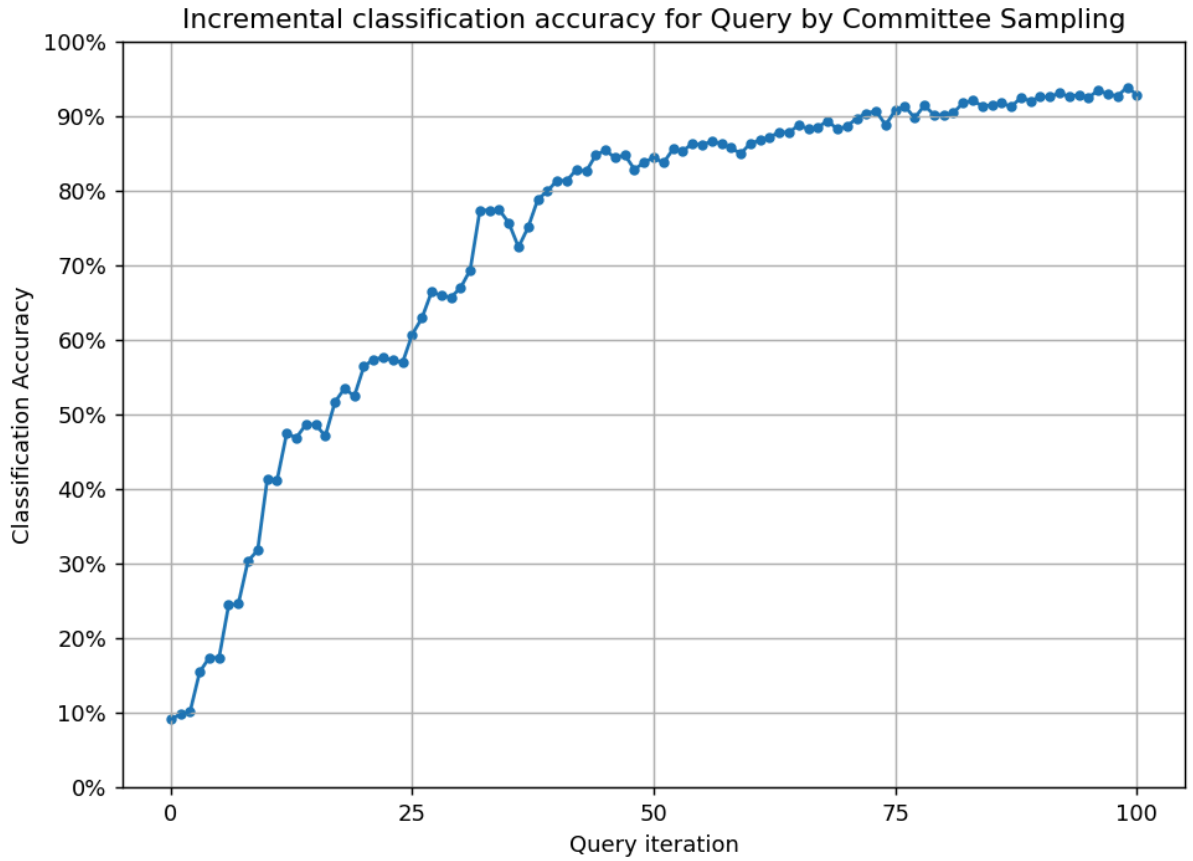
```
Query 91: The accuracy score is 0.9332220367278798.
Query 92: The accuracy score is 0.9354479688369505.
Query 93: The accuracy score is 0.9365609348914858.
Query 94: The accuracy score is 0.9404563160823595.
Query 95: The accuracy score is 0.9376739009460211.
Query 96: The accuracy score is 0.9376739009460211.
Query 97: The accuracy score is 0.9398998330550918.
Query 98: The accuracy score is 0.9432387312186978.
Query 99: The accuracy score is 0.9449081803005008.
Query 100: The accuracy score is 0.9465776293823038.
```



using Query by Committee (QBC):

```
Query 1: The accuracy score is 0.09794101279910963.
Query 2: The accuracy score is 0.10127991096271564.
Query 3: The accuracy score is 0.1547022815804118.
Query 4: The accuracy score is 0.17417918753478018.
Query 5: The accuracy score is 0.17362270450751252.
Query 6: The accuracy score is 0.2442960489705064.
Query 7: The accuracy score is 0.2459654980523094.
Query 8: The accuracy score is 0.30328324986087923.
Query 9: The accuracy score is 0.31775180856983865.
```

```
Query 90: The accuracy score is 0.9254312743461325.
Query 91: The accuracy score is 0.9254312743461325.
Query 92: The accuracy score is 0.9304396215915415.
Query 93: The accuracy score is 0.9254312743461325.
Query 94: The accuracy score is 0.9282136894824707.
Query 95: The accuracy score is 0.9243183082915971.
Query 96: The accuracy score is 0.9343350027824151.
Query 97: The accuracy score is 0.9287701725097385.
Query 98: The accuracy score is 0.9265442404006677.
Query 99: The accuracy score is 0.9382303839732888.
Query 100: The accuracy score is 0.9282136894824707.
```

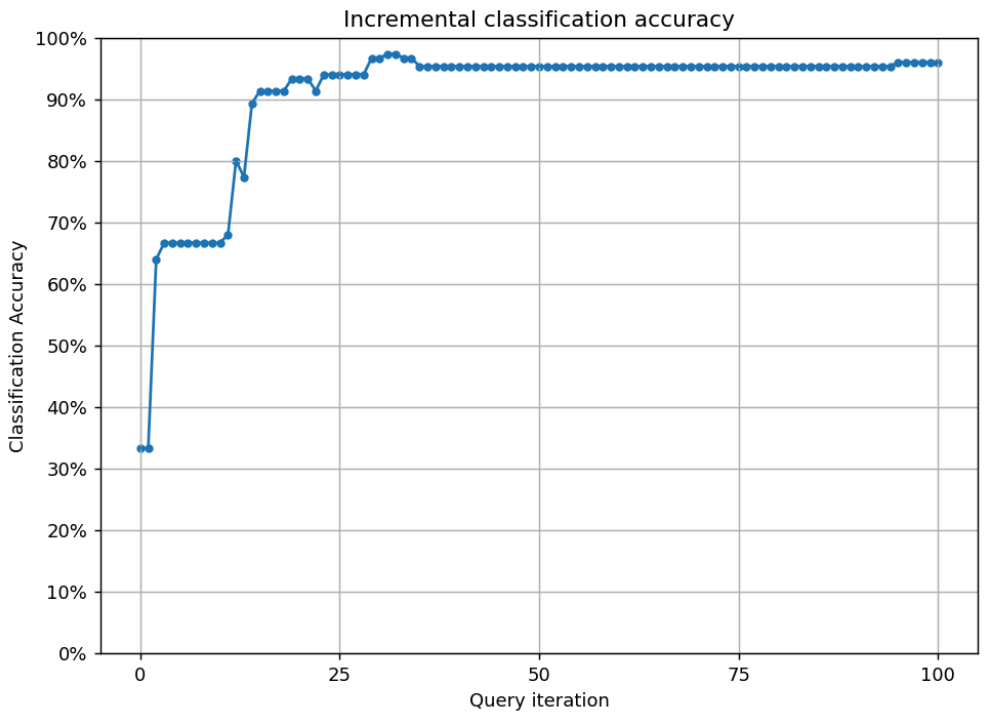


IRIS Dataset Analysis:

we utilised a pool-based sampling technique for training and applied the ParzenWindowClassifier for making predictions. The model attained an accuracy of 0.27 without active learning. Here are the outcomes following its implementation

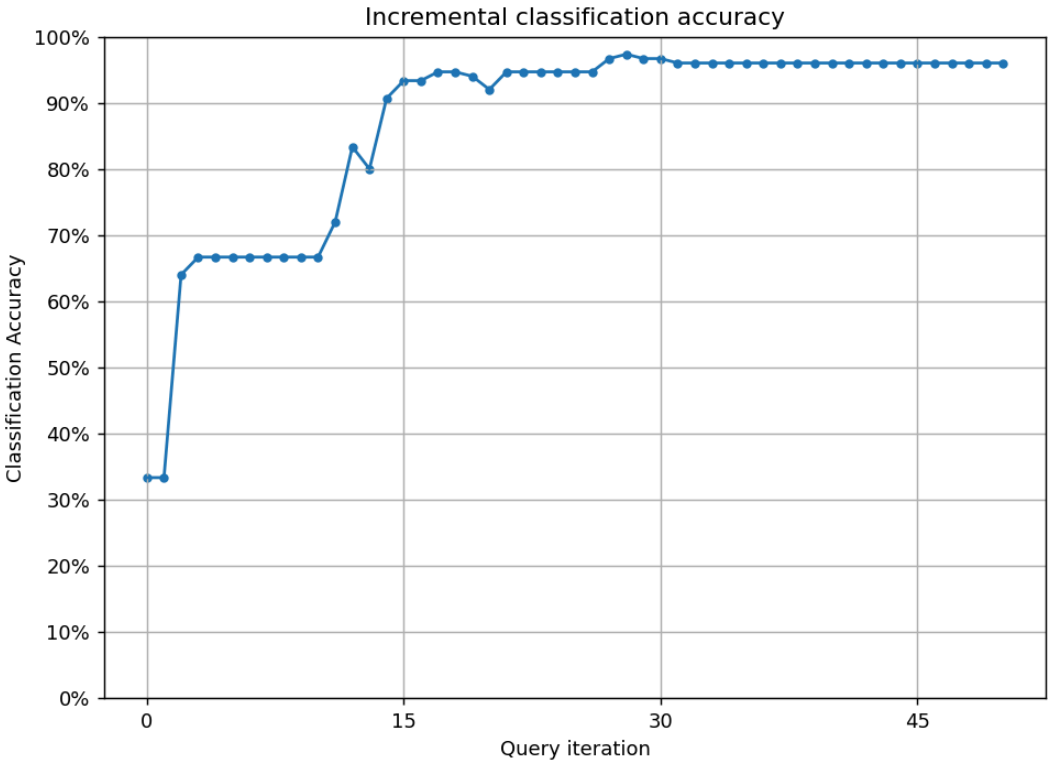
using *uncertainty* sampling strategy:

```
Accuracy after query 1: 0.3333
Accuracy after query 2: 0.6400
Accuracy after query 3: 0.6667
Accuracy after query 4: 0.6667
Accuracy after query 5: 0.6667
Accuracy after query 6: 0.6667
Accuracy after query 7: 0.6667
Accuracy after query 8: 0.6667
Accuracy after query 9: 0.6667
Accuracy after query 10: 0.6667
Accuracy after query 11: 0.6800
Accuracy after query 12: 0.8000
Accuracy after query 13: 0.7733
Accuracy after query 14: 0.8933
Accuracy after query 15: 0.9133
Accuracy after query 16: 0.9133
Accuracy after query 17: 0.9133
Accuracy after query 18: 0.9133
Accuracy after query 19: 0.9333
Accuracy after query 20: 0.9333
Accuracy after query 21: 0.9333
Accuracy after query 22: 0.9133
Accuracy after query 23: 0.9400
Accuracy after query 24: 0.9400
Accuracy after query 25: 0.9400
Accuracy after query 75: 0.9533
Accuracy after query 76: 0.9533
Accuracy after query 77: 0.9533
Accuracy after query 78: 0.9533
Accuracy after query 79: 0.9533
Accuracy after query 80: 0.9533
Accuracy after query 81: 0.9533
Accuracy after query 82: 0.9533
Accuracy after query 83: 0.9533
Accuracy after query 84: 0.9533
Accuracy after query 85: 0.9533
Accuracy after query 86: 0.9533
Accuracy after query 87: 0.9533
Accuracy after query 88: 0.9533
Accuracy after query 89: 0.9533
Accuracy after query 90: 0.9533
Accuracy after query 91: 0.9533
Accuracy after query 92: 0.9533
Accuracy after query 93: 0.9533
Accuracy after query 94: 0.9533
Accuracy after query 95: 0.9600
Accuracy after query 96: 0.9600
Accuracy after query 97: 0.9600
Accuracy after query 98: 0.9600
Accuracy after query 99: 0.9600
Accuracy after query 100: 0.9600
```



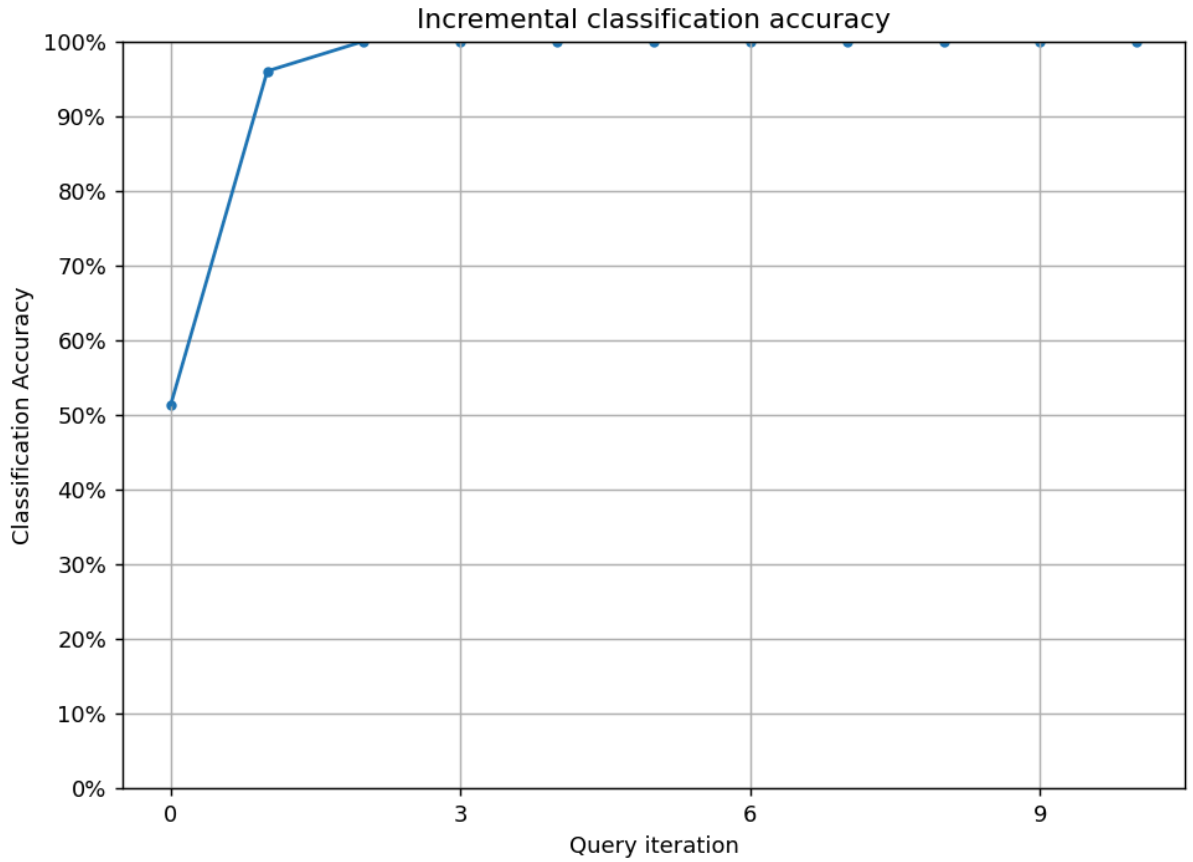
using margin sampling strategy:

```
Accuracy after query 1: 0.3333
Accuracy after query 2: 0.6400
Accuracy after query 3: 0.6667
Accuracy after query 4: 0.6667
Accuracy after query 5: 0.6667
Accuracy after query 6: 0.6667
Accuracy after query 7: 0.6667
Accuracy after query 8: 0.6667
Accuracy after query 9: 0.6667
Accuracy after query 10: 0.6667
Accuracy after query 11: 0.7200
Accuracy after query 12: 0.8333
Accuracy after query 13: 0.8000
Accuracy after query 14: 0.9067
Accuracy after query 15: 0.9333
Accuracy after query 16: 0.9333
Accuracy after query 17: 0.9467
Accuracy after query 18: 0.9467
Accuracy after query 19: 0.9400
Accuracy after query 20: 0.9200
Accuracy after query 21: 0.9467
Accuracy after query 22: 0.9467
Accuracy after query 23: 0.9467
Accuracy after query 24: 0.9467
Accuracy after query 25: 0.9467
Accuracy after query 25: 0.9467
Accuracy after query 26: 0.9467
Accuracy after query 27: 0.9667
Accuracy after query 28: 0.9733
Accuracy after query 29: 0.9667
Accuracy after query 30: 0.9667
Accuracy after query 31: 0.9600
Accuracy after query 32: 0.9600
Accuracy after query 33: 0.9600
Accuracy after query 34: 0.9600
Accuracy after query 35: 0.9600
Accuracy after query 36: 0.9600
Accuracy after query 37: 0.9600
Accuracy after query 38: 0.9600
Accuracy after query 39: 0.9600
Accuracy after query 40: 0.9600
Accuracy after query 41: 0.9600
Accuracy after query 42: 0.9600
Accuracy after query 43: 0.9600
Accuracy after query 44: 0.9600
Accuracy after query 45: 0.9600
Accuracy after query 46: 0.9600
Accuracy after query 47: 0.9600
Accuracy after query 48: 0.9600
Accuracy after query 49: 0.9600
Accuracy after query 50: 0.9600
```



using Query by Committee (QBC) strategy:

```
Accuracy after query 1: 0.9600
Accuracy after query 2: 1.0000
Accuracy after query 3: 1.0000
Accuracy after query 4: 1.0000
Accuracy after query 5: 1.0000
Accuracy after query 6: 1.0000
Accuracy after query 7: 1.0000
Accuracy after query 8: 1.0000
Accuracy after query 9: 1.0000
Accuracy after query 10: 1.0000
```



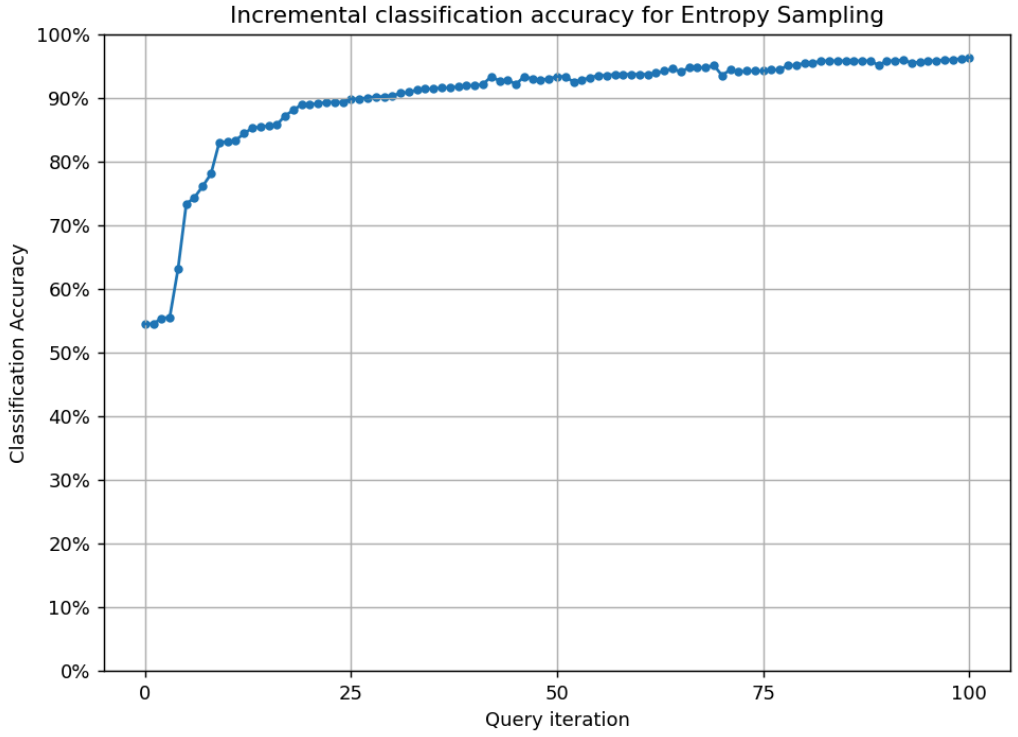
Breast Cancer WisconsinDataset Analysis:

We utilised a pool-based sampling technique for training and applied the ParzenWindowClassifier for making predictions. The model attained an accuracy of 0.54 without active learning. Here are the outcomes following its implementation.

Using Entropy sampling:

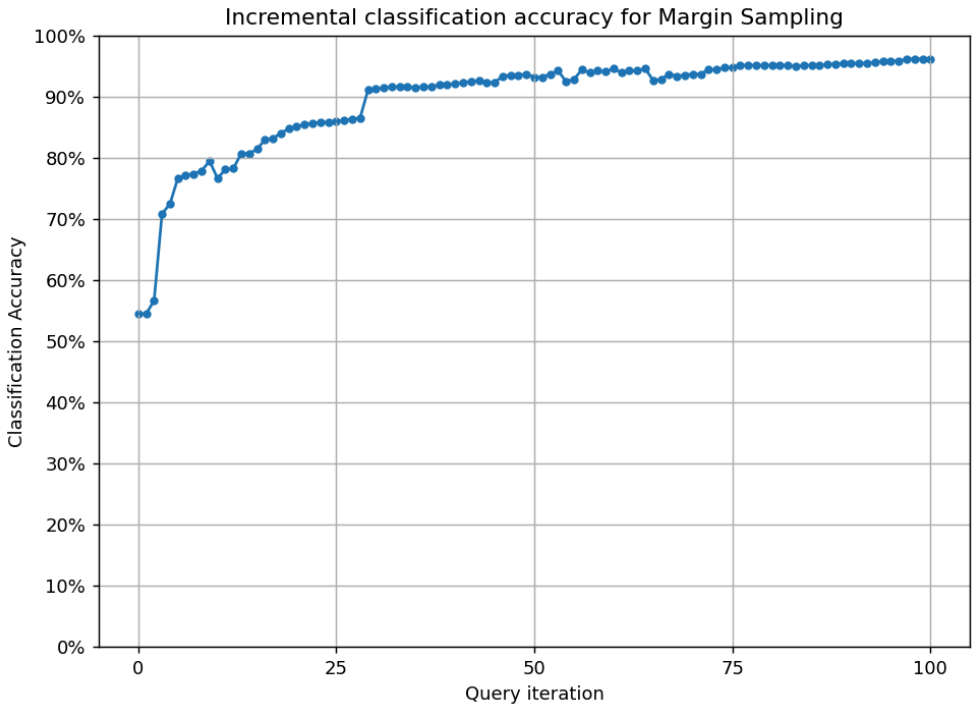
```
Query 1: The accuracy score is 0.5448154657293497.
Query 2: The accuracy score is 0.5536028119507909.
Query 3: The accuracy score is 0.5553602811950791.
Query 4: The accuracy score is 0.6309314586994728.
Query 5: The accuracy score is 0.7328646748681898.
Query 6: The accuracy score is 0.7434094903339191.
Query 7: The accuracy score is 0.7609841827768014.
Query 8: The accuracy score is 0.7803163444639719.
```

```
Query 90: The accuracy score is 0.9578207381370826.
Query 91: The accuracy score is 0.9578207381370826.
Query 92: The accuracy score is 0.9595782073813708.
Query 93: The accuracy score is 0.9543057996485061.
Query 94: The accuracy score is 0.9560632688927944.
Query 95: The accuracy score is 0.9578207381370826.
Query 96: The accuracy score is 0.9578207381370826.
Query 97: The accuracy score is 0.9595782073813708.
Query 98: The accuracy score is 0.9595782073813708.
Query 99: The accuracy score is 0.961335676625659.
Query 100: The accuracy score is 0.9630931458699473.
```



Using Margin sampling:

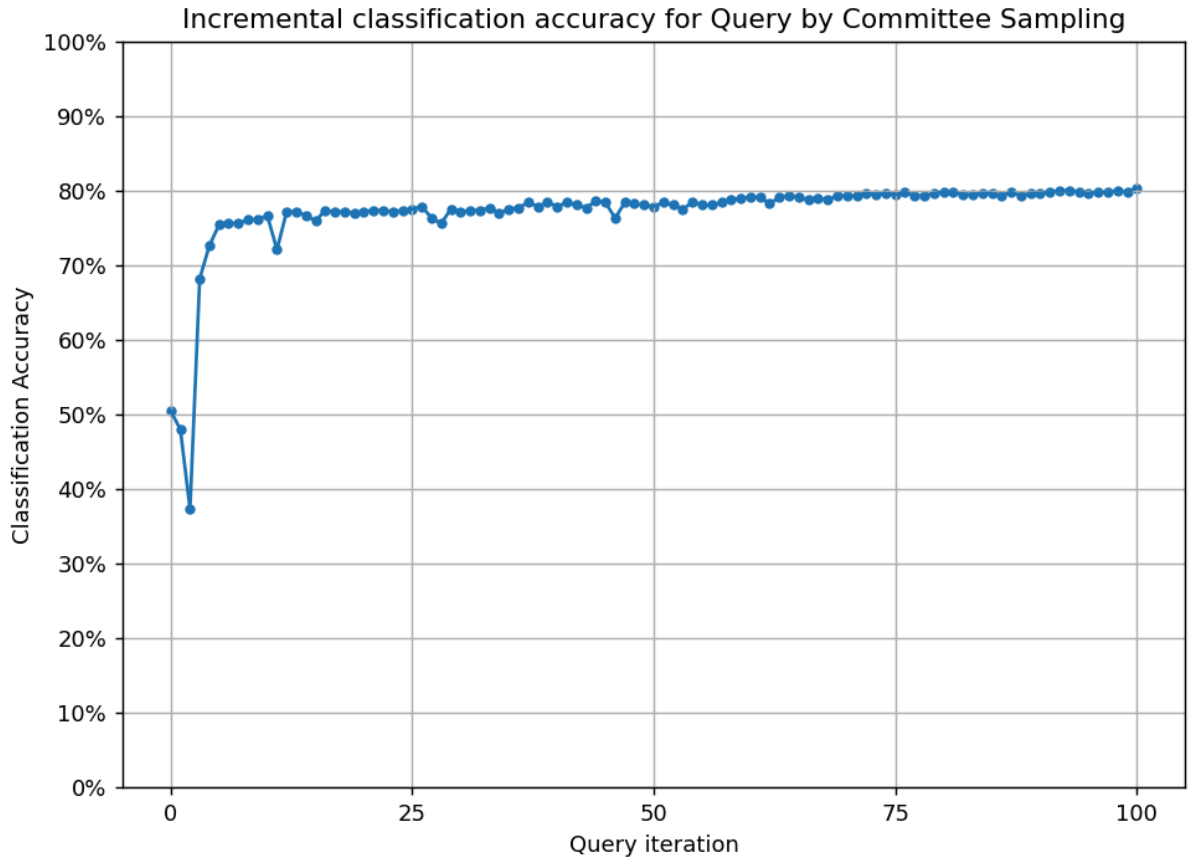
```
Query 1: The accuracy score is 0.5448154657293497.
Query 2: The accuracy score is 0.5659050966608085.
Query 3: The accuracy score is 0.7082601054481547.
Query 4: The accuracy score is 0.7240773286467487.
Query 5: The accuracy score is 0.7662565905096661.
Query 6: The accuracy score is 0.7715289982425307.
Query 7: The accuracy score is 0.773286467486819.
Query 8: The accuracy score is 0.7785588752196837.
Query 9: The accuracy score is 0.7943760984182777.

Query 94: The accuracy score is 0.9578207381370826.
Query 95: The accuracy score is 0.9578207381370826.
Query 96: The accuracy score is 0.9578207381370826.
Query 97: The accuracy score is 0.961335676625659.
Query 98: The accuracy score is 0.961335676625659.
Query 99: The accuracy score is 0.961335676625659.
Query 100: The accuracy score is 0.961335676625659.
```



Using Query by Committee

```
Query 1: The accuracy score is 0.4797891036906854.
Query 2: The accuracy score is 0.37258347978910367.
Query 3: The accuracy score is 0.6818980667838312.
Query 4: The accuracy score is 0.7258347978910369.
Query 5: The accuracy score is 0.7539543057996485.
Query 6: The accuracy score is 0.7557117750439367.
```

```
Query 93: The accuracy score is 0.7996485061511424.
Query 94: The accuracy score is 0.7978910369068541.
Query 95: The accuracy score is 0.7961335676625659.
Query 96: The accuracy score is 0.7978910369068541.
Query 97: The accuracy score is 0.7978910369068541.
Query 98: The accuracy score is 0.7996485061511424.
Query 99: The accuracy score is 0.7978910369068541.
Query 100: The accuracy score is 0.8031634446397188.
```

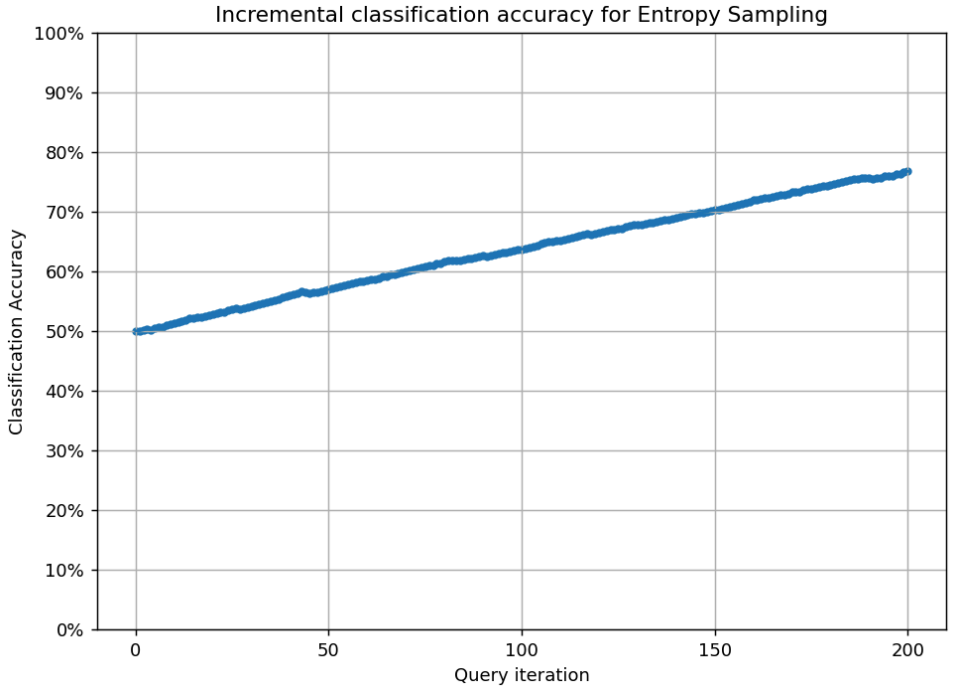


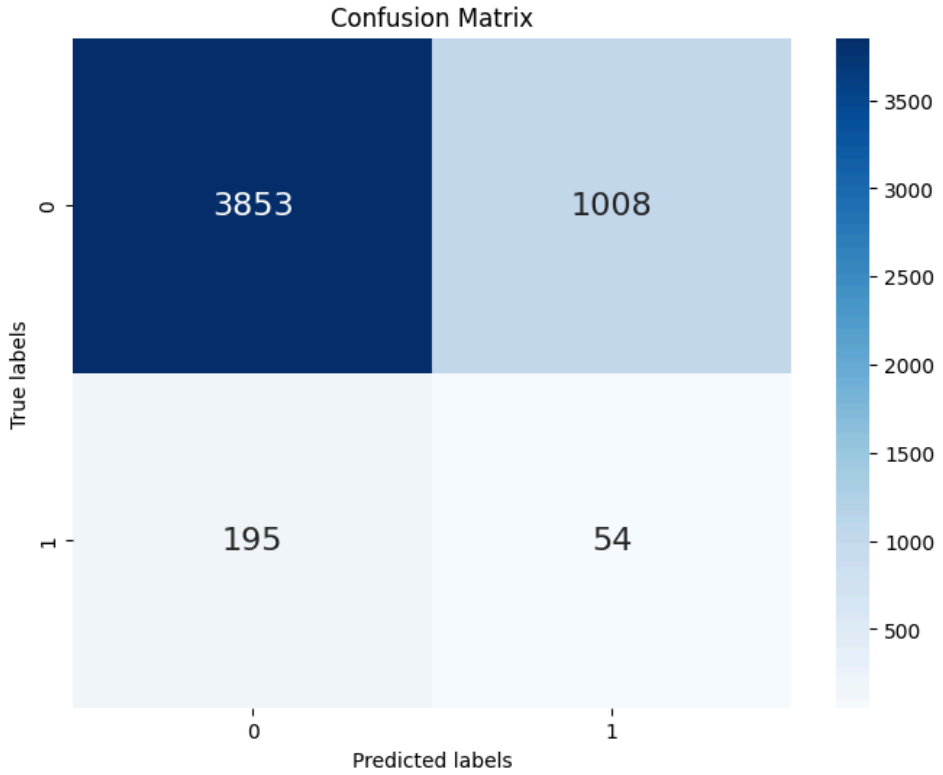
Stroke (imbalanced Dataset) Analysis:

Using Entropy sampling:

```
Query 1: The accuracy score is 0.5001956947162427.
Query 2: The accuracy score is 0.5015655577299413.
Query 3: The accuracy score is 0.5031311154598825.
Query 4: The accuracy score is 0.5021526418786693.
Query 5: The accuracy score is 0.5039138943248532.
Query 6: The accuracy score is 0.5056751467710372.
Query 7: The accuracy score is 0.5070450097847358.
Query 8: The accuracy score is 0.5099804305283757.
Query 9: The accuracy score is 0.5117416829745597.
```

```
Query 188: The accuracy score is 0.7555772994129158.
Query 189: The accuracy score is 0.7561643835616438.
Query 190: The accuracy score is 0.7569471624266145.
Query 191: The accuracy score is 0.7549902152641879.
Query 192: The accuracy score is 0.7553816046966731.
Query 193: The accuracy score is 0.7565557729941291.
Query 194: The accuracy score is 0.7587084148727984.
Query 195: The accuracy score is 0.7596868884540118.
Query 196: The accuracy score is 0.7600782778864971.
Query 197: The accuracy score is 0.762426614481409.
Query 198: The accuracy score is 0.7634050880626223.
Query 199: The accuracy score is 0.7655577299412916.
Query 200: The accuracy score is 0.7671232876712328.
```

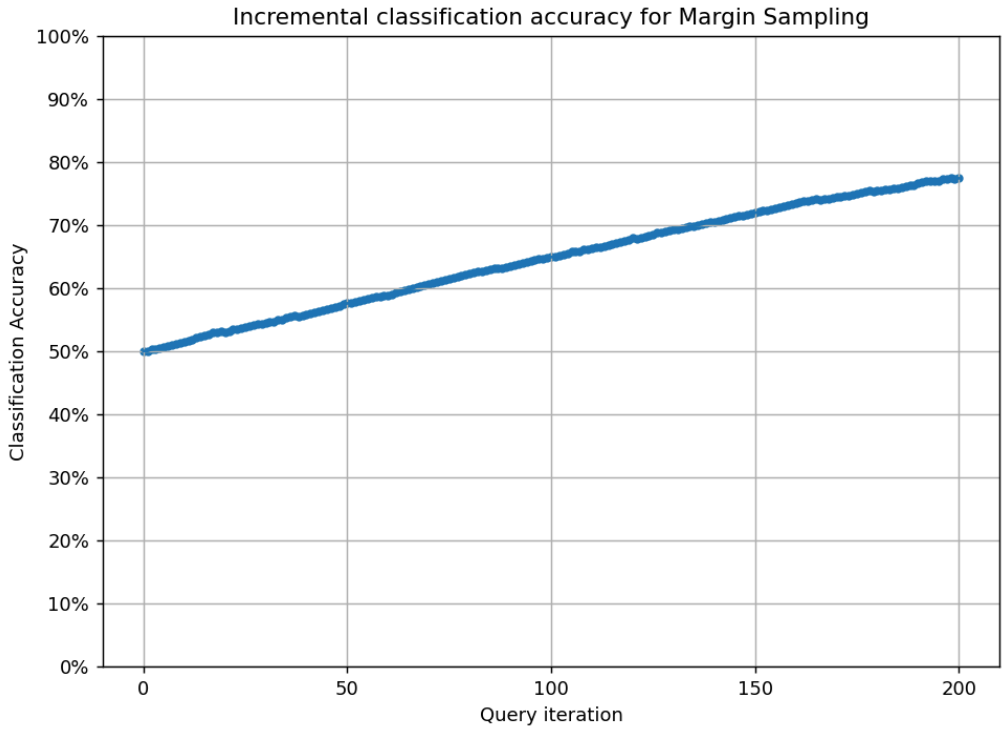


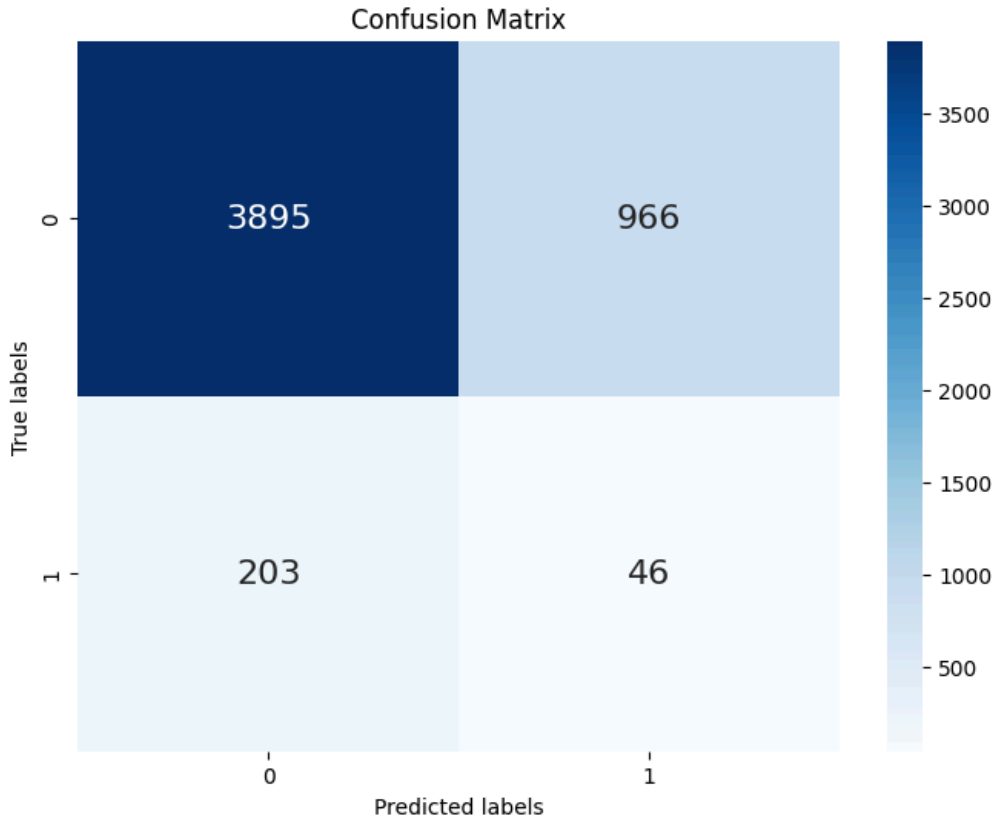


Using Margin sampling:

```
Query 1: The accuracy score is 0.5001956947162427.
Query 2: The accuracy score is 0.5027397260273972.
Query 3: The accuracy score is 0.5037181996086105.
Query 4: The accuracy score is 0.5054794520547945.
Query 5: The accuracy score is 0.5062622309197652.
Query 6: The accuracy score is 0.5084148727984344.
Query 7: The accuracy score is 0.5097847358121331.
Query 8: The accuracy score is 0.510958904109589.
Query 9: The accuracy score is 0.5135029354207437.
```

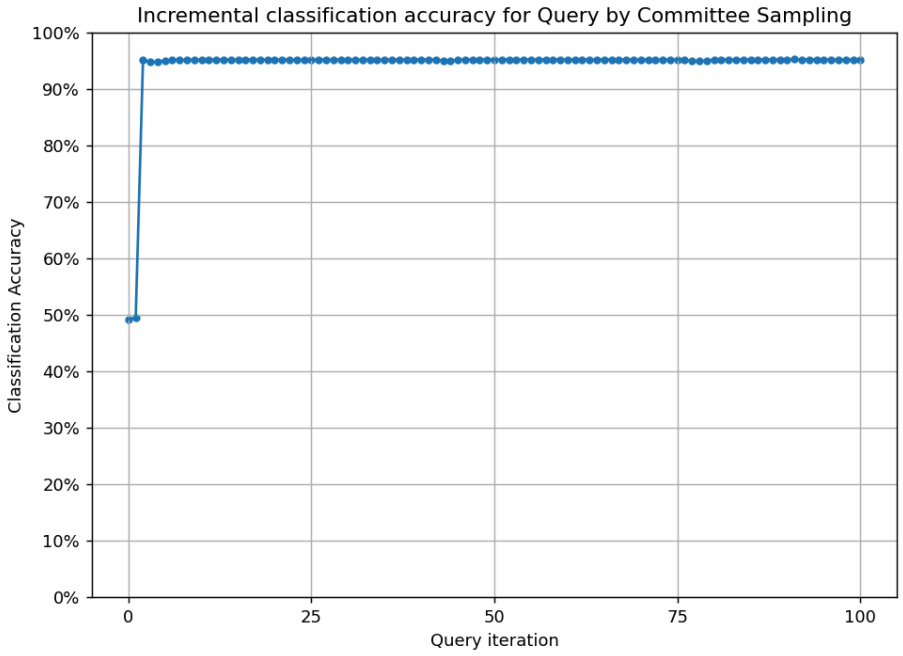
```
Query 187: The accuracy score is 0.7610567514677103.
Query 188: The accuracy score is 0.762426614481409.
Query 189: The accuracy score is 0.763600782778865.
Query 190: The accuracy score is 0.7657534246575343.
Query 191: The accuracy score is 0.7681017612524462.
Query 192: The accuracy score is 0.7694716242661448.
Query 193: The accuracy score is 0.7686888454011742.
Query 194: The accuracy score is 0.7690802348336595.
Query 195: The accuracy score is 0.7702544031311155.
Query 196: The accuracy score is 0.772211350293542.
Query 197: The accuracy score is 0.772211350293542.
Query 198: The accuracy score is 0.775146771037182.
Query 199: The accuracy score is 0.7731898238747554.
Query 200: The accuracy score is 0.7737769080234833.
```

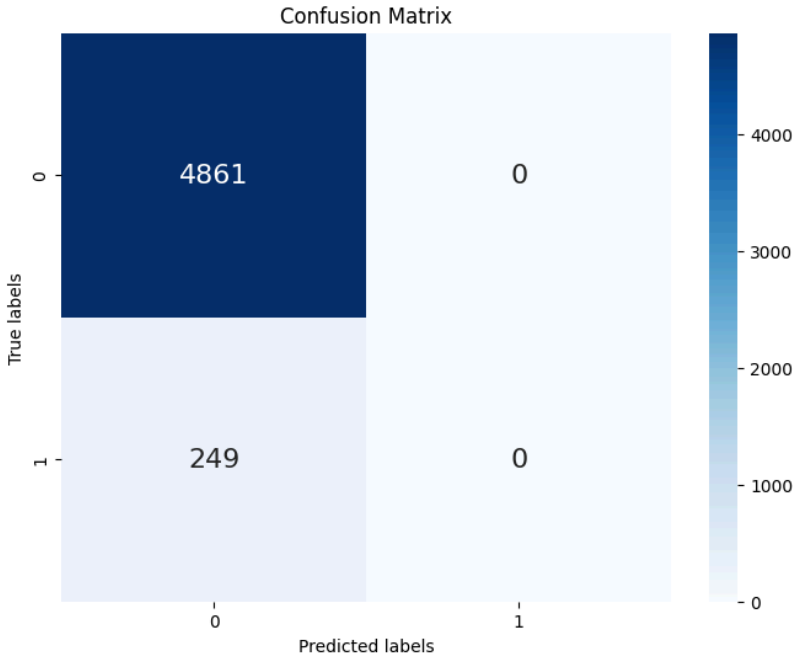




Using Query by Committee:

```
Query 1: The accuracy score is 0.49471624266144815.
Query 2: The accuracy score is 0.9512720156555773.
Query 3: The accuracy score is 0.9473581213307241.
Query 4: The accuracy score is 0.9473581213307241.
Query 5: The accuracy score is 0.9495107632093933.
Query 6: The accuracy score is 0.9512720156555773.
Query 7: The accuracy score is 0.9504892367906067.
Query 91: The accuracy score is 0.9518590998043053.
Query 92: The accuracy score is 0.9516634050880626.
Query 93: The accuracy score is 0.9516634050880626.
Query 94: The accuracy score is 0.95146771037182.
Query 95: The accuracy score is 0.95146771037182.
Query 96: The accuracy score is 0.95146771037182.
Query 97: The accuracy score is 0.9512720156555773.
Query 98: The accuracy score is 0.9512720156555773.
Query 99: The accuracy score is 0.95146771037182.
Query 100: The accuracy score is 0.9512720156555773.
```





In conclusion:

The findings suggest that within the uncertainty family of strategies, there isn't a significant difference. It's difficult to conclusively say that one strategy is consistently better than the others. However, active learning approaches, in general, prove superior in terms of maintaining high scores in evaluation metrics, especially accuracy. Additionally, they excel in reducing computation costs, annotation efforts, and budget requirements. This is because we can operate numerous models with diverse environments simultaneously.

References:

- https://scikit-activeml.github.io/scikit-activeml-docs/
- https://medium.com/@hardik.dave/active-learning-sampling-strategies-f8d8ac7037c8
- https://younsess-elbrag.medium.com/active-learning-approaches-strategies-deep-learning-integration-and-essential-tools-6ff2bdfe5cb