ACTIVE LEARNING PROJECT

Presented by: Our Team

Overview

- *Problem Definition*
- *Objectives*
- *Datasets*
- *Query Strategies*
- *Classifiers*
- *Approaches*
- *Results*
- *Conclusion*
- *Team Members*

Problem Definition

- Machine learning algorithms have shown tremendous success in various domains. However, one common challenge faced by traditional machine learning approaches is the need for large labeled datasets to achieve high performance. Acquiring labeled data can be costly, time-consuming, and sometimes impractical.

- Active Machine Learning (AML) offers a promising solution to address this limitation by actively selecting informative instances for labeling.

Our objectives

In this project, we aim to explore the effectiveness of active machine learning techniques in comparison to traditional approaches on popular benchmark datasets such as Iris, Wine, and Breast Cancer. We will employ a diverse set of classifiers and query strategies to conduct a comprehensive evaluation of the performance before and after the integration of active learning.

Datasets

Breast Cancer

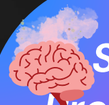
The Breast Cancer dataset is a well-known dataset used for classification tasks in machine learning. It contains features computed from breast mass images, which are used to predict whether a tumor is malignant (cancerous) or benign (non-cancerous)
Number of Instances: 569

IRIS

The Iris dataset is a classic dataset in machine learning and statistics, often used for classification and clustering tasks. It is a small dataset that contains measurements of iris flowers from three different species: Setosa, Versicolor, and Virginica
Number of Instances: 150 (50 instances per class)

WINE

The Wine dataset is another popular dataset commonly used for classification tasks in machine learning. It contains the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars (classes)
Number of Instances: 178



Stroke Prediction

The Stroke Prediction Dataset is a collection of health-related data used for predicting the likelihood of a stroke in individuals based on various attributes It's a valuable resource for exploring various classification algorithms, handling imbalanced datasets, and understanding the factors that contribute to stroke risk.
Number of Instances: 5110

Query Strategies

Random sampling

Random sampling is the simplest and most straightforward query strategy in active learning. It involves randomly selecting instances from the unlabeled dataset for labeling.

Entropy

It measures the entropy of the predicted class probabilities for each instance. Higher entropy indicates higher uncertainty, suggesting that the instance is more informative for labeling.

Margin Sampling

It focuses on instances that lie close to the decision boundary, where the classifier is relatively uncertain. The decision boundary is typically defined by the margin between the probabilities assigned to the top two predicted classes.

Query-By-Committee

Is a query strategy that involves maintaining a committee of diverse classifiers and selecting instances that elicit disagreement among them.

Classifiers

SVM

SVM aims to find an optimal hyperplane that separates the instances of different classes with the maximum margin. It is effective in handling high-dimensional data and can handle both linearly separable and non-linearly separable datasets using kernel functions.

Label Propagation

It leverages the connectivity of instances in a dataset to propagate labels. It assumes that neighboring instances are likely to have similar labels. The algorithm starts with a small set of labeled instances and iteratively propagates the labels to their neighboring unlabeled instances.

Parzen Window

It estimates the probability density function of the underlying data distribution. It assigns a probability to each instance based on the density of the instances in its vicinity. Parzen Window is particularly suitable for modeling complex distributions and can handle datasets with varying densities and overlapping classes.

Mixture Model

Is a probabilistic model that represents the underlying data distribution as a mixture of component densities. It assumes that the dataset is generated from multiple subpopulations, each associated with a specific probability distribution

Approaches

Approach 01

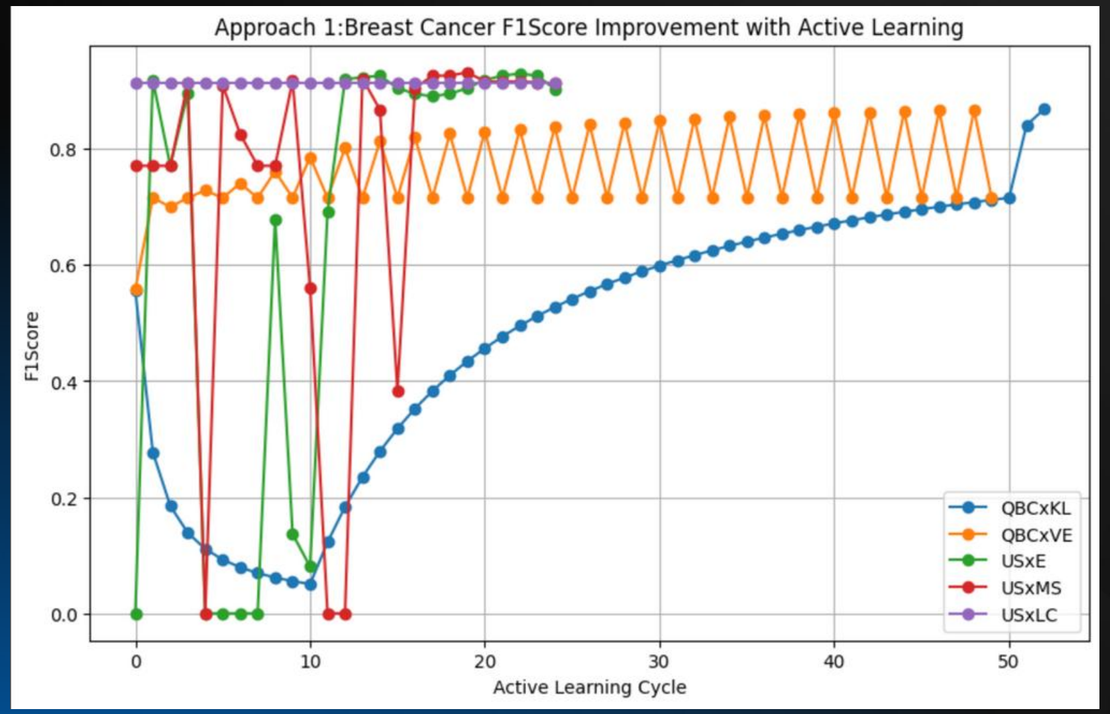
Our first approach to active learning involves training the model exclusively on unlabeled data initially. As we iterate through the process, labeled data is incrementally incorporated based on their true labels. The strategy for selecting which samples to label relies on the methodology employed by the chosen strategy.

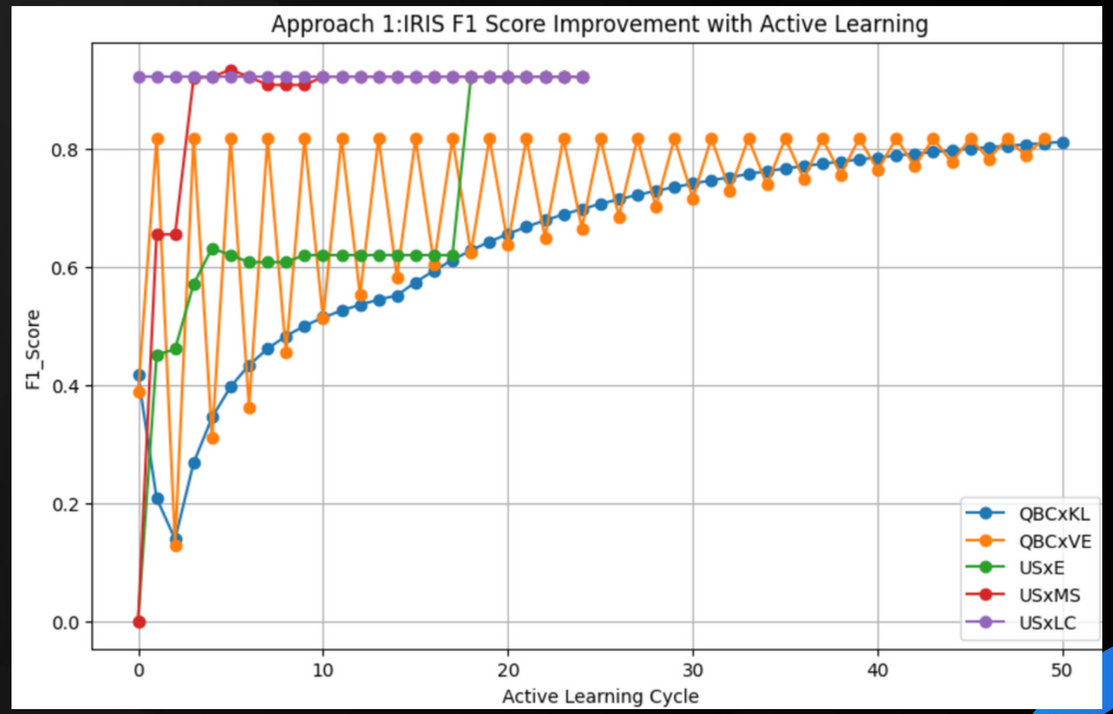
Approach 02

In our second approach to active learning, the model is initially trained on a dataset where a portion of the data is labeled, and the remaining part is left unlabeled. Through iterations, the number of labeled instances increases while the number of unlabeled instances decreases. The selection of samples to be labeled is determined by the chosen strategy, which employs a specific method to make these selections based on its working mechanism.

Results Of The First Approach:

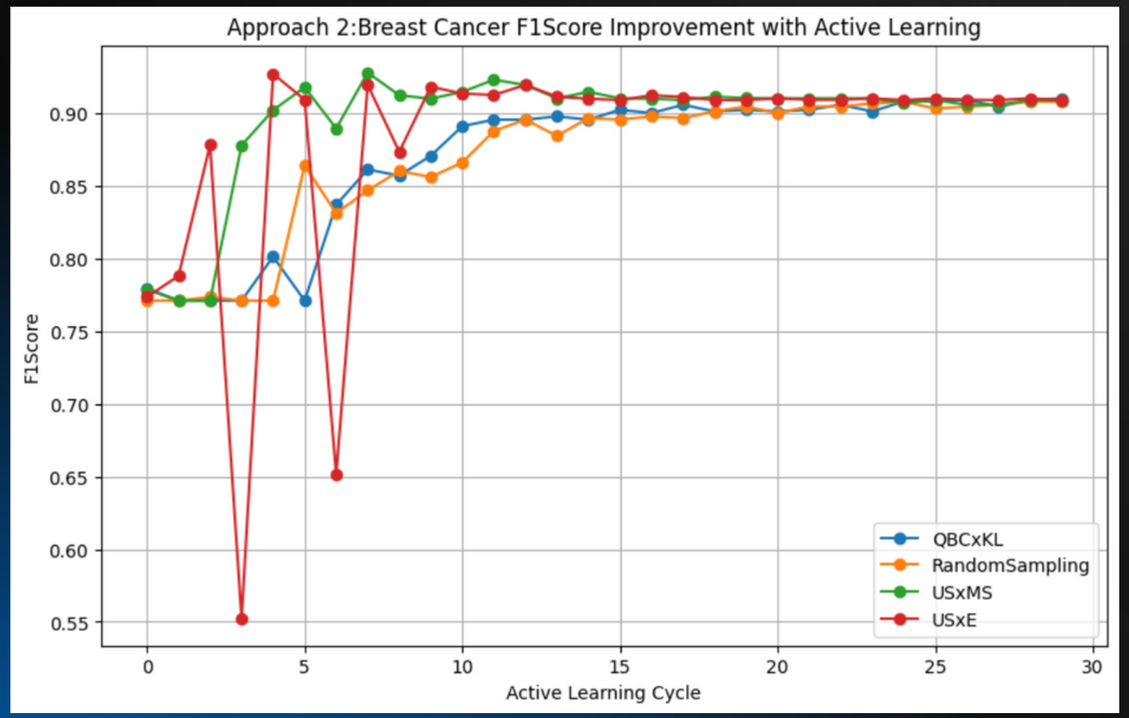
For this approach, we worked on the IRIS and Breast-Cancer datasets

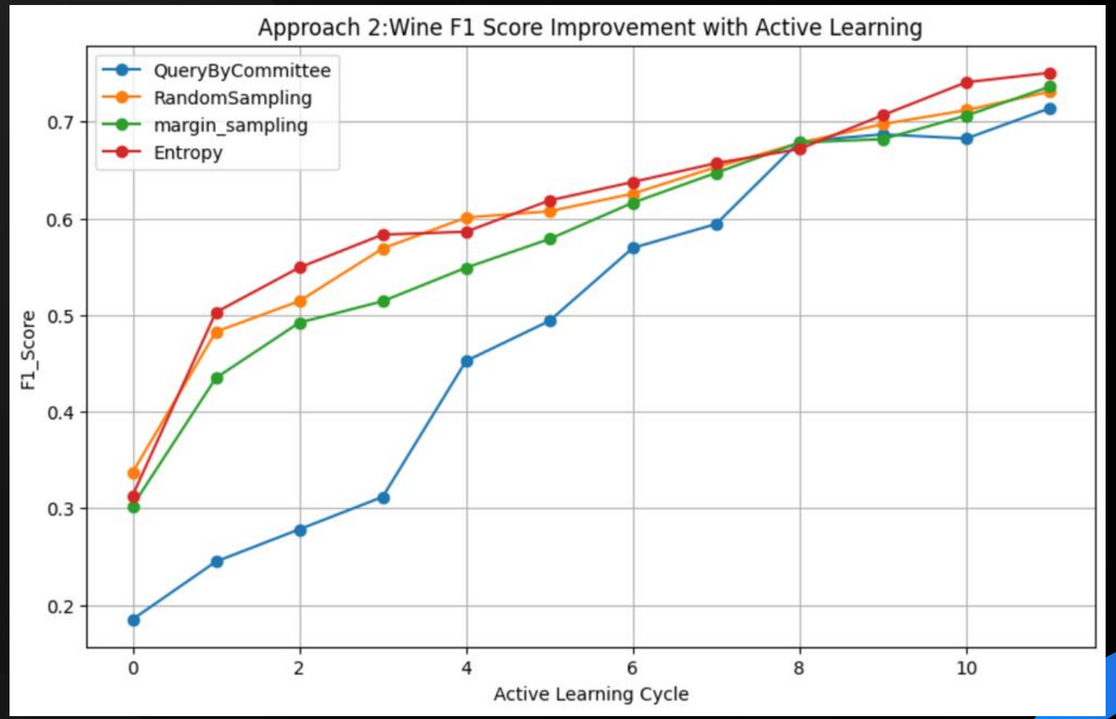




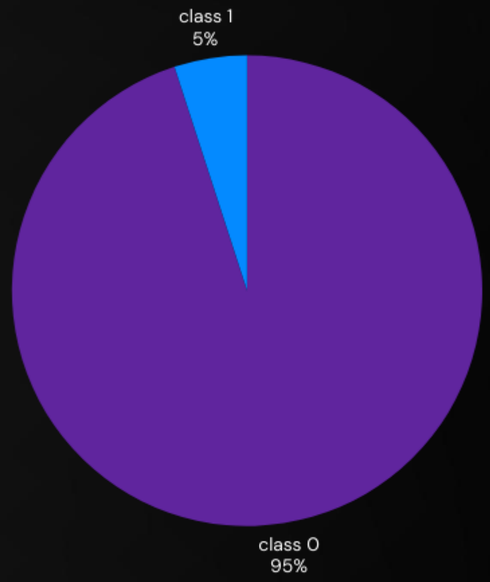
Results Of The Second Approach:

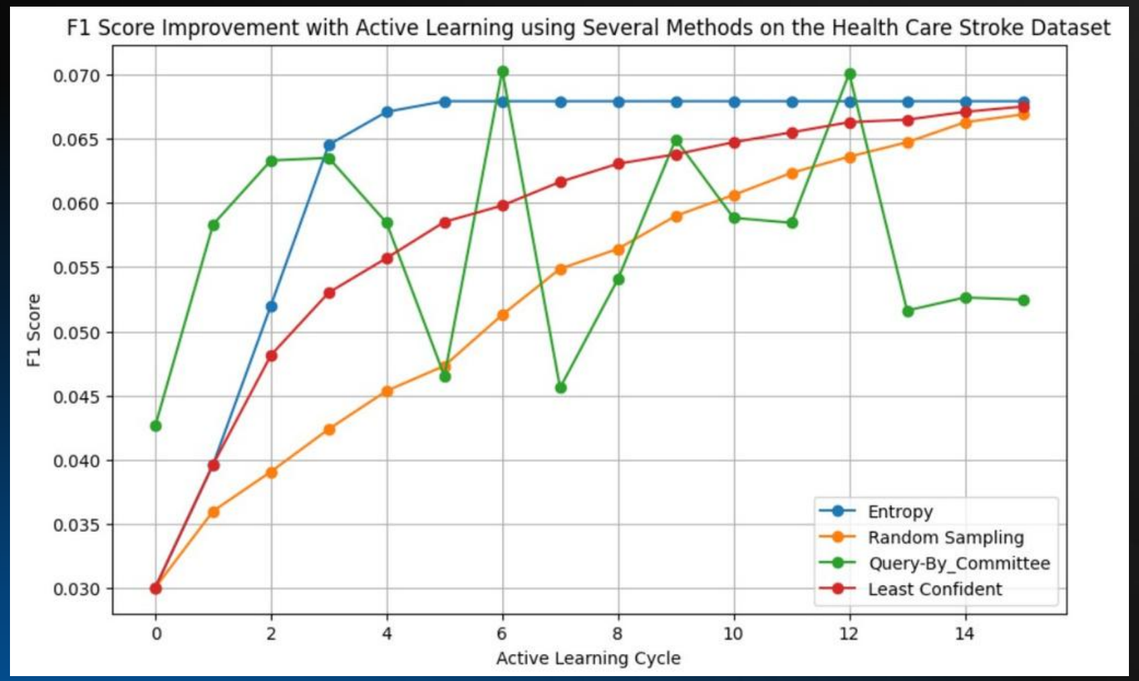
For this approach, we worked on the WINE and Breast-Cancer datasets

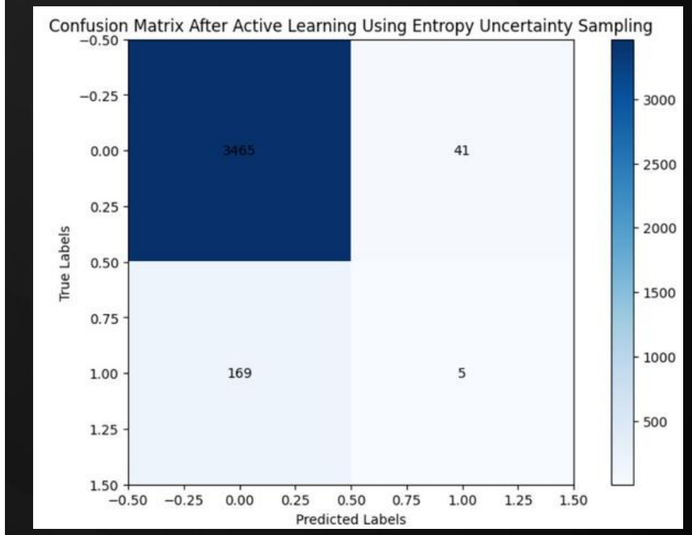




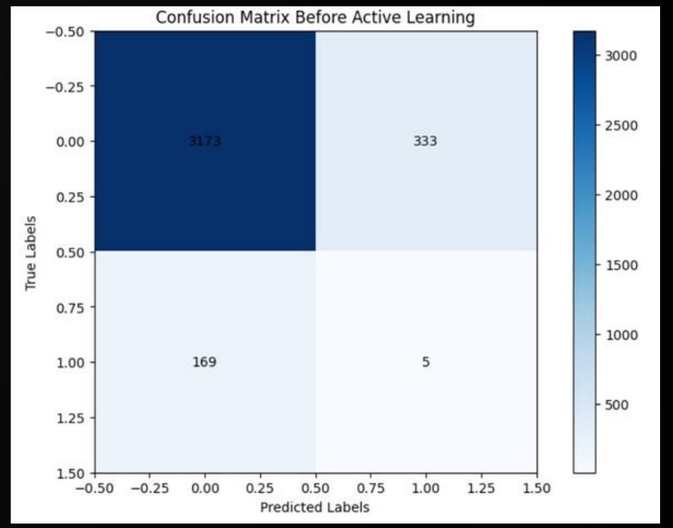
Results Of Imbalanced Dataset:







After Active ML



Before Active ML

Conclusion

Frist Approach

- After careful evaluation, we determined that the IRIS and Breast Cancer datasets exhibited the most significant improvement in classifier accuracy after applying the active learning cycle, making them the top choices for our active learning model. For the IRIS dataset, classifiers like Label Propagation, SVC, or Parzen Window showed promising results, but we favored Label Propagation due to its consistent performance across various strategies. However, in the case of the Breast Cancer dataset, the SVC classifier paired with the Entropy method in the Uncertainty Sampling Strategy emerged as the clear winner, followed closely by the SVC classifier with the Marginal Sampling method in the same strategy.

Second Approach

- After evaluating the performance of different datasets along with various strategies and classifiers, we found that the Wine and Breast Cancer datasets consistently yielded the best performance across different combinations of parameters and techniques For the Breast Cancer dataset, the SVC classifier with the KL divergence strategy emerged as the winning pair, despite tough competition from other strategies. Similarly, for the Wine dataset, the Parzen Window classifier with the Entropy strategy stood out as the winning pair due to its stable and robust performance.

Our Team

Arwa Sallam
20200067

Mahmoud Wael
20200505

Natalie Monged
20200586

Aly Walid
20200336

THANK YOU

For watching this presentation

arwasallam6@gmail.com

NatalieMonged20@gmail.com

aly.walidaly@yahoo.com

mahmoudwael677@gmail.com