2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) | 978-1-6654-2709-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICBASE53849.2021.00010

# Research on predicting heart attack through active learning

Peng Yuan
Computer Science University of Delaware
Newark, Delaware, 19711, USA
ricardoo@udel.edu

***Abstract*—Machine learning has been widely used in medical diagnosis areas, such as detecting disease at its earliest stages, helping doctors diagnose patients more accurately, and providing advice for patients. Active learning is a new machine learning method. Its main goal is to find high-information samples from many data sets to reduce the number of training sample sets, reduce computational complexity, and improve the generalization ability of the classifier. This paper describes how to create a model derived from machine learning techniques and use a pool-based active learning algorithm to analyze and predict whether a patient is at risk of heart disease. The author uses principal component analysis (PCA) to create a framework to extract features and calculate a mathematical model to select relevant features using relevant constraints. This experiment combines two datasets that are analyzed commonly in one dataset. The classification accuracy of this dataset was obtained by k-fold cross-validation. The experiment compared three modeling options, which are Random forest (RF), K Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). The best model tested in the experiment is RF, which achieved an average accuracy of 88%.**

***Keywords: Machine learning, active learning, heart disease prediction, datasets***

## I. Introduction

Cardiovascular diseases are one of the biggest reasons for the death of millions of people worldwide, only second to cancer. A heart attack occurs when a blood clot blocks the blood flow to a part of the heart [9]. I have seen someone have a heart attack in public who suffered much pain. Therefore, it is essential to predict the possibility of a human who will have a heart attack. There are plenty of factors that induce heart diseases, such as alcohol abuse, smoking, high work pressure, improper diet, and genetic factors [11]. These are some significant factors that the public recognizes, but the factors of heart disease are complicated. Education and counseling of older patients at high risk for heart disease is complex but should emphasize atypical symptoms and treatment options [12]. The patient knows some factors are significant, but they usually cannot be sure the real factor causes their heart attack. Patients also compare themselves to other patients in the ward concerning symptoms, age, number of attacks, and perceived progress [4]. A few papers are discussing the Prediction of heart attacks using machine learning. No paper talks about the Prediction of heart attack and active learning at the same time, so this work is about to use datasets of heart attack patients conditions to train an algorithm, and this research hope there is a way to use active learning based on the datasets we already obtain to predicate whether a patient has the probability of getting a heart attack.

## II. Literature review

Machine learning techniques such as support vector machines, decision trees, and logistic regression have been applied to autism-related datasets to construct predictive models. These models claim to improve the clinical ability to provide a firm diagnosis and prognosis for other diseases, for example, Autistic Spectrum Disorder (ASD) [10].

It is generally believed that the more labeled data, the more accurate the labeling, and the more efficient the model trained based on these data. However, while big data provides opportunities, it also brings serious challenges, the most typical of which is low data quality. Therefore, learning an influential model from a small amount of labeled big data is challenging and important [1].

Kavitha and Kannan [5] created a heart disease classification framework that includes PCA for feature extraction. The author points out that reducing the data dimension is to improve the accuracy of the classifier's Prediction and reduce the calculation cost of the Prediction. This can be achieved through feature extraction methods, which create new features derived somehow from the original features, or through feature, selection to obtain a subset of the most relevant features from the data set.

## III. Methodology

The methodology used by this research paper is experimental. An active learning algorithm is the most critical step in the process. The two datasets which contain heart attack data are integrated into a dataset as a training set for the experiment. Before training data, it is essential to apply the Receiver Operating Characteristic curve (ROC) to evaluate whether the classifier is appropriate to our dataset. Then three different modeling classification methods, which are Random forest [6], K Nearest Neighbors [2], and Multilayer Perceptron, will be implemented to the training dataset. It is essential to see how data is distributed; this experiment introduced Principal Component Analysis (PCA) to map multi-dimensional data into a two-dimensional plate. To analyze how active learning algorithms preforms [8]. A random sampling learning

978-1-6654-2709-8/21/$31.00 ©2021 IEEE
DOI 10.1109/ICBASE53849.2021.00010

algorithm is brought out to compare its learning curve with the pool-based active learning algorithm learning curve from 1 to 200 queries.

## IV. Experiment

### A. Dataset

The dataset this experiment used is combined with two datasets [19] [20] that found on the Kaggle. The reason for combining the two datasets is to reserve a larger number of instances and derive a more reliable predictive model through active learning techniques.

The new combine dataset contains 13 features with a predicted result and 598 instances. In particular, the features slp, caa, and thall had quite a few missing values. The author decided to cancel the analysis and modeling of these three features in the experiment.

TABLE I. Experimental dataset

| ***Feature*** | ***Value*** |
|---|---|
| age | 28–77 |
| sex | 0 = female, 1 = male |
| cp (chest pain type:) | Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic |
| trtbps (resting blood pressure) | 92 - 200 |
| chol (serum cholestoral) | in mg/dl |
| fbs (fasting blood sugar) | fasting blood sugar > 120 mg/dl, 1 = true; 0 = false |
| restecg (resting electrocardiographic results) | 0 = normal, 1 = having, ST-T wave abnormality, 2 = showing probable left ventricular hypertrophy |
| thalachh (maximum heart rate achieved) | 71 – 202 |
| exng (exercise induced angina) | 1 = yes; 0 = no |
| oldpeak (ST depression induced by exercise relative to rest) | 0- 6.2 |
| slp (the slope of the peak exercise ST segment) | 1 = up sloping, 2 = flat, 3 = down sloping |
| caa (number of major vessels (0-3) colored by fluoroscopy) | 0-3 |
| thall (Heart rate) | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| output (the prediction value) | 0 = heart disease not present 1 = heart disease present |

There are 14 columns to be determined. Many features of medical datasets are irrelevant or interrelated. In terms of classification, this may lead to overfitting of the predictive model. Many of these features require in-depth patient testing in the real world, and these features may not be available when people need to predict heart disease [5]. These assumptions are evaluated after. For the previous cross-validation graph, this experiment uses all features provided by the dataset exclude ca and thal, missing values.

By input this information, we get the output num which is the predicted attribute, diagnosis of heart disease (angiographic disease status), Value 0: < 50% diameter narrowing, Value 1: > 50% diameter narrowing.

### B. Method: Receiver Operating Characteristic curve

ROC curve is a graph showing the effect of the classification model under all classification thresholds [16]. In this experiment, we use k-folder cross-validation to implement the roc curve corresponding to the experimental data set. In this graph, True Positive Rate (TPR) corresponds to the y-axis, and False Positive Rate (FPR) corresponds to the x-axis. We will analyze the mean of Area under the Curve (AUC) to evaluate the fitness of classifiers.

Method: Principal Component Analysis (PCA)

$$\mathrm{var}(a) = \frac{1}{m}\sum_{i=1}^{m} a_i^2 \qquad (1)$$

The purpose of introducing PCA is to visualize the distribution of data in a two-dimensional space.

### C. Method: Pool-based active learning [17]

Based on the active learning of the data pool, assuming that all unlabeled data has been given, a data pool is formed. The active learning algorithm is iterative. Each time a sample is selected from the unlabeled data pool and the expert queries the label, these newly labeled samples are added to the training set. The model is updated based on the new training set and then enters the next iteration until the iterations are carried out [18].

The termination criterion for this active learning is the number of iterations; after 200 queries, the iteration will terminate automatically.

### D. Results

After the PCA distribution, we can see Figure 1.

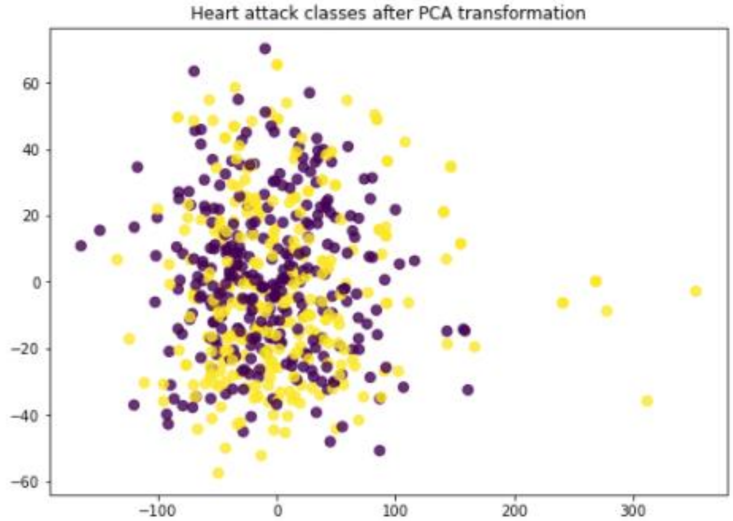

Figure 1. Heart attack classes after PCA transformation

In Figures 2, 3, 4, the ROC curves of all three classifiers are in good shape. The mean AUC of KNN's ROC is 0.86; the mean AUC of MLP's ROC is 0.89; the mean AUC of RF's ROC is 0.94. Random forest gives the best results.

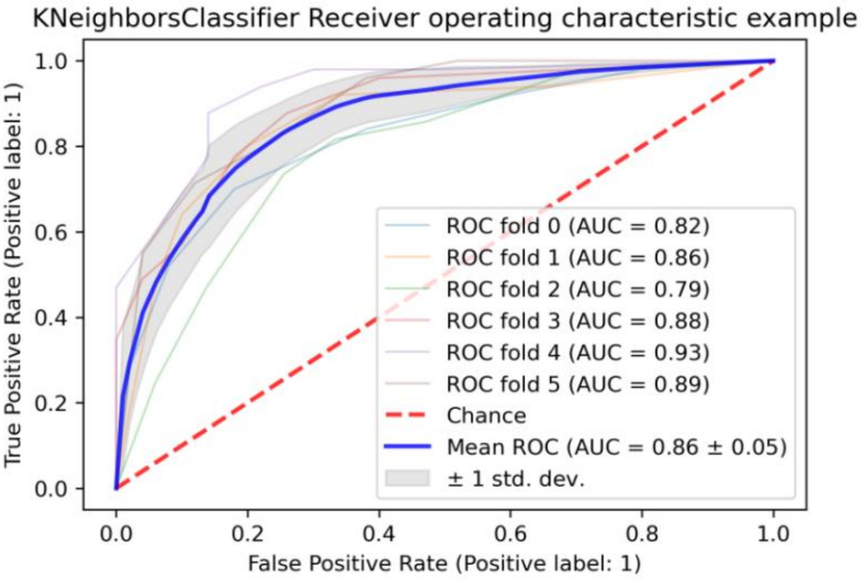

Figure 2. ROC Curve of KNeighbors Classifier

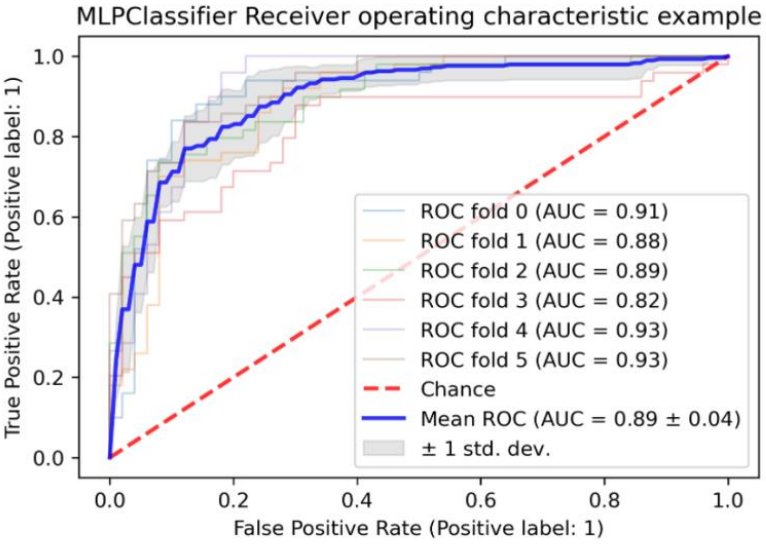

Figure 3. ROC Curve of MLP Classifier

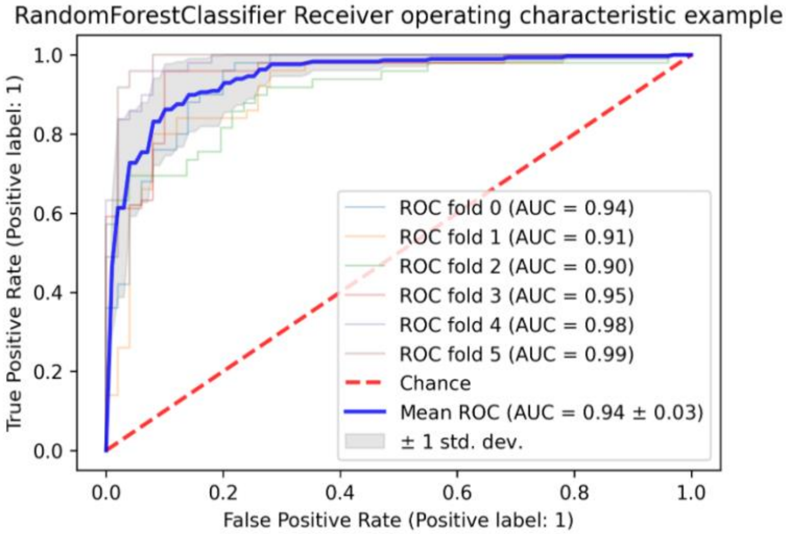

Figure 4. ROC Curve of Random Forest Classifier

Different training models generated different results, K Nearest Neighbors (KNN) modeling with pool-based active learning has an increasing learning curve, and it improves classification accuracy from 0.551 to 0.652 (Figure 5). Because there is no apparent boundary between the positive and negative distribution of heart disease. In the PCA graph (Figure 1), the classification effect of KNN is not very obvious on this dataset. Positive and negative dots spread widely across the entire two-dimension plate. After the pool-based active learning process, the classification accuracy is improved by more than 10 percent. The author compare the score with random sample learning. However, the improvement of the active learning process give is less than using random forest classification. When the value of k gradually increases from 2 to 11, the curve trend does not change.

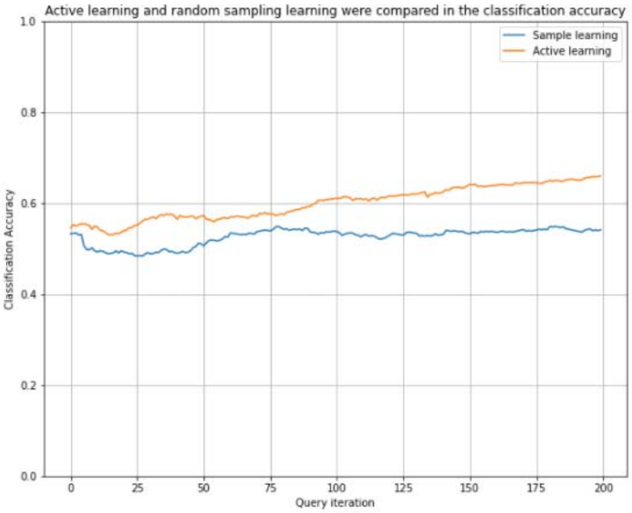

Figure 5. Accuracy Comparison of Active Learning and Sampling Learning (KNN)

Multilayer Perceptron (MLP) learning curve (Figure 6) is a floating curve, and its classification accuracy has only a little increase after 200 iterations. The active learning curve does not have a significant difference from the sample learning curve in this scenario. The image fluctuates greatly, but the overall classification is not significantly improved compared to random classification. Since the hidden layer size is set to 100, making the prediction process more complicated, active learning does not improve the dataset very well.

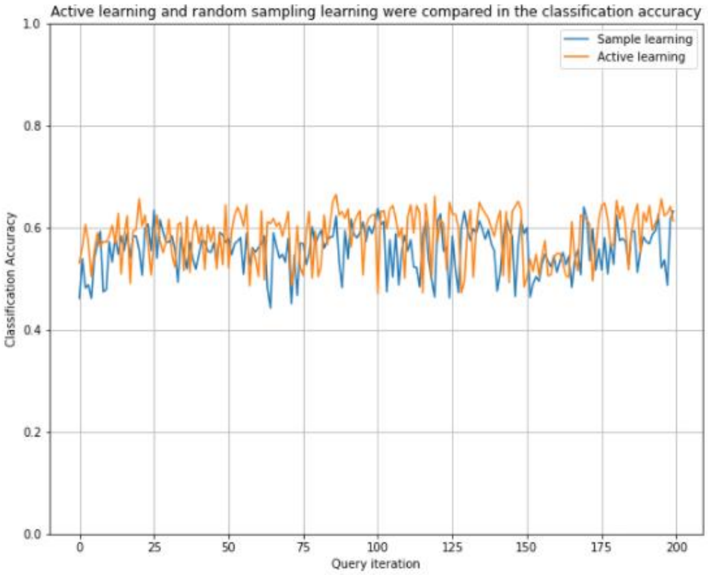

Figure 6. Accuracy Comparison of Active Learning and Sampling Learning (MLP)

The learning curve of Random Forest (RF) is in good shape (Figure 7). Its learning curve keeps increasing, and classification accuracy increase from 0.60 to 0.873. RF builds multiple decision trees and merges them to get a more accurate and stable model. When a sample needs to be predicted, count the prediction results of each tree in the forest for the sample, and then select the final result from these prediction results by voting. Since ten features are selected in total, and the correlation between any two trees in the forest is not high, the classification ability of the entire forest is strong. The improvement of classification accuracy after active learning is the best among the three models.

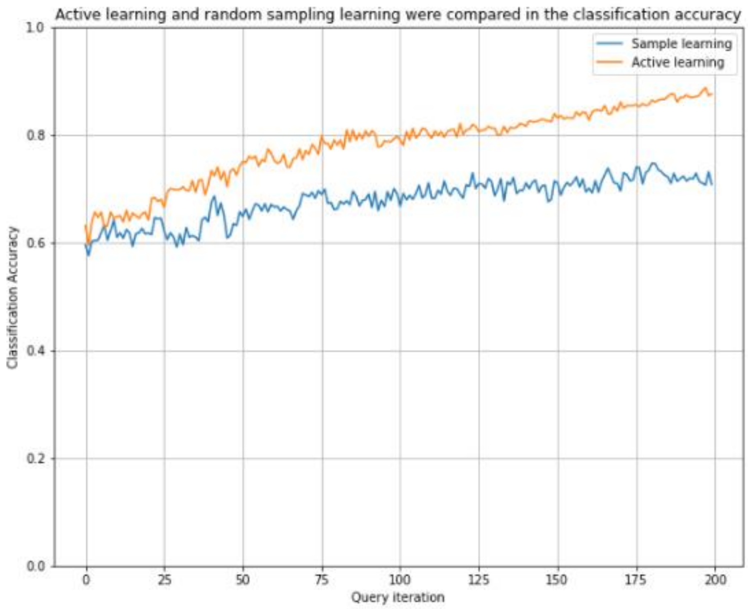

Figure 7. Accuracy Comparison of Active Learning and Sampling Learning (RF)

## V. Discussion

This paper is limited because we only used a pool-based active learning algorithm. It is valuable to use other better prediction models and active learning algorithms to train the data [14]. This work focuses on training the active learning process to learn the dataset features and make predictions as accurate as possible based on different classifiers. There is another guess of why pool-based active learning has a great improvement on random forest modeling. It might be a black box with strong inexplicability, multiple random leads to very good results. According to the three active learning results, we could infer that the pool-based active learning algorithm effectively predicts heart attack. Its improvement varies based on the choice of a classifier. In further research, another way to explore is to use more related features and make more complex models, for example, neural networks [13], to predict more accurate results. Furthermore, inventing a completely new active learning algorithm would make a huge contribution to heart disease prediction.

## VI. Conclusion

Data analysis technology, especially ML, plays an increasingly important role in the medical war with heart disease worldwide. This research investigates the accuracy of pool-based active learning algorithms under three different classification models produced by combined heart-disease datasets.

The results show that an active learning algorithm can train a model to increase its reliability in predicting heart disease. The improvement over using different classification models shows that appropriate classification is necessary as well [15]. This experiment, the most accurate model based on the dataset we used, found that random forest modeling achieved the best accuracy score, around 0.7. Furthermore, the pool-based active learning algorithm improves the accuracy from 0.60 to 0.873 under random forest modeling. The PCA figure shows the overall effect of changes in predicted results. It changes from the fix of correct and incorrect to most of correct. Random forest machine learning significantly improves the accuracy of cardiovascular risk prediction. It increases the number of identified patients who can benefit from prevention and treatment while avoiding unnecessary treatment by others [3].

## Acknowledgement

First and foremost, I appreciate Dr. Robert for helping explain and guide me to this research project, thanks for his teaching, concern and help. Then I would like to thank TA Jiatong who gave his students the greatest care and help, which I admire very much. I can reach Jiatong whenever I need and he always points me out the key of my problem. Finally, I would like to thank my love Rui Zhang, who have been encouraging me, supporting me and warming me with her love during my sleepless nights of writing and revising papers.

## Reference

[1] Exploring feature selection and classification methods for predicting heart disease, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7133070/

[2] Khateeb N and, Usman M. Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. In: Proc Int Conf Big Data Internet Things - BDIOT2017 2017

[3] Can machine-learning improve cardiovascular risk prediction using routine clinical data? Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N PLoS One. 2017; 12(4):e0174944.

[4] Bill Cowie, The cardiac patients perception of his heart attack, Social Science & Medicine (1967), Volume 10, Issue 2, 1976, Pages 87-96, ISSN 0037-7856, https://doi.org/10.1016/0037-7856(76)90059-7.

[5] Kavitha R and, Kannan E. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. In: 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS). 2016.

[6] Jabbar MA, Deekshatulu BL, Chandra P. Prediction of heart disease using random forest and feature subset selection. Advances Intelligent Syst Comp Innovations Bio-Inspired Comp App 2015; 187–196.

[7] Ziasabounchi N, Askerzade I. A comparative study of heart disease prediction based on principal component analysis and clustering methods. Turk J Mathematics Comp Sci 2014, 2: 39–50.

[8] Santhanam T, Ephzibah EP. Heart Disease Classification Using PCA and Feed Forward Neural Networks. Mining Intell Knowledge Exploration Lecture Notes Comp Sci 2013; 90–99.

[9] Akshit Bhardwaj, Ayush Kundra, Bhavya Gandhi, Sumit Kumar, Arvind Rehalia, Manoj Gupta, Prediction of Heart Attack Using Machine Learning, Department of Instrumentation & Control Engineering Bharati Vidyapeeth's College of Engineering Delhi-110063

[10] Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics Health Social Care 2019; 44(3): 278–297.

[11] World Health Organization. Cardiovascular diseases. Available from https://www.who.int/cardiovascular_diseases/en/ (n.d., accessed 9 June 2019)

[12] Tullmann, Dorothy France PhD, RN, CCRN; Dracup, Kathleen DNSc, RN, FAAN Knowledge of Heart Attack Symptoms in Older Men and Women at Risk for Acute Myocardial Infarction, Journal of Cardiopulmonary Rehabilitation: January-February, 2005 - Volume 25 - Issue 1 - p 33-39

[13] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS International Conference on Computer Systems and Applications, 31 March 2008, pp. 108--115.

[14] Gonsalves AH, Thabtah F, Mohammad RMA, et al. Prediction of coronary heart disease using machine learning: an experimental analysis. In: Proc 2019 3rd International Conf Deep Learning Technologies 2019; 51–56.

[15] Kahramanli, H. & Allahverdi, N., Design of a hybrid system for the diabetes and heart disease, Expert systems with applications,35,82-89, 2008.

[16] Zhu, W., Zeng, N. & Wang, N., Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations, 2010.

[17] Yang, Y. Y., Lee, S. C., Chung, Y. A., Wu, T. E., Chen, S. A., & Lin, H. T. (2017). libact: Pool-based active learning in python. arXiv preprint arXiv:1710.00379.

[18] Ganti, R. & Gray, A.. (2012). UPAL: Unbiased Pool Based Active Learning. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, in PMLR 22:422-431

[19] Heart Attack Prediction, https://www.kaggle.com/imnikhilanand/heart-attack-prediction

[20] Heart Attack Analysis & Prediction Dataset, https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset