The Joint Role of Batch Size and Query Strategy in Active Learning-based Prediction - A case study in the Heart Attack Domain

Bruno Faria, Dylan Perdigão, Joana Brás, and Luis Macedo

University of Coimbra, CISUC - Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, Coimbra, Portugal
{brunofaria, dgp, joanabras}@student.dei.uc.pt, macedo@dei.uc.pt

Abstract. This paper proposes an *Active Learning* algorithm that could detect heart attacks based on different body measures, which requires much less data than the passive learning counterpart while maintaining similar accuracy. To that end, different parameters were tested, namely the *batch size* and the *query strategy* used. The initial tests on batch size consisted of varying its value until 50. From these experiments, the conclusion was that the best results were obtained with lower values, which led to the second set of experiments, varying the batch size between 1 and 5 to understand in which value the accuracy was higher. Four query strategies were tested: *random sampling*, *least confident sampling*, *margin sampling* and *entropy sampling*. The results of each approach were similar, reducing by 57% to 60% the amount of data required to obtain the same results of the passive learning approach.

Keywords: Active Learning · Heart Attack

1 Introduction

Artificial Intelligence (AI) and, more specifically, Machine Learning (ML) have highly contributed to the improvement of prediction tasks in various domains, including critical life situations such as those involving the diagnosis of diseases [2,14,12]. Most of the solutions rely on Passive Learning algorithms. The machine is usually trained with large datasets. A function is learned and then used to predict outcomes in new situations.

A different approach, called Active Learning [8,9], can provide similar performance with fewer data, which is relevant for reducing costs in the acquisition of those data[1]. Active Learning operates with a partially or entirely unlabeled dataset, selecting for labeling only the most informative instances/examples from which to learn based on specific measures such as information gain. In order to label those instances, the algorithm asks for help from the experts, such as medical

[1] Active Learning is also used in the ML branch of Reinforcement Learning; in this paper, we are confined to the ML branch of Supervised and Semi-Supervised Learning

doctors (Figure 1). In the end, and in opposition to Passive Learning algorithms, the model is expected to be built with much fewer labeled instances while not losing accuracy in comparison to the Passive Learning counterpart [1].

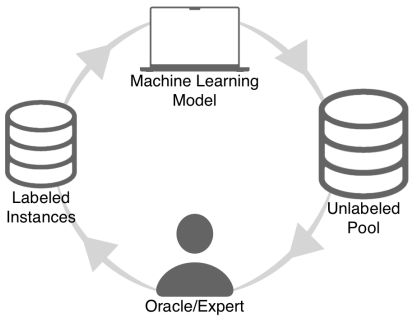


Fig. 1. Active learning cycle.

There are many different scenarios to label data via Active Learning. One of them may consider asking the oracle to label single instances separately or multiple instances simultaneously. For that, we need to define the *Batch Size* that indicates the number of examples that we feed the algorithm at a time. In this case, the dataset is split into multiple batch-sized sets. In a practical scenario, if the dataset were composed of 50 samples and the batch size was set to 5, the data would be split into ten queries, each containing five samples of the original dataset. On the other hand, the *Query Strategy* is the way the oracle chooses the potentially most informative instances of the pool of unlabeled data.

Different performances (efficiency and accuracy) may be achieved with the different query strategies considered for selecting the unlabeled instances to be labeled and with different sizes considered for those sets of unlabelled instances. In this paper, we study the role of this size and the query strategy in the performance of an Active Learning task of predicting heart attacks. The reason for considering our study in this domain is that cardiovascular diseases are the top cause of death globally, and as in other medical domains, acquiring data is a costly task.

The remainder of this paper is organized as follows. Section 2 provides a general review of the state of the art of Active Learning in different application domains and also of the application domain of myocardial infarction. Section 3 explains the tools, dataset, and frameworks used in the experimental tests. Sections 4 and 5 explain the methodologies and the experimental design to achieve the goals proposed previously. Section 6 presents the results obtained which are then discussed in section 7. Finally, section 8 concludes the paper.

2 Related Work

Regarding Active Learning, some progress was made, and researchers found the advantages of using this technique instead of Passive Learning. For example, in [5], a semi-supervised Active Learning approach was used to identify different sounds. After experimenting with both Passive and Active Learning, the best results were obtained with the latter, needing fewer data to achieve a more accurate algorithm.

According to the World Health Organisation [13], 7.9 million people die each year from cardiovascular diseases, an estimated 32% of all deaths worldwide, being 85% of those deaths due to heart attacks and strokes. Heart attack, also known as myocardial infarction, is a disease caused by the interruption of blood circulation in the heart with damage to the heart's muscle. Depending on morphological factors, some symptoms could indicate a risk of myocardial infarction. Considering the statistics previously mentioned, the necessity to quickly and correctly identify if a patient has a heart attack upon arriving at the ICU grows. AI and, more precisely, machine learning can provide a valuable contribution to fast and automatically detecting heart attacks with success. The domain of heart attack diagnostics has been an object of AI. For example, one study done by Chowdhury [3] uses linear classification models to diagnose myocardial infarction in order to prevent road accidents by analyzing electrocardiograms (ECG) waves.

Srinivas's work [11] has some different approaches, beginning with *IF-THEN* branches, turning to more basic machine learning methodologies like ID3 Decisions Trees, Neural Networks, Stochastic Back Propagation Algorithms and Bayesian Networks.

Finding research correlating Active Learning algorithms with heart diseases proved a challenge. However, it is possible to find work with other diseases in which Active Learning was tested. For example, in 2013, Mahapatra [7] studied Crohn's disease detection with Semi-Supervised Learning and Active Learning from magnetic resonances. Semi-Supervised Learning is used to extract from the images' features, and Active Learning classifies each region as "diseased" or "normal".

On that same note, another study was conducted [10] proposing an algorithm that could categorize images in three major groups (images referring to eye fundus, breast and skin cancer). The algorithm also split into two or four sub-categories (the output was binary on images referring to retinal lesions and skin cancer – having a lesion/cancer or not having it – and four categories to distinguish different types of breast cancer – normal, benign, *in situ* and invasive). The researchers concluded that using Active Learning (using a segment of the dataset), they obtained the same accuracy as passive learning (using the whole dataset) with 32 to 40% reduced data. This shows the most significant advantage of using Active Learning rather than the traditional method, which is Passive Learning.

3 Materials

The dataset we used [6] results from merging four databases of clinics and hospitals located in Hungary, Switzerland, and the United States of America. It is composed of 303 samples and 76 attributes. Only 14 of these 76 attributes were used by previous works that relies on the dataset. Those attributes are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and whether the person has thalassemia or not. Previous works have shown that it is possible to predict whether there is a high or low probability of having a heart attack with this relatively small amount of features (more details in Table 1).

ModAL [4] was the selected Active Learning framework. The *ModAL* API allows the user to automate the selection of the best examples in the dataset to constitute a batch. In addition, it is possible to use the different query strategies that will be discussed in section 4.

4 Methods

Active Learning algorithms have various parameters that influence the results according to the dataset. In this paper, the primary focus was on the *query strategy* parameter. We have considered four different query strategies to aid on obtaining the most accurate results with less data: *least confident sampling*, *margin sampling*, *entropy sampling*, and *random sampling*.

Both *least confident sampling* and *margin sampling* take into account the probability of the certainty of the output of a given sample. However, they differ in which query is selected next. Margin sampling subtracts the two highest probabilities and selects the query whose result is the lowest, whereas least confident sampling only considers the highest probability. Another query strategy used was *entropy sampling*. The entropy formula is used on each sample, and the sample with the highest value will be selected as the following query. The last approach tested was *random sampling*. This method implies that the following query is selected at random. The code implemented was adapted from [4] and is expressed in Algorithm 1.

Algorithm 1 Random Sampling

Require: The $classifier$ and the *pool* of examples
Ensure: The index i and the i-th element of the *pool*
1: Let n = length of the *pool*
2: Let i = random integer $\in [0,n[$
3: **return** i, $pool[i]$

Table 1. Dataset Description

Feature	Type	Description
age	numerical	Age in years
sex	categorical	Sex of the individual, 0 represents a female person and 1 represents a male person
cp	categorical	Chest pain type, 1 is typical angina, 2 is an atypical angina, 3 is non-anginal pain and 4 is for asymptomatic people
trestbps	numerical	Resting blood pressure (in mm Hg on admission to the hospital)
chol	numerical	Serum cholesterol in mg/dl
fbs	categorical	Fasting blood sugar, 1 if greater than 120 mg/dl else 0
restecg	numerical	Resting electrocardiographic results, if 0 the patient is normal, if 1 the patient is having ST-T wave abnormality (T wave inversions and/or ST elevation or depression greater than 0.05 mV), if 2 the patient is showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	numerical	Maximum heart rate achieved
exang	categorical	The value is 1 if exercise induced angina, else the value is 0
oldpeak	numerical	ST depression induced by exercise relative to rest
slope	categorical	Represents the slope of the peak exercise ST segment, upsloping if value is 1, flat slope if the value is 2, downsloping if the value is 3
ca	numerical	Number of major vessels colored by fluoroscopy
thal	categorical	If the value is 3 the patient is normal, if the value is 6, that means a fixed defect, and if the value is 7, that means a reversible defect
num	categorical (target variable)	Diagnosis of heart disease (angiographic disease status) the value is 0 if there is less than 50% diameter narrowing else the value is 1

Algorithm 2 represents the code implemented to obtain the Active Learning algorithm used in the project reported in this paper. First, the number of queries n is computed with the number of instances divided by the batch size. Then, random instances with the batch size are chosen from the training set to initialize the process. Next, a pool is constructed with the remaining instances. Finally, the learner is defined with an estimator and a query strategy. We teach the learner with the initial set of instances, and then a first prediction is made. After that, each query picks instances in the pool to teach the learner again.

Regarding the classifier used, it was implemented the *random forest classifier*, which creates multiple decision trees on various sub-samples of the dataset. The estimator then uses the average prediction of the individual trees to improve the algorithm's accuracy and control overfitting.

Algorithm 2 Active Learning

Require: The $X_{\text{train}}/y_{\text{train}}$ and $X_{\text{test}}/y_{\text{test}}$ examples of the dataset, the *estimator*, the *queryStrategy*, and the *batchSize*
Ensure: The predictions y_{pred} of the *learner*
1: Let $n = \left\lfloor \frac{\text{length}(X_{\text{train}})}{batchSize} \right\rfloor$
2: Let idx = *batchSize* random integers $\in$ [0,length(X_{train})[
3: Let $X_{\text{init}}, y_{\text{init}} = X_{\text{train}}[idx]$, $y_{\text{train}}[idx]$
4: Let $X_{\text{pool}}, y_{\text{pool}} = X_{\text{train}}$, y_{train} without $X_{\text{init}}, y_{\text{init}}$
5: Let *learner* = Active Learner instance defined with an *estimator* and a *queryStrategy*
6: Teach the *learner* with X_{init} and y_{init}
7: Let y_{pred} = predictions of the learner for X_{test}
8: Get and Save metrics using y_{test} and y_{pred}
9: **for** $q = 0$ to $n - 1$ **do**
10: Let $qidx$ = *batchSize* indexes of the *learner*'s query
11: Let $X, y = X_{\text{pool}}[qidx]$, $y_{\text{pool}}[qidx]$
12: Teach the *learner* with X and y
13: $X_{\text{pool}}, y_{\text{pool}} = X_{\text{train}}$, y_{train} without X, y
14: y_{pred} = predictions of the *learner* for X_{test}
15: Get and Save metrics using y_{test} and y_{pred}
16: **return** y_{pred}

5 Experiments

The experiments here discussed began by comparing the number of queries and the ideal batch size with *random forest classifier* to obtain results close to Passive Learning. To decide the best batch size to use (amount of data each query contains), the tests contained the batch size value of 1, 5, 7, 10, 20, and 50. However, further examination proved that after a batch size of 5, the accuracy value did not reach 80%. Therefore, the next step consisted of testing batch sizes of 1, 2, 3, 4, and 5. The number of queries varied accordingly to the batch sized, as seen in line 1 of Algorithm 2 until the algorithm analyzed all samples of the dataset (212 random samples of data were used to train the algorithm and the rest for testing, i.e., 70% of the dataset's length used for training). After that, a comparison of the different query strategies was made.

For performance evaluation, several metrics were used, such as accuracy, sensitivity, and specificity:

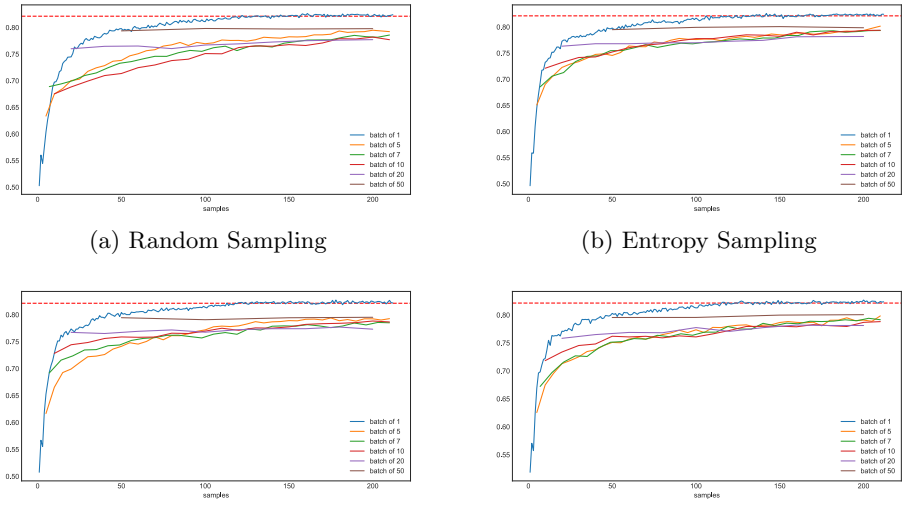
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Sensibility} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

6 Results

In the first experiment, the accuracy for each batch size was compared according to the number of queries. The dashed red line represents the optimal value and the accuracy obtained using the training set with passive learning. That being said, Figure 2a shows the *random sampling* strategy, Figure 2b shows the *entropy sampling* strategy, Figure 2c shows the *margin sampling* strategy, and Figure 2d shows the *least confident sampling* strategy. The results are very similar. After that, we decided to observe behaviours for batch sizes from 1 to 5 (Figure 3). For batch sizes 1 and 2, the accuracy of the plateau is above 80%. For batch sizes greater than 2, the accuracy is between 75% and 80%. The difference lies primarily in the rapid growth of accuracy where *entropy sampling* and *least confident sampling* strategies have a higher slope. The boxplots in Figure 4 show the median and the distribution of the number of samples required to achieve 80% of accuracy. This value increases with the batch size, except for entropy and least confident sampling, where the median decreases for a batch size of 5.



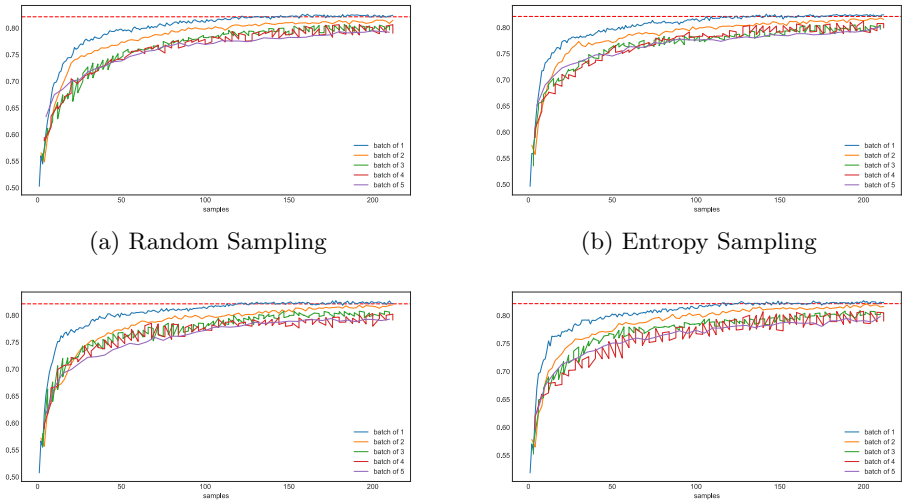
(a) Random Sampling

(b) Entropy Sampling

(c) Margin sampling

(d) Least Confident Sampling

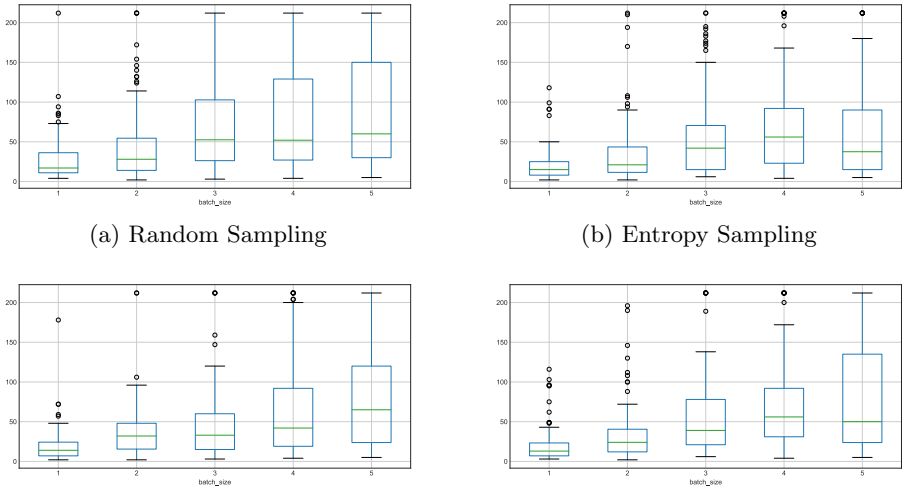
Fig. 2. Accuracy of different query strategies for the *random forest classifier* using batch sizes of 1, 5, 7, 10, 20, 50.



(a) Random Sampling (b) Entropy Sampling

(c) Margin Sampling (d) Least Confident Sampling

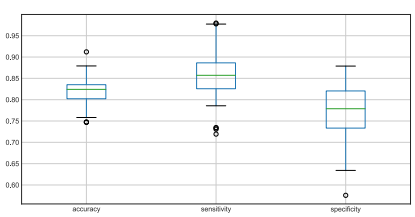
Fig. 3. Accuracy of different query strategies for the *random forest classifier* using batch sizes of 1, 2, 3, 4, 5.



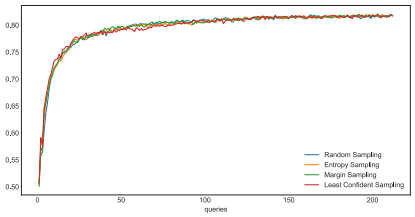
(a) Random Sampling (b) Entropy Sampling

(c) Margin Sampling (d) Least Confident Sampling

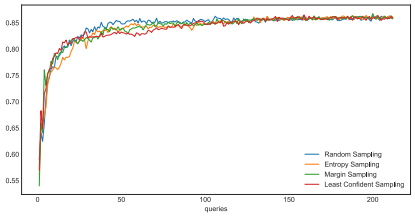
Fig. 4. Comparison of the number of samples required to reach 80% accuracy depending on various values of batch sizes $\{1, 2, 3, 4, 5\}$ for each query strategy used.



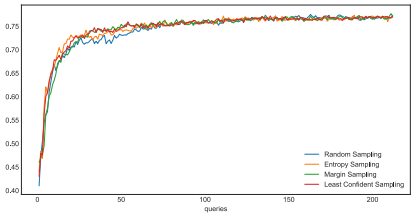
(a) Passive Learning



(b) Accuracy

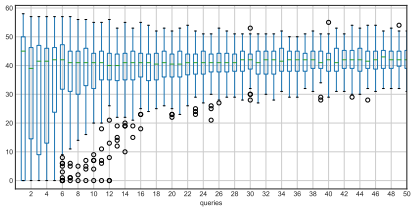


(c) Sensitivity

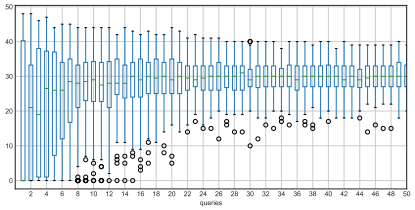


(d) Specificity

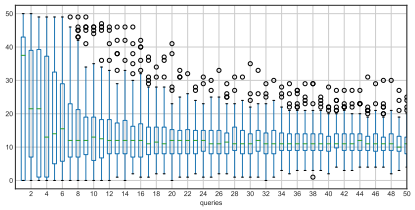
Fig. 5. Results of the accuracy, sensitivity and specificity using the *random forest classifier* and batch size of 1, comparing the four different query strategies of Active Learning (Figures b, c, and d) with Passive Learning (figure a).



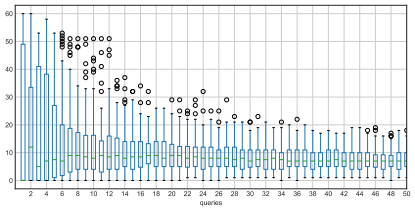
(a) True Positives



(b) True Negatives



(c) False Positives



(d) False Negatives

Fig. 6. Boxplots for true positives, true negatives, false positives, false negatives using *random forest classifier* with *random sampling* strategy and batches of 1.

In the second experiment, the four query strategies were compared with a batch size of 1. The boxplot in Figure 5a shows the distribution of accuracy, sensitivity, and specificity for passive learning. Figure 5b shows the evolution of accuracy along with the number of queries for each query strategy. With 40 queries (a total of 200 samples of data), it is possible to get near 80% of accuracy. Similarly, Figures 5c and 5d show, respectively, specificity and sensitivity for each query strategy.

Figure 6 shows in boxplots the distribution for the *random sampling* strategy of *True Positive*, *True Negative*, *False Positive*, *False Negative* for each number of queries, corresponding respectively to Figures 6a, 6b, 6c and 6d. The number of TP and TN tends to increase. In addition, the number of outliers. decreases. Inversely, cases of FP and FN decrease.

7 Discussion

This paper presented different alternatives to the query strategy and batch sizes in an Active Learning algorithm.

Regarding the batch size, a more general test was made, as described in Section 5. As it was explained, the best accuracies were obtained with batch sizes of 1, 2, 3, 4, 5, so these will be the main focus of this section. Considering the results, it was to be expected that using the value of 1 sample per query would produce the best accuracy since it analyzes the influence of each sample individually on the results according to the query strategy rather than the average of the query. Using this batch size, the dataset can be reduced by around 50% of the training set to obtain the same accuracy of Passive Learning.

Using smaller batch sizes increases the computational complexity since more queries are to run, forcing the processor to take a long time to obtain results. However, the only batch sizes that achieved the same accuracy as the value obtained with Passive Learning were 1 and 2. As such, Figure 4 has the purpose of constituting a middle ground between a good accuracy value and computational time since it indicates that for each batch size ($\{1, 2, 3, 4, 5\}$) and each of the four query strategies, an estimate of the amount of data required to obtain an accuracy of 80%, which all batch sizes ($\{1, 2, 3, 4, 5\}$) achieved. The boxplot corresponding to batch size 1 has the lowest deviation between the lowest and highest quartile across all query strategies, which indicates a lower deviation between data points. Random sampling appears to produce the highest variation between quartiles and, for a batch size of 1, all other three sampling approaches produce similar results. However, when increasing the batch size, there are significant differences between the query strategies, being that least confident sampling has the lowest median (approximately 35 data samples). As such, considering that value, the parameters are chosen previously allow a reduction of the training set by 83%.

Regarding the results obtained using the *random forest classifier*, all the different query strategies seem to obtain similar results across all three metrics analyzed (accuracy, specificity, and sensitivity), with a few discrepancies in the

sensitivity. However, it is still possible to identify in Figure 5 a higher slope when using the query strategy *least confident sampling* in terms of accuracy and specificity. As it was said previously, this can be since a *least confident sampling* considers the higher probability of the output's class. *Least confident sampling*'s main issue is that it does not consider any other probabilities except the highest one. However, since the dataset only has two classes (binary output – 1 if the patient is possibly having a heart attack and 0 if that possibility is low), having one output class with the highest probability automatically indicates that the other class has the lowest probability. Therefore, this query strategy would be expected to obtain the best results.

Lastly, it is essential to mention that Figure 6 represents the number of true positives, true negatives, false positives, and false negatives of each query averaged with the previous queries. Ideally, the sum of true positives and true negatives would be 212 samples, which is the amount of data in the training set, and the false positives and false negatives would be 0. As we can see from Figure 6, the average of the 42 queries is not null for false positives and negatives, although it is a modest number.

8 Conclusions

To summarize, the goal of the work reported in this paper was achieved in some aspects in terms of identifying what query strategies would provide the best results, taking into account also the batch size. The conclusion was that using a batch size of 1 acquires the best accuracy and uses the query strategies least confident and entropy sampling. It was also possible to understand how many samples were required to obtain the same accuracy using Active Learning comparatively with Passive Learning, which is the main focus of using the type of algorithm used in this paper.

The next step would be to use a batch size of 1 preferably and test other parameters, such as the classifier. In this project, the estimator used was *random forest classifier*. However, an idea for future work could be testing shallow or deep neural network-based Active Learning algorithms to increase their accuracy.

Acknowledgements

This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020

References

1. Balcan, M.F., Long, P.: Active and passive learning of linear separators under log-concave distributions. In: Conference on Learning Theory. pp. 288–316. PMLR (2013)

2. Bisdas, S., Topriceanu, C.C., Zakrzewska, Z., Irimia, A.V., Shakallis, L., Subhash, J., Casapu, M.M., Leon-Rojas, J., Pinto dos Santos, D., Andrews, D.M., Zeicu, C., Bouhuwaish, A.M., Lestari, A.N., Abu-Ismail, L., Sadiq, A.S., Khamees, A., Mohammed, K.M.G., Williams, E., Omran, A.I., Ismail, D.Y.A., Ebrahim, E.H.: Artificial intelligence in medicine: A multinational multi-center survey on the medical and dental students' perception. Frontiers in Public Health **9** (2021). https://doi.org/10.3389/fpubh.2021.795284, https://www.frontiersin.org/article/10.3389/fpubh.2021.795284
3. Chowdhury, M.E., Alzoubi, K., Khandakar, A., Khallifa, R., Abouhasera, R., Koubaa, S., Ahmed, R., Anwarul Hasan, M.: Wearable real-time heart attack detection and warning system to reduce road accidents. Sensors (Switzerland) **19**(12) (2019). https://doi.org/10.3390/s19122780, https://www.mdpi.com/1424-8220/19/12/2780
4. Danka, T., Horvath, P.: modAL: A modular active learning framework for Python. CoRR (2018), https://github.com/cosmic-cortex/modAL, available on arXiv at https://arxiv.org/abs/1805.00979
5. Han, W., Coutinho, E., Ruan, H., Li, H., Schuller, B., Yu, X., Zhu, X.: Semi-supervised active learning for sound classification in hybrid learning environments. PloS one **11**(9), e0162075 (2016)
6. Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R.: Heart disease data set. https://archive.ics.uci.edu/ml/datasets/Heart+Disease (2020), (accessed: 03.11.2021)
7. Mahapatra, D., Schüffler, P.J., Tielbeek, J.A., Vos, F.M., Buhmann, J.M.: Semi-supervised and active learning for automatic segmentation of crohn's disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 214–221. Springer (2013)
8. Settles, B.: Active Learning Literature Survey. Machine Learning **15**(2), 201–221 (2010). https://doi.org/10.1.1.167.4245
9. Settles, B.: From theories to queries. In: Guyon, I., Cawley, G.C., Dror, G., Lemaire, V., Statnikov, A.R. (eds.) Active Learning and Experimental Design workshop, In conjunction with AISTATS 2010, Sardinia, Italy, May 16, 2010. JMLR Proceedings, vol. 16, pp. 1–18. JMLR.org (2011), http://proceedings.mlr.press/v16/settles11a/settles11a.pdf
10. Smailagic, A., Noh, H.Y., Costa, P., Walawalkar, D., Khandelwal, K., Mirshekari, M., Fagert, J., Galdrán, A., Xu, S.: Medal: Deep active learning sampling method for medical image analysis. arXiv preprint arXiv:1809.09287 (2018)
11. Srinivas, K., Rani, B.K., Govrdhan, A.: Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE) **2**(02), 250–255 (2010)
12. Tengnah, M.A.J., Sooklall, R., Nagowah, S.D.: A predictive model for hypertension diagnosis using machine learning techniques. In: Telemedicine Technologies, pp. 139–152. Elsevier (2019)
13. World Health Organization: Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases (2021), (accessed: 04.11.2021)
14. Yakar, D., Ongena, Y.P., Kwee, T.C., Haan, M.: Do people favor artificial intelligence over physicians? a survey among the general population and their view on artificial intelligence in medicine. Value in Health **25**(3), 374–381 (2022). https://doi.org/https://doi.org/10.1016/j.jval.2021.09.004, https://www.sciencedirect.com/science/article/pii/S1098301521017411