Abstention and Active Learning: Reducing Labeling Costs and Improving Model Certainty in Healthcare Applications

Filipe Obrist
2024170686
uc2024170686@student.uc.pt

Emanuel Pacheco
2024138898
edp@student.uc.pt

Ivo Simões
2024160048
uc2024160048@student.uc.pt

Universidade de Coimbra

Abstract

Active learning is well known for being an efficient machine learning paradigm in situations where the labeling budget is limited. In real-world scenarios, such as the healthcare system, it is crucial to have an accurate and decisive system that does not incur an expensive labeling cost. To address this challenge, we integrated the concept of abstention — where the model can opt to abstain from making uncertain predictions — with active learning to produce a model with enhanced decision certainty and similar prediction capability to that of a passive learner with a much lower labeling cost. Furthermore, we explored the explainability of our resulting models to evaluate their reasonability and trustworthiness.

1 Introduction

Diabetes is a chronic disease that occurs when insulin is not produced in sufficient quantities, or when the body cannot effectively manage insulin. Diabetes affects millions of people every day, and its prevalence continues to rise, making it one of the most deadly chronic diseases in the world today. Early detection is crucial for managing the disease, as late detection can lead to kidney damage, eye damage, and an increased risk of heart disease or stroke.

The availability of a lot of healthcare data presents an opportunity for machine learning to help efficiently detect whether a patient may have diabetes or not, but despite the success of traditional ML methods in medical applications, most models often function as ”black boxes” (lacking transparency), and acquiring labeled data is often expensive and time-consuming.

We look to address this problem by combining abstention with Active Learning in the prediction modeling process, followed by the integration of Explainable AI methods. For the Active Learning part, we suggest a new approach that would combine abstention with traditional Active Learning to improve the overall decidability of the model and attempt to reduce the need for labeling large datasets. XAI would then provide insights into the models' predictions, allowing us to evaluate the reasonableness and trustworthiness of the learned predictor. With these components we aim to build a more efficient, reasonable, trustworthy, and ethically responsible system for predicting diabetes.

2 Related Work

Both Active Learning and Explainable AI have been extensively debated among the scientific community for several years, with both still being discussed to this day as the demand for efficient, ethical, and interpretable AI systems continues to grow, especially in the healthcare sector. As such, it is possible to find plenty of research regarding the use of Machine Learning in healthcare[6, 14, 5], with some cases applying Active Learning[3, 1] and others covering the ethics, reasonability, and trustworthiness of its use[7, 12].

Apart from the practical applications in healthcare and its impacts, we sought other introductory research that surveyed the methods used both in Active Learning[17, 13] and in Explainable AI[10, 16, 2, 11] to give us some context on the major topics at hand and some of the developed algorithms for Active Learning[15, 18], which inspired the abstention concept we implemented.

3 Data

In this project, we used **Diabetes Prediction Dataset** from Kaggle. This dataset contains medical and demographic data of patients along with their diabetes status, whether positive or negative. The dataset can be utilized to construct machine learning models that can predict the likelihood of diabetes in patients based on their medical history and demographic details. Since we had no convenient way to extract it, new data was not used in this project.

The dataset is described as follows (excerpt taken directly from Kaggle):

- **gender** - Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are three categories in it, male, female and other;

- **age** - Age is an important factor, as diabetes is more commonly diagnosed in older adults. Age ranges from 0-80 in our dataset;
- **hypertension** - Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values of 0 or 1 where 0 indicates they don't have hypertension and for 1 it means they have hypertension;
- **heart_disease** - Heart disease is another medical condition that is associated with an increased risk of developing diabetes. It has values of 0 or 1 where 0 indicates they don't have heart disease and for 1 it means they have heart disease;
- **smoking_history** - Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes.In our dataset we have 5 categories i.e not current, former, No Info, current, never and ever;
- **bmi** - BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese;
- **HbA1c_level** - HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2–3 months. Higher levels indicate a greater risk of developing diabetes. Mostly, more than 6.5% of HbA1c Level indicates diabetes;
- **blood_glucose_level** - Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes;
- **diabetes** - Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

The distribution of the features can be found in the appendix (Figures 18 and 19).

4 Approach

To achieve our desired results, we first needed to analyze and understand our dataset. This was followed by an analysis of the modAL library[8] to identify which parameters would be most useful for our Active Learning model, both with and without abstention.

4.1 Active Learning

The Python library modAL is an active learning framework that allows various parameters to be adjusted, potentially impacting the final results. We adopted a common active learning approach, where the model is initially trained on a small dataset and then updated iteratively with batches of specific sizes until a defined quota is reached.

After reviewing previous works by other students and relevant scientific papers, we decided to focus on two primary parameters: batch size and query strategy. For batch size, we did not have exact values to test initially, but we chose to experiment with sizes of 1, 3, 5, and 10 to analyze how the model evolves across different batch sizes.

Regarding query strategy, we initially explored several approaches, such as uncertainty sampling, margin sampling, and entropy sampling. However, we found, upon further analysis, that the queries performed by these strategies were almost identical, which yielded very similar or even equal results among them, making it redundant to examine all three.

As an alternative, one method that caught our attention was query by committee[4]. This approach involves creating multiple models with variations in certain aspects, such as the initial training set or the type of estimator used. Our hope was that, by having multiple estimators with different initial experience, each one would learn slightly different patterns in the data, thus increasing robustness to overspecification and reducing the risk of overfitting.

4.2 Abstention

The concept of abstention allows the model to refrain from making predictions when it is highly uncertain, particularly during the evaluation of the final test set. By defining a threshold, we aim to prevent the model from making incorrect decisions in critical situations, especially given the sensitivity of the topic.

To implement this, we developed a simple algorithm that evaluates the model's confidence for each instance in the test set. If the model's predicted class probability for a given case falls below the defined threshold, indicating high uncertainty, that case is ignored, signaling that the model abstains from making a prediction for the case in question.

4.3 Explainability

Explainability provides insights into our model, so we can identify and highlight the features it considers most significant when making predictions on instances from the test set. This ensures that the model's decisions are transparent and can be interpreted as reasonable and trustworthy, fostering confidence in its predictions. To achieve this, we considered the use two libraries: LIME and SHAP.

LIME (Local Interpretable Model-agnostic Explanations) focuses on explaining individual instances within a dataset. It provides insights into the model's predictions by showing the contribution of each feature and the corresponding score for each feature. Although we started by using this library for explaining individual predictions, mostly those concerning the cases where a patient has diabetes, we ultimately decided to drop it since SHAP already has the same function and much more.

SHAP (SHapley Additive exPlanations) explains the prediction of an observation by quantifying the contribution of each feature to that prediction through the use of SHapley values. These values are calculated by feeding data into a model explainer and can be used to visualize overall feature importance, dependencies between features and feature contribution in singular or multiple predictions.

5 Experimentation

In this section we'll go through our entire experimentation process, explaining the decisions taken and the derived results across data preprocessing, passive learning, active learning, using entropy sampling and query by committee, with abstention. For both passive and active learning we'll also include the explanations obtained through Explainable AI methods.

5.1 Preprocessing

We began by analyzing and cleaning our dataset. This involved checking for duplicates, missing values, uniqueness, and examining the distribution of the various features. Additionally, we reduced the size of the dataset to improve performance and enhance our results.

We reduced the dataset by 80%, from 100,000 to 20,000 records. After this reduction, we checked for any duplicates or null values, and any instances found were removed.

Next, we examined the distribution of values for each feature in the dataset to determine if they followed a balanced distribution, and to assess whether any adjustments were necessary. We found that the dataset has a large imbalance, with around 91% of the samples representing class 0 (non-diabetic) and the remaining 9% pertaining to class 1 (diabetic). This imbalance is reflected on some features, such as HbA1c_level, blood_glucose_level, heart_disease and hypertension, some of which can be justified by the bias towards non-diabetic patients (e.g. since diabetics are underrepresented, it is expected that there are few cases of high HbA1c levels). We decided to make an exception for the ”other” category in the ”gender” attribute, as it was considered to be an outlier due to severe under-representation and, consequently, removed entirely.

After eliminating what we considered unnecessary records, we proceeded to convert the dataset into a format that could be processed by the model. Specifically, we converted the categorical variables into numerical values. For example, for the ”gender” feature, we mapped ”female” to 0 and ”male” to 1.

Furthermore, at this stage, we tried to decide on train and test split sizes, to ensure equal distribution of examples for the Passive Learning and Active Learning models. Although it was a parameter that suffered a significant number of changes during testing, such as 50/50, 60/40, 70/30, 80/20 train-test splits, a 90/10 split was ultimately decided, as we believe this gives the Active Learning model an extensive pool of samples to learn from (thus combating the class imbalance present in the dataset), as well as being a split used in other problems of similar nature[9].

5.2 Passive Learning

First, we decided to implement a passive learning method using the Random Forest classifier. We tested its performance by varying the number of estimators. As shown in Figure 1, when the number of estimators reached 90-100, we observed that the performance did not change significantly. Therefore, we settled on using 100 estimators.

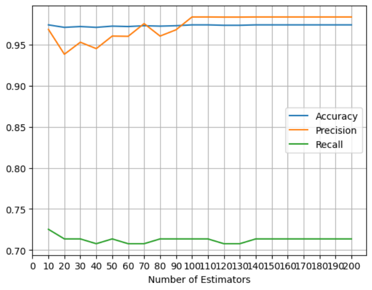

Figure 1: Accuracy, Precision, and Recall over Number of Estimators

Since the results did not differ significantly beyond 100 estimators, we decided to use this number as a baseline for the rest of the project, which generated a model that, after fitting, achieved an accuracy of 97.4%, a precision of 98.4%, and a recall of 71.3% on the test set, as shown in the confusion matrix in Figure 2.

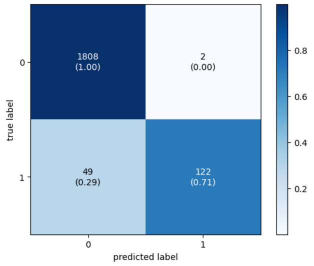

Figure 2: Confusion Matrix of the Passive Learning model with 100 estimators

While the accuracy and precision values are quite high and could be considered ideal, we observed that the recall is relatively low. One of the reasons for this is the imbalance in our dataset, as our model was somewhat biased towards the negative cases (more false negatives than false positives), as seen in Figure 2. We can further analyze the motives behind the model's prediction performance by taking a look at some practical results produced by SHAP when used to explain the test set predictions for positive cases.

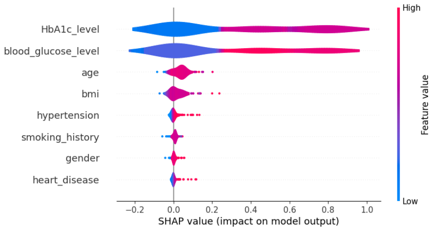

Figure 3: Passive Learner: Feature importance through SHAP values for positive cases

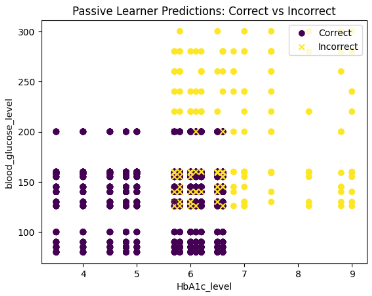

Figure 4: Passive Learner: Correct vs. Incorrect predictions for top 2 features according to SHAP

According to the values produced by SHAP, HbA1c and blood glucose levels are significantly more important factors in determining if a person has diabetes than the other patient characteristics our dataset provides. This was expected as, when describing the dataset's features in Section 3, it was mentioned that high values of HbA1c and/or blood glucose are positive indicators of diabetes, something that can also be inferred from SHAP's explanation. To further observe this, a plot was created to visualize the correct and incorrect predictions according to HbA1c and blood glucose levels in the test set samples. We can see that there's a somewhat clear separation between diabetic (high values) and non-diabetic (low values) patients when comparing the two features and that the model was able to capture this separation. However, there seems to be some overlap of positive and negative cases in samples where blood glucose levels and HbA1c levels range from [125, 200] and [5.7, 6.6], respectively. This is where the false negative cases lie, as the model can't seem to make the distinction between diabetic and non-diabetic when this overlap occurs.

5.3 Active Learning

For Active Learning, we decided to train four separate models, each with various batch sizes: 1, 3, 5, and 10. To perform a more reasonable comparison to the passive learner, we decided to also use Random Forest classifiers with 100 estimators, with the query strategy chosen being entropy sampling. As for the train and test set sizes, we decided to perform the same split as in the passive learning setup, to ensure consistency between the approaches, and an additional split to extract 20 initial labeled training samples, which are given at the start of the training, leaving the rest of the samples in a pool for querying. Furthermore, we limited the model to evaluate a maximum of 210 queries, as further training seemed to either cause some sort of overfitting or an unnecessary/wasteful increase in labeling cost. To implement this, we created a loop where, in each epoch, a specific batch size was given to the model until it reached the 210 query limit.

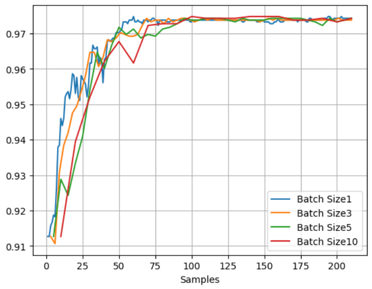

Figure 5: Active Learner: Accuracy over Samples

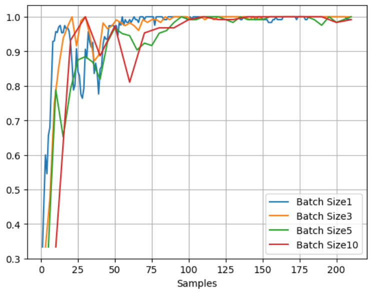

Figure 6: Active Learner: Precision over Samples

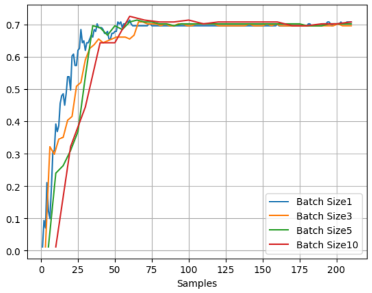

Figure 7: Active Learner: Recall over Samples

By analyzing the graphs, we observe that after providing 75 samples, regardless of batch size, our model achieves very good accuracy and precision, comparable to our passive learning model. However, it also exhibits quite low recall, indicating that while the model performs well with negative cases, it still misses a significant number of positive cases. This similarity in performance between both active and passive learners allows us to infer that the active learner may too be struggling with the classification of cases where HbA1c and blood glucose values are similar between positive and negative examples. To corroborate our theory about the performance of the active learner, we may, once again, turn to SHAP to get explanations for the predictions on test set positive cases.

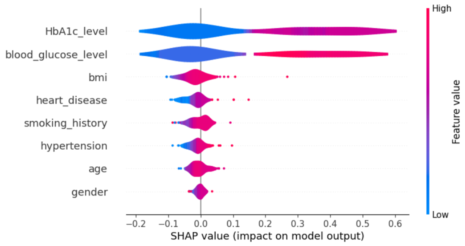

Figure 8: Active Learner (Batch Size 1): Feature importance through SHAP values for positive cases (for other batch sizes, check appendix 8.2)

As expected, both HbA1c and blood glucose levels seem to be the most determinant factors of diabetes in a patient, according to our model and as seen in Figure 8. For the rest of the features, however, our active learner with batch size 1 looks to somewhat disagree with the passive learner, as, for example, it found age to be not as important as knowing if a patient has heart disease, a direct contradiction to the latter.

Like we did for the passive learner, we can also plot the predictions using the two most important features, according to SHAP, to see if the overlapping we mentioned previously is a major cause of recall falloff.

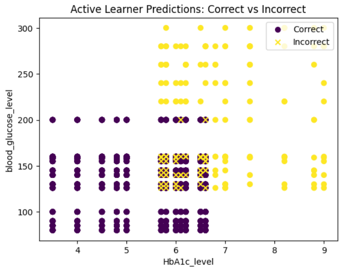

Figure 9: Active Learner (Batch Size 1): Correct vs. Incorrect predictions for top 2 features according to SHAP

Once again, we can observe the model's confusion when trying to classify cases where samples with similar values and different diabetes conditions overlap, shown in Figure 9.

When comparing these results to the passive learning method, we see that the performance did not vary significantly. However, it is important to note that we reduced the labeled dataset by 99%. This demonstrates that active learning allows us to significantly reduce the labeled training set (and the associated costs of obtaining it) while still achieving competitive results compared to passive learning.

5.4 Active Learning with Abstention

However, active learning alone was not enough for us. We wanted to achieve even better results than those obtained with the methods previously used. To this end, we implemented Active Learning with Abstention.

The concept of abstention we applied involves having the model refrain from making predictions when it is uncertain about the outcome. We implemented a simple check where we calculate the probability that the model assigns to each prediction being correct. If this probability is lower than a fixed abstention threshold (set at 0.75), the model will abstain from making a prediction.

We also modified how the queries were selected by introducing a form of "early stopping." Specifically, if the model doesn't encounter a batch of queries where the certainty is below 75%, the training phase would stop. This was implemented because we believe the model is capable of providing accurate answers when the threshold is met. As a result, this approach helps reduce labeling costs even further and may potentially minimize overfitting. Additionally, the hope was that this approach would also give the model the option to not make a decision in a more uncertain scenario, possibly avoiding the overlap situation we saw when trying to explain the learners' decisions.

The following graphs show our results:

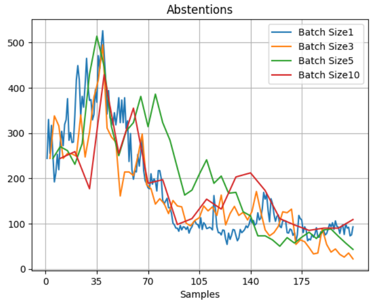

Figure 10: Active Learner with Abstention: Abstentions over Samples

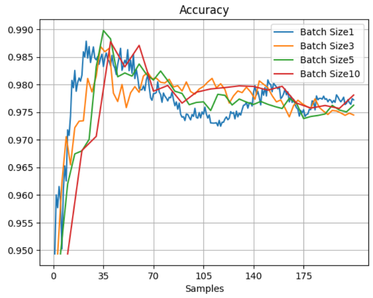

Figure 11: Active Learner with Abstention: Accuracy over Samples

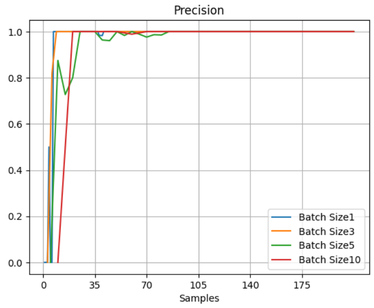

Figure 12: Active Learner with Abstention: Precision over Samples

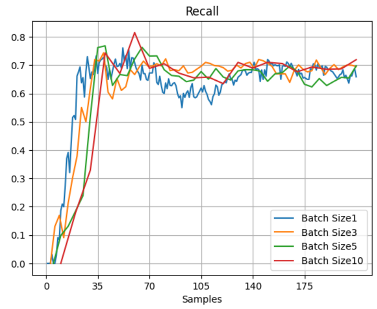

Figure 13: Active Learner with Abstention: Recall over Samples

By analyzing the graphs, we observe that, regardless of batch size, the levels of abstention decrease as we increase the number of samples. This means that the model becomes increasingly less uncertain in its predictions as more samples are fed to it.

The other graphs also show significant differences between active learning and active learning with abstention. In the accuracy graph, we can see that all batch sizes experienced a spike between 98% and 99% with around 35 samples, after which there was a decrease to about 97.5%. This provides a small increase in accuracy compared to the active learning model without abstention.

Precision for all batch sizes experienced an immediate spike to 100%, which remained consistent between both models. As for the recall graphs, we can see a similar pattern in both models, with recall settling at approximately 70%, though not in a very consistent manner.

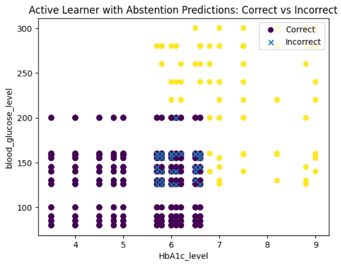

Figure 14: Active Learner with Abstention (Batch Size 1): Correct vs. Incorrect predictions for top 2 features according to SHAP

We have already seen how the active learner perceives feature importance in Section 5.3, particularly in Figure 8, and, since our abstention concept only changes the way predictions are handled, we can check if the prediction plot from the base active learning model to the abstention one has undergone significant changes. Figure 14 shows the predictions made by the active learning model after 210 query training and the application of an abstention threshold of 0.75 and, if we compare it to Figure 9, we can pick up on some slight differences. We can notice some missing data points in the overlapping region and in its surroundings too, specifically in the positive cases region, which tells us that the model mostly abstained from guessing positive cases for diabetes, some of them situated in the overlapping region.

We can conclude that, as the abstention rate kept decreasing, so too did the accuracy and recall, revealing to us that the model was getting more certain of wrong decisions the more it was trained (possible signs of overfitting and class imbalance). A question that derives from this problem is the following: what is considered acceptable or optimal abstention and how can it be achieved?

5.5 Query by Committee with Abstention

We first tested without abstention, but it did not differ significantly from the active learning method. However, the results can be found in Appendix 8.3.

As our final test, we decided to check whether creating multiple active learning models and assigning them different training sets and estimators would enhance our results.

To create a committee, we started by deciding on the number of members (5 in our case). Since query by committee can be computationally expensive, we decided to create a committee of random forests with the following quantities of estimators: [40, 55, 70, 85, 100], each of them having 20 different samples of initial training. The hope here was that each random forest would differ slightly in the decision making process, possibly capturing different patterns in the data.

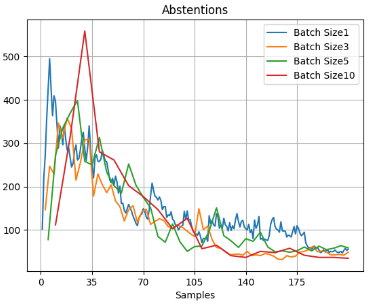

Figure 15: Query by Committee: Abstentions during training

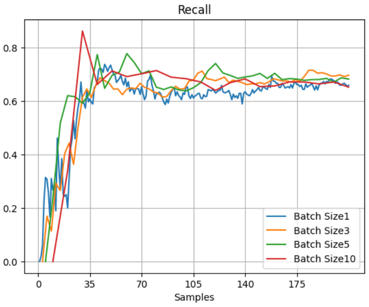

Figure 16: Query by Committee: Recall during training

As we can see, compared to the active learning with abstention model, query by committee has a faster convergence to lower abstention rates than the other methods. This suggests that with multiple models, there is less uncertainty, and the models are more confident in most of the cases presented in the test set.

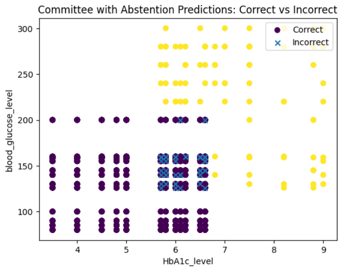

Figure 17: Query by Committee with Abstention (Batch Size 1): Correct vs. Incorrect predictions for top 2 features according to SHAP

Although the convergence to decisiveness is faster in query by committee, we see, once again, the problem of being overconfident, where the model ends up not abstaining on the most difficult cases and wrongfully predicting the instance. Figure 17 shows that overconfidence, with the overlapping zone having more wrong predictions than the active learner in Figure 14, which resulted in a sacrifice of recall score, as seen in the comparison between Figure 16 and Figure 12.

Since the accuracy and precision did not differ significantly, we did not discuss them further. However, the graphs can be found in the appendix (Figures 26 and 27).

6 Conclusions

This project successfully explores the integration of Active Learning (AL) abstention mechanisms and Explainable AI (XAI) to develop a robust and efficient model for predicting diabetes while reducing data labeling costs and addressing ethical challenges. Below, we summarize the main achievements and challenges:

6.1 Main Achievements

- Reduction in Labeling costs - By implementing active learning, we reduced the labeled dataset size by 99% while achieving results comparable to passive learning. This highlights the practical value of AL in healthcare applications, where labeling costs are high;
- Improve Model Confidence - The integration of an abstention mechanism allowed the model to refrain from making predictions in high-uncertainty cases. This enhanced decision reliability and reduced critical errors in sensitive applications such as healthcare;
- Transparency - The use of SHAP for feature importance analysis provided valuable insights into the model's decision-making process, promoting trust and interpretability.

6.2 Main Difficulties

- Class imbalance - The imbalance in the dataset presented challenges, leading to a higher number of false negatives. Addressing this required careful consideration of evaluation metrics and model adjustments;
- Limited Dataset Diversity - The reduced labeled dataset may have constrained diversity, potentially impacting the model's generalization capability in higher contexts.

This project aligns closely with the principles of Human-Centered AI (HCAI) by emphasizing transparency, ethical considerations, and enhancing human decision-making capabilities. The model promotes transparency by explaining its predictions through feature importance analysis. It aligns with ethical principles by using abstention mechanisms to avoid high-risk decisions and providing justifications for those decisions, thus reducing potential harm. By incorporating these elements, the model supports human decision-makers in making more informed and reliable choices.

6.3 Future Work

In Section 5.4, we raised an interesting question: what is an acceptable or optimal abstention-performance trade-off? While we don't have an exact answer, we might be able to make suggestions to find it. In the context of our problem, abstention can be seen as a cost, where the higher the abstention rate is, the higher the cost we incur. In this sense, abstention can be seen as a cost function, and, to obtain a model with optimal abstention, an attempt can be made to optimize the active learning model using this cost function. Similar ideas to the one we propose have already been discussed in some scientific papers.[18, 15].

We have also noticed how the model does not utilize abstention optimally, i.e. it abstains from predicting instances that aren't as clear as others (such as the overlapping case discussed in Section 5.2 and Section 5.4). This problem might be more difficult to solve, as there isn't a clear issue at stake; the model is just predicting based on the patterns it has observed. Possible suggestions to minimize misplaced abstention would be tweaking the cost function mentioned earlier or deeply analyze feature interaction and implement some type of data preprocessing strategy that might alert for presence of this type of noise. These cases could be handled separately from the model, by asking an expert directly to evaluate, raising the labeling cost, and/or by completely excluding them from model training, reducing the chance of performance degradation due to the presence of noise or overfitting.

It should be reinforced that the dataset in this project is very imbalanced, which means that a simple measures to prevent these cases such as undersampling the majority class or oversampling the minority class could also work.

7 Acknowledgements/Disclaiming

LLMs, specifically ChatGPT, assisted us in producing the code for this project, particularly in generating graphs and helping us learn how to work with libraries. ChatGPT was also used to correct grammar throughout the overall structure of this document.

8 Appendix

In this section, we present additional graphs that, while less relevant compared to others, may still be of interest. These provide extra results for some methods.

8.1 Distribution of Features

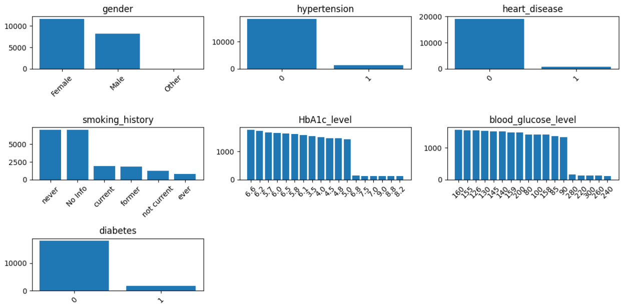

Figure 18: Distribution of Features (without age and BMI)

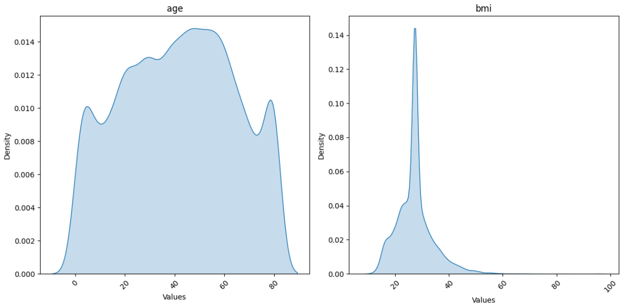

Figure 19: Distribution of Features (age and BMI)

8.2 Active Learner: Feature Importance through SHAP values for positive cases

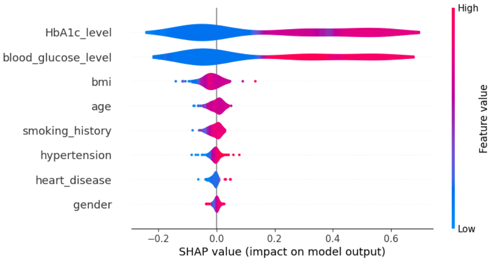

Figure 20: Batch Size 3

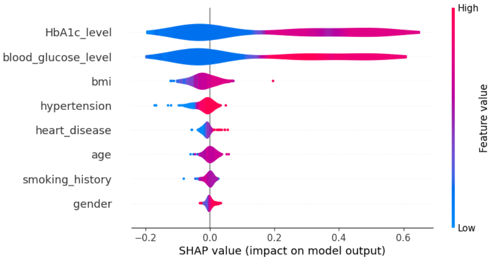

Figure 21: Batch Size 5

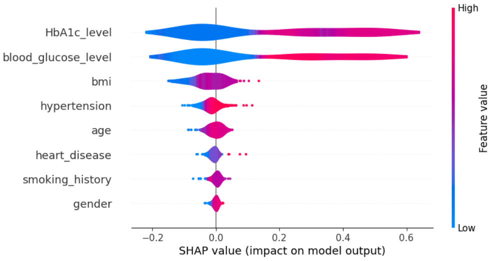

Figure 22: Batch Size 10

8.3 Model Performance Evaluation Query by Committe

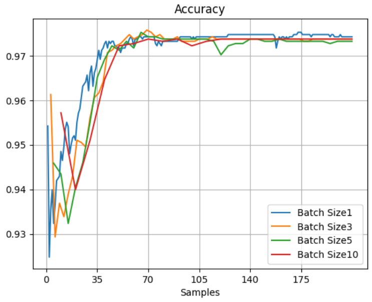

Figure 23: Query by Committe: Accuracy over samples

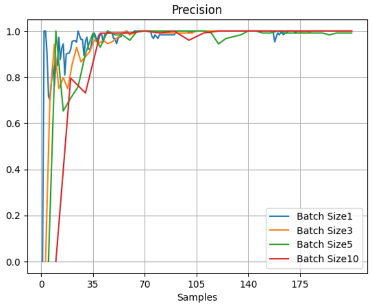

Figure 24: Query by Committe: Precision over samples

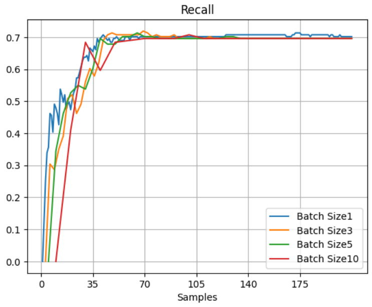

Figure 25: Query by Committe: Recall over samples

8.4 Model Performance Evaluation Query by Committe with Abstention (Accuracy and Precision)

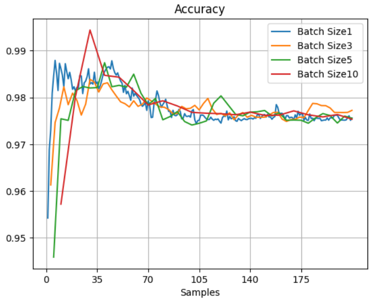

Figure 26: Query by Committe with Abstention: Accuracy over samples

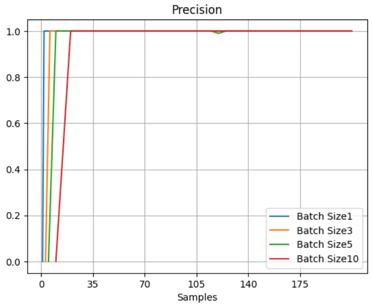

Figure 27: Query by Committe with Abstention: Precision over samples

References

[1] Sidra Abbas, Gabriel Avelino Sampedro, Shtwai Alsubai, Stephen Ojo, Ahmad S. Almadhor, Abdullah Al Hejaili, and Lubomira Strazovska. Advancing healthcare and elderly activity recognition: Active machine and deep learning for fine- grained heterogeneity activity recognition. *IEEE Access*, 12:44949–44959, 2024.

[2] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[3] Rafael S. Bressan, Guilherme Camargo, Pedro Henrique Bugatti, and Priscila Tiemi Maeda Saito. Exploring active learning based on representativeness and uncertainty for biomedical data classification. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2238–2244, 2019.

[4] Robert Burbidge, Jem J. Rowland, and Ross D. King. Active learning for regression based on query by committee. In Hujun Yin, Peter Tino, Emilio Corchado, Will Byrne, and Xin Yao, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pages 209–218, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery.

[6] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.

[7] Giovanni Cinà, Tabea Röber, Rob Goedhart, and Ilker Birbil. Why we do need explainable ai for healthcare, 2022.

[8] Tibor Danka and András Horváth. modal: A modular active learning framework for python. *arXiv preprint arXiv:1805.00979*, 2018.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data, 2017.

[10] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable ai: current status and future directions, 2021.

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models, 2018.

[12] Gajendra Jung Katuwal and Robert Chen. Machine learning model interpretability for precision medicine, 2016.

[13] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers, 2024.

[14] Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, and Constantinos S. Pattichis. Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1837–1857, 2020.

[15] Shubhanshu Shekhar, Mohammad Ghavamzadeh, and Tara Javidi. Active learning for classification with abstention. *IEEE Journal on Selected Areas in Information Theory*, 2(2):705–719, 2021.

[16] Kacper Sokol and Julia E. Vogt. What does evaluation of explainable artificial intelligence actually tell us? a case for compositional and contextual validation of xai building blocks. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM, May 2024.

[17] Li-Li Sun and Xi-Zhao Wang. A survey on active learning strategy. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pages 161–166, 2010.

[18] Yinglun Zhu and Robert Nowak. Efficient active learning with abstention, 2022.