1 2 9 0

UNIVERSIDADE D
COIMBRA

Nelson Filipe de Matos Martins

An Active, Explainable Artificial Intelligence Model for Classification Tasks in Medicine

Dissertação no âmbito do Mestrado de Engenharia Biomédica, no ramo de Informática Clínica e Bioinformática, orientada pelo Professor Luís Miguel Machado Lopes Macedo e apresentada ao Departamento de Física da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Fevereiro de 2024

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

An Active, Explainable Artificial Intelligence Model for Classification Tasks in Medicine

Nelson Filipe de Matos Martins

Thesis submitted to the Faculty of Sciences and Technology of the University of Coimbra
for the degree of Master in Biomedical Engineering with specialization in Clinical Informatics and
Bioinformatics, supervised by Prof. Dr. Luís Macedo

Coimbra, February 2024

This work was developped in collaboration with:

Center for Informatics and Systems of the University of Coimbra

CISUC

Dedication

I want to start by expressing my gratitude to Prof. Dr. Luís Macedo, my advisor, for everything he has done to help me with the development of this thesis. Your advice was crucial to my path, and I sincerely appreciate your kindness, patience, and belief in me. I appreciate you taking the time to guide me.

To my bachelor's and master's friends, with whom I took this journey, your jokes lightened my day and helped me carry on. You helped me feel like I was not alone, and in doing so, gave me strength I would not have been able to find by myself. I only hope I was able to return the favor.

To my best friends Alexandre, João and José, for always being by my side for as long as I can remember. To Adriana, who has not only been a steadfast companion but has also become a cherished friend whom I will carry with me throughout my life. To all my other friends who are present in my daily life and also supported me along the way.

Finally, to my aunt and grandfather, for their unwavering support and encouragement. Their guidance and belief in me have been invaluable throughout my journey, and I am truly fortunate to have them.

Thank you all, from the bottom of my heart.

Resumo

Esta tese explora o panorama da inteligência artificial aplicada aos dados de saúde, extraindo conhecimentos de diversos conjuntos de dados. Os dois primeiros conjuntos de dados, adquiridos no *Kaggle*, centram-se no domínio crítico da triagem de emergência. O terceiro conjunto de dados, proveniente do UC Irvine Machine Learning Repository, centra-se nas doenças cardíacas. Cada conjunto de dados apresenta desafios únicos e exemplifica cenários do mundo real cruciais para aplicação nos cuidados de saúde.

Foi implementada uma estrutura que integra classificadores baseados na incerteza, especificamente regressão logística e *random forest*, individualmente ou combinados. É de notar que esta estrutura obtém resultados notáveis, pois ultrapassa o limiar de precisão de 95% e alcança bons valores em várias métricas de desempenho, quando configurada com parâmetros específicos. A investigação estendeu-se além das métricas convencionais para demarcar a imperatividade da transparência e explicabilidade do modelo. A aplicação de Local Interpretable Model-agnostic Explanations (LIME) fornece informações cruciais sobre os processos de decisão de modelos *machine learning* complexos.

Além disso, desdobra-se um estudo robusto que compara e contrapõe estratégias de aprendizagem passiva e ativa, revelando assim a eficácia da estrutura para atingir um desempenho excecional com apenas 10% do conjunto de dados inicial. Estes resultados, obtidos através de consultas estratégicas, sublinham o potencial da estrutura para ultrapassar os desafios da escassez de dados nos domínios da saúde.

À medida que a inteligência artificial continua a redefinir as práticas de saúde, esta tese contribui com perspetivas detalhadas, oferecendo assim informações para profissionais, investigadores e legisladores.

Palavras-chave: Inteligência Artificial, Aprendizagem Ativa, Explicabilidade, Incerteza na Amostragem, Escassez de Dados, Aplicações nos Cuidados de Saúde

Abstract

This thesis navigates the landscape of Artificial Intelligence (AI) applied to health data, drawing insights from diverse datasets. The first two datasets, acquired from Kaggle, center around the critical domain of emergency triage. The third dataset, sourced from the UC Irvine Machine Learning Repository, focuses on heart diseases. Each dataset presents unique challenges and reflects real-world scenarios that are crucial for healthcare applications.

The framework deployed integrates uncertainty-based classifiers, specifically Logistic Regression (LR) and Random Forest (RF), either individually or combined (LR + RF). Notably, the framework attains remarkable results, surpassing the 95% accuracy threshold and achieving good values across various performance metrics, when configured with specific parameters. The investigation made extends beyond conventional metrics so as to embrace the imperative of model transparency and explainability. The application of LIME provides some insights into the decision-making processes of Machine Learning (ML) models.

Additionally, a robust study comparing passive vs active learning strategies unfolds, revealing the framework's efficacy in achieving exceptional performance with a mere 10% of the initial dataset. These results, which were achieved through strategic queries, underscore the framework's potential to overcome data scarcity challenges in health domains.

As AI continues to redefine healthcare practices, this thesis contributes nuanced perspectives, offering insights for practitioners, researchers, and policymakers alike.

Keywords: Artificial Intelligence, Active Learning, Explainability, Uncertainty Sampling, Data Scarcity, Healthcare Applications

List of Acronyms

ACEP American Collge of Emergency Physicians.

AI Artificial Intelligence.

AL Active Learning.

CTAS Canadian Triage and Acuity Scale.

ED Emergency Department.

ENA Emergency Nurses Association.

ESI Emergency Severity Index.

KTAS Korean Triage and Acuity Scale.

LIME Local Interpretable Model-agnostic Explanations.

LR Logistic Regression.

ML Machine Learning.

MTS Manchester Triage System.

NCD Non-Communicable Diseases.

QBC Query-By-Committee.

RF Random Forest.

SHAP SHapley Additive exPlanations.

XAI Explainable Artificial Intelligence.

List of Figures

List of Tables

Contents

1

Introduction

1.1 Motivation and Context

The integration of AI into the healthcare domain is motivated by the immense potential to revolutionize patient care, medical diagnosis, and treatment. The capabilities of AI, such as rapid data analysis, real-time monitoring, and predictive modeling, can significantly augment the capabilities of healthcare professionals. The collaborative synergy between AI algorithms and human clinicians, embodies a future where intelligent machines assist healthcare providers in making more informed decisions. However, the translation of AI success in other domains to healthcare is impeded by challenges such as the scarcity of data, particularly in critical medical areas. This scarcity accentuates the need for innovative approaches, such as AL, to intelligently select and acquire the most informative data for efficient model training.

In the health domain, where decisions directly impact human well-being, the significance of AI transparency and explainability cannot be neglected. While advanced ML algorithms, including deep learning and RF, exhibit exceptional accuracy, they often lack transparency, hindering their widespread acceptance and deployment in critical healthcare settings. The inability to understand and interpret the decision-making process of these algorithms is a barrier to trust and acceptance. This challenge gains further prominence in light of regulatory frameworks, such as the General Data Protection Regulation (GDPR), which grants individuals the right to question AI systems about their decisions. Thus, the motivation to bridge the gap between AI's power and its transparent, ethical application in healthcare forms the cornerstone of this research.

1.2 Objectives and Approach

The primary objectives of this study are designed to address key challenges in the intersection of ML, healthcare, and model interpretability.

Firstly, an AL framework was designed - and then implemented - to create an innovative tool specifically tailored to healthcare applications. This framework aims to optimize the training process of ML models in health scenarios where labeled data is limited, particularly focusing on emergency triage datasets.

Then, after creating the framework, it was time to test and analyze the performance of machine learning models, particularly LR and RF classifiers, under varying conditions. This includes exploring different estimators' sizes and query strategies within the AL framework.

In order to enhance the transparency and interpretability of the ML models employed in healthcare decision-making, LIME was integrated.

In addition, a comparative study between active and passive learning was approached to elucidate the efficiency and effectiveness of the proposed active learning framework by exploring the percentage of the dataset required for active learning to approximate passive learning benchmarks.

Finally, the developed framework was applied to three healthcare datasets, including not only the emergency triage datasets but also a heart disease dataset, to validate its generalizability and effectiveness across distinct medical domains.

1.3 Contributions

This thesis adds contributions to the field of ML, especially when it comes to model interpretability and healthcare applications, through the combination of different configurations when creating the model. The following sums up the main contributions:

- A framework for AL was created and thoroughly tested, demonstrating that it can greatly improve model performance even with little labeled data. This framework demonstrates the potential to optimize resource utilization while maintaining high predictive accuracy. It is specifically designed to address the challenges posed by data scarcity healthcare.
- A comprehensive analysis of performance metrics was conducted. The results illustrate the robustness and reliability of the active learning framework across various scenarios, emphasizing its adaptability to different datasets and classifier configurations.
- By applying the LIME model into the framework, healthcare professionals and patients can benefit from insights like identifying the most influential features influencing prediction outcomes, gaining confidence in the model's recommendations and facilitating better-informed decision-making.

1.4 Document Structure

This document is organized as follows:

- **Chapter 2 - Background Knowledge and State of Art:** Lays the groundwork by providing essential background knowledge, covering fundamental concepts, theories, and existing practices relevant to the research topic. Simultaneously, explores the current state of the art in the field, reviewing the latest advancements, methodologies, and technologies pertinent to the thesis topic.

- **Chapter 3 - Material and Methods:** Delves into the specifics of the framework design, materials used, and the methodologies employed. Provides a detailed account of datasets, ML algorithms, and the experimental setup.
- **Chapter 4 - Results and Discussion:** Presents the central findings of the research. The discussion section interprets these results in the context of the research questions, highlighting key insights and trends.
- **Chapter 5 - Conclusion and Future Work:** Summarizes key findings and insights from the study, draws conclusions based on the results, and outlines potential avenues for future research.

2

Background Knowledge and State of the Art

This section gives a overview of the history of AI, showing how it progressed from theoretical conceptions to real-world applications across a range of industries, with an emphasis on its importance in healthcare. In addition, it addresses the difficulties and possibilities that come with using AI, such as the lack of data, the interpretability of models, ethical issues, and legal frameworks.

2.1 Artificial Intelligence

The human imagination has always been a source of invention, inspiring us to realize our most improbable fantasies. We have pushed the boundaries of what is possible in our never-ending search for knowledge. At the center of this quest is the unwavering pursuit of understanding and mimicking human-like intelligence in non-biological entities, which gave rise to the field of AI.

The origins of AI can be traced back to the 1940s [1], notably 1942, when American science fiction writer Isaac Asimov released his short story Runaround [2]. The story introduces for the first time these rules:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

In the year after, 1943, the first instance of an AI model was proposed. Warren McCulloch and Walter Pitts introduced a network of artificial neurons, a foundational concept for neural networks [3].

The Turing Test was developed to offer a sufficient operational definition of intelligence. It was first presented by Alan Turing in 1950 [4]. Turing's theoretical construct posits a scenario wherein an evaluator, referred to as "C," interacts with two entities hidden from view—a human, designated as "A," and an artificial agent or machine, referred to as "B." The objective of this exercise is for the evaluator to ascertain which of the two entities is human and which is artificial solely through textual or oral exchanges. Basically, if a human interrogator cannot identify whether the written responses are from a person or a machine after asking some written questions, the computer has passed the test [3].

Critically, the Turing Test pushes beyond the conventional yardstick of machine intelligence, which historically relied on mere computations and algorithmic problem-solving capabilities. Instead, it delves into the realm of understanding, contextual interpretation, and linguistic fluency—attributes indicative of human-like thought processes. In doing so, Turing's Test advances the notion that true artificial intelligence must embody not just logical reasoning but also an ability to exhibit human-like conversational skills.

[5] In 1956, the Dartmouth Conference played a pivotal role in introducing the concept of AI, effectively marking its inception. This period witnessed a surge in international academic interest and collaboration in the field of AI. However, as the 1960s unfolded, predominant paradigms such as connectionism and rule-based systems faced declining popularity, causing a setback in the progress of intelligent technologies. The 1970s saw the emergence of research into backpropagation algorithms. Additionally, the rising cost and expanding computational capabilities of computers presented challenges for the research and implementation of expert systems [5]. Progress in AI encountered challenges, but it continued to advance. By the 1980s, backpropagation neural networks gained widespread recognition, leading to rapid algorithmic advancements in artificial neural networks. Simultaneously, computer hardware capabilities experienced significant growth. However, the proliferation of the Internet during this era posed some challenges for the further evolution of AI, primarily concerning data privacy and security. With the increasing interconnectedness of systems and the exponential growth of online data, safeguarding sensitive information became a paramount concern [5]. The field of AI comprises diverse fields like computer vision, natural language processing, and ML, all of which add to its complex terrain. Therefore, the following subsection introduces the field of ML.

The quick development of AL has made it possible to tackle categorization challenges in a variety of fields, including medicine, with creative methods. Understanding the current state of Explainable Artificial Intelligence (XAI) and AL is essential to developing the proposed framework for classification tasks in medicine as the healthcare sector continues to adopt AL technology.

Chen et al [6] investigated the potential of AL in enhancing the development of high-throughput phenotyping techniques based on supervised MLalgorithms that are generalizable. The main objective of the study is to apply AL to phenotyping algorithms for data from electronic health records. In order to overcome the issue of requiring a large number of annotated samples, which are costly and time-consuming to investigate, this study combines an uncertainty sampling AL approach with support vector machine-based phenotyping algorithms. Using annotated illness cohorts including venous thromboembolism, colorectal cancer, and rheumatoid arthritis, the project's goal is to evaluate AL's performance and compare its effectiveness to a passive learning technique based on random sampling. They aimed to create effective and precise phenotyping algorithms that could be used for a variety of diseases and traits, accelerating the use of data from electronic health records for clinical and translational research and advancing medical research in the process. The results show that AL may efficiently minimize the number of annotated samples required for excellent classification performance, lowering the cost and time required for manual chart inspection.

2.1.1 Machine Learning

ML is a branch of AI that focuses on creating statistical models and algorithms that let computers learn from data in order to become better at a given task without having to be explicitly programmed [7, 8]. Natural language processing, computer vision, speech recognition, and many more industries can all benefit from ML, which is a fundamental technology. It is essential for giving AI systems the ability to carry out activities that call for learning from data and adjusting to new knowledge [7, 8].

2.1.2 Machine Learning Classification models

A key component of data analysis and interpretation are classification models. These models facilitate decision-making in a variety of industries, including healthcare, by predicting the categorical consequence of a given input. The AL system employs the following algorithms to classify instances, similar to a conventional ML system. However, in instances where uncertainty arises, the system solicits guidance or assistance from an oracle.

2.1.2.1 Decision Trees

A decision tree is a sequential model that is used to create regression or classification models in data mining and ML [9]. Decision trees are composed of nodes and branches, where each node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The tree is constructed by recursively splitting the data into subsets based on the values of the attributes until a stopping criterion is met [9, 10].

Because of their transparency and ease of use, decision trees are preferred in machine learning. However, they do not come without obstacles. One common problem is that they are prone to overfitting, which occurs when the model overfits and fails to properly generalize to new data by capturing noise in the data. Furthermore, decision trees may not be as robust when exposed to various data distributions due to their high sensitivity to changes in the dataset. Pruning techniques and ensemble methods are required to overcome these obstacles and improve their performance in real-world scenarios [9, 10].

2.1.2.2 Random Forest

A RF is a collection of decision trees, each of which is trained using a random subset of the training data and a random subset of the features. The combined projections of all the forest's trees lead to the final conclusion [11].

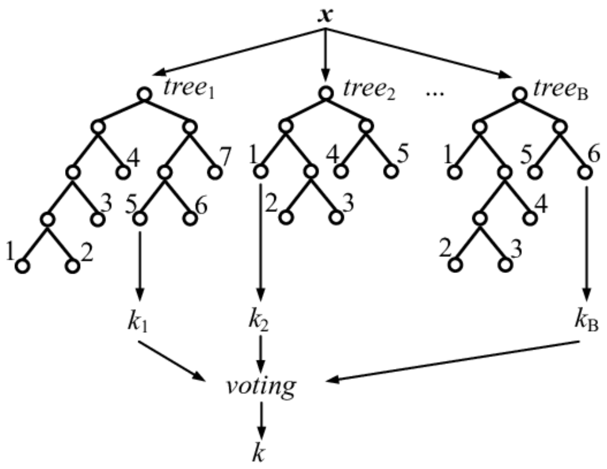

Figure 2.1: A standard RF architecture [12].

Due to the fact that the effects of noisy data in individual trees average out over the forest, RF are resistant to overfitting and noisy data [11]. Since there are not many instances available for training in AL, this is particularly advantageous. RF is also a good classification technique to deal with imbalanced datasets and also outperformed other ML techniques in terms of accuracy and other metrics [13].

2.1.2.3 Logistic Regression

LR classifier is a model that has seen increased use in ML applications over the last decade [14]. The method works by estimating a set of weights that can be used to calculate the probability of the outcome being true for each data point. These weights are learned from the training data using maximum likelihood estimation. Once the weights are estimated, the LR model can be used to predict the probability of the outcome for new data points with similar predictor values [15].

The LR equation 2.1 in AL is given by:

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-(\mathbf{w}^T\mathbf{x}))} \tag{2.1}$$

In this equation, $\mathbf{x}$ represents the input feature vector, $\mathbf{w}$ denotes the weight vector, and $\mathbf{w}^T\mathbf{x}$ represents the dot product of the weight vector and the feature vector. The sigmoid function, denoted by $\frac{1}{1+\exp(-(\mathbf{w}^T\mathbf{x}))}$, maps the linear combination of features to a value between 0 and 1, representing the estimated probability of the positive class.

Only LR and RF models are presented, although there are several others like Support Vector Machines (SVM), working by finding the hyperplane that best separates different classes in the feature space, Naive Bayes Classifier, that are based on Bayes' theorem with the assumption of independence between features, and many others beside those. Decision trees were also mentioned because they are important to understand how RF models work.

2.1.3 Active Learning

In the statistics literature, AL, also known as query learning or optimal experimental design, is an area of ML and, more broadly, AI [16]. AL is a ML approach that involves an iterative process of selecting and labeling the most informative samples from an unlabeled dataset. The goal of AL is to train a model with high accuracy while minimizing the amount of labeled data required [17]. To achieve this, it first determines which examples are the most instructive based on the model's current understanding, then asks a human oracle or an already-existing, labeled dataset for the correct labels.

A large labeled dataset is often required in classical ML to adequately train a model. Obtaining tagged data, on the other hand, might be costly, time-consuming, or difficult in some sectors. Instead of depending entirely on a pre-labeled dataset, AL allows the model to actively query the labels for the most informative occurrences.

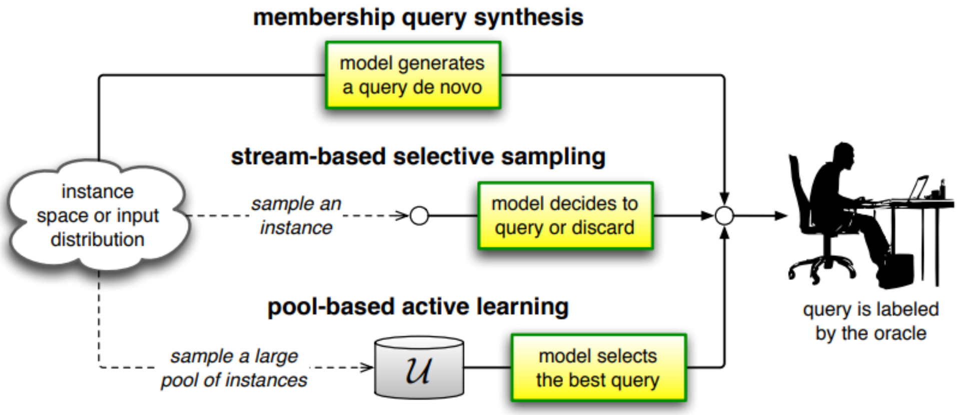
membership query synthesis
model generates a query de novo
stream-based selective sampling
instance space or input distribution
sample an instance
model decides to query or discard
query is labeled by the oracle
pool-based active learning
sample a large pool of instances
$\mathcal{U}$
model selects the best query

Figure 2.2: A schematic representing the three primary AL possibilities. Adapted from [16].

In Figure 2.2, in the first scenario, the learner creates its own instances from the underlying dataset in membership query synthesis so that it can query the oracle of its contents. In terms of stream-based selective sampling, the learner draws each unlabeled instance one at a time and chooses whether or not to query it. In the last scenario, unlike stream-based selective sampling, pool-based AL makes use of the complete pool of unlabeled cases [16]. In this configuration, query instances are selected from the unlabeled pool, and a small number of labeled instances are presumed to be provided with a much larger number of unlabeled instances [18].

In AL, there are numerous query strategies that can be employed, such as uncertainty sampling, diverse sampling.

AL plays a pivotal role in addressing the critical challenge of limited labeled medical data. The scarcity of labeled medical datasets, a common issue in healthcare applications, often hinders the development of accurate ML models. Through this integration of active learning, a substantial improvement is anticipated in the efficiency and cost-effectiveness of medical classification, ultimately benefiting patients and healthcare providers. The interplay of AL in this thesis aligns with the broader trend of harnessing AI techniques to assist medical professionals in making more accurate and timely decisions.

2.1.4 Query Strategies

The techniques used to choose representative and instructive samples from an unlabeled dataset for annotation or labeling are referred to as query strategies in AL. The purpose of query techniques is to deliberately select examples that, when labeled, would most significantly enhance a ML model's performance. AL is inherently designed to make the most efficient use of the limited labeling resources available. The query strategy that is selected has a direct effect on how well the model can learn from a limited number of labeled instances. In a domain where acquiring medical labels can be resource-intensive and time-consuming, optimizing the selection of instances for labeling is crucial.

In AL, there are numerous query strategies that can be applied, including predicted model change, query-by-committee, uncertainty sampling, diversity sampling and others.

2.1.4.1 Uncertaintly Samping

Uncertainty sampling is AL query approach that selects instances based on how confident the model's predictions are. Since they can produce the most informative data for improving the model's performance, the most ambiguous or confusing instances of the model are given precedence for labeling. This query approach seeks to focus on challenging examples that the model is least certain about by selecting instances with significant uncertainty. These occurrences have the ability to provide useful information for improving the model's performance in uncertain parts of the feature space. Some of the approaches to uncertainty sampling are the following.

Least Confidence sampling

Least confidence sampling refers to the distinction between the highest level of confidence in a prediction and complete certainty, denoting the degree of uncertainty in a given scenario. Simply the probability of the label's highest level of confidence makes up the fundamental equation:

$$\phi_{LC}(x) = P_\theta(y^* \mid x) \tag{2.2}$$

Instead of relying solely on confidence levels, it proposes converting uncertainty scores to a 0–1 scale, with 1 indicating the highest uncertainty. A normalization procedure is used to do this: the number of labels is multiplied by the confidence score, which is then subtracted from 1, and the result is divided by the number of labels minus 1. When all labels have an equal expected confidence, this normalization assures that the minimal confidence value is never less than 1 divided by the number of labels. This method, which is referred to as least confidence sampling with a 0–1 range, makes it easier to assess uncertainty more precisely [19].

$$\phi_{LC}(x) = (1 - P_\theta(y^* \mid x)) \frac{n}{n-1} \tag{2.3}$$

Entropy Sampling

AL frequently employs the query technique known as entropy sampling, in which cases with the highest entropy are chosen for labeling. In the context of AL, entropy, a measure of uncertainty, represents the amount of information or unpredictability connected to a label for an occurrence.

A sample's entropy is determined using the model's expected probabilities for various classes. A model that has higher entropy is a good candidate for querying since it means the model is unsure of the proper label for that particular instance. AL prioritizes the labeling of samples that are most likely to enhance the performance of the model by choosing instances with high entropy.

When entropy is applied to a probability distribution, the negative sum of each probability's logarithmic multiples is obtained [19]:

$$\phi_{ENT}(x) = -\sum_{y} P_\theta(y|x) \log_2 P_\theta(y|x) \tag{2.4}$$

and it is also possible to divide the entropy by the log of the number of predictions (labels) to convert it to a 0–1 range:

$$\phi_{ENT}(x) = \frac{-\sum_y P_\theta(y|x) \log_2 P_\theta(y|x)}{\log_2(n)} \tag{2.5}$$

Margin Sampling

Another popular query technique in AL is margin sampling, which chooses instances depending on the degree of confidence between the model's two top predicted classes. Instances near the decision border, where the model is unsure of the proper label, are prioritized.

To determine how much more confident the label that the model predicted versus the next-most-confident label, the following formula is used, already on a range of 0 to 1 [19]:

$$\phi_{MC}(x) = 1 - (P_\theta(y_1^* \mid x) - P_\theta(y_2^* \mid x)) \tag{2.6}$$

2.1.4.2 Diversity Sampling

There are various approaches when using diversity sampling, like Model-based outlier, Cluster-based sampling, Representative sampling, and others. Those approaches focus on enriching the training dataset with diverse examples to improve model generalization.

2.1.4.3 Query by Committee

The Query-By-Committee (QBC) framework [16] for AI involves maintaining a committee $C = \{\theta(1), \ldots, \theta(C)\}$ of models that have all been trained on the current labeled set while simultaneously representing possibly opposing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The degree of disagreement among the committee is used to determine how unsure an instance is that has not been labeled. The instance in which they disagree the most is considered to be the most informative query [16, 18].

The committee size can have an impact on the results of the QBC algorithm. The generalization error lowers and the asymptotic information gain rises with committee size. Beyond a certain point, there is a diminishing return on expanding the committee [20].The literature is divided on the ideal committee size to employ, which can actually differ depending on the model class or application. On the other hand, it has been demonstrated that small committee sizes, such as two or three, actually work well [16].

Two major methods have been suggested for calculating the degree of disagreement [18]: vote entropy and disagreement margin. Vote entropy [21], which can be considered a generalization of entropy-based uncertainty by committee, represents degrees of disagreement on class estimate across a committee of classifiers to evaluate uncertainty. When the class probability estimates are equally likely, it displays the greatest value; when the estimates are unanimous, it displays the smallest value. To calculate the vote entropy equation 2.7 is used:

$$fVE(x) = -\sum_{y \in Y} \frac{VC(y,x)}{|C|} \log \frac{VC(y,x)}{|C|}, \tag{2.7}$$

where y represents a class label, $VC(y,x) = P_{\theta \in C} I[h_\theta(x) = y]$ signifies the count of votes received by label y for instance x from the hypotheses within the committee C, and $|C|$ denotes the size of the committee. By converting each committee's votes to class probability estimates, Settles [17] produced a more softly defined vote entropy:

$$fSVE(x) = -\sum_{y \in Y} PC(y|x) \log PC(y|x), \tag{2.8}$$

where $PC(y|x) = \frac{1}{|C|} \sum_{\theta \in C} P_\theta(y|x)$, represents the average probability that the label of instance x is predicted as y by the committee members.

The difference in votes between the most confident class estimate and the second most of an instance is measured by the disagreement margin [18, 22]:

$$fDM(x) = VC(y^*, x) - \max_{y \in (Y - \{y^*\})} VC(y_0, x), \tag{2.9}$$

where $\arg\max_{y \in Y} VC(y,x) = y^*$. Similarly, one can utilize the difference between the probabilities of the two largest classes. Disagreement margins, in contrast to vote entropy measures, indicate the confidence of the class estimate, therefore, their values suggest the opposite. The margins represent the highest value when all class probability estimates are in agreement, and the lowest value when the probability estimates are evenly distributed among classes. Consequently, when a disagreement margin is employed, an additional evaluation function may be necessary. This function would rank instances higher for selection when the classifiers are less certain about the class of that instance [18].

On another disagreement measure, Settles [17] points the Kullback-Leibler (KL) divergence:

$$x^*_{KL} = \arg\max_x \frac{1}{|C|} \sum_{\theta \in C} KL\left(P_\theta(Y|x) \middle\| PC(Y|x)\right), \tag{2.10}$$

where KL is defined to be:

$$KL\left(P_\theta(Y|x) \middle\| PC(Y|x)\right) = \sum_y P_\theta(y|x) \log \frac{P_\theta(y|x)}{PC(y|x)}, \tag{2.11}$$

where the aim is to measure disagreement as the mean divergence between the predictions made by each committee member, represented by θ, and the consensus predictions of the committee C.

2.2 AI in Healthcare

The pursuit of knowledge and innovation has always been a commitment in the field of medicine and healthcare, where every choice made can mean the difference between life and death. The advent of AI has signaled a paradigm shift for humanity as it has struggled to understand the complexities of medical conditions and the complexity of the human body [23]. The recent successful uses of AI in healthcare have been made possible by the expanding availability of healthcare data and the quick development of big data analytic techniques. Vast amounts of data contain clinically relevant information that can be unlocked by powerful AI techniques guided by pertinent clinical questions [23].

There are various applications of AI in the medical field, as substantiated by the findings in [24]. Among those are:

- Providing online services to patients and checking patients without visiting clinics/hospitals.
- Planning treatment to achieve better results
- Developing targeted treatments, uniquely composed drugs, and personalized therapies.
- Guiding surgeons during medication, treatment, and operation
- Tracking, detecting, investigating, and controlling infections in hospitals
- Improving clinical decision-making to enhance patient outcomes

- Acquiring information using neural networking, advanced imaging, and natural language processing

2.3 Responsible and Trustworthy AI

2.3.1 Ethics

The application of AI in healthcare carries a number of ethical considerations. Transparency is one of the hardest problems to solve because a lot of the AI algorithms used in image analysis are so hard to understand or interpret [25]. Furthermore, it's likely that AI systems will make mistakes when diagnosing and treating patients, and it might be challenging to hold them accountable [25]. In a recent review [26], suggests that in order to ensure that healthcare systems are held responsible for providing high-quality, equitable, and safe care, appropriate governance models can be developed.

As we explore the exciting world of AI in healthcare, it's critical to acknowledge and resolve a number of important issues that come with this development in technology.

2.3.1.1 Data Privacy

Protecting patient data is a top priority in the field of healthcare AI, which has led to a number of data privacy issues. The data is collected by privacy-invasive social media platforms, smartphone apps, and Internet of Things devices with countless sensors [27].

The effective operation of AI systems necessitates access to vast quantities of patient data, giving rise to concerns regarding the collection, storage, and utilization of said data. Specifically, there is a chance that private patient information could be accessed or misused, either by criminals or by well-meaning but inadequately qualified medical personnel [26] .

Even with the availability of privacy-friendly methods, data privacy is still greatly concerned about the gathering and use of personal information in AI systems [27].

2.3.1.2 Bias and Fairness

Hidden within the algorithms and data that power AI systems is the potential for bias, which can fundamentally affect fairness and equity in healthcare outcomes.

Racial bias in predictive policing algorithms is one well-researched instance of reported bias in AI. According to the study reported in [28], there exists prejudice in the algorithms towards specific racial groups, which may result in unjust treatment of persons inside the criminal justice system. In addition, there are concerns about the accountability of firms for mistakes made by their algorithms, as well as the need for a moral codex of AI engineers [1].

One solution to these issues could be to develop commonly accepted requirements regarding the training and testing of AI algorithms, possibly in combination with some form of warranty, comparable to consumer and safety testing procedures for tangible goods [1]. AI may fulfill its greatest potential by assisting in better diagnosis and prediction while safeguarding patients if bias is addressed [29].

2.3.1.3 Transparency

In AI, transparency pertains to the capacity to comprehend and clarify the decision-making and action-taking processes of AI systems. It involves allowing users, stakeholders, and regulators to see and comprehend the workings and results of AI systems. Since transparency may promote ethical standards, accountability, and trust in the creation and application of AI systems, it is regarded as a crucial component of trustworthy AI.

2.3.2 Explainable Learning

XAI refers to the development of artificial intelligence systems that can provide clear and understandable explanations of their decision-making processes to human users. The goal of XAI is to make AI more understandable, trustworthy, and transparent so that people can have more confidence in the decisions made by AI systems [30]. Nevertheless, while developing explanations, it is necessary to strike a balance between technical correctness and understandability, as various users will require varying levels of complexity [31]. In addition to the well-known LIME, several other prominent XAI models have emerged, each with its unique approach to enhancing model interpretability. For example, SHapley Additive exPlanations (SHAP) uses cooperative game theory to distribute feature importance fairly [32]. By creating instances with modified results, Counterfactual Explanations helps users comprehend how modifications to input data affect model predictions [33]. On the other hand, Anchors provide consistent, rule-based explanations for ML models [34].

In the XAI field, Khishigsuren Davagdorj et al [35] presents an XAI framework for early prediction of Non-Communicable Diseases (NCD)s. It incorporates DeepSHAP-based feature selection and it addresses the need for interpretable models in healthcare decision-making, providing clinicians with transparent and understandable insights to facilitate informed decisions. The framework's ability to offer explanations at both global and local levels enhances its effectiveness in supporting clinicians in predicting NCDs.

In another study, Khodabandehloo et la. [36] delves into the application of collaborative and transparent AI to facilitate the early detection of cognitive decline among elderly individuals. The research introduces a sensor-driven system named HealthXAI, focused on recognizing activities, which leverages pervasive healthcare technology to identify cognitive decline symptoms and enable timely interventions. HealthXAI is rooted in established clinical parameters and aims for scalability and personalization. The authors conducted a user study involving healthcare professionals and observed that HealthXAI outperforms existing AI systems in aiding clinicians in diagnosis. Moreover, the system's explanations were regarded as comprehensible, valuable, and sufficiently detailed. Nevertheless, a notable constraint of the system is the absence of a comprehensive clinical model, a point highlighted by the authors. It was referenced that future research should concentrate on enhancing the system's precision by addressing subtler inefficiencies and considering alternative sensor infrastructures.

Also on the XAI field, this time using LIME as well, Magesh et la. [37] developed a machine learning model for the early diagnosis of Parkinson's Disease using DaTSCAN images, while simultaneously providing transparent and interpretable insights into the decision-making process of the model. LIME is specifically used to highlight aberrant or diminished characteristics in the putamen and caudate areas of non-PD patients, providing valuable insights into classification judgments. This interpretability provided a deeper understanding of the model's decision-making process, enabling medical practitioners to obtain significant insights into the model's visual markings on the forecasts. Ultimately, the use of LIME helped to achieve the research's primary objective of providing an interpretable solution for the early diagnosis of Parkinson's Disease, allowing healthcare professionals to make informed decisions based on the model's transparent and explainable predictions.

2.3.3 LIME

LIME is a novel explanation technique that provides interpretable and accurate explanations for any classifier. LIME works by using an interpretable model to locally approximate the predictions of a black-box model. In order to achieve this, a set of perturbed instances is created around the original instance, and these perturbed instances are used to train a local interpretable model. The prediction made by the initial black-box model for that instance can then be explained using the resulting interpretable model [38].

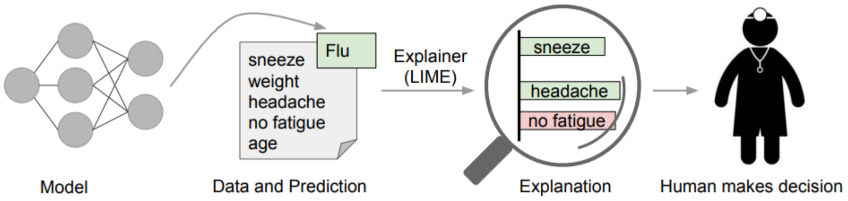

Figure 2.3: LIME Explanation adapted from [38].

Figure 2.3 depicts the process of explaining individual predictions, which is a crucial aspect of getting humans to trust and use machine learning models effectively. The figure illustrates an example of a doctor using a machine learning model to predict whether a patient has a certain disease. The model provides an explanation in the form of a list of symptoms with relative weights, where symptoms that contribute to the prediction are highlighted in green, and symptoms that are evidence against it are highlighted in red.

The model being explained is denoted as $f : \mathbb{R}^d \rightarrow \mathbb{R}$, where $\mathbb{R}^d$ represents the d-dimensional real space and $\mathbb{R}$ represents the real numbers. The probability (or binary indicator) that x belongs to a particular class is expressed as $f(x)$ in classification. As a result, the model f produces a real number that indicates the likelihood that an input instance, x, belongs to a particular class given the input instance.

The features that are employed in datasets are frequently difficult for humans to comprehend, as the authors defend. Therefore, in the quest for interpretability, the authors proposed a binary representation of feature attributes. In text-based datasets, this binary representation denotes if a word is present or not, while in image-based datasets, it represents a super-pixel. Thus, a binary vector was introduced $x' \in \{0,1\}^{d'}$ as an interpretable representation of x. In simpler terms, for any instance x that needs an explanation, there is an associated x'.

The explanation model, characterized by a domain $\{0,1\}^{d'}$, is represented by $g \in G$, where G encompasses a set of possible XAI models. To gauge the complexity of an explanation produced by g, Ribeiro et al. [38] introduce $\Omega(g)$.

The model randomly samples instances around x' represented as $z' \in \{0,1\}^{d'}$. These perturbed instances makes it possible to recover the original values of z in R^d. These recovered values, in turn, help determine $f(z)$, which represents the probability of an instance belonging to a specific class. These samples are assigned weights based on $\pi_x(z)$, serving as a measure of the similarity between the original instance x and the sampled instance z. By introducing these perturbations, the model aims to observe changes in predictions and, subsequently, identify which attributes have the greatest impact on the model's classifications.

It is necessary to minimize $L(f, g, \pi_x)$ while keeping $\Omega(g)$ low enough for human interpretation in order to guarantee both interpretability and local fidelity.

The following Formula 2.12 provides the LIME-produced explanation:

$$\xi(x) = \arg\min_{g \in G} \left[L(f, g, \pi(x)) + \Omega(g)\right] \tag{2.12}$$

2.4 Performance Evaluation

Different assessment criteria have been used to evaluate ML classification models, and in order to choose the best one, it is important to consider the main objectives of the hospitals and healthcare facilities where the models will be used. Additionally, it is essential to explore and comprehend the data. For the type of data on which this project is working, there is a lot of class imbalance. As a result, the measures selected must be appropriate for a proper assessment in this situation.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2.13}$$

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2.14}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{2.15}$$

Balanced accuracy was also calculated due to the fact that is a machine learning error metric for binary and multi-class classification models. It is an improvement over the common accuracy measure in that it has been tweaked to work better on unbalanced datasets, which is one of the main compromises when employing the accuracy metric.

The imbalance problem occurs when one class in a dataset with two classes is significantly underrepresented in comparison to the other class. As many techniques presuppose that the problem classes share comparable prior probabilities, this can have a substantial impact on learning approaches. However, this presumption is frequently broken in real-world problems, and the ratios of prior probabilities between classes can be seriously skewed [39]. This is crucial in situations where misclassifying examples from the minority class would be expensive. Therefore, in order to enhance the performance of learning systems on these kinds of datasets, the imbalance problem needs to be solved [39]. Even in cases when the classifier is not biased in favor of the class that is more frequently observed, balanced accuracy can be employed as a general safeguard against presenting an optimistic accuracy estimate [40].

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{2.16}$$

The harmonic mean of accuracy and recall for the minority positive class is used to determine the F1 score, sometimes referred to as the F-measure or balanced F-score, which is an error metric used to evaluate the performance of models.

As it offers accurate results for both balanced and imbalanced datasets and takes into account the precision and recall ability of the model, it is a common statistic to employ for classification models.

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.17}$$

2.5 Triage Scales

Triage scales in AI in healthcare refer to a set of tools and algorithms that help prioritize and categorize patients based on the severity of their medical conditions, the urgency of their care needs, and the available healthcare resources. These scales guarantee that patients with the most serious conditions get the attention they need right away, which helps medical professionals make quick decisions, especially in emergency and high-demand situations [41].

Emergency Department (ED) triage has undergone significant evolution since its introduction in the 1950s in the United States, with the development of various triage scales designed to guide the decision-making process for assessing patient acuity. Over time, different triage scales, such as the Emergency Severity Index (ESI), the Manchester Triage System (MTS), and the Canadian Triage and Acuity Scale (CTAS), have been developed and tested for reliability and validity. The five-level triage system, developed by the American Collge of Emergency Physicians (ACEP) and the Emergency Nurses Association (ENA) in 2005, represents the latest iteration of ED triage. The need to guarantee that patients receive immediate and appropriate therapy and to prioritize patient care depending on the severity of their condition has motivated the development of ED triage [41].

Emergency triage is a systematic process that classifies patients into distinct categories or priority levels based on the urgency of their medical condition. Below are some of the typical categories used in emergency triage [42]:

1. **Critical**: Patients in this category require immediate medical attention. They often have life-threatening conditions or severe injuries that demand rapid intervention. Examples include cardiac arrest, severe trauma, or respiratory distress.

2. **Emergent/Urgent**: These patients have serious medical conditions that require immediate attention. Delays could cause their diseases to worsen, even though they may not be immediately life-threatening. Chest pain, fractures, or serious infections are a few examples.

3. **Non-Urgent/Non-Emergent**: Patients in this group have less severe conditions that can be managed with less urgency. They typically experience less discomfort or distress. Some of the examples are minor injuries, rashes, or mild respiratory infections.

2.5.1 CTAS Triage Scale

CTAS is a widely adopted triage system used in ED across Canada and abroad. It is designed to prioritize patient care based on the severity of their condition and the resources required to manage it. The CTAS assigns patients to one of five triage categories based on a set of criteria, including the patient's presenting complaint, vital signs, level of consciousness, and other clinical indicators. The five triage categories range from level 1 (resuscitation) to level 5 (non-urgent), with each level corresponding to a specific set of clinical criteria and recommended response times. In order to help nurses assign appropriate acuity levels while accounting for aspects like age, mobility, and pain, the CTAS also includes a set of modifiers [43].

2.5.2 KTAS Triage Scale

The Korean Triage and Acuity Scale (KTAS) triage scale, which has five levels, was created in 2012 and is based on the CTAS but has been adjusted to fit the needs of South Korean healthcare. KTAS is a classification system focused on symptoms that is used to assess patients [44, 45]. Emergency nurses evaluate the primary considerations, such as vital signs, pain score, hemorrhagic disease, mechanism of injury, and secondary concerns, such as blood glucose level and degree of dehydration, to establish the triage level following the critical first-look evaluation [44, 45]. The patient waiting period before being evaluated by an ED doctor and the appropriate medical treatment price are both determined by the KTAS score given by the emergency nurses (emergency, KTAS 1-3 level, or non-emergency, KTAS 4-5 level) [44, 45].

2.5.3 Challenges in Triage and the Role of AI

Manual triage plays a crucial role in the initial assessment and prioritization of patients in emergency departments, providing the foundation for timely and appropriate medical care. This essential component of emergency medicine doesn't come without difficulties, though. Because it depends on the assessment and classification of patients by medical professionals according to their perceived level of acuity, the manual triage process is fundamentally subjective. Subjectivity can lead to differences in how triage levels are assigned, which may have an impact on patient outcomes. Bijani et al. [46] identified several challenges that affect the quality of triage in emergency departments, including a lack of clinical competency and psychological capabilities among triage nurses, inadequate staffing, high workload, inadequate training, and poor communication and collaboration among healthcare providers.

Additionally, the exigencies of time constraints in busy emergency settings demand rapid decision-making, leaving limited opportunity for thorough assessment. For example, the implementation of AI virtual assistants in healthcare systems globally [47] presents both potential benefits and challenges. On the one hand, such technology has the potential to reduce costs and improve access to healthcare in resource-poor settings. However, great care must be taken to ensure that algorithms are fair and generalize to different subsets of the population. In particular, it is important to ensure that the AI system is able to service the needs of patients in diverse regions and segments of the population, as explained before at 2.3.1.2.

AI in emergency triage has the potential to improve decision-making processes' speed and accuracy, resulting in the prompt delivery of critical care to the most vulnerable. This combination of AI and emergency medical services not only highlights AI's adaptability but also suggests a possible paradigm shift in the way urgent medical interventions are handled.

According to their investigation, Marta Fernandes et al. [48], the goal of the study was to create models that predict the likelihood of ICU admission during emergency department triage in two hospitals, one in Portugal and one in the US. The patient's primary complaint and a variety of clinical characteristics were used to create the models. The study employed machine learning methods such as decision tree boosting with random undersampling, random forest regression, and regularized logistic regression.

In their study, Dror Zmiri et al. [49] examined the viability of grading the severity of emergency department patients using data mining techniques. The goal of the project was to develop Naive Bayes and C4.5 classifiers from patient data into severity grades. The outcomes of the classifiers were then compared with the medical professionals' opinions, a classifier chosen at random, and a classifier that chose the class with the highest prevalence.

Jamie Miles et al. [50] did a systematic review that evaluates the efficacy of machine learning techniques for first triaging incoming patients according to their level of acuity. Four phases of screening based on inclusion and exclusion criteria were used in the study selection process, and the included papers were too heterogeneous to allow for a thorough meta-analysis. As discrimination was the most frequently reported summary statistic of model performance, a narrative synthesis was carried out. The included models were divided into subgroups based on the method and the outcome, and the median and IQR were used to show how widely distributed the C-statistics were for each method.

In another study, Hong et al. [51] aimed to develop a machine learning model to predict hospital admission at emergency department ED triage using patient history and triage information. The study used retrospective data from three EDs covering the period of March 2013 to July 2017, including a level I trauma center, a community hospital-based department, and a suburban, free-standing department. The study employed gradient boosting and deep neural networks, two of the best performing algorithms in classification tasks, to model the nonlinear relationships among the variables. The study also identified variables of importance using information gain as the metric and presented a low-dimensional model amenable to implementation as clinical decision support. The results showed that the model achieved high accuracy, sensitivity, and specificity in predicting hospital admission at ED triage. However, the study acknowledged the limitations of using the ED provider's prior decision as the true label and the need for further research into a gold-standard metric for hospital admission. The study also highlighted the importance of analyzing methods of implementation and their effect on patient outcomes. Overall, this study provides valuable insights into the use of machine learning models for predicting hospital admission at ED triage and lays the groundwork for future research in this area.

3

Material and Methods

3.1 Pipeline Overview

Data selection, pre-processing, defining query strategies, and performance assessment are examples of steps that are commonly included in a typical AL framework. In addition to these steps, this thesis also includes an application of the LIME model, in order to apply explanations into the framework. To asset the performance evaluation various performance metrics were calculated to each combination of classifiers, query strategies and number of estimators.

The experimental framework is designed to systematically explore and enhance the performance and interpretability of the ML model. The process goes through a number of crucial phases:

- **Data Pre-processing:** The initial phase involves enhancing the quality of the data through preprocessing steps. This includes tasks such as cleaning, normalization, and feature engineering to ensure the dataset is suitable for subsequent modeling.
- **AL Model:** The core of the framework involves the application of ML techniques. An ensemble model is formed using learners with different configurations, classifiers like LR, RF, and their combination. These models are iteratively queried and updated, incorporating uncertainty measures from diverse perspectives.
- **XAI Model:** This step involves generating explanations for the model's predictions, providing insights into the decision-making process of the AI system.
- **Performance Evaluation:** The final step focuses on evaluating the performance of the AL model. Various evaluation metrics, including accuracy, sensitivity, specificity, balanced accuracy, and F1 score, are employed to comprehensively assess the model's effectiveness in the classification task.

3.2 Datasets

3.2.1 KTAS Dataset

The first dataset was used to identify emergency department triage accuracy using the KTAS and evaluate the causes of mistriage.

The data for this cross-sectional retrospective analysis came from 1267 carefully chosen records of adult patients admitted to two emergency rooms between October 2016 and September 2017. Chief complaints, vital signs recorded in the first nursing records, clinical outcomes, and other factors were among the twenty-four that were evaluated. The actual KTAS was determined by three triage experts: a qualified emergency nurse, a KTAS provider and instructor, and a nurse who was highly recommended based on her exceptional emergency department experience and skill [52].

Table 3.1: Features of the KTAS dataset.

Variables	**Description**
Sex	Sex of the patient
Age	Age of the patient
Arrival mode	Type of transportation to the hospital
Injury	Whether or not the patient is hurt
Chief Complain	The patient's complaint
Mental	The mental state of the patient
Pain	Whether the patient has pain
NRS Pain	Nurse's assessment of pain for the patient
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
HR	Heart Rate
RR	Respiratory Rate
BT	Body Temperature

3.2.1.1 Preprocessing of KTAS dataset

A number of operations were carried out to enhance and change the dataset during the preprocessing stage. To make the data more understandable, categorical variables were reclassified to create more relevant labels. Data cleaning methods were used to manage '?'-denoted invalid entries, preserving the consistency of the data. To develop new features that accurately represented crucial components of the data, including age groups, risk levels, and disease classifications, feature engineering approaches were used. These additional features helped arrange the data for the following experiments and offered insightful information about various danger levels.

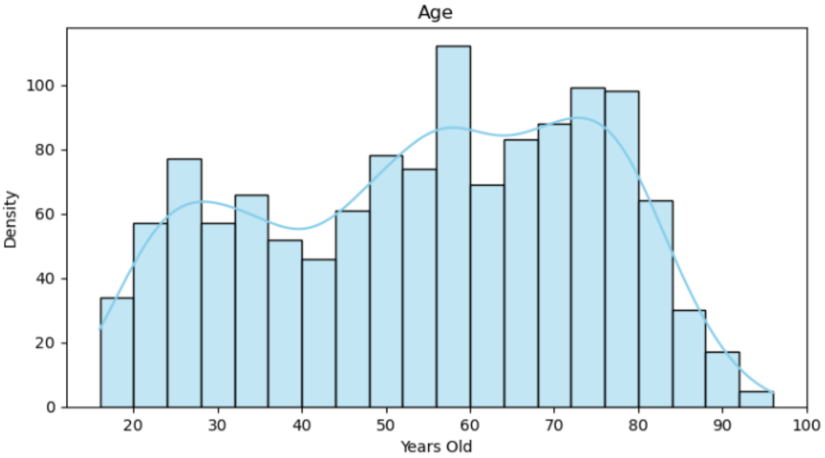

Figure 3.1: Age distribution of the dataset's patients

3.2.2 Patient Priority Dataset

The Patient Priority dataset is a comprehensive collection focused on the urgent prioritization of medical care based on patient symptoms. The dataset amalgamates three distinct datasets, each addressing urgent illnesses or injuries, and combines their features to discern the most critical symptoms for each condition. The primary objective is to identify severe symptoms and allocate patients to predesignated care areas, ensuring a prioritized and swift initiation of diagnostic and therapeutic measures [53].

Triage serves as a critical mechanism for efficient emergency response, necessitating the rapid and accurate prioritization of patients based on their medical conditions. The dataset incorporates the following categories for prioritization:

- **Red:** Immediate attention for critical, life-threatening injuries or illnesses.
- **Yellow:** Serious injuries requiring prompt attention, often prioritized for transport due to their potential for recovery.
- **Green:** Less serious or minor injuries that are non-life-threatening, allowing for delayed transport.
- **Black:** Deceased or mortally wounded, signifying a lower priority for those beyond help.
- **White:** No injury or illness (optional category, not universally used).

Table 3.2: Features of the Patient Priority Dataset.

Feature	**Description**
age	Age of the individual
gender	Gender of the individual
chest pain type	Type of chest pain
blood pressure	Blood pressure of the individual
cholesterol	Serum cholesterol level
max heart rate	Maximum heart rate achieved
exercise angina	Exercise-induced angina
plasma glucose	Plasma glucose level
skin thickness	Skin thickness
insulin	Insulin level
bmi	Body Mass Index (BMI)
diabetes pedigree	Diabetes pedigree function
hypertension	Presence of hypertension
heart disease	Presence of heart disease
Residence type	Type of residence
smoking status	Smoking status
triage	Triage level

3.2.3 Heart Disease Dataset

The Heart Disease dataset is a compilation of data from four databases: Cleveland, Hungary, Switzerland, and the VA Long Beach. This multivariate dataset is characterized by its relevance to the field of health and medicine, specifically tailored for classification tasks related to heart diseases [54].

Table 3.3: Features of the Heart Disease Dataset.

Variable Name	Description
age	Age of the patient
sex	Sex of the patient
cp	Chest pain type
trestbps	Resting blood pressure (on admission to the hospital)
chol	Serum cholesterol
fbs	Fasting blood sugar (> 120 mg/dl)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise induced angina
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment
ca	Number of major vessels colored by fluoroscopy
thal	Thalassemia type
num	Diagnosis of heart disease

The purpose of this final dataset [54] is to examine how the framework performs in a different medical context by stepping back from the patient emergency triage area slightly.

3.3 Project's Framework

Algorithm 1 Active Learning Framework

```
Require: $X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}}$
Convert $X_{\text{train}}$ and $X_{\text{test}}$ to NumPy arrays
for all $n_{\text{estimators}}$ in $\{1, 2, 3, 4, 5, 10, 50\}$ do
  for all Learner configurations do
    for all Query strategies in {entropy_sampling, margin_sampling, uncertainty_sampling}
    do
      Define learner with specified classifiers and estimators
      Initialize ActiveLearner with learner, query strategy, and training data
      Query and update the model iteratively
      for all Number of Queries do
        Evaluate and store performance metrics
        Remove queried instance from the pool
      end for
      Generate and save explanations for model predictions
    end for
  end for
end for
return =0
```

Algorithm 1 provides the AL framework aimed for ML model improvement. The approach is designed to take advantage of AL principles, with the goal of improving model performance through the careful selection of cases for annotation. The training and testing datasets, labeled as $Xtrain$, $Xtest$, $ytrain$, and $ytest$, are converted into NumPy arrays for computational efficiency during initialization. The framework is designed to be adaptable, allowing for the investigation of alternative learner structures and query methodologies. The framework systematically explores learner configurations across a range of estimators sizes ($n\_estimators$), examining combinations of LR, RF, and a hybrid of both, building learners compositions. While LR is not traditionally considered when applying different sizes for estimators, LR was used in this framework for consistency in testing different query strategies. In each setting, the system evaluates multiple query strategies, such as least confidence, margin, and entropy sampling. By using these techniques, the ActiveLearner consistently finds examples that need to be annotated, systematically enhancing the training set. The core of the framework is the ActiveLearner, initialized with the designated number of estimators, query strategy, and training data. Through iterative querying, the model adapts by incorporating newly annotated instances. Key performance metrics, encompassing accuracy, sensitivity, specificity, balanced accuracy, and F1-score, are systematically assessed and stored. Furthermore, to improve interpretability at the instance level, the framework integrates LimeTabularExplainer to produce local explanations for model predictions.

3.3.1 Model Parameters

The variance in the number of estimators is intended to investigate the relationship between estimators size and the AL framework's efficacy. As was previously mentioned, research has demonstrated that smaller estimators' size typically result in higher model accuracy and efficiency, which can be linked to better generalization and lower complexity. A comparative analysis is made possible by the addition of a larger number of estimators, 50 in this case, which sheds light on the scalability and possible advantages of a larger group of estimators.

Table 3.4: Learner Configurations for Active Learning

Learner Configuration	Classifiers	Number of Estimators
`learner1`	Logistic Regression	NA
`learner2`	Random Forest	1, 2, 3, 4, 5, 10, 50
`learner3`	Logistic Regression + Random Forest	1, 2, 3, 4, 5, 10, 50

This table 3.4 outlines the different learner configurations employed in the ALframework. The framework utilizes three distinct learner configurations, each consisting of specific classifiers and varying numbers of estimators for the RF and combined classifier.

Hyperparameters for Logistic Regression Classifier:

- **random_state**: This is a seed for the random number generator. It ensures reproducibility of results. Reproducibility is crucial for scientific rigor and to be possible to recreate the exact experiments, setting a random seed helps achieve this.
- **max_iter**: LR is an iterative algorithm, and max_iter specifies the maximum number of

iterations for the solver to converge. In this case, LR might not converge within the default number of iterations, so increasing max_iter helps ensure convergence.

Hyperparameters for Random Forest Classifier:

- **n_estimators**: : This parameter defines the number of trees in the RF. Increasing the number of trees generally improves performance up to a point, as it helps the model generalize better.
- **max_iter**: Similar to LR, it ensures that the randomization in the RF is reproducible.

3.3.2 LIME Model

In the context of this thesis, LIME is integrated into the AL framework to provide transparent and interpretable insights into the decision boundaries of the ML models. This integration allows for a deeper understanding of the AL process by visualizing and analyzing the regions of the feature space that significantly influence model predictions.

To implement it, the framework employs the function LimeTabularExplainer from the LIME library [55]. Using a perturbation-based methodology, this function modifies the input features of a given instance slightly to produce locally faithful approximations of the model's decision boundaries. Then, using the perturbed instances as a basis, it builds a surrogate interpretable model, which is typically a linear model, to explain the model's behavior around the instance in question.

Hyperparameters for LIME:

- **mode**: This parameter sets the mode of the explainer, and in this case, it's set to 'classification' because the framework is dealing with a classification problem. LIME is an explainability technique, and in the context of this thesis, it helps in understanding why the model is making certain predictions. This interpretability is crucial in healthcare applications, where decisions need to be explainable and justifiable.
- **feature_names**: LIME requires the names of features to provide meaningful explanations. In this thesis, where medical data is being used, having feature names is crucial for interpreting and explaining model decisions.

4

Results and Discussion

4.1 KTAS Dataset Results

4.1.1 Active Learning Results

Table 4.1: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 1 estimator

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.952
	RF	0.815	0.815	0.830	0.409	0.822
	LR + RF	0.866	0.866	0.863	0.511	0.864
Margin	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.815	0.815	0.830	0.409	0.822
	LR + RF	0.858	0.858	0.857	0.484	0.856
Least confidence	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.815	0.815	0.830	0.409	0.822
	LR + RF	0.874	0.874	0.880	0.564	0.877

In Table 4.1, where a single classifier was employed, LR exhibited commendable results across all query strategies. Under the entropy query strategy, LR demonstrated an impressive accuracy of 0.961 and sensitivity/specificity of 0.961, although the balanced accuracy was comparatively lower at 0.659. This suggests that LR excelled in making accurate classifications but faced challenges in handling imbalanced data. Conversely, the RF classifier exhibited lower performance, with an accuracy and sensitivity/specificty of 0.815, showcasing its struggle with the inherent characteristics of the dataset. LR + RF, a combination of classifiers, only marginally improved over the individual RF, with an accuracy of 0.866.

When LR switched to the margin query approach, it kept up its high levels of sensitivity, specificity and accuracy (0.969), with a minor improvement in balanced accuracy (0.754) over the entropy technique. The fact that RF's performance stayed close to that of the entropy approach suggests that the margin sampling technique had little effect on RF's performance.

With a estimators' size of 1, LR + RF once more produced results similar to RF, indicating that there was no discernible benefit to merging classifiers when using the margin technique.

As the best classifier under the Least confidence query approach, LR achieved the greatest balanced accuracy of 0.754 out of all the classifiers. With 0.815 accuracy and sensitivity, RF's performance was in line with the trends shown in the entropy and margin strategies. With the greatest accuracy (0.874) and balanced accuracy (0.564), LR + RF notably demonstrated some improvement in performance, demonstrating the combination's efficacy, when used in conjunction with the Least confidence sampling method.

The F1 Score for every query strategy and classifier showed excellent precision and recall balance in the estimators' size 1 scenario, when each classifier functioned independently. LR had the greatest F1 Score (0.965) under the uncertaintly and margin query technique, demonstrating a strong ability to strike a balance between precision and recall.

Table 4.2: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 2 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.951
	RF	0.878	0.878	0.881	0.454	0.863
	LR + RF	0.945	0.945	0.950	0.622	0.939
Margin	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.878	0.878	0.881	0.454	0.863
	LR + RF	0.965	0.965	0.963	0.752	0.961
Least confidence	LR	0.969	0.969	0.7969	0.754	0.965
	RF	0.878	0.878	0.881	0.454	0.863
	LR + RF	0.961	0.961	0.957	0.705	0.956

estimators' size 2 added a collaborative element that changed the performance dynamics from estimators' size 1, when classifiers functioned independently. As it can be seen in Table 4.2 LR values don't change, because the estimators' size doesn't affect this classifier. Meanwhile, the combined classifier, LR + RF, where the number of voters affect the outcome results, showed an increase in accuracy, confirming the benefit of cooperation. RF showed an improvement when comparing estimators' size 1, suggesting the possible advantages of a larger group of estimators.

After switching to the margin query approach, even though RF outperformed the entropy technique, its accuracy rate remained at 0.878. Surprisingly, LR + RF performed equally to the LR classifier, attaining an almost similar accuracy (0.965) and balanced accuracy (0.752).

LR remained dominant under the Least confidence query approach, achieving 0.969 accuracy and 0.969 sensitivity. RF performed consistently with varying the number of estimators, attaining an accuracy rate of 0.878. Under Least confidence, LR + RF outperformed solo classifiers with the highest accuracy (0.961) and balanced accuracy (0.705), confirming the value of collaboration.

Table 4.3: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 3 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.951
	RF	0.929	0.929	0.919	0.597	0.921
	LR + RF	0.953	0.953	0.955	0.698	0.947
Margin	LR	0.969	0.969	0.961	0.754	0.965
	RF	0.929	0.929	0.918	0.597	0.921
	LR + RF	0.965	0.965	0.965	0.719	0.959
Least confidence	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.929	0.929	0.918	0.597	0.921
	LR + RF	0.969	0.969	0.969	0.720	0.963

Adding another level of classifier collaboration to estimators' size 3 has resulted in some notable performance variations when compared to Estimators of Size 1 and 2, which can be seen in Table 4.3. The LR stays the best performing classifier but is worth mentioning that the RF classifier improved every accuracy to results above 0.9 across all classifiers. This indicates that the overall prediction accuracy is positively impacted by the additional classifiers. LR + RF achieved better results for accuracy under the entropy and Least confidence methods, which was a modest improvement above estimators' size 2's 0.945 accuracy. All F1 scores were high in this experience, with RF classifier results being the most noticeable improvement.

Table 4.4: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 4 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.950
	RF	0.906	0.906	0.894	0.572	0.895
	LR + RF	0.965	0.965	0.966	0.706	0.959
Margin	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.933	0.933	0.935	0.574	0.920
	LR + RF	0.972	0.972	0.973	0.768	0.969
Least confidence	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.937	0.937	0.938	0.622	0.928
	LR + RF	0.972	0.972	0.973	0.768	0.969

Using a bigger ensemble of classifiers in estimators' size 4, the AL framework shows interesting results under different query tactics. In Table 4.4, it shows that LR is no longer the best classifier. When using entropy sampling as the query strategy, RF got worst results than before. The major improvement in this estimators' size 4 was LR+RF, achieving an accuracy of 0.972 on both margin and Least confidence sampling. The balance accuracy was also improved to 0.768 as well as the f1 score of 0.969, which means this is the best combination so far in this framework results. estimators' size 4 continues the smaller estimators number trend where LR + RF performs better than individual classifiers. The efficacy of the collaborative approach is demonstrated by its accuracy, balanced accuracy, and F1 Score, underscoring its appropriateness for AL situations with a larger number of estimators.

Table 4.5: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 5 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.951
	RF	0.933	0.933	0.911	0.553	0.919
	LR + RF	0.969	0.969	0.969	0.754	0.965
Margin	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.961	0.961	0.962	0.659	0.951
	LR + RF	0.969	0.969	0.969	0.720	0.963
Least confidence	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.933	0.933	0.927	0.574	0.921
	LR + RF	0.972	0.972	0.973	0.768	0.969

In estimators' size 5 results, shown in Table 4.5, RF performs consistently but marginally worse than LR, which retains good accuracy, sensitivity, and specificity under all query strategies. The best accuracy, balanced accuracy, and F1 Score are constantly achieved by the ensemble LR + RF, outperforming individual classifiers and highlighting the effectiveness of ensemble techniques in active learning environments. The major values for this experience is the accuracy obtained by LR+RF of 0.972 when using Least confidence sampling as a query strategy, matching the previous results of estimators' size 5.

Table 4.6: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 10 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.951
	RF	0.945	0.945	0.948	0.637	0.936
	LR + RF	0.969	0.969	0.969	0.754	0.965
Margin	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.965	0.965	0.966	0.673	0.956
	LR + RF	0.969	0.969	0.969	0.754	0.965
Least confidence	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.961	0.961	0.962	0.659	0.951
	LR + RF	0.972	0.972	0.973	0.768	0.969

For estimators' size 10 in Table 4.6, RF classifier improved in terms of balanced accuracy but still falls short when comparing with the other two classifiers. The ensemble LR + RF consistently outperforms individual classifiers, but maintaining the same pinnacle in accuracy 0.972 and balanced accuracy 0.768 under the Least confidence sampling strategy. This suggests that, beyond a certain estimators' size, there might be diminishing returns in terms of classification performance. This implies that there may be diminishing returns in terms of categorization performance over a particular estimators' size.

Table 4.7: KTAS Dataset Performance Metrics' Results for all classifiers and query strategies using 50 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.961	0.961	0.962	0.659	0.951
	RF	0.976	0.976	0.977	0.782	0.973
	LR + RF	0.969	0.969	0.969	0.754	0.965
Margin	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.972	0.972	0.973	0.768	0.969
	LR + RF	0.969	0.969	0.969	0.754	0.965
Least confidence	LR	0.969	0.969	0.969	0.754	0.965
	RF	0.976	0.976	0.977	0.782	0.973
	LR + RF	0.972	0.972	0.973	0.768	0.969

The ensemble of classifiers in estimators' size 50 performs reliably and consistently when utilizing various query methodologies. On Table 4.7, LR results finally are surpassed making it the worst classifier in these terms. Interestingly, using either the entropy or Least confidence query approach, the RF classifier shines out with a notable improvement, obtaining an astounding

0.976 accuracy, 0.976 sensitivity, 0.977 specificity and 0.782 balanced accuracy. This indicates a significant improvement in RF's ability to record both favorable and unfavorable cases, particularly when the estimators' size is higher. Consistent performance is demonstrated by the LR + RF ensemble, which sustains a high degree of performance across various query strategies. Robust and balanced precision are maintained by LR, highlighting its stability in a variety of settings, but when the estimators' size got larger, other classifiers managed to surpass its performance. The LR + RF ensemble maintains its competitiveness, highlighting the fact that the notable improvements seen are more particular to the features of the RF classifier.

To summarize, the data indicates that while RF performs significantly better with bigger estimators' sizes, LR consistently performs well. The results show that the ensemble technique (LR + RF) performs better than any single classifier, highlighting the usefulness of this strategy in combining several classifiers to improve predictive power. Different classifiers have different effects from estimators' size, which emphasizes how crucial it is to take into account both estimators' size and classifier type for the best possible model performance.

4.1.2 Passive vs Active Study

In order to determine the relative effects of the Passive Learning and AL approaches on model performance, both were compared in this next study. Initially, the best-performing classifier from the prior AL results was trained using the complete labeled dataset through the use of passive learning. This made it easier to set a standard for accuracy by which the Active Learning framework that followed could be measured. By methodically applying the Active Learning framework to different proportions of the dataset, performance metrics at various phases of data use may be insightfully compared. By taking this approach, it was hoped to determine the point at which the accuracy obtained by Passive Learning is approximated or converged to by AL, providing insight into the relative effectiveness of both methods within the framework of this healthcare-related categorization job.

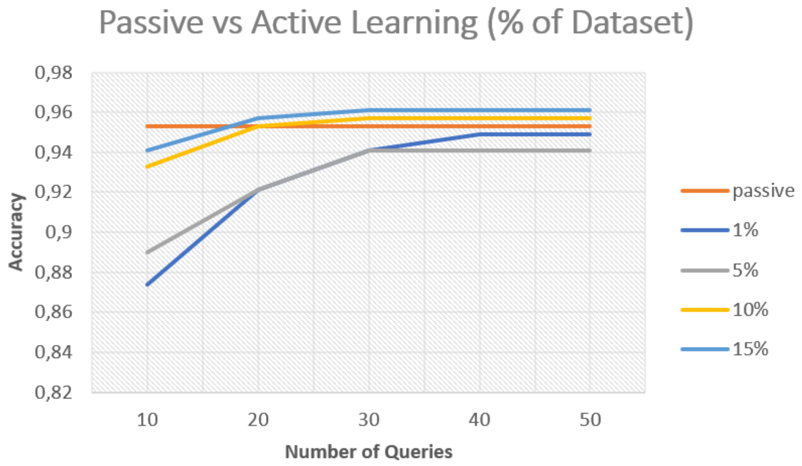

Figure 4.1: Accuracy for each percentage of the KTAS Dataset and the number of queries used in the Active Learning Framework for the LR classifier

Figure 4.1 illustrates that the LR classifier's acquired passive learning accuracy was 0.953. The Active Learning Framework was run with four different initial dataset sizes, three estimators, and LR as the classifier. It can be seen that by utilizing only 10% of the dataset and about 20 queries, so approximately 140 instances in total, it is possible to achieve the same accuracy as the passive learning method. Using small percentages of 1% and 5%, the framework falls short of the passive threshold.

This observation is especially relevant in the context of healthcare, where there may be critical information shortages in some areas. Access to labeled data in the health domain is frequently restricted because of data sensitivity, privacy concerns, or the sheer difficulty of obtaining expert annotations. Reduced annotation burden for domain experts can be achieved by using only 10% of the dataset to achieve accuracy levels in passive learning. Such efficiency gains can speed up the creation of ML models for health applications in situations where expert time is a valuable and scarce resource.

4.1.3 Explainable Results

This section proceeds with an analysis of LIME results obtained from one of the best classifier identified within the AL framework, offering insights into the interpretability of the models in the context of emergency triage using the KTAS dataset.

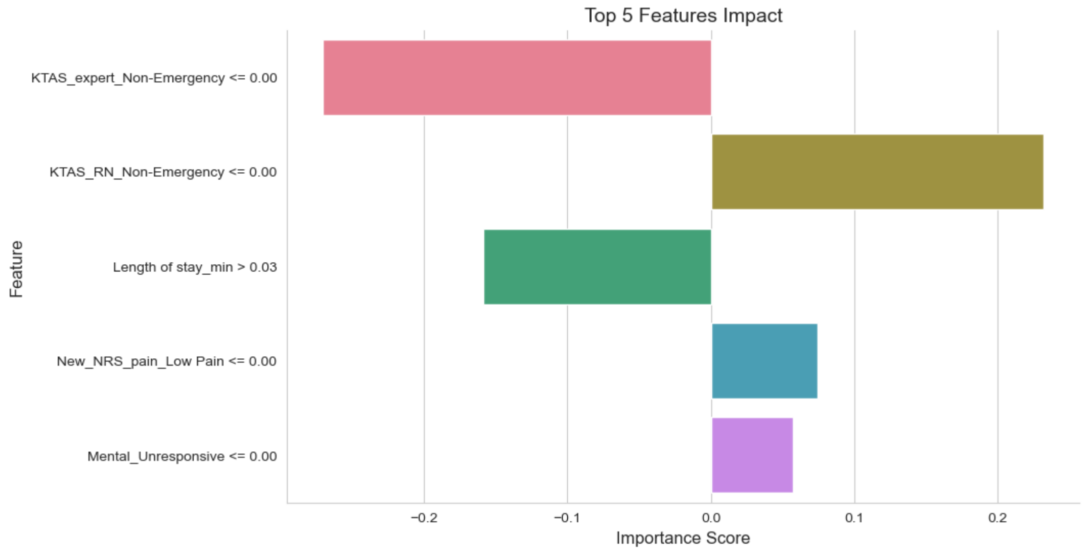

Figure 4.2: LIME graphic for random instance from the KTAS dataset using LR classifier, two estimators and margin sampling

When analyzing Figure 4.2, one possible explanation for the graphic is that the first two features, KTAS_expert_Non-Emergency and KTAS_RN_Non-Emergency, indicate that the expert and RN KTAS scores for non-emergency cases are both very low. This suggests that the patient is likely to be triaged as non-emergency. The third feature, Length of stay_min >0.03, indicates that the patient has been in the emergency department for more than 3 minutes. This could be a sign that the patient is more likely to be triaged as high acuity, as they may have a more complex condition or require more resources. The fourth feature, New NRS pain_Low Pain, indicates that the patient's new NRS pain score for low pain is very low. This suggests that the patient is not experiencing significant pain, which is a factor that is considered in triage. The fifth feature, Mental_Unresponsiv, indicates that the patient is not mentally unresponsive. This is a good sign, as it suggests that the patient is alert and able to communicate.

4.2 Patient Priority Dataset Results

4.2.1 Active Learning Results

Table 4.8: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using 1 estimator

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.937	0.937	0.931	0.714	0.932
	RF	0.954	0.954	0.957	0.913	0.955
	LR + RF	0.981	0.981	0.981	0.951	0.981
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.954	0.954	0.957	0.913	0.955
	LR + RF	0.960	0.960	0.960	0.904	0.960
Least confidence	LR	0.940	0.940	0.935	0.727	0.935
	RF	0.954	0.954	0.957	0.914	0.955
	LR + RF	0.964	0.964	0.965	0.804	0.965

In Table 4.8, the results for a single estimator are presented. The LR classifier demonstrates strong performance across all query strategies, with accuracy ranging from 0.937 to 0.940. It maintains a good balance between sensitivity and specificity, as indicated by the balanced accuracy values ranging from 0.714 to 0.738. The F1 Score, a measure of precision and recall balance, is also consistently high for LR.

The RF classifier consistently performs well, with accuracy ranging from 0.954 to 0.964. It excels in balanced accuracy, with values exceeding 0.9 for all query strategies. The LR + RF ensemble consistently outperforms individual classifiers, achieving accuracy values ranging from 0.960 to 0.981. The balanced accuracy and F1 Score for LR + RF are also notably high, indicating the effectiveness of the ensemble approach.

These results suggest that, even with a single estimator, the ensemble LR + RF is beneficial, showcasing improved performance compared to individual classifiers.

Table 4.9: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using 2 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.936	0.936	0.931	0.713	0.932
	RF	0.975	0.975	0.976	0.826	0.974
	LR + RF	0.979	0.979	0.979	0.865	0.978
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.975	0.975	0.976	0.826	0.974
	LR + RF	0.987	0.987	0.986	0.925	0.986
Least confidence	LR	0.940	0.940	0.935	0.727	0.935
	RF	0.975	0.975	0.976	0.826	0.974
	LR + RF	0.984	0.984	0.985	0.908	0.984

The results are shown in Table 4.9 for a estimators' size of 2. These results suggest that, even with a estimators' size of 2, the LR + RF ensemble maintains its effectiveness in improving performance compared to individual classifiers. The RF classifier continues to show robust performance.

Table 4.10: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using using 3 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.936	0.936	0.931	0.713	0.932
	RF	0.994	0.994	0.995	0.982	0.994
	LR + RF	0.986	0.986	0.985	0.912	0.985
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.994	0.994	0.995	0.982	0.994
	LR + RF	0.987	0.987	0.987	0.908	0.974
Least confidence	LR	0.940	0.940	0.935	0.723	0.935
	RF	0.994	0.994	0.995	0.982	0.994
	LR + RF	0.986	0.986	0.986	0.926	0.985

In Table 4.10, the results for a estimators' size of 3 are presented. The RF classifier exhibits exceptional performance with accuracy consistently at 0.994 for all query strategies. It excels in balanced accuracy, with values consistently at 0.982 or higher, showcasing its ability to handle imbalanced data, surpassing easily the LR classifier results. The LR + RF ensemble maintains strong performance, showcasing accuracy values ranging from 0.986 to 0.987. All F1-scores are high for every classifier but the RF achieves the highest. The RF classifier demonstrated amazing results for this estimators's size.

Table 4.11: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using 4 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.937	0.937	0.931	0.714	0.932
	RF	0.990	0.990	0.990	0.938	0.989
	LR + RF	0.979	0.979	0.891	0.891	0.978
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.990	0.990	0.990	0.914	0.989
	LR + RF	0.987	0.987	0.986	0.929	0.986
Least confidence	LR	0.941	0.941	0.935	0.726	0.935
	RF	0.992	0.992	0.992	0.949	0.992
	LR + RF	0.987	0.987	0.987	0.918	0.986

In general, when looking at the table 4.11, all accuracies are high and above 0.9 for this estimators's size. In terms of balanced accuracy, the LR classifier continues to fall short in comparison to both RF and LR+RF classifiers. RF continues to be the best classifier in terms of every performance metric and using any of the query strategies, except when using margin sampling, the balance accuracy was slightly lower than the combined classifier LR+RF.

Table 4.12: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using 5 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.937	0.937	0.931	0.713	0.923
	RF	0.995	0.995	0.995	0.985	0.994
	LR + RF	0.983	0.983	0.983	0.903	0.982
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.991	0.991	0.992	0.959	0.991
	LR + RF	0.987	0.987	0.987	0.932	0.987
Least confidence	LR	0.940	0.940	0.935	0.726	0.935
	RF	0.996	0.966	0.966	0.973	0.996
	LR + RF	0.985	0.985	0.985	0.923	0.984

For this results, when looking at the Table 4.12, it seems there were no major changes to the previous commmittee's size of 4. RF stays the best classifier, consistently achieving very high accuracy scores, reaching 0.995 for the entropy query strategy and 0.996 for the uncertainty query strategy. This indicates that RF performs exceptionally well in correctly classifying instances. Also maintains a high balance accuracy, meaning it can provide a comprehensive evaluation of classification performance.

Table 4.13: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using 10 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.936	0.936	0.931	0.714	0.932
	RF	0.996	0.996	0.996	0.985	0.996
	LR + RF	0.982	0.982	0.982	0.903	0.982
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.997	0.997	0.998	0.997	0.997
	LR + RF	0.989	0.989	0.989	0.935	0.989
Least confidence	LR	0.941	0.941	0.935	0.727	0.936
	RF	0.997	0.997	0.998	0.997	0.997
	LR + RF	0.988	0.988	0.988	0.932	0.988

Going into the results presented in Table 4.13, it is noticeable that the RF classifier consistently outperforms the LR classifier in terms of accuracy, reaching scores of 0.996 and 0.997, and especially balance accuracy. So there is no considerable chances on this estimators's size either, the RF maintains the status of best classifier, and LR+RF continues to get good results slightly inferior to the RF classifier.

Table 4.14: Patient Priority Dataset Performance Metrics' Results for all classifiers and query strategies using 50 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.936	0.936	0.931	0.714	0.932
	RF	0.997	0.997	0.997	0.997	0.997
	LR + RF	0.984	0.984	0.984	0.899	0.984
Margin	LR	0.939	0.939	0.935	0.738	0.935
	RF	0.997	0.997	0.998	0.997	0.997
	LR + RF	0.981	0.981	0.981	0.900	0.980
Least confidence	LR	0.940	0.940	0.935	0.727	0.935
	RF	0.997	0.997	0.998	0.997	0.997
	LR + RF	0.982	0.982	0.982	0.893	0.981

For the final estimators's size of 50 as seen in Table 4.14, the same conclusion can be applied to this large number on the estimators's voters. The LR continues to be the worst of the classifiers and RF stands at the top.

Because there was no major changes on the framework results when using this large estimators' size, it can be concluded that adding a larger number of estimators doesn't change much past a certain number, this case being 3 or 4, for this dataset. The LR classifier obtained good results when using this dataset and testing the multiple query strategies, but falls short to the other two classifiers. RF was the classifier that got the best results when the estimators number was bigger than 2, so for this dataset, it is possible to say that for low number of estimators, the best option is the combined classifier LR+RF, but past two estimators, the RF classifier get the best performance metrics results.

4.2.2 Passive vs Active

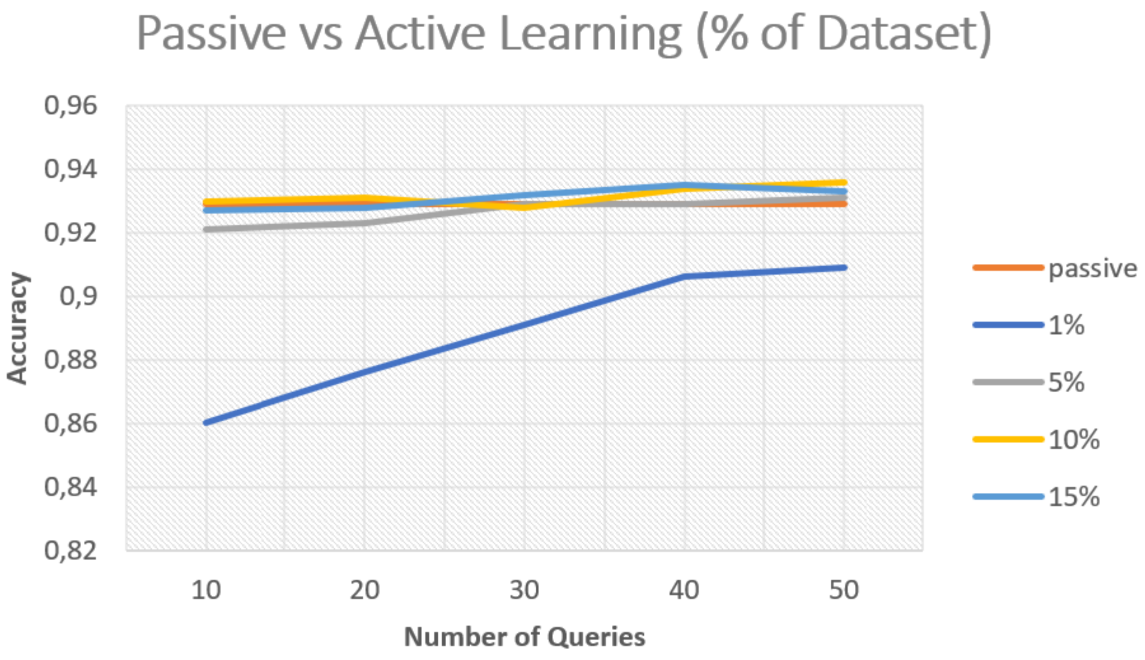

Figure 4.3: Accuracy for each percentage of the Patient Priority Dataset and the number of queries used in the Active Learning Framework for the RF classifier

Figure 4.3 illustrates that the RF classifier's acquired passive learning accuracy was 0.929. The Active Learning Framework was run with four different initial dataset sizes, a single estimator, and RF as the classifier, as the KTAS dataset. It can be seen that by utilizing only 5% of the initial dataset and about 30 queries, making it around 357 instances in total, it is possible to achieve the same accuracy as the passive learning method. Using a small percentage of 1% , the framework doesn't reach the passive threshold. Compared to the initial KTAS dataset, this dataset contains a larger total number of cases, therefore, although the beginning data used (5%) is smaller in percentage terms than the prior dataset, the total amount of data used is larger. It is still a good achievement in order to deal with the problem of scarcity of data.

4.2.3 Explainable Results

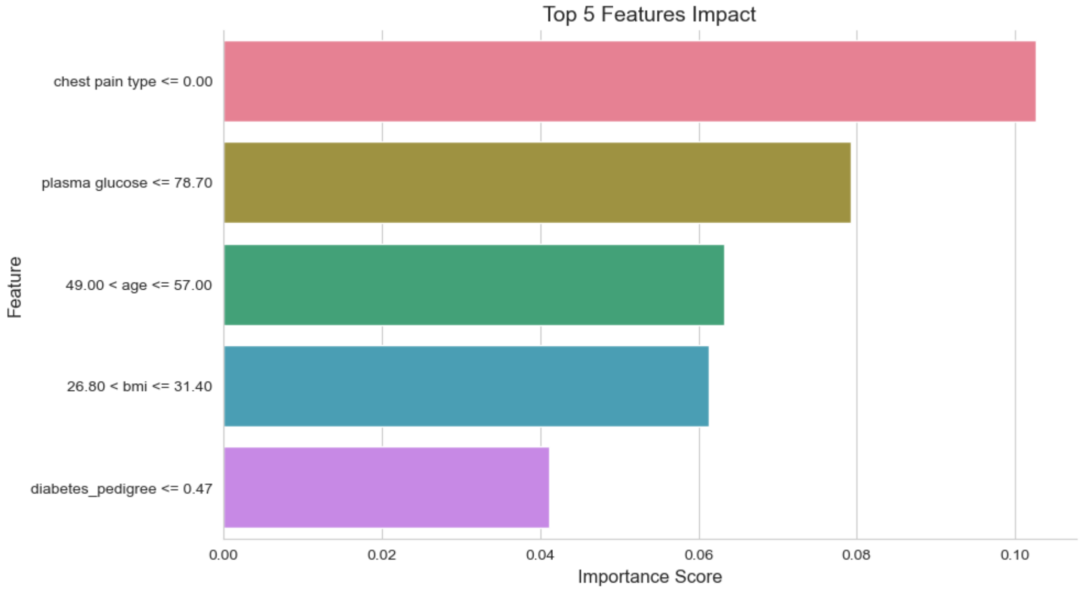

Figure 4.4: LIME graphic for random instance from the Clustering for Triage dataset using RF classifier, five estimators and margin sampling

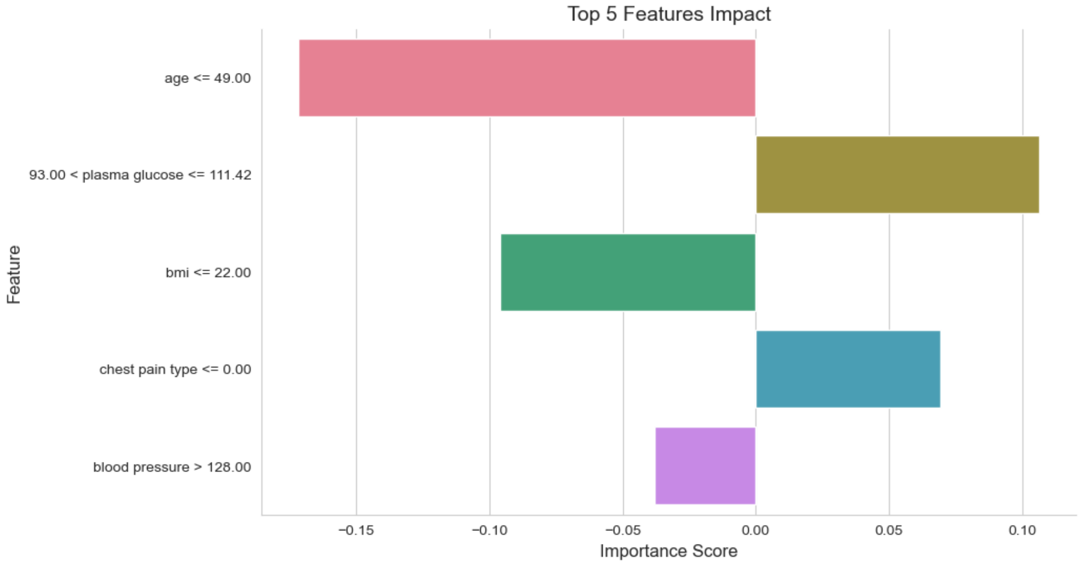

Figure 4.5: LIME graphic for random instance from the Clustering for Triage dataset using LR+RF classifier, a single estimator and entropy sampling

A possible analysis for the Figure 4.4 is: If the patient has a chest pain type of 0.00 or less, it means they don't have any chest pain. This indicates that the patient is less likely to experience a heart attack, which is encouraging. A plasma glucose level that falls within the normal range is 78.70 or lower., which can indicatethat the patient is less likely to have diabetes. The range of middle age is defined as 49.00 to 57.00. Although it is a risk factor for heart disease, this risk factor is not as important as others, like high blood pressure or cholesterol. A person is deemed overweight or obese if their BMI is between 26.80 and 31.40, which poses a serious risk for heart disease. A diabetes pedigree with a score of 0.47 or less is regarded as low. This implies that the patient has a lower chance of developing diabetes, another heart disease risk factor.

According to the results for Figure 4.5, patients under the age of 49 are generally less likely than older patients to have heart diseases, because that heart disease risk factors, like high blood pressure and cholesterol, often get worse with time. A plasma glucose reading within the range of 93.00 to 111.42 is deemed prediabetic, which indicates that this individual is more likely to develop diabetes, increasing the risk of heart disease. This individual appears to have a normal BMI, less than 22.0, but if underweight, people in this condition are at a slightly higher risk of heart disease than people with a normal BMI. A chest pain type of 0.00 indicates that the patient is experiencing no chest pain. This is a positive sign, because it suggests that the patient is less likely to have a heart attack. For last, a blood pressure of 128.00 or higher is considered to be elevated, being a significant risk factor for heart disease.

With this results, it is possible to say that this AL framework was successfully used in addressing the challenge of the restricted availability of data in the emergency triage area.

4.3 Heart Disease Dataset Results

4.3.1 Active Learning Results

Table 4.15: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 1 estimator

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.794	0.631	0.794
	RF	0.550	0.550	0.549	0.383	0.544
	LR + RF	0.833	0.833	0.853	0.837	0.839
Margin	LR	0.800	0.800	0.792	0.618	0.787
	RF	0.550	0.550	0.549	0.383	0.544
	LR + RF	0.850	0.850	0.859	0.704	0.841
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.550	0.550	0.549	0.383	0.544
	LR + RF	0.883	0.883	0.878	0.767	0.876

The performance metrics for a single estimator in relation to the Heart Disease dataset are shown in Table 4.15. The best reliable classifier in the assessment of the Heart Disease dataset with a estimators' size of one is the LR + RF ensemble. This ensemble achieves substantial accuracies and consistently beats individual classifiers under a variety of query strategies. This means that the LR classifier won't be the best option to this dataset when working, because the results of the other two classifiers tend to improve when the using 1 estimator's size increases.

The LR + RF ensemble, for example, obtains an outstanding accuracy of 0.883 under the Least confidence-based query method, demonstrating its usefulness in managing the difficulties of heart disease prediction. Conversely, the RF classifier encounters difficulties with accuracies as low as 0.55. This disparity demonstrates how the model's performance is sensitive on the classifier and approach selected.

Table 4.16: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 2 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.793	0.632	0.794
	RF	0.850	0.850	0.848	0.754	0.846
	LR + RF	0.916	0.916	0.911	0.868	0.908
Margin	LR	0.800	0.800	0792.	0.618	0.787
	RF	0.850	0.850	0.848	0.754	0.846
	LR + RF	0.950	0.950	0.953	0.906	0.945
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.850	0.850	0.848	0.754	0.846
	LR + RF	0.883	0.883	0.878	0.833	0.877

Table 4.16 displays the results of the performance metrics for a estimators' size of 2. The LR + RF ensemble continues to demonstrate its dominance in the evaluation of the framework with a estimators' size of two, consistently outperforming individual classifiers. The ensemble is able to identify intricate patterns linked to heart illness, as evidenced by its remarkable accuracy of 0.950 using the margin-based query technique. This significant improvement over the estimators' size of one implies that greater predictive performance can be achieved by collaborative AI with a larger number of estimators. The RF classifier improved considerable by adding only one more estimator, reaching an accuracy of 0.850 and other decent results across all the metrics. Comparing the results between estimators' sizes, it is evident that increasing the estimators' size positively impacts the overall results.

Table 4.17: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 3 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.793	0.631	0.794
	RF	0.916	0.916	0.913	0.892	0. 914
	LR + RF	0.933	0.933	0.930	0.899	0.929
Margin	LR	0.800	0.800	0.791	0.618	0.787
	RF	0.916	0.916	0.913	0.892	0. 914
	LR + RF	0.950	0.950	0.953	0.906	0.945
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.916	0.916	0.913	0.892	0. 914
	LR + RF	0.950	0.950	0.953	0.906	0.945

When evaluating the results of Table 4.17 with three estimators as opposed to two, we see a steady trend of improvement in a number of performance parameters. Remarkably, the LR + RF ensemble still performs better than the others; accuracy under the uncertaintly-based query technique reaches 0.950. This demonstrates how well the group collaborates to identify patterns in the heart disease dataset, outperforming the accuracy attained with a estimators' size of two. The RF classifier exhibits substantial improvement achieving an accuracy of 0.916 across all query strategies. This indicates that the inclusion of more estimators contributes to a more robust decision-making process, enhancing the overall predictive capabilities of the classifiers.

Table 4.18: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 4 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.793	0.631	0.794
	RF	0.983	0.983	0.983	0.968	0.982
	LR + RF	0.933	0.933	0.930	0.899	0.929
Margin	LR	0.800	0.800	0.792	0.618	0.787
	RF	0.850	0.850	0.859	0.732	0.851
	LR + RF	0.950	0.950	0.953	0.906	0.945
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.933	0.933	0.933	0.889	0.931
	LR + RF	0.933	0.933	0.939	0.878	0.928

Examining the results for a estimators' size of 4 of Table 4.18 in the evaluation of the Heart Disease dataset, several notable trends emerge compared to the estimators' size of 3. The RF classifier demonstrates a substantial leap in performance under the entropy-based query strategy, achieving an accuracy of 0.983. This substantial improvement in accuracy signifies the effectiveness of the RF classifier in discerning patterns within the heart disease dataset when supported by a larger estimators number. With an accuracy of 0.983, the RF classifier shows significant progress, especially when using the entropy-based approach. This suggests that having more estimators increases the decision-making process's robustness and improves the classifiers' overall predictive power. Comparing the results between estimators' sizes 3 and 4, we observe a noteworthy improvement in the RF classifier's accuracy under the entropy strategy, but decreasing when using the margin sampling. This suggests that the introduction of more diverse perspectives and decision-making processes within the learner enhances the model's ability to handle complex patterns in the data.

Table 4.19: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 5 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.793	0.631	0.794
	RF	0.933	0.933	0.939	0.929	0.907
	LR + RF	0.950	0.950	0.954	0.909	0.948
Margin	LR	0.800	0.800	0.792	0.618	0.787
	RF	0.950	0.916	0.947	0.945	0.924
	LR + RF	0.967	0.967	0.968	0.937	0.964
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.966	0.966	0.967	0.940	0.966
	LR + RF	0.933	0.933	0.939	0.878	0.928

Analyzing the performance metrics for a estimators' size of 5 present in Table 4.19, it is possible to see that the RF classifier continues to exhibit strong performance, maintaining an accuracy of 0.933. This continues to suggest that the RF model benefits from the increased number of estimators, capturing more nuanced patterns in the dataset. The LR + RF ensemble also sees a slight improvement in accuracy, reinforcing the collaborative advantage of combining LR and RF classifiers.

When using 10 estimators for the evaluation of the Heart Disease dataset, in Table 4.20, the RF classifier maintains its robust performance, achieving a 0.997 accuracy using the uncertainty query strategy. The LR continues to show no improvement when adding numbers to the estimators' size. For the combined classifier, it also seems to reach its maximium performance, due to the fact that show little to no improvement in this and the previous estimators' size tests. These findings suggest that, for the Heart Disease dataset, a estimators' size of 10 might doesn't yield substantial gains for all classifiers, and the effectiveness depends on the specific characteristics of the classifier and the chosen query strategy.

To complete the AI study for this dataset, using a estimators' size of 50, as it can be seen in

Table 4.20: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 10 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.793	0.631	0.794
	RF	0.950	0.950	0.948	0.930	0.948
	LR + RF	0.933	0.933	0.939	0.878	0.948
Margin	LR	0.800	0.800	0.792	0.618	0.787
	RF	0.983	0.983	0.983	0.968	0.982
	LR + RF	0.950	0.950	0.953	0.906	0.945
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.997	0.997	0.997	0.997	0.997
	LR + RF	0.950	0.950	0.953	0.906	0.945

Table 4.21: Heart Disease Dataset Performance Metrics' Results for all classifiers and query strategies using 50 estimators

QS	Classifier	Accuracy	Sensitivity	Specificity	BA	F1 Score
Entropy	LR	0.816	0.816	0.793	0.631	0.794
	RF	0.983	0.983	0.983	0.968	0.982
	LR + RF	0.950	0.950	0.953	0.906	0.945
Margin	LR	0.800	0.800	0.792	0.618	0.787
	RF	0.950	0.950	0.953	0.906	0.945
	LR + RF	0.950	0.950	0.953	0.906	0.945
Least confidence	LR	0.816	0.816	0.796	0.649	0.802
	RF	0.997	0.997	0.997	0.997	0.997
	LR + RF	0.950	0.950	0.953	0.906	0.945

Table 4.21, no major changes occur on the results when using a a larger number of estimators in the learner.

Examining the results across all seven tables, a notable trend emerges. The RF classifier consistently improves as the number of estimators increases, indicating its capability to leverage additional diverse perspectives. As it was expected in the beginning of the showcase for the results of this dataset, using the LR classifier as an individual voter, made that the results were not good enough to compete with any estimators' size used for the other two classifiers. The LR + RF combination consistently produces good results across different estimators' sizes and query strategies, highlighting the effectiveness of combining classifiers to achieve robust performance. In conclusion, these findings suggest that the optimal number of estimators depends on the classifier and query strategy employed. While RF benefits from larger estimators' size, LR exhibits consistent performance. The LR + RF combination emerges as a promising approach, offering a balance between accuracy and stability across different scenarios.

4.3.2 Passive vs Active

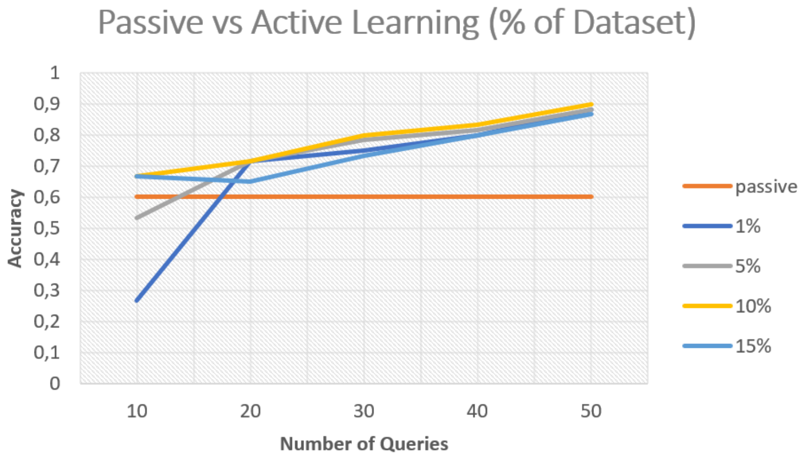

Figure 4.6: Accuracy for each percentage of the Heart Disease Dataset and the number of queries used in the Active Learning Framework for the RF classifier

Figure 4.6 illustrates that the RF classifier's acquired passive learning accuracy was 0.600. The Active Learning Framework was run with four different initial dataset sizes, a estimators' size of three, and RF as the classifier. By utilizing only 1% of the initial dataset and about 15 queries, so approximately 18 instances in total, it is possible to achieve the same accuracy as the passive learning method. This is an amazing result, because of the fact that this a small dataset, which aligns with the problem of scarcity of data in the medicine area.

4.3.3 Explainable Results

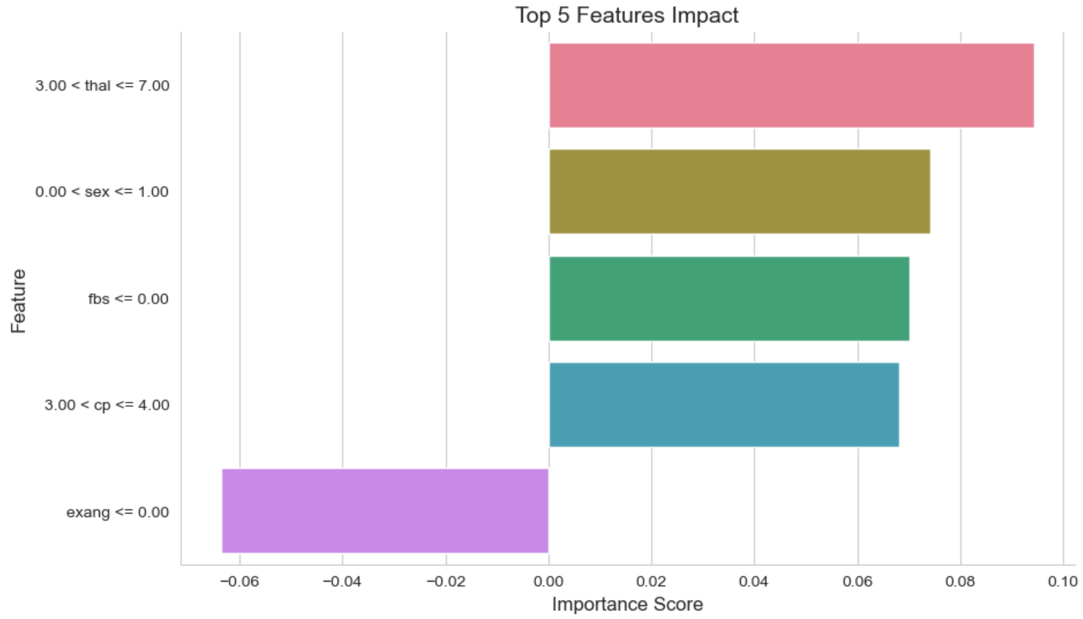

Figure 4.7: LIME graphic for random instance from the Heart Disease Dataset using RF classifier, five estimators and entropy sampling

According to the results for Figure 4.7, a thal of 7, wihch means thalassemia reversable defect, has the most impact on the prediction. The patient is a man and also has a big impact. For the fbs value, the patient doesn't have a value superior to 120 mgdl and also shows no symptons for chest pain because the value of cp is 4. The only feature not contributing for a non heart disease result is exang, that represents exercise induced angina, meaning pain in the chest that comes on with exercise, stress, or other things that make the heart work harder.

5

Conclusion and Future Work

In conclusion, this thesis conducted an investigation into AL approaches concerning emergency triage and heart disease. This study sought to solve important issues in the application of AI to healthcare by utilizing a variety of healthcare related datasets.

The AL framework, augmented by the combination of different learners composed of various classifiers and number of estimators, exhibited promising outcomes. Through meticulous variations in query strategies and number of estimators, a significant understanding of the interaction between classifiers and medical domains emerged. The distinct behaviors of individual LR, RF, and their amalgamation in the form of LR+RF underscored the need for tailored approaches in different medical scenarios. Also, the framework employed in this study has proven to be a success. The experimental results demonstrated that, even with a minimal percentage of the dataset and a small number of queries, the AL model surpassed the performance of passive learning. This not only attests to the efficiency of the AL strategy in selecting and utilizing the most informative data points but also underscores its potential for optimizing the learning process in scenarios where data scarcity is a significant challenge. The achieved performance metrics, including accuracy, sensitivity, specificity, balanced accuracy, and F1 score, consistently surpassed acceptable thresholds, affirming the robustness of the proposed AI framework. This success helps future exploration and application of AI methodologies in addressing data scarcity issues in critical domains such as healthcare. Moreover, this study delved into the imperative realm of model explainability, leveraging LIME for interpretation. By elucidating the features and factors contributing to individual predictions, LIME enhances the interpretability of the AI model, rendering it more accessible and trustworthy for healthcare practitioners and end-users.

Bibliography

[1] M. Haenlein and A. Kaplan, “A brief history of artificial intelligence: On the past, present, and future of artificial intelligence,” *California management review*, vol. 61, no. 4, pp. 5–14, 2019.

[2] I. Asimov, “Runaround,” *Astounding science fiction*, vol. 29, no. 1, pp. 94–103, 1942.

[3] S. J. Russell, *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[4] C. Machinery, “Computing machinery and intelligence-am turing,” *Mind*, vol. 59, no. 236, p. 433, 1950.

[5] C. Zhang and Y. Lu, “Study on artificial intelligence: The state of the art and future prospects,” *Journal of Industrial Information Integration*, vol. 23, p. 100224, 2021.

[6] Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu, “Applying active learning to high-throughput phenotyping algorithms for electronic health records data,” *Journal of the American Medical Informatics Association*, vol. 20, no. e2, pp. e253–e259, 2013.

[7] E. Alpaydin, *Machine learning: the new AI*. MIT press, 2016.

[8] K. Kersting, “Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines,” 2018.

[9] S. B. Kotsiantis, “Decision trees: a recent overview,” *Artificial Intelligence Review*, vol. 39, pp. 261–283, 2013.

[10] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, “Decision trees: an overview and their use in medicine,” *Journal of medical systems*, vol. 26, pp. 445–463, 2002.

[11] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[12] A. Gelzinis, A. Verikas, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hållander, V. Uloza, and E. Padervinskis, “Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders,” in *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pp. 125–132, IEEE, 2014.

[13] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, and M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: a machine learning approach," *International journal of cardiology*, vol. 288, pp. 140–147, 2019.

[14] Y. Yang and M. Loog, "A benchmark and comparison of active learning for logistic regression," *Pattern Recognition*, vol. 83, pp. 401–415, 2018.

[15] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: an evaluation," *Machine Learning*, vol. 68, pp. 235–265, 2007.

[16] B. Settles, "Active learning literature survey," 2009.

[17] B. Settles, "Active learning, volume 6 of," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, pp. 1–114, 2012.

[18] S. Kee, E. Del Castillo, and G. Runger, "Query-by-committee improvement with diversity and density in batch active learning," *Information Sciences*, vol. 454, pp. 401–418, 2018.

[19] R. M. Monarch, *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Simon and Schuster, 2021.

[20] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

[21] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Machine Learning Proceedings 1995*, pp. 150–157, Elsevier, 1995.

[22] N. Abe, "Query learning strategies using boosting and bagging," in *International Conference on Machine Learning, 1998*, pp. 1–9, 1998.

[23] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[24] A. Haleem, M. Javaid, and I. H. Khan, "Current status and applications of artificial intelligence (ai) in medical field: An overview," *Current Medicine Research and Practice*, vol. 9, no. 6, pp. 231–237, 2019.

[25] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[26] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi, "The ethics of ai in health care: a mapping review," *Social Science & Medicine*, vol. 260, p. 113172, 2020.

[27] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and machines*, vol. 30, no. 1, pp. 99–120, 2020.

[28] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[29] R. B. Parikh, S. Teeple, and A. S. Navathe, "Addressing bias in artificial intelligence in health care," *Jama*, vol. 322, no. 24, pp. 2377–2378, 2019.

[30] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[31] J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the need for explainable artificial intelligence (xai)," *arXiv preprint arXiv:2012.01007*, 2020.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[33] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *International Conference on Parallel Problem Solving from Nature*, pp. 448–469, Springer, 2020.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[35] K. Davagdorj, J.-W. Bae, V.-H. Pham, N. Theera-Umpon, and K. H. Ryu, "Explainable artificial intelligence based framework for non-communicable diseases prediction," *IEEE Access*, vol. 9, pp. 123672–123688, 2021.

[36] E. Khodabandehloo, D. Riboni, and A. Alimohammadi, "Healthxai: Collaborative and explainable ai for supporting early diagnosis of cognitive decline," *Future Generation Computer Systems*, vol. 116, pp. 168–189, 2021.

[37] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery," *Computers in Biology and Medicine*, vol. 126, p. 104041, 2020.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[39] V. García, R. A. Mollineda, and J. S. Sánchez, "Index of balanced accuracy: A performance measure for skewed class distributions," in *Iberian conference on pattern recognition and image analysis*, pp. 441–448, Springer, 2009.

[40] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*, pp. 3121–3124, IEEE, 2010.

[41] N. Farrohknia, M. Castrén, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, and K. E. Göransson, "Emergency department triage scales and their components: a systematic review of the scientific evidence," *Scandinavian journal of trauma, resuscitation and emergency medicine*, vol. 19, no. 1, pp. 1–13, 2011.

[42] G. FitzGerald, G. A. Jelinek, D. Scott, and M. F. Gerdtz, “Emergency department triage revisited,” *Emergency Medicine Journal*, vol. 27, no. 2, pp. 86–92, 2010.

[43] M. J. Bullard, B. Unger, J. Spence, E. Grafstein, C. N. W. Group, *et al.*, “Revisions to the canadian emergency department triage and acuity scale (ctas) adult guidelines,” *Canadian Journal of Emergency Medicine*, vol. 10, no. 2, pp. 136–142, 2008.

[44] S.-H. Moon, J. L. Shim, K.-S. Park, and C.-S. Park, “Triage accuracy and causes of mistriage using the korean triage and acuity scale,” *PloS one*, vol. 14, no. 9, p. e0216972, 2019.

[45] J. Park and T. Lim, “Korean triage and acuity scale (ktas),” *Journal of The Korean Society of Emergency Medicine*, vol. 28, no. 6, pp. 547–551, 2017.

[46] M. Bijani and A. A. Khaleghi, “Challenges and barriers affecting the quality of triage in emergency departments: a qualitative study,” *Galen medical journal*, vol. 8, p. e1619, 2019.

[47] A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Butt, A. DoRosario, and S. Johri, “A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis,” *Frontiers in artificial intelligence*, vol. 3, p. 543405, 2020.

[48] M. Fernandes, R. Mendes, S. M. Vieira, F. Leite, C. Palos, A. Johnson, S. Finkelstein, S. Horng, and L. A. Celi, “Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing,” *PloS one*, vol. 15, no. 3, p. e0229331, 2020.

[49] D. Zmiri, Y. Shahar, and M. Taieb-Maimon, “Classification of patients by severity grades during triage in the emergency department using data mining methods,” *Journal of evaluation in clinical practice*, vol. 18, no. 2, pp. 378–388, 2012.

[50] J. Miles, J. Turner, R. Jacques, J. Williams, and S. Mason, “Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review,” *Diagnostic and prognostic research*, vol. 4, pp. 1–12, 2020.

[51] W. S. Hong, A. D. Haimovich, and R. A. Taylor, “Predicting hospital admission at emergency department triage using machine learning,” *PloS one*, vol. 13, no. 7, p. e0201016, 2018.

[52] “Emergency service - triage application.” `https://www.kaggle.com/datasets/ilkeryildiz/emergency-service-triage-application`, 2021.

[53] “Patient priority for clustering.” `https://www.kaggle.com/datasets/hossamahmedaly/patient-priority-classification/data`, 2022.

[54] S. W. P. M. Janosi, Andras and R. Detrano, “Heart Disease.” UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

[55] LIME Documentation Contributors, “Lime 0.2.0 documentation.” https://lime-ml.readthedocs.io/en/latest/lime.html, 2023. Accessed 03 05 2023.