UNIVERSIDADE D
COIMBRA

Sarah Luísa Jenny Martins Holm

AL-DLIME - Active Learning-Based Deterministic Local Interpretable Model-Agnostic Explanations: A Comparison with LIME and DLIME in the Field of Medicine

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

July 2023

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Sarah Luísa Jenny Martins Holm

AL-DLIME - Active Learning-Based Deterministic Local Interpretable Model-Agnostic Explanations: A Comparison with LIME and DLIME in the Field of Medicine

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

July 2023

This work was developed in collaboration with:

Center for Informatics and Systems of the University of Coimbra

This research was supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

Acknowledgements

I would like to begin by thanking my advisor, Prof. Dr. Luís Macedo, for all that he has done for me during the elaboration of this thesis. Your guidance was fundamental to my journey, and I am so grateful for the patience, kindness and faith in me that you have shown. Thank you for taking the time to help me along the way.

To my 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 family, of course, who are an endless source of support in every aspect of my life. To my mother, especially, without whom I would not be where I am today. Your kindness and faith in me pushed me to always do the best that I can, and your teachings have shown me how to be gentle with myself when I feel that it is not enough.

Finally, to Pedro, who has always seen the best in me, and has helped me see the same even during the hardest days. Thank you for being there for me, through both the bad, as well as the good moments. I would also like to thank my cat, Kira. Though he has no way of knowing, the simple fact of his presence never failed to draw a smile from me.

Thank you all, from the bottom of my heart.

"The best way out is always through"

Robert Frost

Abstract

Artificial intelligence has garnered significant interest since its inception due to its vast potential benefits and applications. However, numerous cases have put into question the ethics of artificial intelligence, with topics of privacy, data protection, bias, and even safety becoming more and more prevalent in regards to this technology. To address these issues, many authors believe that we must place a larger focus on responsible artificial intelligence in order to ensure a safer future; in other words, artificial intelligence that is ethical and, therefore, trustworthy. Despite their proven accuracy, the models that are currently most prevalent (such as neural networks), are inherently black boxes. In order to preserve the accuracy of these models while maintaining the transparency and, thus, the trustworthiness of simpler models such as decision trees, the field of explainable artificial intelligence was created.

One of the most cited explainable artificial intelligence models is LIME (Local Interpretable Model-agnostic Explanations). However, its non-deterministic nature signifies that explanations regarding the same instance may vary. This may lead to some tension in sensitive areas of application such as medicine, where the end-users do not understand the underlying technology and, thus, may doubt its efficiency. The authors of DLIME (Deterministic Local Interpretable Model-agnostic Explanations) hoped to assuage this issue by creating a deterministic model based on LIME. While the authors of DLIME provided a comparison between their model and LIME, there is still scope for experimentation and, hopefully, improvement within the framework.

The goal of this thesis is twofold. Firstly, it aims to introduce a novel explainable artificial intelligence model that integrates active learning into the DLIME framework: Active Learning-based Deterministic Local Interpretable Model-agnostic Explanations (AL-DLIME). Secondly, it aims to perform a detailed comparison of LIME, DLIME, and AL-DLIME for medical diagnosis applications, with a focus on assessing the impact of DLIME and AL-DLIME's deterministic behavior on their overall performance.

For the purposes of this study, four datasets were selected within some of the areas of medicine that are considered to have the least accuracy in terms of diagnosis, oncology and cardiovascular diseases. In terms of the underlying black box model, random forest was selected due to its popularity and overall good performance. The decision tree model was also selected to address the accuracy-explainability tradeoff, more specifically, if the use of a black box model is strictly necessary. The performance of each model was evaluated using several metrics, including accuracy and F1-score for both the machine learning and explainable artificial intelligence models. Regarding solely the explainable artificial intelligence models, the metrics of faithfulness to the black box model, stability of the model through Jaccard's distance, and single and incremental deletion were selected.

The results show AL-DLIME outperformed random forest on several occasions, achieving the best overall results for accuracy and F1-score among the explain-

able artificial intelligence models. However, LIME obtained the overall highest scores of faithfulness to random forest, with results consistently above 60%. Finally, random forest outperformed decision tree on accounts of both accuracy and F1-score across both experiments, with its highest score of accuracy, 99%, being on par with other state of the art machine learning models. The study provides insights into the strengths and weaknesses of each explainable artificial intelligence model and their suitability for medical diagnosis applications. Further research may expand upon these findings by evaluating the models with a larger array of metrics, as well as through the use of other machine learning models.

Keywords

Local interpretability, Local explainability, Active learning, LIME, DLIME, XAI metrics, Trustworthiness, Faithfulness, Accuracy, Random Forest, XAI, Medicine

Resumo

A inteligência artificial tem suscitado um interesse significativo desde a sua criação devido aos seus vastos potenciais benefícios e aplicações. Contudo, múltiplos acontecimentos acabam por colocar em causa a ética da inteligência artificial, com tópicos de privacidade, proteção de dados, *bias*, e até segurança a tornarem-se cada vez mais prevalentes em relação a esta tecnologia. Para abordar estas questões, muitos autores defendem que se deve dar maior ênfase a uma inteligência artificial responsável de forma a garantir-se um futuro mais seguro; ou seja, uma inteligência artificial éticamente correta e, consequentemente, confiável. Apesar da sua comprovada *accuracy*, os modelos que atualmente são mais prevalentes (como, por exemplo, redes neuronais) são fundamentalmente caixas pretas. Assim, de modo a preservar a *accuracy* destes modelos sem perder a transparência, e por isso, também a confiabilidade de modelos mais simples, como árvores de decisão, foi criada a área de inteligência artifical explicável.

Um dos modelos de inteligência artificial explicável mais citados é o LIME. No entanto, a sua natureza não-determinística significa que as explicações sobre uma mesma instância podem variar. Isto pode gerar alguma tensão em áreas de aplicação mais sensíveis, como a medicina, onde os utilizadores não têm necessariamente que compreender a tecnologia subjacente e, portanto, podem duvidar do seu desempenho. Os autores do DLIME esperavam resolver este problema ao criar um modelo determinístico com base no LIME. Apesar deste autores fornecerem uma comparação entre o seu modelo e o LIME, existe, contudo, a possibilidade de melhoria dentro do sistema.

O objetivo desta tese é duplo. Em primeiro lugar, pretende introduzir um novo modelo de inteligência artificial explicável, chamado AL-DLIME, que integra aprendizagem ativa no sistema DLIME. Em segundo lugar, visa realizar uma comparação detalhada entre LIME, DLIME e AL-DLIME em aplicações de diagnóstico médico, com foco na avaliação do impacto do comportamento determinístico de DLIME e de AL-DLIME no seu desempenho geral.

Para efeitos deste estudo, foram selecionados quatro *datasets* dentro de algumas das áreas da medicina consideradas de menor precisão em termos de diagnóstico: oncologia e doenças cardiovasculares. Em termos do modelo de caixa preta subjacente, o *random forest* foi selecionado devido à sua popularidade e bom desempenho geral. O modelo árvore de decisão também foi selecionado de modo a abordar o compromisso entre *accuracy* e explicabilidade, mais especificamente, averiguar se o uso de um modelo de caixa preta é estritamente necessário. O desempenho de cada modelo foi avaliado usando várias métricas, incluindo *accuracy* e F1-*score* para ambos os modelos de aprendizagem computacional e de inteligência artificial explicável. Para os modelos de inteligência artificial explicável, foram selecionadas as métricas de fidelidade ao modelo de caixa preta, a estabilidade do modelo, e exclusão única e incremental.

Os resultados mostram que o AL-DLIME obteve melhor desempenho que o *random forest* em várias ocasiões, alcançando os melhores valores gerais para *accuracy* e F1-*score* entre os modelos de inteligência artificial explicável. No entanto,

o LIME obteve os maiores valores gerais de fidelidade comparativamente ao *random forest*, com resultados consistentemente acima de 60%. Por fim, o *random forest* superou a árvore de decisão em termos de *accuracy* e F1-*score* em ambos os testes experimentais, atingindo o resultado mais alto de 99% para *accuracy* e estando assim ao mesmo nível de outros modelos de aprendizagem computacional de última geração. Este estudo fornece informação sobre os pontos fortes e fracos de cada modelo de inteligência artificial explicável e a sua adequação para aplicações no âmbito de diagnóstico médico. Futuros estudos podem complementar estes resultados ao avaliar os modelos com uma maior seleção de métricas, bem como por meio do uso de outros modelos de aprendizagem computacional.

Palavras-Chave

Interpretabilidade local, Explicabilidade local, Aprendizagem ativa, LIME, DLIME, Métricas de inteligência artificial explicável, Confiabilidade, Fidelidade, *Accuracy*, *Random Forest*, Inteligência artificial explicável, Medicina

List of Figures

List of Tables

Contents

Acronyms

AHC Agglomerative Hierarchical Clustering.

AI Artificial Intelligence.

AI HLEG High Level Expert Group on AI.

AL Active Learning.

AL-DLIME Active Learning-based Deterministic Local Interpretable Model-agnostic Explanations.

BCD Breast Cancer Dataset.

CVD Cardiovascular Disease.

DLIME Deterministic Local Interpretable Model-agnostic Explanations.

DT Decision Tree.

HDD Heart Disease Dataset.

IARC International Agency for Research on Cancer.

kNN k-Nearest Neighbors.

LIME Local Interpretable Model-agnostic Explanations.

LR Logistic Regression.

ML Machine Learning.

NCD Non-Communicable Disease.

OCD Ovarian Cancer Dataset.

PCD Pancreatic Cancer Dataset.

PDR Predictive, Descriptive and Relevant.

RF Random Forest.

SVM Support Vector Machine.

WHO World Health Organization.

XAI eXplainable Artificial Intelligence.

Chapter 1

Introduction

In this chapter, Section 1.1 will detail the overall motivation behind this thesis, as well as the context surrounding it. Following this, Section 1.2 will explore the main goals. Section 1.3 will detail the approach this thesis had in order to achieve the main goals, with Section 1.4 providing a list of the contributions this work has made. Finally, Section 1.5 will provide an outline of the structure of this document.

1.1 Motivation and Context

In recent years, the field of Artificial Intelligence (AI) has evolved tremendously, garnering significant interest and expanding to include numerous subfields, such as Machine Learning (ML), deep learning, natural language processing, and many more [1] [pp. 16-29]. From autonomous cars, search and recommendation algorithms, chatbots, and resumé screening for employment, to the integration of AI in the process of medical diagnoses, there seems to be no lack of imagination in regards to how this technology can be used in day-to-day life. Nevertheless, as with any other facet of technology, it is of utmost importance to put into question the consequences AI may have on society, whatever its application may be [2].

Many concerns have been raised over the ethics of AI, especially when considering its most sensitive areas of application [2–4]. Some of these concerns, such as those related to privacy, data protection, bias, and safety, are rooted in real-life events that have shown the potential danger of AI, while others, such as the "awakening" of AI, are issues that some authors fear may come to pass in the near future [2–7]. While it is possible that these issues may derive from a flaw in the overall AI design, or from biased data, in other cases, the problem may not be so apparent, warranting a more exhaustive investigation [8]. Whatever the cause, it has become most evident that if we are to continue with this pace in AI evolution, there is a large need for AI regulation.

Reliability, accountability, and traceability are all topics of interest when it comes to discussions of how AI could be regulated. The common thread among all of these core concepts is transparency, and through transparency, there is hope to

achieve trust in AI [8]. In a study published in 2019 by the High Level Expert Group on AI (AI HLEG) [8], they determined that in order to achieve trustworthy AI, the technology should follow four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. While the first three are self-explanatory, serving as a clear reflection of fundamental human rights, the final concept is something more inherent to AI. When it comes to applications that may have catastrophic consequences given a single instance of misclassification, or a flaw in the algorithm's overall design, it is of utmost importance that we understand the origin of any potential issues. Given the current popularity of algorithms which are, for all intents and purposes, black boxes, this issue becomes more complex. Therefore, explainability has been proposed to mitigate these concerns, and thus, the subfield of eXplainable Artificial Intelligence (XAI) has formed around it.

There are many different fields where XAI would be an important asset, however, one of the most critical areas of application is perhaps in medicine, where any single mistake may cost the life of a patient [9]. Despite the potential benefits of AI in healthcare being vast, good performance is insufficient in regards to garnering approval for use in real life scenarios – or, more specifically, trust [10–13]. Thus, the potential explainability could offer to such areas is unparalleled.

1.2 Research Problem, Aims, Objectives, and Questions

One of the most cited XAI models is LIME [14], due to its impressive performance and faithfulness to the underlying black box model. However, despite its prevalence, the non-deterministic nature of the algorithm means that there may be differences between explanations regarding the same instance. For applications such as medical diagnoses, where the end-users have no obligation to understand how these algorithms work, this may cause tension and, therefore, a loss of trust. It is for this reason that Zafar and Khan proposed DLIME [15], a deterministic model based on LIME. In their original paper [16], they demonstrated the stability of their proposed model through the use of Jaccard's distance among a select array of medical-based datasets. However, the lack of evaluation metrics in their original paper limited the comparison between DLIME and LIME. To address this gap, they published a more comprehensive comparison, which showed DLIME outperforming LIME in terms of stability and classification quality. However, LIME performed better in terms of faithfulness. There is still scope for experimentation within the DLIME framework, however, and despite a lack of metrics to assess the quality of explanations, DLIME's potential for improving medical diagnoses should not be dismissed. As such, in this thesis I propose AL-DLIME, a novel XAI model that implements Active Learning (AL) within the DLIME framework. By proposing and evaluating AL-DLIME in this study, I hope to contribute to the development of XAI models that may be applied in sensitive and critical domains, such as medicine, with improved performance and transparency.

While DLIME is a modified version of LIME that avoids non-determinism and thereby promotes a more stable generation of explanations, AL-DLIME is, in turn, a modified DLIME. With the AL-DLIME algorithm, the clustering stage of the original DLIME is substituted by an AL stage, thereby selecting the most informative instances to train the surrogate model (ridge linear regression in this case). Thus, AL-DLIME not only offers the benefit of determinism, an important feature for critical domains such as medicine, but also the possibility of training the surrogate model with a limited number of instances, specifically those that provide the most valuable information. The latter property, in particular, is also of extreme importance in domains where data exists in large quantities, though it is mostly, if not completely, unlabeled, as is the case with medicine.

In order to evaluate AL-DLIME, the following metrics were selected: the faithfulness metric from Ribeiro et al. [14], the metric of stability from Zafar and Khan [15], the single deletion metric, and the incremental deletion metric [17–20]. These measures provide a comprehensive evaluation of the quality of the explanations generated by both XAI models, including their accuracy, consistency, and faithfulness. The results obtained with these metrics by AL-DLIME in four datasets from the field of medicine are confronted with those of the XAI models of LIME and DLIME.

Having defined the research problem and summarise the research aims and objectives, this thesis attempts to answer the following research question:

Can XAI models be developed and applied in sensitive and critical domains, such as medicine, with improved performance, stability, and transparency, while requiring less data for training?

This research question can then be split into the following sub-research questions:

1. Can LIME-like XAI models be trained with less data while not losing performance (while not losing the quality of the generated explanations, including their accuracy, consistency, and faithfulness)?

2. Can a deterministic modified version of LIME be developed while not losing the quality of the generated explanations, including their accuracy, consistency, and faithfulness?

3. Can such deterministic modified version of LIME be applied to critical domains such as medicine, while not losing the quality of the generated explanations, including their accuracy, consistency, and faithfulness?

1.3 Approach

The answer to the aforementioned questions rely strongly on the proposal of AL-DLIME, as well as on a comprehensive comparison between the XAI models of LIME, DLIME, and AL-DLIME for use in the medical field.

In short, the rationale for AL-DLIME is twofold: (i) avoiding the non-determinism

of LIME, and thus ensuring a more stable generation of explanations, and (ii) training the surrogate model with a restricted selection of instances, which is especially helpful in domains where the labelled data is scarce.

In order to achieve the aforementioned goals, I have selected four publicly available datasets, three of which are related to the classification of different types of cancer[1] [2] [3], and the last, to heart disease classification[4].

Two experiments are performed for the purposes of this study: the first, with standardized versions of the dataset, and the second, with non-standardized data. The necessity for these two experiments stems from the use of the ridge linear regression classifier in both the LIME and DLIME frameworks, which works best with standardized data [21][p.82].

Several metrics are used for the purposes of comparing not only LIME, DLIME, and AL-DLIME, but also Random Forest (RF) and Decision Tree (DT). These metrics include: accuracy, F1-score, faithfulness to the black box model from Ribeiro et al. [14], single and incremental deletion, and stability of the model using Jaccard's distance from Zafar and Khan [15].

1.4 Contributions

The expected contributions of this thesis are the following:

- AL-DLIME, a deterministic modified version of LIME which maintains the quality of the generated explanations, including their accuracy, consistency, and faithfulness;
- Providing a more comprehensive comparison between DLIME and LIME, and also AL-DLIME, in regards to the performance of the models themselves, as well as their faithfulness to the underlying black box models;
- A novel XAI model which utilizes AL for training the surrogate model;
- A study of the application of XAI models to critical domains such as medicine
- Application of several different XAI metrics
- Focus on improving the trustworthiness of XAI models, by enhancing their determinism

Additionally, from this thesis resulted an article [22] which has been accepted for publication and for presentation at the 1st World Conference on XAI: S. Holm and L. Macedo, "The accuracy and faithfulness of AL-DLIME - Active Learning-based

[1] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
[2] https://data.mendeley.com/datasets/th7fztbrv9
[3] https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer
[4] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Deterministic Local Interpretable Model-Agnostic Explanations: a comparison with LIME and DLIME in medicine." Proceedings of the 1st World Conference on Explainable Artificial Intelligence, July 2023.

1.5 Document Structure

This document is organized as it follows:

- Chapter 2 presents background information related to AI, the ethical dilemmas related to its current use, XAI, and medical knowledge related to the selected datasets
- Chapter 3 showcases studies related to the current paper, as well as a more in-depth analysis of LIME and DLIME, and the evaluation metrics that were chosen for this thesis
- Chapter 4 describes the methods employed throughout this study, including pre-processing of all datasets and the process of model optimization, as well as the materials selected, such as the datasets, ML and XAI models, and evaluation metrics
- Chapter 5 explores the results obtained throughout this study, and provides an in-depth discussion of them
- Chapter 6 presents the final conclusions of this thesis
- Chapter 7 addresses possible topics of interest for future work within the same frame of work

Chapter 2

Background Knowledge

In this chapter, Section 2.1 includes an overview of the topic of AI, a discussion of the fields of ML and AL, as well as a short description of the ML models, DT and RF, which will be of special focus in this study. Some of the ethical concerns related to AI will be detailed in Section 2.2, followed by a description of what trustworthy AI should consist of. In Section 2.3, XAI will be explored in greater detail. All aspects regarding medical knowledge that will be relevant during this thesis are discussed in Section 2.4. Finally, Section 2.5 will deliver a short summary of all topics covered throughout this chapter.

2.1 Artificial Intelligence

There is something to be said about the boundless imagination of mankind, and how it leads to the birth of great inventions, how it drives us to act, to create, to innovate. We dreamed of flight, and so came the first planes; we dreamed of space, and set foot on the moon. What separates us from other living beings has long since been defined as our degree of intelligence and, in some cases, our sentience. As far as we know, there is no other form of intelligent life in the universe, which begs the question: what if we were to create something new, something that does not occur naturally and yet possesses the same cognitive abilities as humans?

In 1921, author Karel Čapek introduced the term "robot" in his play R.U.R. (Rossum's Universal Robots) in order to to describe artificial people created from inorganic flesh and blood by the titular character, Rossum [23]. They are used as workers in Rossum's factory and, despite their initial happiness, they end up revolting and ultimately cause the downfall of mankind. While this idea of "robot" diverges from the concept we are familiar with, instead lending itself to what we would describe as an "android", it is an important milestone in the journey of what we conceptualize AI as.

Nearly two decades later, Isaac Asimov would introduce the "Three Laws of Robotics", which serve to prevent any harm robots may cause to their creators, and humans in general. His short story Runaround [24], published in 1942, was

the first to explicitly state 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.

Though R.U.R and Runaround are two works of fiction which predate the introduction of AI in real-life settings, it is interesting to note that they both echo concerns related to the ethics of AI which remain relevant even today. This topic, alongside the concept of trustworthy AI, will be explored in detail in Section 2.2.

Čapek, Asimov, and many others explored the nuances and possibilities related to artificial, intelligent life in their stories. However, the prospect of AI existing outside the realm of fiction would not be explored until 1943, when the first instance of an AI model was proposed by Warren McCulloch and Walter Pitts: a network composed of artificial neurons, of which each single neuron could turn "on" or "off" according to stimuli received by neighboring neurons [25]. The first computer with neural network logic would be built seven years later, in 1950, by students Marvin Minsky and Dean Edmonds [1] [p.16]. In that same year, Alan Turing published his paper "Computing Machinery and Intelligence" [26].

Can machines think? It is this question which Turing explored in his 1950 paper [26], with the proposal of the now famous Turing Test. Initially, we are asked to consider a scenario consisting of three participants: a man (A) and a woman (B) who are in separate rooms, and an interrogator (C). The function of C is to ask a series of questions through a computer (or, as Turing initially suggested, a teleprompter), during which they will attempt to correctly identify which of the players is the man and the woman, as they are known to C only as X and Y, respectively. However, as an added twist, A must pretend to be the woman. Turing then proposed the substitution of the human A for a machine, or, an artificially intelligent agent. The objective, therefore, shifts to the agent's ability to convince the judge that it is, indeed, player B. Thus, if successful, can it then think, as Turing questioned? According to Hoffman [27], perhaps not.

Despite its fame and large presence in the world of AI, the Turing Test is not without fault. Hoffman [27] raises the question of what the Turing Test actually seeks to evaluate, and settles on humanity, in lieu of intelligence. It stands to reason that certain questions are nigh impossible for a human to answer quickly, such as complicated calculations, or complex world issues, while an AI agent may take only a fraction of a second. Therefore, in order to truly pass the Turing Test, the machine must emulate how we think, and the faults that may bring.

In the time between its birth and now, AI as a whole has experienced many breakthroughs, with the field expanding with concepts such as machine learning, deep

Table 2.1: The four categories of AI classifications. Adapted from [1][p.2][27].

	Rationality-Based	**Human-Based**
Thinking-Based	Systems that think rationally	Systems that think like humans
Behaviour-Based	Systems that behave rationally	Systems that behave like humans

learning, natural language processing, and many more. Thus, it comes as no surprise that AI remains without a definition that is universally agreed upon [28]. Russel and Norvig [1] [p.1-5] found that, in general, AI has historically been defined along two dimensions: the first concerns itself with whether or not the objective is to approximate rationality, or humanity, while the second focuses on the goals of thought processes and reasoning versus behavior. Table 2.1 demonstrates the four quadrants that may be derived from the aforementioned dimensions. For the purposes of this study, I consider AI belongs to the category of "thinking rationally", and put forth the definition from Kaplan and Haenlein [29] as a good representation of this line of reasoning:

> [Artificial Intelligence is] a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation.

Following the definition of AI, it is equally important to explore some of the core concepts that will be dealt with and discussed throughout this work. These concepts are ML, AL, DTs, and RF classifiers.

2.1.1 Machine Learning

In a general sense, ML is a subfield of AI in which the algorithms focus on the task of learning, whether that be from examples, definitions, behaviors, or from being told [30; 31]. ML permits the iterative analysis of extensive datasets, which, oftentimes, reveal complex patterns and insights that otherwise would remain hidden for humans. As such, ML algorithms are apt for tasks such as dimensionality reduction, regression, clustering, prediction, and, of course, classification.

2.1.2 Active Learning

Active learning is a subfield of ML that focuses on minimizing the amount of labeled instances that is necessary to achieve a good performance [32]. It does so by first selecting the most informative instances according to the model's current state of knowledge, followed by a query to a human oracle or a pre-existing, labeled dataset for the correct labels. This is especially appealing for fields in which there is a higher cost related to labeling data, as is the case for the areas of speech recognition, for example, or any specialized area with a reduced number of specialists.

There are three main scenarios for the queries made by the AL models: membership query synthesis, stream-based selective sampling, and pool-based AL

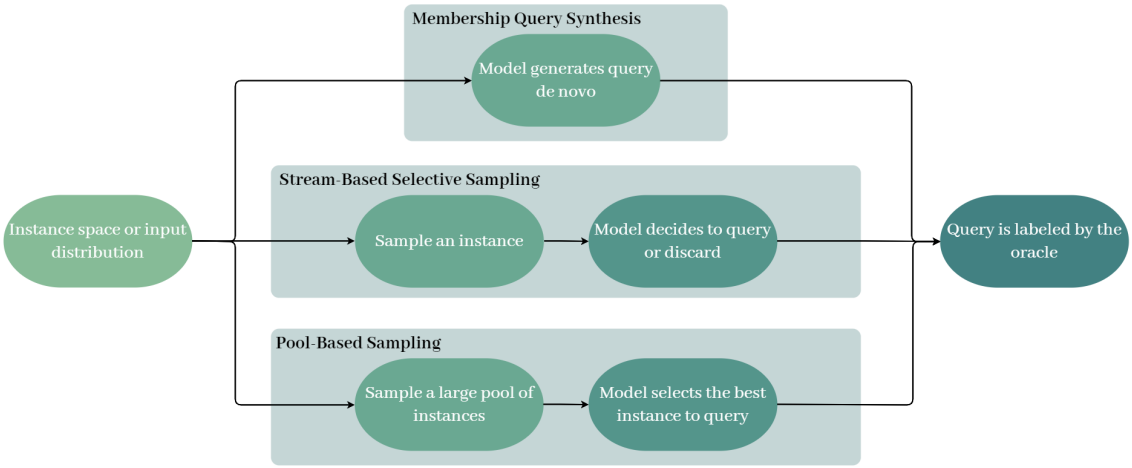


Figure 2.1: Diagram illustrating the three main AL scenarios. Adapted from [32].

[32]. These scenarios are depicted in Figure 2.1. In membership query synthesis, the learner generates its own instances from the underlying dataset in order to query the oracle of its contents (e.g., for a dataset with pictures of dogs and cats, the learner could clip an image around an appendage and query the oracle of whether that appendage belongs to a dog or a cat). As for stream-based selective sampling, the learner decides whether or not to query unlabeled instances while drawing them one at a time. This process is aided through the use of, for example, an "informative measure", "query strategy", or even the computation of an "explicit region of uncertainty" . Finally, pool-based AL utilizes the entire pool of unlabeled instances, as opposed to stream-based selective sampling. Then, in a similar fashion to the previous scenario, some measure is utilized to find the most informative instance for the learner to query.

There are a variety of query strategies that may be used in AL, such as query-by-committee, expected model change, and so forth. However, for the purposes of this thesis, it is perhaps more pertinent to focus on uncertainty sampling. The idea behind this measure is simple: query the instances where the learner is most uncertain on how to label them. For probabilistic models in binary settings, this approach is quite straightforward, as the learner may focus on instances which present a probability of 50% of belonging to a determined class. Other uses of this strategy may rely on measures such as entropy, for example, or the measure of least confidence.

2.1.3 Decision Trees

A DT reaches its outputs through a series of tests related to the dataset's features [33–35]. These tests (or nodes, as they are most commonly called) may be thought of as questions, as they query each input relative to the possible values of a certain feature. The node then branches into two outcomes, a positive and a negative, and the process repeats until a classification is reached, or, in other terminology, a leaf (also known as decision node, or terminal node). In the case of categorical features, the nodes are composed of 'yes' or 'no' questions, while nodes which

utilize numeric features are based on a threshold, for example, "Is the value >k?" [34; 35].

The structure of a DT forms a hierarchy which is highly intuitive for humans to understand [33]. This is why DTs may also be categorized as white box models, for their internal decision making processes, in most cases, are simple to understand. However, this level of simplicity in understanding may also come at a cost in accuracy; most commonly, DTs are outperformed by other, more complex algorithms, such as ensemble methods [36; 37]. These models take the collective output of numerous DTs and combine them to reach a classification, thereby achieving increased predictive performance and decreased risk of overfitting [37; 38].

2.1.4 Random Forest

Bootstrap aggregating (also known as bagging) is one of the three most commonly used ensemble techniques, with RF serving as its main representative [34; 37; 38]. The construction of each DT within the RF model is achieved through the use of bootstrap samples taken from the original training data. These samples are selected randomly, and utilize approximately two-thirds, or 64%, of the dataset. Upon reaching an output, the DTs then perform a majority vote in order to decide which label to use for the classification, with each tree entitled to one vote. In other words, the label with the most votes will then be used as the classification for that instance.

While DTs are intuitive for most people, this may not be the case for RF models. The quantity of DTs that go into each RF model create a process that is much more difficult to follow. Thus, while DTs may be considered as "white box" models, RF are "black box" due to their more complicated nature, and inaccessibility to laymen. Therefore, despite an increase in accuracy, this opaqueness has, in turn, generated some distrust with end-users. With AI gaining more and more popularity throughout the years, this topic, as well as many others, have been raised, leading the way for many discussions surrounding the ethics of AI.

2.2 Ethics and AI

As the field of AI grows and develops, so do the concerns over how this technology can be used, as well as the many possible consequences its use may imply. As many as thirty-nine of these issues were proposed by Stahl [2], ranging from questions regarding privacy and data protection, which have already been breached by AI, to scenarios that may occur in the near future, such as the "awakening" of AI. The most common ethical concerns related to AI that are raised, however, reveal a vast array of different topics that must be addressed if AI is ever to be commonplace and widespread.

2.2.1 Privacy and Data Protection

In the digital era, it has become increasingly difficult to guarantee the protection of personal data [3; 4]. With the introduction of AI, the matter becomes even more complex. A popular example of the infringement on privacy through the use of AI is facial recognition, whether it be from pictures or videos [3]. While the technology may have certain benefits, such as the identification of someone who has committed a serious crime, it nonetheless remains a topic of serious concern regarding the right to privacy [4; 5].

The continuous collection of data is facilitated by our online habits – for instance, social media. Even in situations where this information may seem, for all intents and purposes, harmless, AI possesses the unique ability to identify patterns that may breach rights to privacy and data protection [2; 4]. Moreover, it may even be used to manipulate our behavior through the use of advertisements, or political propaganda [3].

Reliability is, therefore, a vital question to be addressed when considering the threat AI may pose to privacy and data protection [2]. Without reliability, how can individuals trust that the integrity of their personal data will be maintained?

2.2.2 Bias

Like with any other facet of technology, AI was created by humans, and thus bears the capacity to echo our biases, whether intentionally, or inadvertently [2; 4]. Several real-life instances may be cited, such as the Correctional Offender Management Profiling for Alternative Sanctions system, an algorithm which predicts the probability a given defendant has of re-offending given a wide array of features pertaining to the individual, as well as their past criminal record [4; 6]. While none of the features included information about the defendants' race, the system nonetheless was found to generate a higher number of false positives and lower number of false negatives in the case of black individuals, contrary to white individuals [6]. Other examples include Amazon's recruitment tool, which skewed heavily in favor of male applicants, as well as popular image databases, which often propagate old-fashioned ideals of women commonly being in the kitchen, or men being found hunting [3; 4].

Thus, in much the same way privacy concerns beg the need for reliability, issues regarding bias require transparency [2–4]. Accountability can only be achieved through transparency – it is an absolute necessity to know why these issues occur, whether due to historical bias present in datasets, or bias from the developer. Elsewise, they will remain ingrained in AI, leading to discrimination as well as unfairness.

2.2.3 Safety

For any application of AI in the physical world, safety is usually a point of concern. In the case of autonomous vehicles, this seems obvious: ceding all control of a powerful machine to an algorithm may not be the best course of action, if the algorithm has not been tested vigorously. However, although there have been incidents involving autonomous vehicles[1] (including at least one fatal crash in 2018[2]), it is necessary to highlight the fact that the most common causes of vehicular accidents are due to driving over the speed limit, not keeping a safe distance, and many other examples of reckless driving [3]. Thus, it is not incorrect to assume that, given autonomous vehicles are programmed to follow the rules already stipulated in existing driving laws, there should at least be a reduction in human-caused accidents.

Another common example of issues regarding safety is AI applied to healthcare, more specifically, the case of misdiagnosis [2; 7]. The matter of black box models only exacerbates the issue, and thus, while the proposed models may perform well, this barrier complicates the integration of AI into healthcare [10–13].

When discussing safety, reliability and traceability are both necessary. Individuals need to feel that the AI is reliable, in the sense that they are reassured it will perform its duties safely. At the same time, traceability is crucial in order to hold the responsible parties accountable, for example, in the case of a malfunction.

In light of these concerns, as well as many others, the focus shifts to how they may be assuaged. Reliability and traceability are two possible avenues of interest, though the question of how they might be reinforced remains. This matter is addressed in a 2019 paper published from the AI HLEG [8], wherein they propose a number of guidelines for the development of trustworthy AI.

2.2.4 Trustworthy AI

Before advancing to a more detailed description of the guidelines, it is important to broach the topic of trust, and why so many deem it imperative going forward for the development of AI. There is no single definition for trust, as it is a concept dealt with in several different subjects, such as sociology, philosophy, psychology, and even economics [39]. In general, however, it can be thought of as a bidirectional relationship between two parties, A and B, wherein A believes B will act in A's best interest, thereby accepting vulnerability to B's actions [40]. Should B fail to meet A's expectations, A may feel betrayed. However, the status of "trustworthy" does not signify an inherent trust in the subject, while trust may also be placed in someone who is not "trustworthy". Additionally, there is the argument that AI should and cannot be trusted, but rather relied upon [40; 41]. Ultimately, and to reiterate the point made at the beginning of Section 2.2, the most vital aspect at this point in time is to ensure that any ethical issues are addressed

[1] https://www.nytimes.com/2022/06/15/business/self-driving-car-nhtsa-crash-data.html

[2] https://www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html

throughout the entire developmental process – including deployment and, afterwards, maintenance. Transparency must also be provided for accountability, should it ever be needed. Thus, whether the term used be reliability or trustworthiness, we as humans will most certainly have less resistance to accepting AI into our lives if these conditions are met.

According to the AI HLEG [8], a trustworthy AI consists of three components, all of which should preferably work in harmony: law, ethics, and robustness. In other words, a trustworthy AI must follow all laws and regulations in an ethical manner, while guaranteeing it maintains its integrity on both a technological and social standpoint. However, the application of all three components may not always be viable; for instance, in the case of facial recognition used to identify a high-profile criminal through CCTV footage, the individual's right to privacy – an ethical principle – should be overridden in favor of their apprehension [5; 8]. Moreover, while there are several ethical values that are universal, oftentimes ethics is a subjective topic, prone to change and evolve throughout time and even between cultures, and thus, this component may sometimes be difficult to define [42].

The basis of these core components should be fostered upon four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability [8]. Despite differences in opinion over what might constitute as ethical or moral, the AI HLEG [8] have defined these principles as a reflection of the fundamental rights that we, as human beings, must always be afforded. Such rights include respect for human dignity, freedom, equality, non-discrimination, and solidarity. They also refer to respect for democracy, justice, and the rule of law, which encompasses citizens' rights, though this point is mentioned separately.

Human autonomy is, indisputably, a core value for human dignity and freedom. Despite many different definitions, Prunkl [43] defends that autonomy can be thought of as two aspects: authenticity, and agency. In the context of AI, the preservation of authenticity implies humans must not be subject to "external manipulation or distorting influences" on behalf of AI, meaning any and all decisions made by humans are authentic to themselves. Respect for agency, on the other hand, signifies AI must allow humans full control over decisions, guaranteeing that they are able to act "on the beliefs and values they hold". Both of these definitions are in line with what the AI HLEG describes, and thus this principle intends to safeguard humans' place in the world, with AI serving to "augment, complement and empower human cognitive, social and cultural skills".

The second principle is perhaps less ambiguous, with a clear link to the issue of safety that was explored previously. As a general guidance, the AI HLEG states that no harm should be inflicted upon any living beings, or the natural environment on behalf of AI systems; in the specific case of humans, they include both the mental and physical aspect of harm for this definition. Beyond this, they stress the need for robust AI, in the sense of maintaining technical integrity against potential malicious use. Finally, special care should be taken for applications of AI regarding more vulnerable people, such as the elderly, as well as situations in which power imbalances might be present (as examples, they suggest "between employers and employees, businesses and consumers or governments and citi-

zens").

Fairness, akin to autonomy, is a somewhat nebulous concept, with numerous interpretations that may not always coincide. In the case of the AI HLEG, they consider fairness to be composed of two facets: substantive, and procedural. The first involves the development of AI systems without bias, discrimination, or stigmatisation through an "equal and just distribution of both benefits and costs". However, this definition may still be seen as relatively superficial; what may be considered as "just distribution"? Moreover, there is an overlap with the first principle, in that substantive fairness is also equated to the absence of manipulation or coercion. Procedural fairness addresses the need for accountability, seeing as how it encapsulates an individual's right to appeal "decisions made by AI systems and by the humans operating them". Beyond accountability, it is here that the AI HLEG introduces a new concept: explicability, or, in other words, an explanation of the decision-making processes of AI systems.

Thus, we are led to the final principle, which is also one of the core concepts of this thesis. To enhance the general description mentioned previously, explicability, or explainability, is fundamental for building trust between humans and AI. For many algorithms (more specifically, those we call "black box"), the process between input and output is, to all intents and purposes, an enigma for those who have little to no knowledge regarding the area of AI. They are therefore less likely to trust, or rely on, these results – especially in the case of high risk applications, such as hospitals. This topic will be broached in further depth in Section 2.3.

While these principles provide further context as to what the AI HLEG believes a trustworthy AI should act like, they remain nebulous. It is for that exact reason that they establish seven requirements regarding how to implement these same principles. They are as follow:

1. **Human agency and oversight:** Before an AI system's development, it should be assessed for any possible infringement upon humans' fundamental rights. During development, human autonomy should be taken into consideration for each step. Finally, there should be some level of involvement on behalf of humans regarding the decision-making process, whether it be mandatory for each decision, or optional.

2. **Technical robustness and safety:** AI systems should primarily be technically robust, as aforementioned. However, in the event of a security breach, there should be plans already in place to circumvent these issues. Additionally, AI should be accurate, reliable, and reproducible. In this context, the AI HLEG describes a reliable AI as "one that works properly with a range of inputs and in a range of situations".

3. **Privacy and data governance:** All data provided to AI, as well as any data that may be generated, should be protected and remain private, save for a select number of official and qualified personnel. The quality of this data should also be verified, and its integrity maintained.

4. **Transparency:** AI should be traceable, in the sense that all decisions and functionalities should be well documented to improve accountability. AI should also be explainable. Moreover, users should have the choice to choose whether or not they interact with AI in place of humans, and thus, AI agents should be labeled as such in order to prevent deception.

5. **Diversity, non-discrimination and fairness:** All biases should be avoided through careful consideration of the datasets that are being used, as well as further requirements and assessments regarding the developmental process of the system. The use of AI systems should also be intuitive and accessible to all users. Finally, stakeholders should solicit feedback for the entire duration of the AI life-cycle.

6. **Societal and environmental well being:** The impact of AI on society, the environment, and democracy must be monitored closely. Though not explored, another ethical issue related to the use of AI is its environmental impact, be it due to large consumption of energy, or waste generated due to an increase in demand of electronic goods.

7. **Accountability:** Internal and external audits of AI should be performed, with the results being made available for consultation. This is especially important in the event of negative impacts, which should be well documented and shared for the purpose of avoiding similar incidents in the future. Those affected by such events should also be fairly compensated.

The manner in which these requirements are linked to the four ethical principles can be explored through Figure 2.2.

While all of the listed requirements broach important topics for building and maintaining trust in AI, explainability will henceforth be the focus of this thesis. Namely, explainability as a means to ease the integration of AI in healthcare.

2.3 Explainable AI

When discussing the topic of XAI, the concepts of interpretability and explainability tend to be of interest. There are conflicting opinions concerning how to define these notions; while some authors consider them to be interchangeable [44; 45], others defend that this is not the case [46–48], which is the stance of this thesis. Gilpin et al. [46], for example, argue that explainability is an improvement upon interpretability – a next step towards the ultimate goal of trustworthy AI. In this context, they define interpretability as the ability to "describe the internals of a system in a way that is understandable to humans". The success of interpretability should therefore be judged on the ability of a system to make use of vocabulary that is meaningful to the user in order to produce a description that is simple to understand. However, they pose that an interpretable system alone is lacking. Explainability, on the other hand, would provide a means to not only answer questions from users, but also defend outputs, as well as permit auditability. Thus, Gilpin et al. defend that explainability may be seen as a subcategory

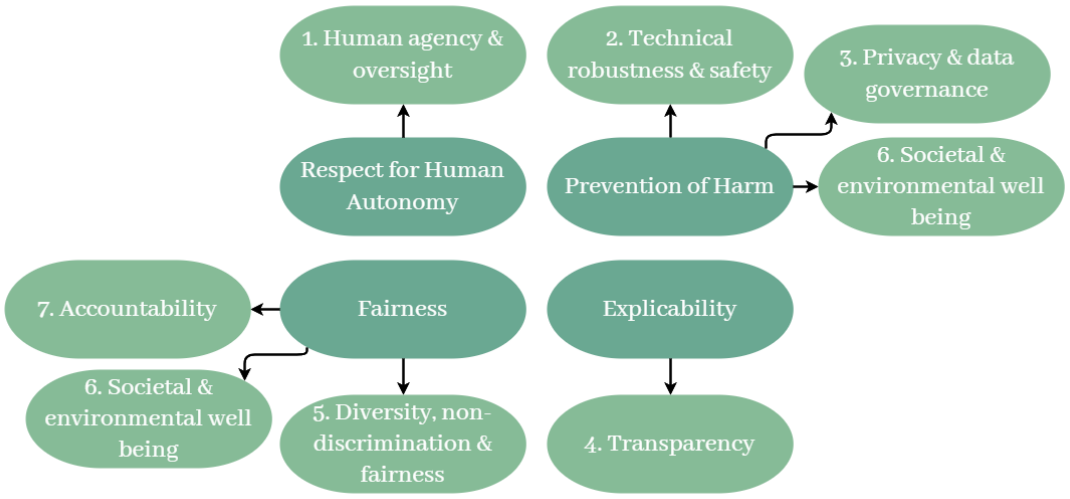


Figure 2.2: Diagram illustrating how the four core ethical principles relate to the seven requirements defined by the AI HLEG.

of interpretability – while the description of explainability implies an inherent interpretability, a model that is interpretable is not necessarily explainable.

Linardatos et al. [47], on the other hand, propose that interpretability is linked to the outputs of an AI system, while explainability focuses on the internal workings of the system. Therefore, an interpretable system is defined by its ability to allow users to understand the relationship between an AI system's inputs and outputs, while an explainable system permits a better understanding of the internal workings of the system. Contrary to the definitions put forth by Gilpin et al., this implies no correlation between either concept.

As a third point of view, Barredo Arrieta et al. [48] view interpretability and explainability as passive and active characteristics of AI systems, respectively. In other words, a system that is inherently understandable to the user is interpretable, while one that actively provides clarification regarding actions or procedures is explainable.

Why is explainability so sought after, however? From the previous definitions, we may conclude that XAI will permit a better understanding of how AI algorithms work. Thus, the question evolves: should we not already know how these algorithms reach their outputs? In the case of many current AI models, such as deep neural networks and random forest, high accuracy comes at a price: an elevated complexity which causes the inner workings to become opaque. This is commonly referred to as the black box problem, and arguably the primary catalyst for the origin of XAI. After all, and to reiterate a point made in Section 2.2, we as humans are less likely to trust something we cannot understand – especially when applied to high-risk areas. However, building trust is merely one of the reasons given for those in favor of XAI. The ability to explain unexpected decisions allows for auditability, as well as the ability to enforce some degree of fairness in systems such as these [48; 49], both of which have been discussed in Section 2.2 as desirable characteristics for trustworthy AI. Furthermore, other au-

Table 2.2: Some of the many available definitions of interpretability and explainability applied to the area of XAI.

Author(s)	Definition of Interpretability	Definition of Explainability
Gilpin et al. [46]	A system's ability to describe its internal workings in a way that is understandable to humans	A system's ability to answer questions from users, defend outputs, and permit auditability
Linardatos et al. [47]	A system's ability to allow users to identify the cause--and-effect relationships within the system's inputs and outputs	A system's ability to allow users to understand the internal procedures that take place while the system is training or making a decision
Barredo Arrieta et al. [48]	A passive characteristic of a model referring to the level at which a given model makes sense for a human observer	An active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions

thors defend that XAI will also facilitate the improvement of these same models: as our understanding of the model and its decision-making processes increases, our ability to detect possible flaws within it increases as well [44; 45; 49]. Following this reasoning, XAI may also permit a more thorough maintenance of black box models, thereby preventing errors or flaws which otherwise may not have been caught [44; 49]. The ability to detect causal relationships between data (or, in other words, causality) is another common goal for XAI, as well as an example of just how these models may help us learn [44; 48]. Finally, there is the issue of accessibility for the end users who do not have in-depth knowledge of this area, and how XAI may ease this barrier [48].

The challenge of constructing explanations is not restricted solely to translating complex computational processes to something that the user can understand, however. It stands to reason that different users may have varying degrees of expertise and, therefore, require explanations of varying complexity [49–52]. Thus, though an explanation may be technically correct (or, in other words, complete), that does not guarantee that it is understandable to the user and, consequently, may not be considered a "good" explanation [46; 53]. Gilpin et al. [46] note that the tradeoff between these two concepts must be done with some caution. The risk of creating a system that is understandable to users, but over-simplified, may be considered unethical in the sense that it is sacrificing a better understanding of the algorithm in order to be trusted. Thus, that trust, rather than be earned, is fabricated through manipulation.

In a comprehensive article, Miller [45] draws on theories from the social sciences in order to delineate a clearer idea of what a good explanation consists of. Firstly, he found that contrastive explanations are considered more insightful. In other

words, an explanation of "Why did event P happen instead of event Q?" is preferable over "Why did event P happen?". This leads us to Miller's second finding: explanations are selected. An exhaustive list of all the causes that may have led to the presented output is unappealing and oftentimes overwhelming to the user. Explanations that are simpler (those that present fewer causes) and more general (those that may be applied to more events) are far more preferable. As Miller states, "Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation". However, this selection is not without its biases; explanations that align with a user's prior beliefs are far more likely to be accepted than ones that do not. Miller also found that explanations based solely on probabilities are unsatisfying to users, with the use of concrete causes rather than statistical relationships regarded as more desirable.

From these criteria, and per Miller's own conclusions, we can safely assume explanations are heavily influenced by context. Thus, for now there is no consensus as to what a "good" explanation consists of. The analysis of inherently transparent models (or glass box models), however, may be of some interest for the construction of explainable models.

Presently, three levels of transparency may be used to describe glass box models: simulatability, decomposability, and algorithmic transparency [48; 54]. Simulatibility characterizes the ability for the user to understand the entire model at once, including all parameters and calculations made to reach the output. This implies that the model must be simple, as the user must be able to understand the entire process, from input to output, in a short amount of time. The second level of transparency describes models which present an intuitive explanation for each of its parts (input, parameter, and output), without the need for the user to be able to contemplate the entire model at once. Finally, algorithmic transparency centers on the explainability of a system's learning algorithm. These three levels may also be thought of as a hierarchy, as demonstrated in Figure 2.3, in the sense that a simulatable model is both decomposable and algorithmically transparent, however, a model that is only algorithmically transparent is neither simulatable nor decomposable.

In juxtaposition with inherent transparency, there is post-hoc explainability, or, in other words, XAI models. Currently, they may be characterized along two broad dimensions [49; 55]: the scope of the explanations, and the scope of the model itself. As for the scope of explanations, there are two subcategories [44; 47; 49; 55]: global explainability, with the objective of making the general process of the underlying black box model understandable, and local explainability, which focuses instead on explaining each individual prediction. Similarly, there are two subcategories regarding the scope of the model [47; 49; 55]: model-specific, which may only be applied to a specific class of AI, and model-agnostic, which may be applied to any class of AI.

As the field evolves, so too will the approaches to constructing XAI models. At this stage, however, it is not erroneous to believe that there is a promising future for the integration of XAI in healthcare. A diagnosis that is missed, delayed, or wrong may have deadly consequences; nonetheless, these errors are always a possibility that primary care physicians face when dealing with their patients [9].

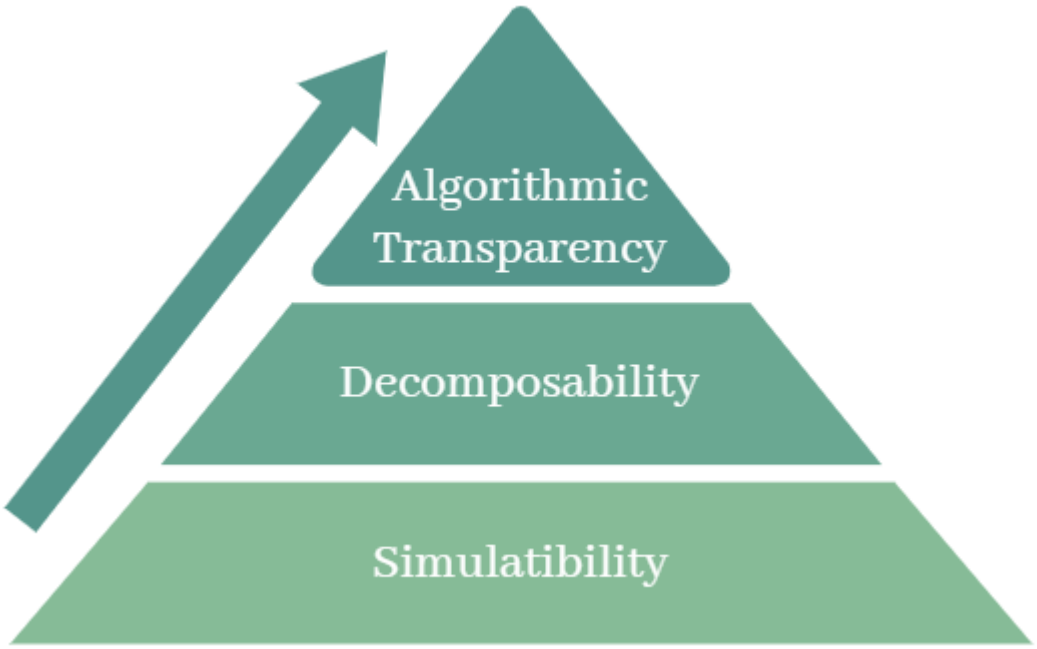


Figure 2.3: Diagram illustrating the hierarchy between the three levels of transparency.

Figure 2.4: Diagram illustrating the general pipeline of intrinsic and post-hoc XAI models.

Uncertainty on behalf of healthcare professionals, high patient volume, symptoms which may appear benign in early stages of disease, and rare diseases may all impact the accuracy and timeliness of a diagnosis, thereby causing diagnostic errors. In an attempt to diminish the occurrence of such errors, Singh et al. [9] proposed several potential improvements relative to system-based (i.e., the diagnostic process) and cognitive-based dimensions. Among these improvements, the authors point to information technology as a possible source of support, for example, through the use of algorithms. Indeed, countless AI models have been made over the years for this very purpose [56–60], with the vast majority able to reach an accuracy of over 80%. However, good performance alone, particularly in the presence of issues surrounding privacy, data protection, bias, and security, is insufficient to foster trust among healthcare professionals [10–13]. It is for this reason that the potential of XAI to mitigate such concerns is recognized as fundamental if we are to continue using AI in such applications [10; 11].

2.4 Medical Knowledge

Non-Communicable Diseases (NCDs) may be divided into five main groups: Cardiovascular Diseases (CVDs), cancers, respiratory diseases, diabetes, and, most recently, mental disorders [61]. These physical and mental conditions share amongst themselves several common threads, the most relevant being the fact that they are not caused by infectious agents, thus the term "non-communicable" [61]. However, some cancers and CVDs have been known to occur due to either viral or bacterial infections [62–64]. Furthermore, some may also find the description of NCDs too restrictive, owing to the fact that it does not include other disorders and diseases which may also be considered non-communicable, such as visual and hearing disorders, and musculoskeletal diseases [64].

Beyond the link of non-communicability, NCDs also share several risk factors, the most common being tobacco, alcohol, obesity, lack of physical activity, and an unhealthy diet. Due to the nature of these risk factors, NCDs also present a high degree of preventability [61; 64; 65]. In fact, the World Health Organization (WHO) released an action plan in 2016 which stated that approximately 80% of all cases of heart disease, stroke and diabetes, as well as 40% of cancers could be prevented, given an appropriate approach to tackling the main risk factors [65].

Adding to this, most NCDs are also chronic, meaning that they either develop over a long period of time due to sustained exposure to the aforementioned risk factors, or the nature of the disease is persistent and long-lasting. Moreover, the presence of two or more chronic conditions in the same individual (or, in other words, multimorbidity) is especially common for those of whom have been diagnosed with a NCD [61].

According to WHO, in 2016 around 71% of all deaths globally (41 million) were caused by NCDs, making it the leading cause of death worldwide, and therefore a great cause of concern [63]. Of these deaths, CVDs were responsible for about 43.6% (17.9 million), and cancer, for 21.9% (9.0 million deaths) [63]. With such a high mortality rate, it stands to reason that the decrease in incidence of these two groups are of special interest, especially in regards to ensuring the healthy aging of the global population. It is due to this reason that the datasets for this paper concern three different types of cancer, as well as CVDs.

2.4.1 Cancer

The term "cancer" is used to describe a large group of diseases which are characterized by their ability to trigger the rapid proliferation of abnormal cells anywhere in the body [63; 66; 67]. By metastasizing, the cancerous cells may spread to other organs and tissues, a process through which it becomes an invasive tumor, or, in other words, malignant. If the abnormal cells remain in the tissue wherein they were created, it is considered in situ cancer, or benign [66; 67]. Either case is the result of genetic or environmental factors influencing cells during their development, and thus leading to the acquisition of various aberrant characteristics (such as evasion of apoptosis, tissue invasion and metastasis, for exam-

ple) [68]. However, most commonly it is the metastases which pose the biggest threat to cancer patients' lives; while a benign tumor remains in the same place, the spread of malignant tumors throughout the body further complicates their removal [66; 67]. This, in conjunction with the fact that there are over 600 types of cancer, causes this disease to require a wide variety of different, and often unique, forms of diagnostic procedures and treatments [63].

While the world progresses towards better conditions for the decline of mortality from infectious diseases, as well as for the prevention of CVDs, cancer remains a cause for concern. According to the WHO Report on Cancer published in 2020, the year 2018 marked 18.1 million new cases of cancer, as well as 9.6 million deaths [63]. They estimated that by 2040, we will see these numbers nearly double, and, in fact, in 2020 there was a 6.2% increase for new cases (19.3 million), as well as a 4% increase in the number of deaths (10 million) [63; 69]. Furthermore, the World Cancer Report published by the International Agency for Research on Cancer (IARC) in 2020 named cancer as a leading threat to healthy aging due to its role in regards to premature death, or, in other words, death between the ages 30 to 69 [62]. Out of 183 countries, cancer is listed as the first or second leading cause of premature death in 134, and the third or fourth cause in an additional 45 [62]. Figures 2.5 and 2.6 offer a more detailed view on the incidence and mortality rates of the top 10 most common cancers in 2020.

As aforementioned, cancer is a NCD, and therefore may be preventable in nearly half of the cases. There are several groups of risk factors to be considered, all of which can be seen depicted in Figure 2.7. Among these groups, tobacco, alcohol consumption, and obesity are widely recognized as the major global risk factors [62; 63]. Moreover, the IARC has identified over 100 carcinogens throughout the years, providing further knowledge on how best to reduce the risk of developing cancer [62]. Be that as it may, however, it is important to note that there are other underlying factors which take part in the trends of cancer prevalence and mortality, most importantly, perhaps, being socioeconomic differences between and within countries [63].

Later stages of cancer mark a bigger development or spread of the tumors, which, in turn, leads to a worse prognosis. In other words, a low long-term survival rate. Furthermore, cancer is a very complex disease, oftentimes requiring a careful approach to treatment which frequently surpasses the intricacy of other disease management [62; 63]. Thus, in the many cases where cancer is non-preventable, screening and early detection may be the next best option [63]. Screening concerns programs, oftentimes headed by public health sectors, that focus on the target groups are the most likely to develop some specific form of cancer [63]. Therefore, the goal is to detect the disease at a pre-invasive state through precancerous lesions, or at an already invasive state, in which the patient has not yet suffered symptom onset (i.e., pre-clinical or asymptomatic cancer).

In regards to early detection, the target population is symptomatic people in the early stages of invasive cancer. Through early detection, the main objective is to diagnose cancer at a stage where it has most likely not grown, metastasized, or developed, consequently increasing the patient's long-term survival rate, and even quality of life [63]. However, for this approach to be successful it is neces-

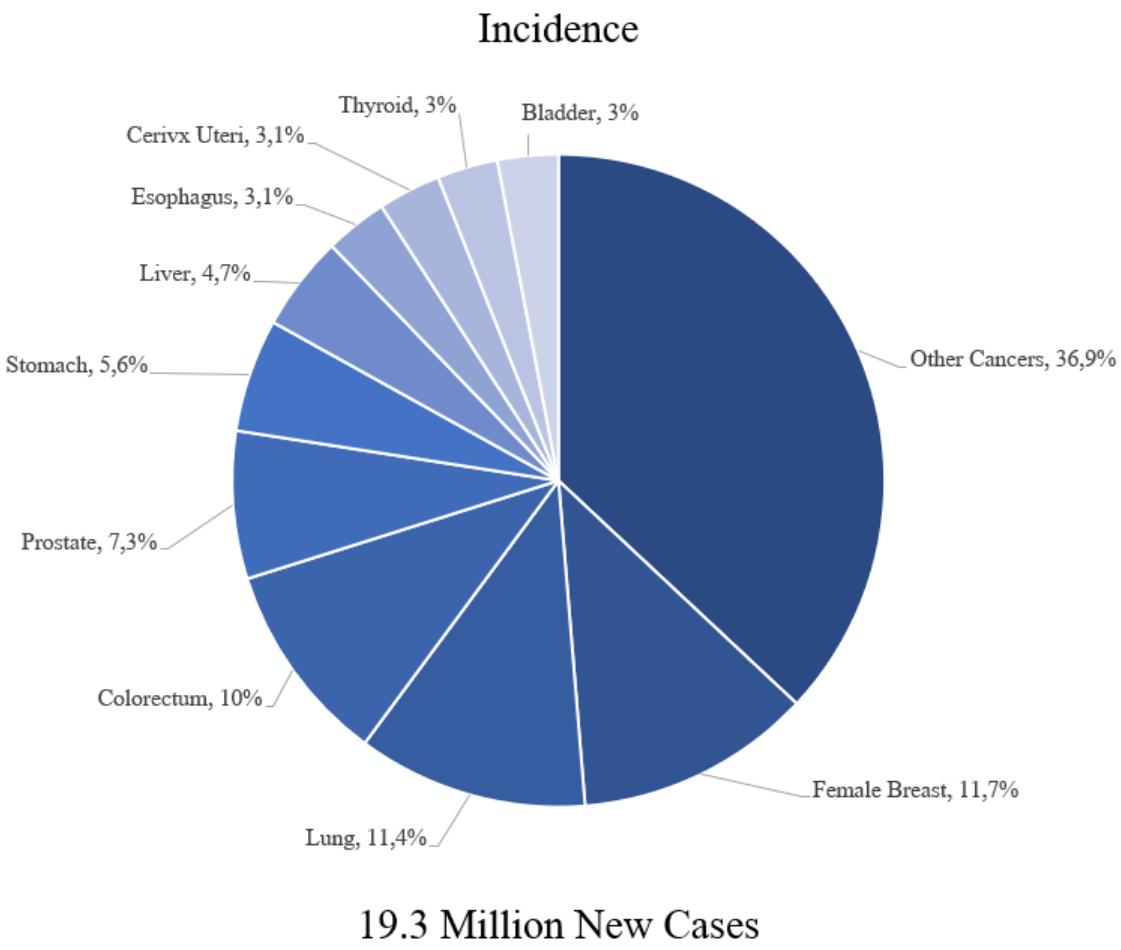


Figure 2.5: Distribution of cases for the top 10 most common cancers in 2020 for both sexes. For each sex, the area of the pie chart reflects the proportion of the total number of cases; non-melanoma skin cancers (excluding basal cell carcinoma for incidence) are included in the "other" category. Adapted from [69].

sary for not only health care providers to recognize early signs and symptoms of cancer, but also the general public. Moreover, delays amid the healthcare pipeline may further hinder the earliest possible diagnosis, as well as lack of financing for pathology and diagnostic capacity [63]. Thus, there is room – and urgency – for improvement, which AI may come to provide, as mentioned in Section 2.3.

For the purposes of this study, three of the four selected datasets concern cancer of the breast, ovaries, and pancreas. Both breast and pancreatic cancer are among the deadliest types, as pictured in Figure 2.6, with breast cancer accounting for around 6,9% of total deaths in 2020, and pancreatic cancer, 4,7% [69]. In other words, they are the fifth and seventh deadliest cancers, respectively. Moreover, breast cancer was the most prevalent type of cancer in 2020, corresponding to 11,7% of all registered cases. Finally, despite ovarian cancer only corresponding to approximately 1,6% of cases in 2020, it is no less lethal – of those cases, approximately 66,0% resulted in death [69]. Thus, due to the impact of all three types, in both prevalence and mortality, I found it apt to include datasets regarding them.

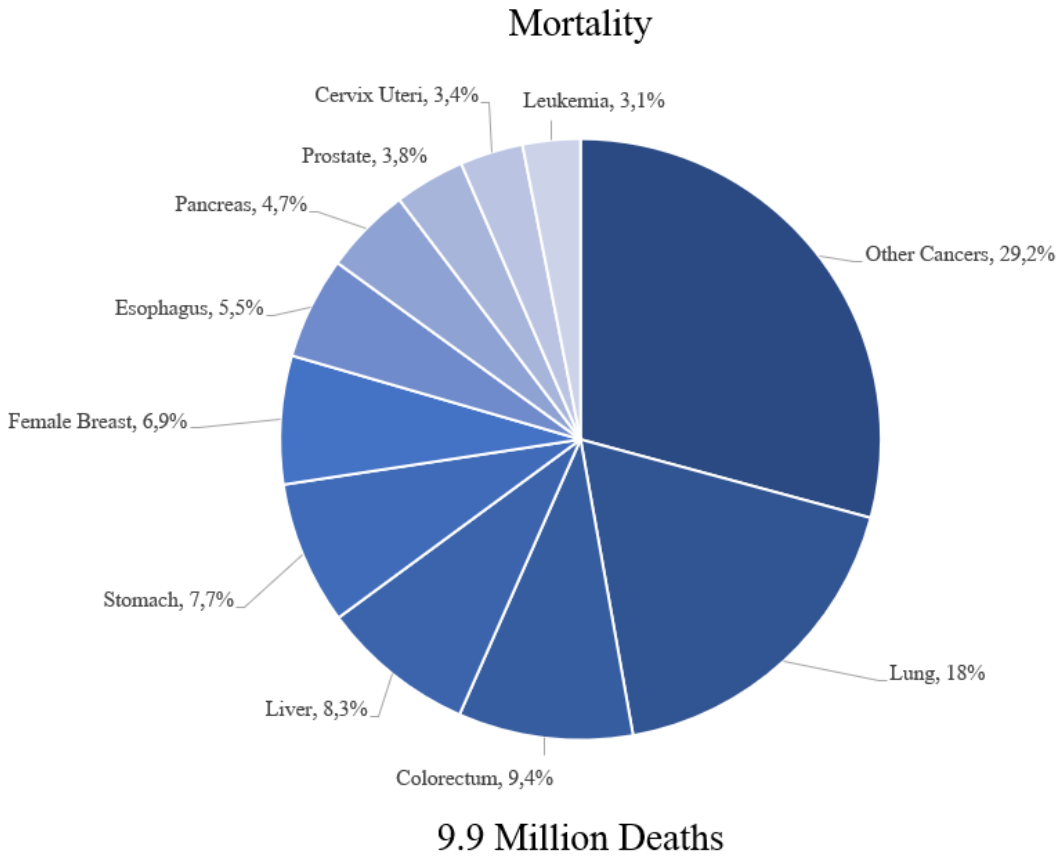


Figure 2.6: Distribution of deaths for the top 10 most common cancers in 2020 for both sexes. For each sex, the area of the pie chart reflects the proportion of the total number of deaths; non-melanoma skin cancers (excluding basal cell carcinoma for incidence) are included in the "other" category. Adapted from [69].

2.4.2 Cardiovascular Diseases

The umbrella term of CVD encompasses all afflictions concerning the heart, vasculature of the brain, and blood circulatory system [70] [p.13] [71]. Many CVDs are caused by atherosclerosis, including cerebrovascular disease, peripheral vascular disease, and coronary artery disease (otherwise known as ischaemic heart disease, or simply heart attack) [72]. Other types of CVD include heart failure, congenital heart disease, rheumatic heart disease, arrhythmias, cardiac valvulopathies, and cardiac myopathies [70][p.13][71]. Thus, much like cancer, CVDs describe a vast array of diseases, all of which are deadly in their capacity to cause complications in the event of ineffective treatment.

According to a study published by Roth et al. [73], in 2019, there was a total number of 523 million cases of CVDs worldwide - an increase of nearly 100% since 1990. CVDs were also responsible for 12.6 million deaths globally, which marked yet another increase in comparison to the statistics given by WHO for 2016. Ischemic heart disease was by far the most lethal CVD in 2019, accounting for the majority of deaths (49.2%), followed by ischemic stroke (17.7%), intracere-

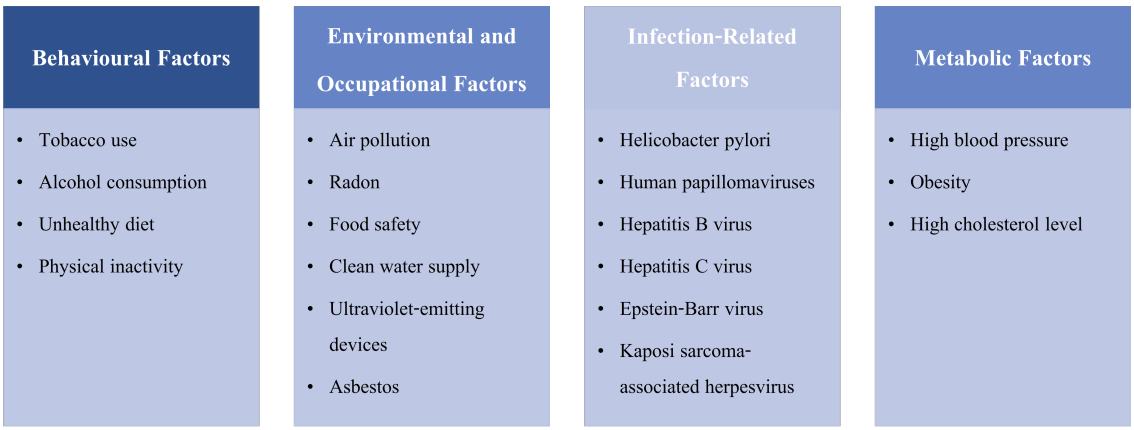


Figure 2.7: Risk factors associated to the development of cancer. Adapted from [63].

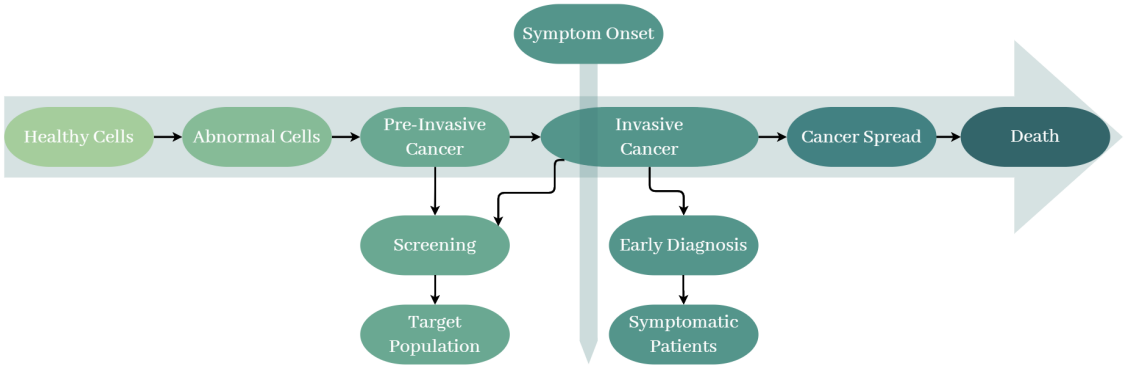


Figure 2.8: Diagram depicting the pipeline of screening in comparison to early diagnosis. Adapted from [63].

bral hemorrhage (15.5%), and hypertensive heart disease (6.2%) (Figure 2.9).

As aforementioned, prevention of CVDs is achievable due to the nature of their most impactful risk factors [70] [p.8,18,26-43] [65; 71; 74]. Use of tobacco, over-consumption of alcohol, physical inactivity, and an unhealthy diet (for example, excessive intake of sodium and saturated fats) are all behavioral factors that result in alterations at a metabolic level, namely raised blood pressure, LDL cholesterol, blood sugar, and obesity. In turn, these metabolic risk factors accelerate the process of atherosclerosis. Thus, the first line of prevention for CVDs should rely on the control of these risk factors - especially for those at high risk, such as people with diabetes, people above a certain age, and so forth. Additionally, there are non-modifiable factors, such as air pollution, genetic predisposition, gender, age, and infections. Alcohol consumption and poor nutrition during pregnancy may also result in congenital heart defects of the infant.

Similarly to the process of selecting datasets for various cancer types, the choice of a dataset for classifying a specific CVD was guided by the consideration of prevalence and mortality rates associated with these diseases. Thus, a dataset regarding ischemic heart disease was chosen, due to the extremely high proportion

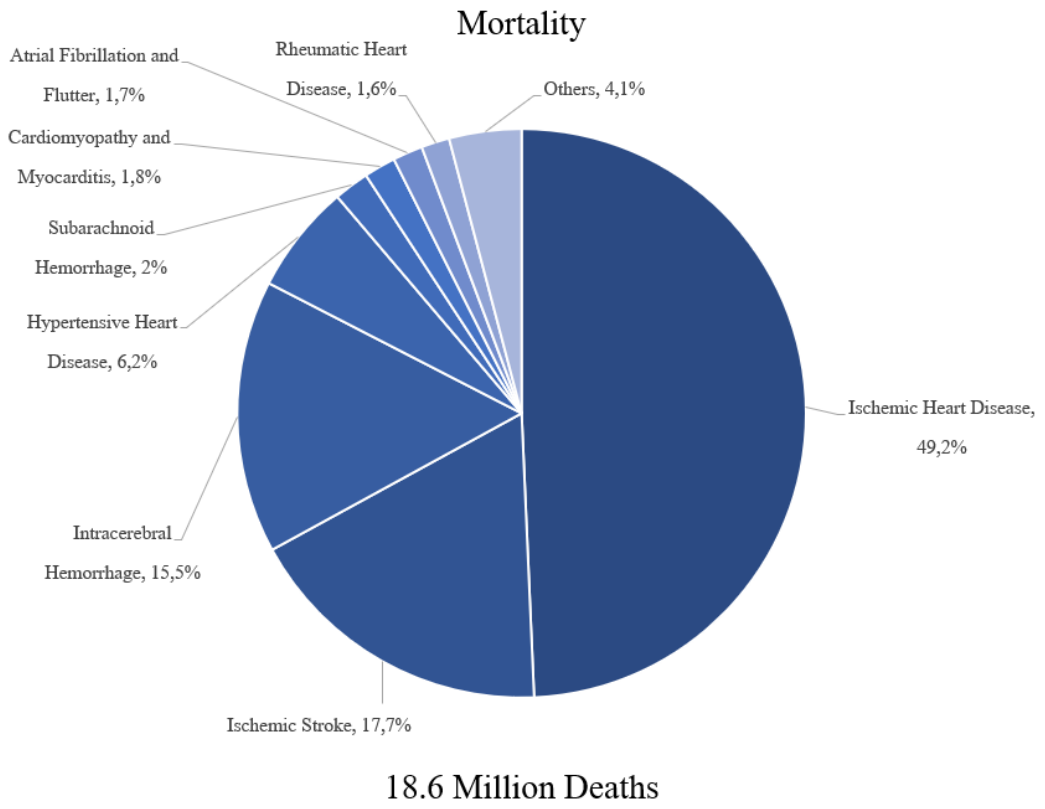


Figure 2.9: Proportion of CVD deaths by cause in 2019. Adapted from [73].

of deaths in comparison to other CVDs.

Early detection is key to improving the longevity and quality of life of both CVDs and cancer. However, human error, whether on behalf of the patient, or the physician, has proven to be one of several difficult obstacles to overcome. Hopefully, with the introduction of XAI models, there will be a greater possibility of forming trust between medical professionals and AI models, thus creating a harmonious relationship which may permit many benefits, including, of course, the global quality of life.

2.5 Summary

In this chapter, the topics of AI, ethics related to AI, XAI, cancer, and CVDs were explored.

2.5.1 Artificial Intelligence

The concept of AI has been explored in both fiction and reality since as early as 1921, with concerns about its ethical implications raised early on [23; 24]. Despite significant breakthroughs in AI throughout the years, there is still no universally agreed-upon definition [28]. However, Russel and Norvig [1][p.1-5] found that AI

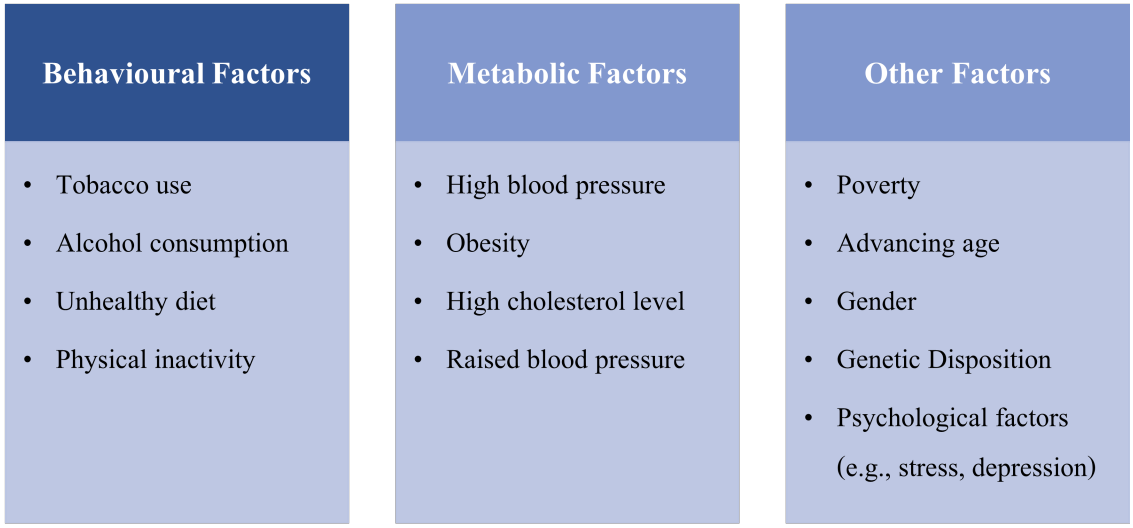


Figure 2.10: Risk factors associated to the development of CVDs. Adapted from [70].

may be categorized based on its objective of approximating rationality or humanity, as well as the goals of thought processes and reasoning versus behavior. This study adopts the definition proposed by Kaplan and Haenlein [29], which emphasizes the system's ability to interpret external data, learn from it, and achieve specific goals through flexible adaptation.

Related to AI, and within the realm of this thesis, there are the concepts of ML, AL, and DT and RF classifiers.

ML is a subfield of AI that focuses on learning from examples, definitions, behaviors, or being told [30; 31]. AL, in turn, is a subfield of ML that minimizes the need for labeled instances by selecting informative examples based on the model's current knowledge [32]. AL utilizes query scenarios like membership query synthesis, stream-based selective sampling, and pool-based AL. Related to stream-based selective sampling and pool-based AL there is the use of a determined query strategy. In the case of this thesis, uncertainty sampling was chosen. This involves the learner selecting the instances it is most uncertain about, which may be achieved through different measures.

DTs use tests related to the dataset's features in order to reach its outputs [33–35]. The tests may be thought of as questions with two possible answers, resulting in different branches of the tree. While DTs are intuitive, "white box" models, they have been known to be outperformed by more complex algorithms like ensemble methods [36; 37]. RF, for instance, is a popular ensemble technique where multiple DTs are combined, resulting in a model that, overall, performs better than a single DT [34; 37; 38]. However, this performance comes at a cost of larger complexity and, therefore, less accessibility for laymen. The opaqueness of RF and other black box models has sparked discussions on the ethics of AI, namely regarding trust in the models.

2.5.2 Ethics and AI

With the development of the field of AI, concerns about the potential consequences from its use have been raised. Some of the most common ethical concerns which also align with the topic of this thesis are: privacy and data protection, bias, and safety [2; 4].

From these concerns, stems the topic of trust in AI. In order for AI to continue to be applied in real-life situations, trustworthiness is an essential trait moving forward if we wish to circumvent the issues posed previously. In response to this issue, the AI HLEG proposed four ethical principles to guide the development of trustworthy AI: respect for human autonomy, prevention of harm, fairness, and explicability [8]. Furthermore, the AI HLEG also defined seven requirements as a means of providing more concrete guidelines on how to implement an AI that follows the previously stated ethical principles. These requirements are: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; accountability.

Explainability, the focus of this thesis, plays a vital role in building trust in AI, particularly in healthcare integration.

2.5.3 Explainable AI

The black box problem motivates the need for XAI: complex AI models can be opaque, which may compromise the trust of those who do not understand their inner workings. In this sense, the goal of XAI is to offer a solution by making AI algorithms understandable to laymen, which, in turn, will lead to trust, auditability, and fairness [48; 49]. However, constructing explanations must balance technical correctness and understandability, as different users will require varying levels of complexity [49–52].

White box models may also be referred to as intrinsic XAI models. XAI models that are applied to black box models, however, may also be called Post-Hoc XAI models [49; 55]. These models may be defined along the scope of their explanations (whether they explain the black box model's entire process, or individual predictions), and the scope of the model (whether they may only be applied to a single class of black box models, or multiple).

XAI models in healthcare have a promising future as the field evolves [9]. Diagnostic errors pose significant risks, and primary care physicians constantly face the possibility of missed, delayed, or incorrect diagnoses. Factors such as uncertainty, high patient volume, benign early-stage symptoms, and rare diseases contribute to diagnostic errors. Despite numerous AI models with good performance having been developed for diagnostic purposes over the years, trust among healthcare professionals requires more than just good performance. This is due to concerns surrounding privacy, data protection, bias, and security. Recognizing XAI's potential to mitigate these concerns is crucial for the continued use of AI in healthcare applications [10; 11].

2.5.4 Medical Knowledge

NCDs encompass cardiovascular diseases, cancers, respiratory diseases, diabetes, and mental disorders [61]. In 2016, NCDs accounted for 71% of global deaths, with cardiovascular diseases causing 43.6% and cancer causing 21.9% of these deaths [63]. The high mortality rates highlight the need to reduce the incidence of these diseases, particularly to ensure the healthy aging of the global population.

The term "cancer" refers to a group of diseases characterized by the rapid proliferation of abnormal cells throughout the body [63; 66; 67]. Genetic and environmental factors contribute to the development of cancer, which presents a significant challenge due to its various forms and diagnostic and treatment requirements. Globally, cancer is a major concern, with millions of new cases and deaths each year, and its incidence is projected to increase substantially. Risk factors such as tobacco, alcohol, and obesity play a significant role in cancer prevention, as well as socioeconomic disparities between and within countries. Early detection through screening programs and public awareness is crucial for improving long-term survival rates. The complexity of cancer necessitates advanced approaches, including AI, to aid in diagnosis and treatment.

CVDs encompass various conditions affecting the heart, brain vasculature, and blood circulatory system [70] [p.13] [71]. In 2019, there were 523 million cases of CVDs globally, resulting in 12.6 million deaths, with ischemic heart disease being the most lethal form [73]. Prevention of CVDs involves addressing behavioral factors like tobacco use, excessive alcohol consumption, physical inactivity, and unhealthy diet [70] [p.8,18,26-43] [65; 71; 74]. Non-modifiable factors like genetics, age, and pollution also contribute. Similarly to cancer, early detection is crucial for improving outcomes, and the introduction of XAI models holds promise in enhancing trust and collaboration between healthcare professionals and AI systems to benefit global quality of life.

Chapter 3

State of the Art

This chapter provides an overview of the current state of the art in AI and XAI applied to medicine. Firstly, the current state of AI classifiers in medicine will be explored in section 3.1, followed by a discussion of the innerworkings of the most relevant XAI models involved in this study, LIME and DLIME, in section 3.2. Thereafter, section 3.3 involves a review of XAI evaluation metrics, including those which were selected for this thesis. Finally, section 3.4 includes a brief summary of the topics discussed in this chapter.

3.1 AI and Healthcare

With the ability to detect and exploit underlying relationships in vast amounts of data that would otherwise go unnoticed by the human eye, AI is an excellent candidate to be applied to the area of medicine. This becomes especially clear in time-sensitive cases such as cancer and CVDs, where even the smallest details may count towards the survival of the patient. As such, many ML models have been developed over the years with the goal of aiding in the diagnostic process. Considering the ML models chosen for this thesis in conjunction with the selected datasets, this section will focus on DT and RF classifiers applied to one of the following areas: cancer of the breast, ovaries, and pancreas, as well as CVDs.

Lu et al. [75] chose DT for the prediction of ovarian cancer due to the model's simplicity and ease of interpretability, as well as its adeptness to the authors' dataset, with its ability to process both categorical and numerical data. Furthermore, ROMA, a mathematical algorithm built for calculating the probability of ovarian cancer, as well as Logistic Regression (LR), were utilized for comparison. The authors found that the DT model outperformed both ROMA and the LR model, with an accuracy, sensitivity, and specificity of 0.921, 1, and 0.899, respectively, on the testing data.

In a study done by Setiawan et al. [76], LR and DT were yet again compared, this time in the context of pancreatic cancer diagnosis. In this study, the DT model outperformed LR once more in all evaluation metrics, which were accuracy, sensitivity, recall, F1-score, specificity, and balanced accuracy.

When discussing AI models for the purpose of classification, white-box models are, in general, thought of as less powerful than black-box models, as discussed in Section 2.3. This is one of the justifications used for XAI, as there would be no need to explain these complex models if they performed similarly, or even underperformed, in comparison to white-box models. However, in some cases, this may occur. Osmanović et al. [77], for example, compared two different DT models with a multilayer perceptron for the purpose of ovarian cancer detection. The DT models used in this paper, J48 and LMT, were selected from the WEKA tool. As the authors explain, this is a data mining tool which encompasses a wide variety of ML algorithms. Despite being unable to achieve an accuracy of over 80%, both DTs were still able to outperform the multilayer perceptron. For future improvement of performance, Osmanović et al. proposed the collection of more data.

Kawakami et al. [78] employed seven different ML algorithms for ovarian cancer classification: gradient boosting machine, Support Vector Machine (SVM), RF, conditional RF, naïve bayes, neural network, and elastic net. Overall, they found that the ensemble methods (RF and conditional RF) performed best, with RF obtaining the best overall results out of all the models.

Hassan et al. [79] also compared several ML models, among them being LR, k-Nearest Neighbors (kNN), XGBoost, SVM, stochastic gradient boosted tree, naïve bayes, neural network, DT, radial bias function, RF, and multi-layer perceptron. The objective of this study was CVD detection or, more specifically, coronary heart disease detection. Among all of the utilized models, RF achieved the best performance, obtaining a score of 0.960 for accuracy.

Shan et al. [80] explored the use of DT, artificial neural networks, SVM, and RF models for the diagnosis of breast cancer. In this study, RF outperformed the DT algorithm, while the SVM achieved a higher value for area under the curve.

Shekar and Dagnew [81] selected five datasets related to ovarian cancer, as well as implementing the grid search method in order to optimize their selected model, RF. They achieved an accuracy of over 0.970 on all datasets.

Rustam et al. [82] compared both LR and RF on a dataset pertaining to pancreatic cancer. Of the two, RF achieved the highest score of accuracy, managing 0.994 in comparison to the LR's best value of 0.965.

Simegn et al. [83] also included RF among a select few models for the purpose of developing their web-based application intent on aiding CVD diagnosis. Their platform is devised of three modules: the first, for processing ECG data, the second for predicting heart disease, and, lastly, a multiclassification module for the different types of CVD. With it, Simegn et al. demonstrate how AI could be integrated into the area of healthcare in a user-friendly manner. Furthermore, such a platform could just as easily integrate some form of explainability alongside the prediction results.

Of the two classification modules, RF managed to outperform all other selected ML models, with these being a convolutional neural network, and XGBoost.

DT, SVM, and XGBoost were once again compared against RF in a paper pub-

lished by Ahamad et al. [84], alongside other models, such as LR, gradient boosting machine, and light gradient boosting machine. The authors employed the use of three different datasets, over which they concluded all classifiers performed well, with the RF, gradient boosting machine, and light gradient boosting machine achieving the best results for accuracy, sensitivity, and area under the curve. Moreover, it is interesting to note the calculation of feature importance on behalf of the authors for each of the ML models. Despite the article not being labeled as such, this may be viewed as a kind of global explainability, as it permits an overall look at one of the most important aspects the ML algorithms used to reach their predictions.

In a similar manner, Massafra et al. [85] also focused on feature importance. However, as opposed to Ahamad et al., three different models (RF, SVM recursive feature elimination, and neighborhood component analysis) were used before the task of classification in order to identify which features would be most useful. In doing so, the authors were able to extract different sets of features from each feature importance model, to then be used to train and test the selected ML models for classification. Of these, Massafra et al. chose RF, SVM, and naïve bayes.

The objective of their paper was the prediction of breast cancer recession amid two scopes: within 5 years, or 10 years. In both situations, the authors found that the RF model performed the best overall.

These articles all demonstrate a common thread of promising results for the application of AI in healthcare, and a general comparison between them may be found in Table 3.1. However, only the work of Ahamad et al. presented an approximation of explainability. Furthermore, despite all algorithms having been created for the purpose of clinical support, none of the authors attempted to evaluate their models in terms of their target audience: healthcare professionals. Thus, it remains uncertain whether or not such solutions would be welcome in their focus area. From the concerns explored in Section 2.2 and Section 2.3, it would would perhaps be more prudent to assume otherwise.

3.2 Explainable AI

Many different XAI models have been proposed along the years. As discussed in Section 2.3, they may be categorized along two different axes: in relation to which models they may be applied to, and in relation to the application of their explanations. For the purpose of this thesis, the focus lays solely on post-hoc, model-agnostic and locally explainable models. In other words, XAI algorithms which may be applied to any existent ML model, and whose explanations focus on the individual outputs of the underlying model, in place of explaining the global behaviour. Within this category, several different models may be of interest, such as SHAP [86], Anchors [87], LIME [14], and so forth.

To the best of my knowledge, there is currently only one other AL-based XAI model that is both model agnostic and locally explainable, the UnRAvEL (Uncertainty driven Robust Active Learning Based Locally Faithful Explanations) model

Table 3.1: General comparison between the articles explored in Section 3.1

Author(s)	Dataset(s) Focus	Model(s) Used	Overall Best Model(s)	Accuracy for Best Model(s)
Lu et al. [75]	Ovarian cancer	DT, LR, ROMA	DT	0.872
Setiawan et al. [76]	Pancreatic cancer	DT, LR	DT	0.100
Osmanović et al. [77]	Ovarian cancer	J48 (DT), LMT (DT), Multilayer Perceptron	J48, LMT	0.772
Kawakami et al. [78]	Ovarian cancer	Gradient Boosting Machine, SVM, RF, Conditional RF, Naive Bayes, Neural Net, Elastic Net	RF	0.924
Hassan et al. [79]	CVD	LR, kNN, XGBoost, SVM, Stochastic Gradient Boosted Tree, Naive Bayes, Neural Network, Radial Bias Function, RF, DT, Multilayer Perceptron	RF	0.960
Shan et al. [80]	Breast cancer	DT, Artificial Neural Networks, SVM, RF	RF	0.785
Shekar and Dagnew [81]	Ovarian cancer	RF	RF	>0.970
Rustam et al. [82]	Pancreatic cancer	RF, LR	RF	0.994
Simegn et al. [83]	CVD	RF, Convolutional Neural Network, XGBoost	RF	>0.933
Ahamad et al. [84]	Ovarian cancer	DT, SVM, RF, XGBoost, LR, Gradient Boosting Machine, Light Gradient Boosting Machine	RF	>0.880
Massafra et al. [85]	Breast cancer	RF, SVM, Naive Bayes	RF	>0.775

from Saini and Prasad [88]. As the authors of this algorithm describe, it generates surrogate data through uncertainty sampling, as well as gaussian process regression. They compared their model against LIME [14] and BayLIME [89] (a Bayesian extension of the LIME model) using a select few metrics, with these being uncertainty regarding explanations, and the stability of explanations. UnRAvEL was able to outperform LIME and BayLIME in both categories, demonstrating stable explanations as well as less uncertainty. However, it may be of interest in the future to implement other features, such as, for example, faithfulness to the underlying black box model, in order to achieve a more comprehensive comparison.

3.2.1 LIME

With the overall purpose of facilitating trust between humans and black box algorithms, Ribeiro et. al [14] proposed the novel explanation technique of LIME and its extension, SP-LIME. In their article, they tackled two issues: the first being trust in individual explanations, of which LIME aimed to address as a locally explainable model, and the second, trust in individual black box models as a whole, which was provided to be the goal of SP-LIME, a globally explainable algorithm. Beyond this, both methods are model agnostic, and provide support for text and image-based datasets, therefore providing a large realm of application.

Ribeiro et. al denote the model being explained as $f : \mathbb{R}_d \rightarrow \mathbb{R}$, with $x \in \mathbb{R}_d$ serving as the representation of an instance to be explained. However, as the authors defend, oftentimes the features that are used in datasets are not understandable, or graspable, to humans, with word embeddings given as an example.

Therefore, they proposed a binary representation of such features, which may denote the "presence or absence" of a word, in the case of text-based datasets, or of a super-pixel, in the case of image-based datasets. Thus, $x' \in \{0,1\}^{d'}$ was put forward to serve as a binary vector for an interpretable representation of x. In other words, for a given instance x to be explained, there is, consequently, a corresponding instance of x'.

The explanation model, with a domain of $\{0,1\}^{d'}$, is described through $g \in G$, wherein G represents a collection of possible XAI models. As a means to measure the complexity of an explanation generated through g, Ribeiro et. al define $\Omega(g)$.

The model samples instances around x' uniformly at random, which are denoted by $z' \in \{0,1\}^{d'}$. From these perturbed instances, the original values of $z \in R^d$ are obtained, which are then used to determine $f(z)$, the probability of that instance belonging to a certain class. These samples are weighted by $\pi_x(z)$, which serves as a measure of the proximity between the original instance, x, and the sampled one, z. Through these perturbations, the model aims to observe changes in prediction and, consequently, determine which attributes contribute the most to the model's classifications.

The objective, therefore, becomes the minimization of the measure $\mathcal{L}(f, g, \pi_x)$, which serves to assess g's unfaithfulness in approximating f in the locality defined by π_x, while maintaining a value of $\Omega(g)$ that is low enough so that g may be interpretable. This is described through Formula 3.1.

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3.1}$$

This general framework provides leeway for future authors to try different approaches. For example, regarding the explanation families, fidelity functions, or complexity measures. In the case of LIME, the authors chose the family of sparse linear models for their explanations, with $g(z') = w_g \cdot z'$. For the proximity measure, we have, therefore, $\pi_x(z) = exp(-D(x,z)^2/\sigma^2)$, wherein D represents a generic distance function, and σ, the width. For text-based datasets, the authors appoint the cosine distance for D, and for images, the $L2$ distance. Finally, the authors assign Formula 3.2 to the measure of complexity, with K existing to represent a limit upon the number of words or superpixels used in the explanation.

$$\Omega(g) = \infty \mathbb{1}[\|w_g\|_0 > K] \tag{3.2}$$

Since its proposal, LIME has gained a vast popularity among the XAI landscape. However, there is, of course, room for improvement. Such is the case of their choice of G, which, as they explain, cannot boast high faithfulness for models which are highly non-linear, even in the locality of the prediction. Moreover, a poor choice in parameters may lead to a model that is unable to isolate the most important features [90].

3.2.2 DLIME

Despite LIME's far reach in success and popularity, it is necessary to evaluate how it might be perceived in different situations. Namely, for the purpose of this work, in a medical setting. Due to its non-deterministic nature, in the sense that the surrogate model is trained on randomly perturbed data points, there is always the possibility that LIME may present different explanations for the same instance. This, of course, could place the relationship between the XAI model and health professionals in a precarious situation, as consistency is key to any well-grounded explanation. It is due to this reason that Zafar and Khan [15] proposed DLIME, a deterministic approach to the LIME model.

There are two main differences between the LIME and DLIME frameworks, the first being that, initially, DLIME utilizes Agglomerative Hierarchical Clustering (AHC) to partition the training dataset into clusters. Originally, all data points correspond to distinct clusters; then, each of these N clusters are merged until only a C number of clusters remain, with C corresponding to the number of classes the original dataset presents (i.e., for a binary dataset, $C = 2$).

Subsequently, instead of selecting samples in the local proximity to the test instance, x, through random perturbations, DLIME uses the kNN algorithm. In other words, the euclidean distance between x and the surrounding instances is computed, from which the k-nearest instances are selected. Of the instances that are selected, the DLIME algorithm then determines the most prevalent cluster among them, and the samples with the majority cluster label are used to train the chosen regression model. In this case, the authors chose linear regression.

Once the selection of samples has been completed, LIME and DLIME follow the same path of weighting the samples and training the interpretable surrogate model on those same weighted samples.

3.3 Evaluation Metrics

From the topics addressed in Section 2.3, we may assume that there is still much to be explored in the field of XAI, namely, evaluation metrics. To reiterate one of the points made previously, the "goodness" of an explanation is highly subjective, which lends itself to an equally high difficulty in determining how to evaluate such constructs. Thus, as of yet, there is no gold standard for evaluation metrics regarding XAI models [91–94]. Recent strides have been taken to formulate some semblance of a baseline for such purposes, however.

As a more popular example, we have the proposal put forth by Doshi-Velez and Kim [92]. In their article, they define three possible avenues through which XAI evaluation may take place: application-grounded, human-grounded, and functionally-grounded. For application-grounded approaches, the main goal is to utilize a domain expert to determine the quality of an explanation. In other words, these experts are requested to perform some experiment which encompasses the end-task of the XAI model (e.g., the diagnosis of cancer), while having

at their disposal XAI-generated explanations. Thus, if the explanations are able to aid the experts in any way, whether that be through "identification of errors, new facts, or less discrimination", it becomes easier to determine whether or not they may be considered "good". Furthermore, Doshi-Velez and Kim stress that the impact of human-formulated explanations in the same context should be considered as a baseline for the evaluation of XAI models.

While both application-grounded and human-grounded approaches are centered on the use of human subjects, human-grounded evaluations involve laymen, rather than domain experts. These experiments should serve as a more general evaluation of the quality of the explanations, regardless of whether or not the output is correct. For such, Doshi-Velez and Kim provide three examples: binary forced choice, wherein the human must choose which explanation they prefer over a series of options; forced simulation/prediction, wherein the human is tasked with correctly simulating the model's output while only being given the input and the output's explanation; counterfactual simulation, wherein the human is tasked with perturbing the input so as to alter the output, with the input, output and explanation at their disposal. Furthermore, the use of laymen, rather than domain experts, allows for more ease in obtaining a larger sample size.

Overall, while human-grounded approaches appear to be more appealing to those with less resources, both categories involving humans will always be costly. This may show through tangible expenses to compensate for their involvement, or time spent explaining the experiments, as well as time spent performing them. As a result, Doshi-Velez and Kim propose their third and final category, functionally-grounded. These evaluation metrics base themselves on formal definitions of explainability, with the ultimate goal of approximating various characteristics through proxies in order to evaluate them. For example, the simplicity of an explanation may be approximated through the number of rules, or even features, used to reach the output.

In a similar vein, Murdoch et al. [94] present the Predictive, Descriptive and Relevant (PDR) framework. Likewise to Doshi-Velez and Kim, this framework presents three categories – predictive accuracy, descriptive accuracy, and relevancy – through which XAI models may be evaluated. Furthermore, the PDR framework serves also as a guide for the selection and construction of such models.

Predictive accuracy is described as a measure of faithfulness for the approximation of underlying data relationships with a black box model. In this sense, the objective is to evaluate how well the base algorithm performs, before it is used for the purpose of XAI. If its performance is lacking, then, consequently, any explanation derived from it cannot be considered trustworthy.

The PDR framework's second category, descriptive accuracy, aims to gauge how well the XAI model is able to approximate the base model. Thus, it is most apt to evaluate post-hoc methods.

Finally, Murdoch et al. introduce relevancy as their third and final desiderata for XAI. In this case, relevancy refers to the information presented in the model's explanation – thus, its purpose lies in determining how well-suited the explanations

Table 3.2: The Co-12 explanation quality properties, grouped by their most prominent dimension. Adapted from [95].

Dimension	Name	Description
Content	Correctness	Faithfulness of the explanation to the black box
	Completeness	How much of the black box's behavior is included in the explanation
	Consistency	How deterministic and implementation-invariant the explanation method is
	Continuity	How continuous/generalizable the explanation function is
	Contrastivity	How discriminative the explanation is with other events or targets
	Covariate Complexity	How complex the features in the explanation are
Presentation	Compactness	The size of the explanation
	Compositionality	The format and organization of the explanation
	Confidence	The presence and accuracy of probability information in the explanation
User	Context	How relevant the explanation is to the user and their needs
	Coherence	How accordant the explanation is with prior knowledge and beliefs
	Controllability	How interactive or controllable an explanation is for a user

are to their application. Similarly to Doshi-Velez and Kim's application-grounded and human-grounded approaches, Murdoch et al. defend that relevancy may only be evaluated through human involvement.

A comprehensive survey published by Nauta et al. [95] provides an in-depth view of quantitative evaluation metrics. As a result of an extensive review of 361 papers, the authors define 12 categories for XAI evaluation metrics, which they denominate as the Co-12 properties. The articles surveyed either introduced a new XAI method or evaluated an existing model, and were all published between 2014 and 2020. Of the proposed categories, Nauta et al. align them along three different dimensions, these being content, presentation, and user. However, it is important to emphasize that the authors recognize that trade-offs may have to be realized, as some categories oppose one another. Therefore, not unlike other dimensions of XAI, researchers must carefully consider what aspects are most important to the end-goal of their models, and evaluate them appropriately. All Co-12 properties may be viewed in Table 3.2, alongside a general description of them.

As for concrete metrics, those proposed by the authors of LIME and DLIME are perhaps of most interest. Thus, they will now be explained in-depth.

3.3.1 LIME Evaluation Metrics

In order to evaluate their newly proposed model, Ribeiro et al. [14] implemented both human and functionally-grounded evaluation metrics.

Faithfulness to the Model: In the case of surrogate models such as LIME and DLIME, the question of faithfulness to the black box model becomes necessary to address. Any explanation that is generated becomes meaningless if the instance it is explaining has no correlation to the original classifier. Thus, Ribeiro et al. [14] proposed their own metric for this purpose.

By training an interpretable classifier (such as, for example, a DT, or sparse linear regression algorithm) with a maximum of 10 features to be used for any individual instance, the authors are left with a gold standard of features for that model. After applying LIME on the test set, they compute the fraction of gold standard features that are used in the XAI model's explanations.

F1 of Trustworthiness: For this metric, Ribeiro et al. first perform a random selection of 25% of the dataset's features to be deemed untrustworthy, with the presumption that users can also identify them as such. If, by removing all "untrustworthy" features, the model's classification changes, then the trustworthiness oracle will consider the prediction as untrustworthy. This is accomplished through the use of the F1-score, with the original set of predictions used as the ground truth.

Model Selection with Simulated Users: In order to simulate a situation wherein a user must choose between two models with similar values of accuracy and validation data, Ribeiro et al. introduced 10 artificially noisy features in the test, training and validation datasets. However, these features were only applied to 20% of instances of one class, and 10% of the other class for training and validation data, while for testing data, the features were applied equally in both classes, at 10%. Thus, introducing a scenario where the model will use both features with meaning, as well as purposefully noisy features.

Using the validation data, the simulated users must then choose the more trustworthy explanation between both models, thereby choosing the overall better model.

Model Selection with Real Users: For this experiment, Ribeiro et al. employed real humans from Amazon Mechanical Turk in order to see whether or not an explanation would facilitate the process of choosing the better model between two options. To achieve this, they first trained the same classifier on two different datasets: one that had been altered manually to include only the most relevant features (the "cleaned" version), and the other, unaltered (the original version). In their example, the "cleaned" dataset performed better overall, however, it demonstrated lower accuracy on the test data. Thus, if one were to base a decision solely off of accuracy, the chosen classifier might not be the best one.

Given these conditions, the human subjects are then asked to choose the best model, with the prediction, explanation, and raw data used for the classification at their disposal. By previously identifying the best classifier, Ribeiro et al. are

therefore able to discern the percentage of correct choices made by the users, and, in doing so, may then gauge whether or not the explanations were at all helpful.

Improving the Classifier: Using the classifier deemed worse from the previous experiment, Ribeiro et al. asked the users to eliminate the features they considered to be least useful to the classifier, with the overall objective of improving its accuracy. This experiment was carried out through three rounds of interaction, beginning with 10 subjects in the first round, then 5 for the second and third rounds, and resulting in a total of 250 versions of improved classifiers (one for each subject across the three rounds). Finally, the accuracy is averaged over all versions of the final, improved classifier. It is interesting to note, as well, that this may be considered an example of one of the experiments proposed by Doshi-Velez and Kim for their human-grounded metrics.

Insightfulness: Oftentimes, models identify erroneous correlations present in the underlying relationships from the data they are trained on. This may be difficult to identify solely through the predictions and raw data, as Ribeiro et al. defend, and thus this experiment serves to judge whether or not an explanation may help such issues.

First and foremost, Ribeiro et al. purposefully introduce incorrect correlations through a carefully selected dataset. In this case, the classifier was made to classify images of wolves and huskies, where all instances of wolf images used during training had snow in the background. In doing so, the model was trained to classify any image as "wolf", granted there was snow present in the background.

Graduate students with a minimum amount of knowledge in ML were then presented with a balanced set of 10 predictions, without explanations. Of these predictions, 2 were selected to induce an incorrect prediction: one image contained a husky with a snowy background, while the other contained a wolf with a different colored background. The students were then asked to answer three questions: firstly, if they trusted the algorithm; secondly, why or why not; thirdly, how they think the algorithm distinguishes between classes. After collecting the responses, the students were then presented with the same images, their corresponding predictions, and LIME's accompanying explanation, and were asked the same questions as previously.

3.3.2 DLIME Evaluation Metrics

With the aim of measuring the stability of their model, Zafar and Khan [15] used Jaccard's distance. This metric quantifies the similarity between two finite sets of data points, S_1 and S_2. Jaccard's coefficient, shown in Formula 3.3, describes the fraction between the intersection of S_1 and S_2 and their union. Thus, a result of $J(S_1, S_2) = 1$ would imply high similarity between S_1 and S_2, while a result of $J(S_1, S_2) = 0$ would imply the exact opposite, meaning no similarities may be found between the sets.

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{3.3}$$

From Jaccard's coefficient we may reach Jaccard's distance, shown in Formula 3.4. Based on the previous metric, we may extrapolate that the higher the distance, the more dissimilar S_1 and S_2 are. Therefore, in the case of DLIME, the objective lies in achieving a result of $J_{distance} = 0$ for explanations surrounding the same instance.

$$J_{distance} = 1 - J(S_1, S_2) \tag{3.4}$$

Beyond Jaccard's distance, the authors of DLIME also employed the use of a metric of faithfulness of the XAI algorithm to the underlying black box model. Contrary to Ribeiro et al., Zafar and Khan first obtain a set of true predictions ($e = [e_1, e_2, ..., e_N]$) from their baseline model (linear regression), then compute the cosine similarity score between these predictions, and the predictions obtained from the XAI model ($e' = [e'_1, e'_2, ..., e'_N]$), shown in Formula 3.5. The final quality score, q_s, is obtained through the average of all cosine similarity scores (Formula 3.6).

$$q_i = \frac{e_i \cdot e'_i}{||e_i|| \cdot ||e'_i||} \tag{3.5}$$

$$q_s = \frac{1}{n} \sum_{i=1}^{n} q_i \tag{3.6}$$

As for the quality of DLIME's classifications, Zafar and Khan used precision, recall, F1-Score, accuracy, and balanced accuracy. Beyond this, the authors also aimed to provide a comprehensive justification for the use of AHC, kNN, and linear regression in their model. Therefore, they constructed several models based on the DLIME framework: DLIME-KM, which uses k-means in place of AHC, thereby demonstrating the superiority of AHC in finding local subspaces; DLIME-NN, which forgoes the use of AHC, in order to once again prove the usefulness of AHC; DLIME-Tree, which utilizes tree regression in place of linear regression in an attempt to compare their ability of generating explanations.

3.4 Summary

This chapter discussed some of the most relevant AI models in healthcare related to the topics dealt with in this thesis, as well as providing a more in-depth explanation of how LIME and DLIME function. Some of the most relevant evaluation metrics were also explored, with a special focus on those used by the authors of LIME and DLIME.

3.4.1 AI and Healthcare

There are a wide number of existent ML methods used in the areas of breast cancer, ovarian cancer, pancreatic cancer, and CVD diagnosis. Of those discussed in this section, it was shown that RF consistently outperformed other ML models, including DT. However, when isolated, DT also proved to be robust. Moreover, the lack of human involvement in all of the presented studies was notable.

3.4.2 Explainable AI

LIME is a post-hoc, model agnostic, locally explainable model. By randomly perturbing data points, it views the underlying model's differences in prediction in order to determine feature importance. Once the model has chosen a set of samples, these are then weighted, and then used to train the surrogate model. Then, for each explanation, the surrogate model performs feature selection to determine the most important features used for that determined classification, and generates an interpretable representation.

DLIME was created in order to explore the possibility of a deterministic version of LIME. It differs from LIME in two aspects: firstly, it partitions data into clusters using AHC, leaving as many clusters as there are classes in the dataset; secondly, in place of random perturbation, it uses kNN in order to determine the k-nearest neighbors to a determined instance, and the neighbors with the cluster label corresponding to the most prevalent cluster within the vicinity are selected. Following this, it proceeds in much the same way as LIME.

3.4.3 Evaluation Metrics

As of yet, there is no "gold standard" when selecting metrics to evaluate XAI models. However, several authors have attempted to define some guidelines for this purpose. Doshi-Velez and Kim, for example, proposed three different categories of metrics: application-grounded, which uses domain experts in order to evaluate the XAI models; human-ground, which uses laymen, or humans with simple knowledge regarding the subject manner; functionally-grounded, which implements proxies of formal definitions regarding explanations.

Murdoch et al. proposed the PDR framework, which consists of: predictive accuracy, which aims to evaluate the faithfulness of the XAI model in approximating the underlying data relationships the black box has formed with its classifications; descriptive accuracy, which aims to evaluate the faithfulness to the underlying model; relevancy, which aims to evaluate the information presented in the XAI model's explanations.

Nauta et al. performed a survey of XAI models, and constructed their Co-12 framework. Their objective was to provide a more comprehensive categorization of XAI evaluation metrics, from which they derived twelve different categories. Moreover, these categories may be grouped along three different dimensions:

content, presentation, and user.

The authors of LIME defined several different metrics for the purpose of evaluating their model. Of these metrics, they included both human and functionally-grounded approaches. They evaluated several different characteristics of their model, including: its faithfulness to the underlying black box model; its trustworthiness in regards to its behaviour among untrustworthy features; the ease through which users could select the better model between two choices; the ease through which users could improve the model; the ease through which users could identify errors in the model.

In contrast, DLIME was evaluated solely through functionally-grounded metrics. Of these, the authors proposed Jaccard's distance in order to measure explanation stability. Furthermore, Zafar and Khan also applied a metric to evaluate the faithfulness of the model to the underlying black box model, precision, recall, F1-score, accuracy, and balanced accuracy.

Chapter 4

Material and Methods

This chapter serves to provide a description of the steps taken for the development of the XAI pipeline. First, an overview of the general pipeline is presented in Section 4.1, followed by a description of all datasets used for the purpose of this thesis in Section 4.2, and the methods involved in the pre-processing of such data in Section 4.3. Thereafter, the choice of XAI and ML models is explored in Section 4.4, and the process of ML model optimization in Section 4.4.4. The metrics chosen to evaluate both types of models will be explained in Section 4.5, followed by a description of the general methodology undertaken during this thesis in Section 4.6. Finally a summary of this chapter may be found in Section 4.7.

4.1 Pipeline Overview

A standard ML framework typically includes stages such as data extraction, data pre-processing, classification, and performance evaluation. In addition to these steps, this thesis also includes model optimization and the application of the post-hoc models LIME, DLIME, and AL-DLIME to the RF algorithm, as well as an evaluation of their performance. Moreover, two experiments were conducted in this thesis: the first involved the standardization of each dataset, which permitted the addition of two evaluation metrics (single and incremental deletion); as for the second, there was no process of standardization in the pre-processing stage, which, therefore, left only accuracy, F1-score, faithfulness, and Jaccard's distance for evaluation metrics.

Figure 4.1 presents a flowchart which illustrates the pipeline of this thesis in greater detail.

Summarily, ensuing the selection of datasets, the steps for either experiment may be described as follows:

- Pre-processing: Improvement of the quality of the data;
- Model Optimization: Selection of the optimal set of hyperparameters for both the RF and DT models;

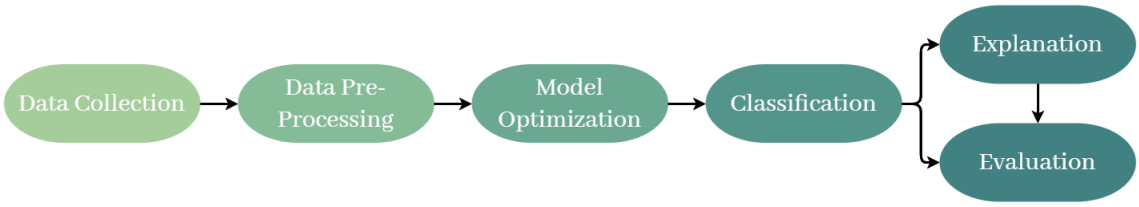


Figure 4.1: Flowchart which depicts the general pipeline of this study.

- Classification: Training and testing the optimized ML models;
- Explanation: Application of LIME, DLIME, and AL-DLIME on the RF model;
- Evaluation: Application of evaluation metrics to gauge the performance of the ML and XAI models.

The subsequent sections of this chapter will provide a more thorough explanation of each step.

4.2 Datasets

Once the problem definition was completed for this thesis, it was necessary to choose which datasets would be most apt. Owing to the high mortality and misdiagnosis rate of both cancer and CVDs, I thought it best to procure datasets related to either group. For this purpose, I used the online platform of Kaggle. This platform allows users to publish datasets, as well as vote on the quality of others. Furthermore, once published, Kaggle will calculate a metric of "usability", which exists on a scale of 0 to 10. For a dataset to achieve a high "usability" rating, it must comply with a series of requirements defined along three categories: completeness, which relates to whether or not the user has provided a subtitle, cover image, tag, or description regarding the published dataset; credibility, which includes whether or not the user has provided a source for the data, if the dataset is public or not, and the frequency in updates to the data; compatibility, which relates to the license information of the data, the file (or files) format, and whether or not the user has provided a description of the files, and of the columns (which, in this context, would correspond to the features).

Given these criteria, I applied the search term of "cancer classification" and selected the highest rated datasets which also presented a high "usability" score. For CVDs, I applied the terms of "heart failure classification", "cardiovascular disease classification", and "heart disease classification", and used the same selection criteria as I had previously with the cancer datasets. Beyond this, I prioritized larger datasets in order to avoid overfitting, in addition to removing any image-based datasets, as DLIME is currently only capable of working with tabular datasets. This left four datasets: three related to cancer, and one, to CVDs. They will now be discussed in greater detail, with Table 4.1 providing a general comparison between them.

Table 4.1: General comparison between the datasets utilized in this study.

Dataset	Type	Instances	Features	Classes	Class Labels	Class Distribution
Breast Cancer	Tabular	569	30	2	0: Benign 1: Malignant	0: 63% 1: 37%
Ovarian Cancer	Tabular	349	47	2	0: Benign 1: Malignant	0: 49% 1: 51%
Pancreatic Cancer	Tabular	590	12	3	1: Healthy 2: Benign 3: Malignant	1: 31% 2: 35% 3: 34%
Heart Disease	Tabular	918	11	2	0: Healthy 1: Heart Disease	0: 45% 1: 55%

4.2.1 Breast Cancer Dataset

The Wisconsin breast cancer dataset is commonly used for the purpose of evaluating ML classifiers. Available through the UCI ML repository[1], it comprises 569 instances and 30 features, with a binary target (benign, or malignant). As shown in Table 4.1, this is the most unbalanced dataset, with a class distribution of 63% to 37% for the benign and malignant classes, respectively. This is addressed in Section 4.3, which details the steps taken during pre-processing.

A fine needle aspirate was taken of a breast mass, from which the authors computed 10 core features from each cell nucleus present in the digitized image. These core features are the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Following this, the authors then calculated the mean, standard error, and worst of each core feature, resulting in a total of 30 features. In other words, for each core feature, there exists three different variations (for example, there is a mean radius, standard error radius, and worst radius). For further details, refer to Table A.1.

4.2.2 Ovarian Cancer Dataset

This dataset[2], published alongside the study carried out by Lu et al. [96], was constructed with the purpose of ovarian cancer classification, and has a total of 349 instances and 47 features.

There are five files provided by Lu et al., titled "Supplementary Data" 1 through 5. Among these files, we may find the entirety of the raw data, a description of each feature, all raw training data with the exception of the biomarker CA72-4, and a division of data for the purpose of training and testing, with the former containing 235 instances, and the latter, 114.

The majority of features consist of biomarkers, with the addition of demographic information, such as age, and whether or not the patient has gone through menopause. Once again, the dataset is binary, with classes differentiating between benign

[1]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
[2]https://data.mendeley.com/datasets/th7fztbrv9

ovarian tumors, and malignant ones. More in-depth details may be found in Table A.2.

4.2.3 Pancreatic Cancer Dataset

Debernardi et al. [97] collected samples from multiple sources[3] (more specifically, from Barts Pancreas Tissue Bank, the University College London, the University of Liverpool, the Spanish National Cancer Research Center, Cambridge University Hospital, and the University of Belgrade), resulting in a dataset containing 590 instances and 12 features. Similarly to the previous dataset, several demographic features are included, such as age and sex, with the remaining features corresponding to various biomarkers (such as, for example, creatinine).

This is the sole multiclass dataset, with three classes to consider: healthy patients, patients with non-cancerous pancreatic conditions (such as pancreatitis), and, finally, patients with pancreatic cancer, (more specifically, pancreatic ductal adenocarcinoma). In Table A.3, further details regarding this dataset may be found.

4.2.4 Heart Failure Dataset

Finally, I selected the following dataset for CVD classification[4], or, heart disease classification. It is composed of five different heart disease datasets that have been published in the UCI ML repository (all avaliable under the index of heart disease datasets[5]), with a total of 918 unique instances, and 11 features.

The data includes categorical and continuous values, with demographic information regarding the patients, such as age and sex, as well as information resulting from routine tests, such as chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, ST segment measurement, and ST segment slope type. Once more, the dataset is binary, with a class for healthy individuals, while the other serves to describe patients with some sort of heart disease. A more detailed view of this dataset may be found in Table A.4.

4.3 Data Pre-Processing

After selecting the datasets, pre-processing was necessary in order to ensure the data was ready to be applied to the ML models. As the majority of the data had been used previously in scientific studies, only minimal pre-processing was required.

[3] https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer
[4] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
[5] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Firstly, each dataset was analyzed for the purpose of identifying missing values. This proved to be the case for the ovarian cancer dataset (henceforth referred to as the OCD), as well as the pancreatic cancer dataset (henceforth referred to as the PCD). Regarding the OCD, several features (more specifically, the biomarkers AFP, CA125, CA19-9, CA72-4, and CEA) had multiple missing values. In order to preserve the number of instances, these features were excluded from the dataset. Thereafter, all patients with missing values were also removed, thus leaving the OCD with 212 instances, and 44 features. With respect to the PCD, four features were found to contain a majority of null values: the biomarkers plasma CA19-9 and REG1A, stage, and benign sample diagnosis. Considering the fact that each of these features contained information regarding solely to one class, I found it pertinent to remove them, thus leaving the dataset with eight features.

In terms of data cleaning, it was necessary to convert categorical values into numerical values for the PCD, the breast cancer dataset (henceforth referred to as the BCD), and the heart disease dataset (henceforth referred to as the HDD). For such, I attributed different integers for each unique category, beginning with a scale of one. As an example, we have the feature of sex, in which female would be attributed to 1, and male, to 2.

Data balancing was another point of interest during this stage. Before altering the datasets in any way, it was necessary to first verify whether or not they were reasonably balanced. As we may observe through Table 4.1, and as has previously been mentioned, the largest disparity between classes was observed with the BCD, which corresponded to a difference of around 26% between class 0 and class 1. On the other hand, among all datasets, the OCD and the PCD were the most balanced, with disparities no larger than 5% between each class. However, due to the fact that instances were removed from the OCD, it was necessary to verify if the data had become skewed, which was not the case.

Each dataset was then split into sets for training and testing the ML models that would be used posteriorly. For such, the standard split of 80-20 was chosen for training and testing, respectively, for all datasets. Following this, a copy of the processed datasets were saved, and should be henceforth known as the non-standardized datasets. Afterwards, another copy of the datasets was made, with these copies then undergoing the process of standardization (or, normalization) for use in the first experiment. For such, the *StandardScaler* class from the *sklearn* library[6] was used, which conducts a simple process of standardization. In other words, for each data point, the mean is subtracted, and consequently divided by the standard deviation. This is necessary for some models which perform best on data that behaves like standard normally distributed data, as is the case for ridge regression [21] [p.82].

[6]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

4.4 Models

Beyond datasets, the selection of appropriate models is a crucial aspect in attaining optimal performance in any analysis. For the purpose of this thesis, with its main focus relying on explainability in the area of medicine, it was necessary to not only select an underlying black box model, but also a XAI model. In this case, I have opted to conduct a thorough comparison between two existing XAI models, as well as propose a novel model.

4.4.1 AL-DLIME

One of the main contributions of this study is the proposal of AL-DLIME. As aforementioned in Chapter 1, this model is based on the DLIME framework, with one key difference: the use of AL in place of AHC and kNN. This alteration preserves the original model's determinism while exploring an entirely different branch of AI, one which I hope will improve the transparency and interpretability of complex ML models or, at the very least, lend some interest towards future research involving the same concepts.

As mentioned in Section 2.1.2, AL focuses on minimizing the amount of labeled instances that is necessary to achieve a good performance [32]. In the case of the area of medicine, where there is a common problem of large datasets being mostly (or completely) unlabeled due to the medical professionals' lack of time [98; 99], AL could plausibly present a more attractive choice over other ML models.

For the purposes of AL-DLIME, pool-based sampling was selected due to its popularity [100]. However, in cases of limited memory or processing power, stream-based or membership synthesis scenarios are the more preferable options [32]. As for the query strategy, uncertainty sampling was selected.

The basis of AL-DLIME, therefore, is quite simple. Firstly, a logistic regression model is trained on a small set of labeled data. Following this, the probabilities for class distribution are obtained on a larger, unlabeled set of data through the LR model. A range of uncertainty is selected, which may vary depending on how strict the process of sampling is determined to be. In the case of this thesis, a range of 47% to 53% was used, as I believed it to be an acceptable representation of the degree of uncertainty I wished to explore. All instances from the previously collected set of probabilities which belonged to this range were then selected as the most informative instances. This process may be viewed in Algorithm 1.

Once AL-DLIME has selected a set of instances with the highest level of uncertainty, it proceeds in much the same way as DLIME and LIME (Figure 4.2). These instances are then weighted and labeled using the chosen black box model (in this case, RF), thereafter being used to train a weighted and interpretable white box model, the surrogate model (in the case of this study, ridge linear regression). Feature selection is then achieved through the newly trained white box model, and an interpretable representation is generated.

Algorithm 1 Selection of the Most Informative Instances

Input: Dataset D_{train}, Dataset D_{test}, Labels L_{test}
Initialize $S \leftarrow \{\}$
Initialize $S_{labels} \leftarrow \{\}$
Initialize $ind \leftarrow \{\}$
C_{lr} = Create new Logistic Regression model
Train C_{lr} with D_{test}
y = Probabilities regarding all instances from D_{train} to belong to class 0
$p = 0.47$ (uncertainty interval from 0.47 to 0.53)
for i from 0 to (number of rows in D_{train})-1 **do**
    if instance i from y is between p and $(1 - p)$ inclusive **then**
        $ind_i \leftarrow \{i\}$
    end if
end for
$S \leftarrow$ Get instances from D_{train} with indices ind
$S_{labels} \leftarrow$ Get instances from L_{train} with indices ind
return S, S_{labels}

4.4.2 Baseline Explainable Models

For this thesis, I first selected LIME [14] as a model for explainability due to its robustness and high popularity in the area of XAI. However, to reiterate a point made in Section 3.2.2, Zafar and Khan [15] highlighted the non-deterministic nature of LIME as a XAI model. Given one of the main objectives of this thesis, which is to evaluate the performance of XAI in the medical field, it is imperative to have a deterministic approach in order to maintain the trust of medical professionals. Therefore, I have also employed DLIME to allow for a more comprehensive comparison between the two models, in addition to the metric of stability used in the original paper by Zafar and Khan [15].

4.4.3 Baseline Black and White Box Models

The choice of ML model was made after careful consideration of the limitations of the procured datasets – namely, their size. Algorithms such as neural networks were deemed unfeasible, as they require substantial amounts of data in order to perform effectively. Furthermore, it is important to note that image-based datasets are commonly larger than tabular datasets; however, due to the fact that DLIME currently only supports tabular data, my selection was restricted. Additionally, while selecting the datasets for this thesis, I prioritized quality over size. This resulted in datasets that, while acceptable in size, could not feasibly be expected to lead to good performance with neural networks, for example.

The RF algorithm was selected primarily due to its superior performance against different models in a number of different articles, as discussed in Section 3.1, as well as its ability to achieve such results with smaller datasets. Additionally, a DT model was utilized as a control for the black box model. Specifically, the DT

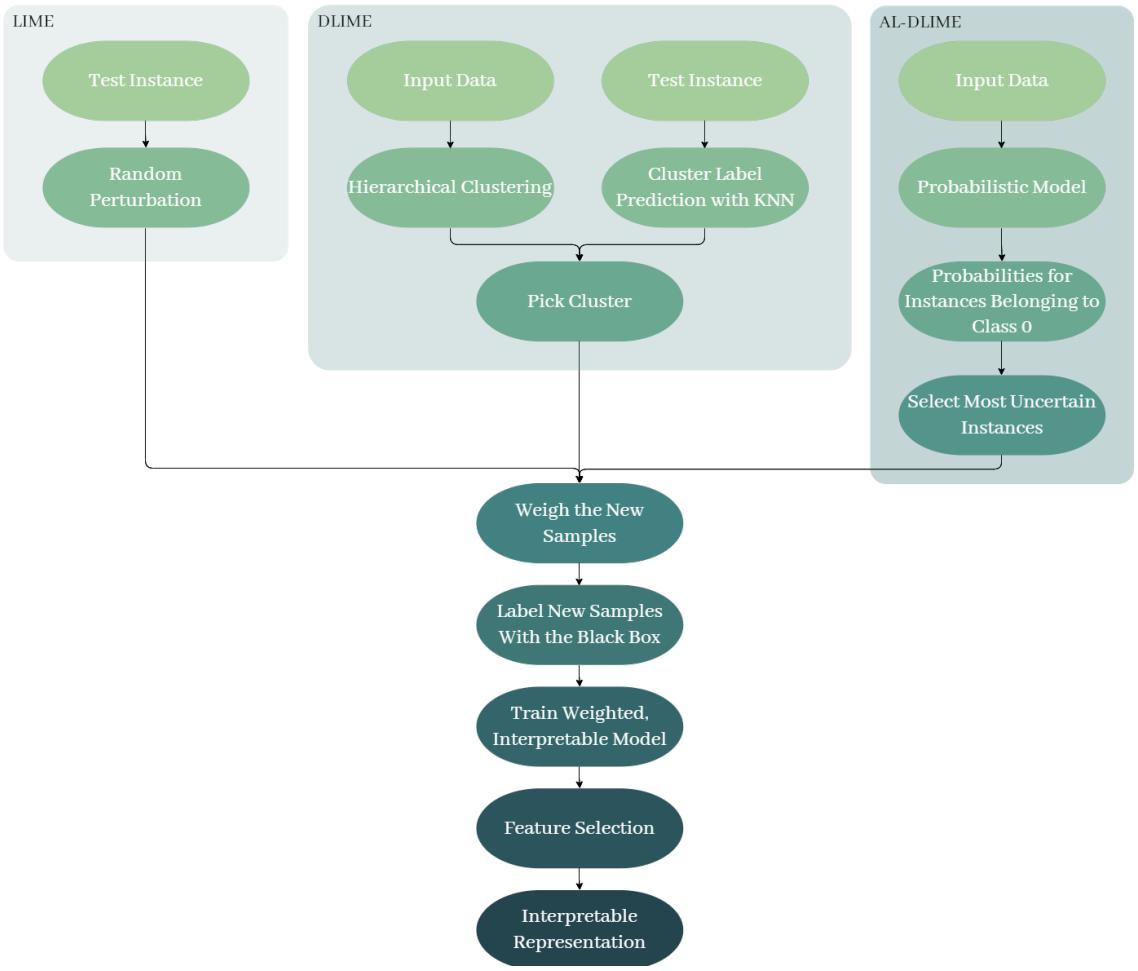


Figure 4.2: Flowchart which depicts the general pipeline of all XAI models used in this study, and how they relate to one another.

model serves to verify the necessity of utilizing a black box model, by comparing its performance against the RF model with the evaluation metrics which will be discussed in Section 4.5.

4.4.4 Model Optimization

During the construction of ML models, it is vital to take into account the impact each hyperparameter may have on model performance, and adjust them accordingly. The process of adjusting hyperparameters in order to optimize the performance of the model is known as model optimization [101] [p.98-99] [102]. One of the most widely used methods is the Grid Search method, which determines the optimal set of hyperparameters through an exhaustive search, or, in other words, brute force. More specifically, the Grid Search method constructs a model for every possible combination of hyperparameters within a given grid of hyperparameters to test, and compares them against each other with a certain metric, such as accuracy. This process can be time-consuming and computationally intensive, depending on the number of combinations to consider.

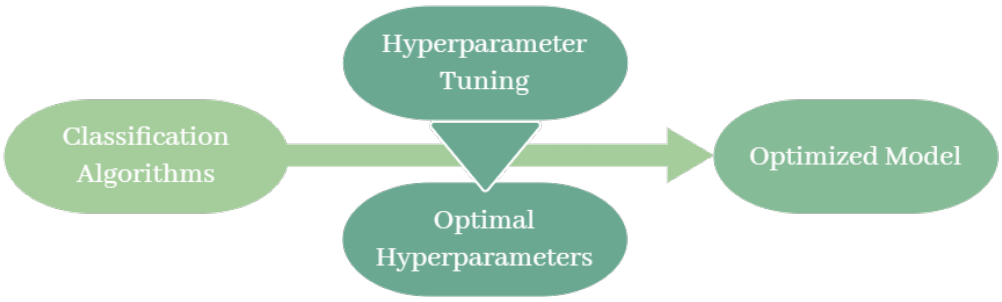


Figure 4.3: Flowchart which depicts the general pipeline of the grid search method.

In the case of this thesis, the Grid Search method was used to find the optimal set of hyperparameters for both the RF and DT models, in relation to both the standardized and non-standardized datasets. A comprehensive list of all parameters that were tested may be found in Table 4.2, and the final set of parameters that were selected for both versions of the datasets in Tables 4.3 and 4.4.

Table 4.2: Hyperparameter values considered during the Grid Search method.

Algorithm	Parameters	Values
Random Forest	n_estimators	[5, 10, 15, ..., 100]
	criterion	[gini, entropy]
	min_samples_leaf	[1, 2, 3]
	min_samples_split	[2, 3, 4, 5, 6, 7]
	max_features	[sqrt, log2]
Decision Tree	criterion	[gini, entropy]
	min_samples_leaf	[1, 2, 3]
	min_samples_split	[2, 3, 4, 5, 6, 7]
	max_features	[sqrt, log2]

4.5 Evaluation Metrics

In order to evaluate an XAI algorithm, we need to consider not only the performance of the underlying model, but also the quality of the explanation. This has proven to be difficult to achieve, for reasons explored in Section 3.3. However, although there may not currently exist a gold set of metrics used to evaluate XAI, there is also no lack of proposed metrics to choose from.

The performance of the DT and RF algorithms was evaluated through accuracy and F1-score. Both metrics rely on the concepts derived from confusion matrices: true positives (TP), which are the number of positive instances (class 1) correctly predicted; true negatives (TN), which are the number of negative instances (class 0) correctly predicted; false positives (FP), which are the number of positive instances incorrectly predicted; false negatives (FN), which are the number

Table 4.3: Optimized hyperparameters for RF and DT on the standardized datasets.

		Parameters				
Model	**Dataset**	nEstimators	criterion	minSamplesLeaf	minSamplesSplit	maxFeatures
Random Forest	**BCD**	20	gini	1	3	sqrt
	OCD	95	gini	1	4	sqrt
	PCD	90	entropy	2	3	log2
	HDD	85	entropy	2	6	log2
Decision Tree	**BCD**	-	entropy	2	6	log2
	OCD	-	entropy	3	3	sqrt
	PCD	-	entropy	1	3	log2
	HDD	-	gini	1	6	sqrt

Table 4.4: Optimized hyperparameters for RF and DT on the non-standardized datasets.

		Parameters				
Model	**Dataset**	nEstimators	criterion	minSamplesLeaf	minSamplesSplit	maxFeatures
Random Forest	**BCD**	40	gini	1	3	sqrt
	OCD	15	gini	1	7	sqrt
	PCD	55	entropy	3	6	log2
	HDD	75	gini	2	4	log2
Decision Tree	**BCD**	-	entropy	1	5	sqrt
	OCD	-	entropy	3	5	sqrt
	PCD	-	entropy	1	6	sqrt
	HDD	-	gini	1	4	sqrt

of negative instances incorrectly predicted. It is important to note, however, that the aforementioned descriptions are accurate only in the case of binary datasets. In the case of multiclass datasets, such as the PCD, the positive instances aforementioned refer to the class you are currently considering, while the negative instances refer to all other classes.

Accuracy (Formula 4.1) is the percentage of correct predictions achieved by an algorithm. In and of itself, oftentimes it is not sufficient to properly gauge performance. This is due to the fact that, for unbalanced datasets, it may depict algorithms in an unfaithful manner. For example, a model trained on a dataset which is composed 80% of class 0 need only predict each instance as class 0 in order to achieve an accuracy of 80%. Therefore, despite the fact that all selected datasets are reasonably balanced, to overcome this limitation I have chosen to use the F1-score measure in conjunction with accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Unlike accuracy, F1-score (Formula 4.3) represents the harmonic mean between precision (Formula 4.1), a measure that depicts the amount of positive predictions that were correct, and recall (Formula 4.2), a measure that depicts the amount of true positive instances that were correctly predicted. By considering both precision and recall, the F1-score is able to provide a more accurate perspective of the algorithm's performance, regardless of the imbalance in the dataset.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.4}$$

Beyond the evaluation of the DT and RF classifiers, I found it pertinent to measure LIME and DLIME's performances with the same metrics. Due to the nature of both LIME and DLIME as post-hoc models, it is important to determine if the surrogate models are comparable to the RF algorithm. Thus, through the use of accuracy and F1-score, it is possible to assess the effectiveness of LIME and DLIME in approximating the RF classifier.

To evaluate the quality of the explanations provided by the XAI models, several metrics were employed, including: the faithfulness metric from Ribeiro et al. [14], the metric of stability from Zafar and Khan [15], and the single and incremental deletion metrics. These measures provide a comprehensive evaluation of the quality of the explanations generated by both XAI models, including their consistency and faithfulness.

In regards to faithfulness, Ribeiro et al. [14] originally suggested using a maximum of ten features for the gold set. This was possible for all datasets except the PCD, which, as previously stated, had only eight features following preprocessing of the data. Thus, in this instance, I set the number of gold standard features to six.

The single deletion metric [17; 18] aims to evaluate the accuracy of an XAI algorithm's approximation of a black box model. This metric works by removing a feature in order to see the degree of perturbation caused in the algorithm's classification, and replacing it into the dataset in order to remove another feature. Generally, the features that are deemed most important by the black box model have priority, meaning that the most important feature is removed and then replaced, followed by the second most important feature, and so forth. If the predictions made by the XAI algorithm change drastically as a result, it can be assumed that the approximation is faithful to the original algorithm. Additionally, this metric may also begin with removing the least important features, with the objective being that there should be no change in the predictions made by the XAI model. In this study, I have elected to remove the most important features first; furthermore, the features are removed two at a time in order to preserve computational power.

Juxtaposed to the single deletion metric, there is incremental deletion [19; 20; 103; 104]. The basis of this metric is very similar to single deletion, in that it involves the removal of features in order to determine how the XAI model reacts. However, in this case, the removal of features is successive. In other words, once a feature has been removed, it will not be replaced. This would normally imply

an even larger perturbation in the resulting classifications. Once again, if the XAI model demonstrates this behavior, it may be deemed faithful to the black box. Similarly to the single deletion metric, the most important features are removed first, and the removal is made with two features at a time.

Finally, it is important to address the lack of both application and human-grounded metrics, despite the very real need for the involvement of humans in an application such as medicine. Unfortunately, due to a lack of monetary resources, as well as a difficulty in procuring medical professionals with time and interest for such experiments, the evaluation of the XAI models in this thesis relied primarily on functionality-grounded metrics.

4.6 Experimental Methodology

Following the optimization of the ML models on both the standardized and non-standardized datasets, both experiments were carried out in a similar manner, with the exception of the single and incremental deletion metrics. Furthermore, either experiment was repeated across the four datasets, with the following order: BCD, OCD, PCD, and, finally, the HDD.

The first experiment included the standardized datasets, beginning with the training of the ML models. From the trained RF, ten features were extracted based on their order of importance to the model, thus constructing the "gold set" to be used posteriorly for the faithfulness metric. In the case of the PCD, only six "gold set" features were selected due to its limited size. The trained models were then evaluated on the test set, obtaining the scores for accuracy and F1-score for RF and DT.

Once the evaluation metrics for the ML models were obtained, AL-DLIME, DLIME, and LIME were all constructed and trained on the training set. During the classification phase, all XAI models were programmed to only utilize the number of features contained in the "gold set" determined by the RF.

In order to obtain the metric of stability, which utilizes Jaccard's distance, a random instance (or, patient) was chosen across all datasets. Following this, the features used during the classification of this instance were extracted from all XAI models across ten iterations. Confusion matrices were then computed from the resulting stability scores.

Classifications for all instances were then obtained for the ML and XAI algorithms, in addition to the set of features used to reach each classification. This data – more specifically, the information regarding each model's classifications – was then used to calculate the accuracy and F1-score. Faithfulness was measured using the sets of features used across the classifications of all instances, in comparison to the "gold set" determined previously through RF.

In relation to the second experiment, the procedure was nearly identical to the first, with the exception of the additional metrics of single and incremental deletion. For these metrics, it was necessary to create additional copies of the current

datasets, as aforementioned.

Both the single and incremental deletion metrics were carried out across five rounds, due to the removal of two features at a time, as well as the pre-determined number of features to be used during classification, which was ten. In regards to the PCD, only three rounds were possible due to the restricted amount of six features. Following the removal of features, all models, be them ML or XAI, were retrained on the smaller datasets. Then, classifications for all instances were obtained. Once all rounds had been concluded, the similarity score between the classifications obtained during each round and the original set of classifications obtained previously was computed.

4.7 Summary

This chapter explored all aspects related to the methodology of this thesis, as well the the materials that were employed.

4.7.1 Pipeline Overview

This thesis employed the standard ML framework, including steps of data collection, data pre-processing, model optimization, classification, and evaluation. In addition to this, there is also the use of XAI models, which signifies an additional step of explanation, as well as evaluation of their performance.

4.7.2 Datasets

As the focus of this thesis is the medical domain, it was important to consider which areas would be most relevant to apply AI. I elected to consider datasets related to cancer and CVDs, as they are both diseases with a high amount of misdiagnoses.

Four publicly available datasets were selected, with three involving a different type of cancer, and one related to CVDs. They are all tabular due to DLIME's restrictions, and all except one are binary, with the final dataset containing three classes.

4.7.3 Data Pre-Processing

Following the selection of datasets, I proceeded with the step of pre-processing. Minimal processing was required due to the selected datasets being previously used in other scientific studies.

Data balancing was required for two datasets, with the removal of features with excessive missing data. Data cleaning was also required, in this case to convert

categorical information into numerical information. All datasets were split into sets for training and testing using the standard 80-20 split. Finally, copies of the datasets were made in order to be standardized for use in one of the two experiments conducted in this study. The original copies, or the non-standardized data, would be used for the other experiment.

4.7.4 Models

This thesis proposes a novel XAI model, AL-DLIME, which utilizes the basic framework of DLIME. The main difference is the implementation of AL in place of AHC and kNN. This model uses pool-based sampling and uncertainty sampling in regards to its AL component. Once the most uncertain samples are selected, the AL-DLIME algorithm behaves in the same manner as LIME and DLIME.

As for the baseline XAI models, I selected LIME due to its popularity, followed by DLIME due to LIME's non-deterministic nature. The chosen baseline black box model, which would serve as the underlying black box model for the XAI algorithms, was RF due to its good performance and popularity. Finally, DT was chosen in order to compare RF to a white box model, and address the explainability-accuracy trade-off.

The selected ML were optimized on both sets of datasets, using the grid search method.

4.7.5 Evaluation Metrics

In order to evaluate the performance of the ML models, the metrics of accuracy and F1-score were selected. The F1-score in particular was selected due to the fact that results of accuracy may be biased in cases of unbalanced datasets.

As for the XAI evaluation metrics, accuracy and F1-score were selected in order to evaluate the approximation of the XAI models' surrogate model to RF. In regards to the quality of the explanations, Ribeiro et al.'s metric of faithfulness, Zafar and Khan's metric of stability, and the metrics of single and incremental deletion were selected.

4.7.6 Experimental Methodology

Two experiments were carried out for this thesis. Both involved training the ML models, and extracting ten "gold features" from the trained RF. The ML models were then evaluated using accuracy and F1-score.

The features used to reach the classification across ten rounds for each XAI model were compiled in order to calculate the stability metric. As for the faithfulness metric, the features used to reach classifications of all instances were obtained

across all XAI models. For accuracy and F1-score, the classifications of all instances in the test set were used.

The second experiment included the addition of the single and incremental deletion metrics, which involved the removal of features and, thus, re-training all algorithms on the reduced datasets. The similarity score was computed from the classifications obtained across all rounds of feature removal.

Chapter 5

Results and Discussion

In this chapter, the results of this thesis will be presented alongside their interpretative analysis and discussion. Section 5.1 will focus on the results obtained through the standardized datasets, Section 5.2 on the results obtained through the non-standardized datasets, and Section 5.3 will provide a comparative analysis between both sets of results. Finally, Section 5.4 will summarize the contents of this chapter.

5.1 Standardized Datasets

As explained in Section 4.3, this first experiment involved an additional step in pre-processing, standardization. Each dataset was then applied to the optimized RF and DT models, as well as the XAI models of LIME, DLIME, and AL-DLIME. Results regarding accuracy and F1-Score are compiled in Table 5.1, while results regarding XAI-specific performance may be found in Tables 5.2, 5.3 and 5.4. Figure 5.1 shows examples of the confusion matrices generated through the scores of Jaccard's distance for a random instance of the BCD.

Table 5.1: Accuracy and F1-Score results for all models on the standardized datasets. Best results for accuracy and F1-Score for each dataset are highlighted through bold text.

	Accuracy				**F1-Score**			
Model	**BCD**	**OCD**	**PCD**	**HDD**	**BCD**	**OCD**	**PCD**	**HDD**
RF	**0.982**	0.767	**0.576**	0.870	**0.977**	0.722	**0.574**	0.896
DT	**0.982**	0.558	0.534	0.766	0.929	0.537	0.517	0.804
LIME	0.956	0.767	0.492	0.647	0.940	0.737	0.415	0.775
DLIME	0.947	0.767	**0.576**	0.870	**0.977**	0.762	**0.574**	0.896
AL-DLIME	**0.982**	**0.791**	**0.576**	**0.875**	**0.977**	**0.780**	**0.574**	**0.900**

Beginning with the results of accuracy and F1-Score shown in Table 5.1, it is clear, first and foremost, that AL-DLIME performed the best on each dataset across all models, including RF. In fact, AL-DLIME even outperformed RF in both accuracy and F1-score with the OCD and HDD. However, it is necessary to state that

both instances were only small improvements upon the black box model's performance. In contrast, DLIME performed slightly better than RF on only one account, the F1-score obtained with the OCD. Finally, LIME proved to be the least faithful to RF in terms of classifier performance, obtaining the highest discrepancy between results through the PCD and HDD, with the largest difference being the value of accuracy with the HDD (0.869 for RF, and 0.647 for LIME). LIME's performance was most similar to the RF model on two accounts: the BCD, and the OCD. Results of accuracy and F1-score of LIME did not differ beyond 0.040 in comparison to the values obtained by RF, in these instances. Moreover, LIME even outperformed RF in the case of the F1-score on the OCD dataset, albeit a slight difference (0.722 for RF, and 0.737 for LIME).

Additionally, the RF classifier outperformed the DT model in all accounts, across each dataset. The greatest difference in performance was observed in the case of the OCD, where the accuracy and F1-Score of the RF model were 0.767 and 0.722, respectively, while those of the DT model were 0.558 and 0.536. However, the worst results in performance were noted on the PCD, with neither model able to achieve accuracy of F1-Score values above 0.600. On the contrary, both models performed exceptionally well on the BCD, scoring above 0.900 in accuracy and F1-Score.

Overall, the XAI models all performed most similarly to RF through the BCD. Moreover, RF outperformed DT on all accounts across all datasets, which was to be expected.

Table 5.2: Faithfulness of the XAI models on the standardized datasets. Best results for each dataset are highlighted through bold text.

	Faithfulness			
Model	**BCD**	**OCD**	**PCD**	**HDD**
LIME	**0.600**	**0.700**	**0.833**	0.900
DLIME	0.272	0.300	0.667	0.900
AL-DLIME	0.349	0.219	0.668	**0.935**

As aforementioned, Table 5.2 shows the results of faithfulness between each XAI model, and the RF classifier. In general, LIME boasts the best results, with all values being above 0.600. It is especially interesting to note that LIME performed the best with the PCD and HDD, with its best result being 0.900 with the HDD. This trend is reflected with DLIME and AL-DLIME, though in terms of the BCD and OCD, they performed significantly worse than LIME, with all values below 0.400. Between the two, DLIME obtained the worst results, though it performed equally to LIME with the HDD. Additionally, AL-DLIME also obtained its best result of faithfulness with the HDD, which also accounts for the best result regarding this metric among all XAI models, 0.935.

In Table 5.3, the results for the single deletion metric are shown. As for the PCD, there are no results for rounds 4 and 5 due to the reduced number of features available, compounded with the fact that two features were removed at a time. From all models, LIME obtained the overall best results, consistently across round 1 of removal. To reiterate a previous point, low results for similarity are more

Table 5.3: Results for single deletion across all five rounds. Lowest values across all rounds for each dataset are highlighted through bold text.

Model	Dataset	Round 1	Round 2	Round 3	Round 4	Round 5
RF	**BCD**	0.991	0.991	0.982	0.969	0.991
	OCD	0.930	0.930	0.930	0.930	0.860
	PCD	0.873	0.763	0.780	-	-
	HDD	0.902	0.967	0.967	0.973	0.973
LIME	**BCD**	1.000	0.991	0.991	0.991	0.982
	OCD	**0.721**	0.884	0.930	0.907	0.884
	PCD	**0.508**	0.797	0.669	-	-
	HDD	**0.446**	0.984	0.598	1.000	0.951
DLIME	**BCD**	0.991	0.991	0.982	**0.965**	0.982
	OCD	0.930	0.930	0.930	0.930	0.860
	PCD	0.873	0.763	0.780	-	-
	HDD	0.902	0.967	0.967	0.973	0.973
AL--DLIME	**BCD**	0.991	0.991	0.982	**0.965**	0.982
	OCD	0.930	0.907	0.930	0.930	0.860
	PCD	0.873	0.763	0.780	-	-
	HDD	0.886	0.973	0.973	0.967	0.967

Table 5.4: Results for incremental deletion across all five rounds. Lowest values across all rounds for each dataset are highlighted through bold text.

Model	Dataset	Round 1	Round 2	Round 3	Round 4	Round 5
RF	**BCD**	0.965	0.982	0.991	0.965	0.956
	OCD	0.884	0.744	0.884	0.884	0.674
	PCD	0.907	0.500	0.619	-	-
	HDD	0.891	0.625	0.690	0.821	**0.054**
LIME	**BCD**	0.982	0.991	1.000	1.000	0.991
	OCD	0.721	0.721	0.791	0.791	0.721
	PCD	0.788	0.449	**0.407**	-	-
	HDD	0.495	0.435	0.500	0.582	0.582
DLIME	**BCD**	0.965	0.982	0.991	0.965	**0.947**
	OCD	0.884	0.744	0.884	0.884	0.674
	PCD	0.907	0.441	0.619	-	-
	HDD	0.891	0.625	0.690	0.821	**0.054**
AL--DLIME	**BCD**	0.965	0.982	0.991	0.965	**0.947**
	OCD	**0.651**	0.907	0.860	0.860	0.721
	PCD	0.907	0.441	0.619	-	-
	HDD	0.875	0.625	0.690	0.804	0.065

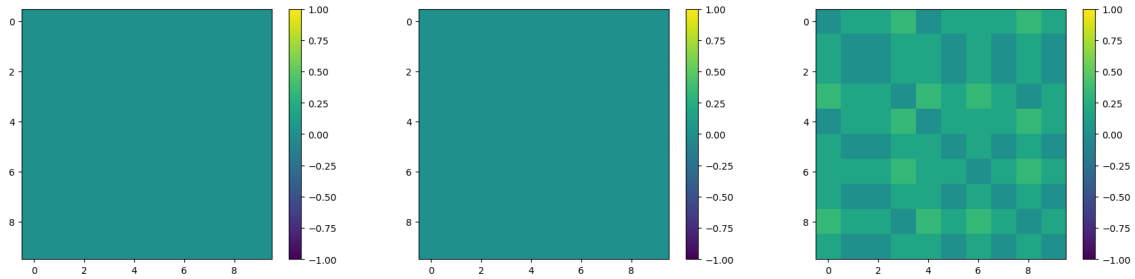
desirable, as that would signify the model is faithful to the choice of the most important features. As I began with removing the most important features, it was to be expected that the lowest results would be found in round 1. However, LIME was the only model which obtained values below 0.600. In fact, the results from RF are nearly all above 0.900, with the exception of the values obtained through the PCD, which range from 0.780 to 0.873, and a single instance obtained through the OCD, with a value of 0.860. Both DLIME and AL-DLIME performed similarly to RF, with the lowest results of similarity being through the PCD. However, DLIME and AL-DLIME also achieved the lowest value in regards to the BCD, despite this result being relatively high, regardless (0.965).

Finally, Table 5.4 shows the results for the incremental deletion metric. Once again, there are no results for the PCD for rounds 4 and 5 due to the reduced number of features Overall, it appears that DLIME and AL-DLIME both achieved the lowest results for two datasets (BCD and HDD, and BCD and OCD, respectively), while LIME and RF achieved the lowest results for the PCD and HDD, respectively. Furthermore, there was further variability between results, which was to be expected to the nature of the metric – successively removing features from the datasets. However, as for the BCD, no value below 0.940 was achieved across all rounds, from all models. Finally, the lowest results that were obtained were nearly all from round 5, which further aligns with the predicted behaviour.

For this first experiment, the results for accuracy and F1-score from the ML models were satisfactory, with the exception of those obtained with the PCD. As aforementioned, RF outperformed DT on every occasion which was consistent with the findings in Section 3.1. This provides a more robust defense for the use of black box models in this area of application, though a more comprehensive study may be executed in the future, with a larger array of both black box and white box models. Within the scope of this paper, the results of the model optimization through the grid search method lead me to believe that RF is preferable over DT among the selected datasets. When compared to performance of AL-DLIME, however, the question remains: are black box algorithms strictly necessary? Despite the the fact that the instances in which AL-DLIME outperformed RF were merely slight improvements, the fact remains that its performance exceeded RF. This could be another potential avenue for future research.

In terms of faithfulness to the black box model, the results for both DLIME and AL-DLIME were surprisingly low, though all XAI models performed well on the HDD. Beyond this, there is an especially interesting correlation between the results of accuracy and F1-score, and faithfulness. All models performed the best in terms of accuracy and F1-score on the BCD and OCD, while, in stark contrast, obtaining their worst results in regards to faithfulness. Moreover, LIME's results of accuracy and F1-score showed the biggest discrepancy to those of RF through the HDD, while it obtained its best result for faithfulness on that same dataset. This would seem to suggest that a higher faithfulness to the black box model comes at a cost of performance. However, in the case of AL-DLIME, the opposite occurred: on the HDD, not only did it outperform RF, it also obtained the best result for faithfulness out of all XAI models, from all utilized datasets.

The trend of high results for similarity regarding the single deletion metric was



(a) DLIME (b) AL-DLIME (c) LIME

Figure 5.1: Results for Jaccard's distance across ten iterations on a single, random instace from the standardized BCD, presented in confusion matrices.

yet another point of interest. It is important to note that the overall lowest results were obtained through the PCD, the dataset with the smallest amount of features and, thus, the smallest "gold set", consisting of only six features. The removal of two out of six of these features would, in theory, have a greater impact, which could explain the low results. As for the results of the incremental deletion metric, a larger discrepancy was noted, though still less than expected. The lowest results for similarity were obtained through the PCD, as well as the HDD, while the results for the BCD were overall the highest. Once again, this outcome may be due to the difference in terms of number of features among the datasets. The BCD and OCD, with the largest amount of features, are more likely to have multiple features with similar values of importance, than the PCD and HDD, both of which contain under fifteen features. However, despite the results not being in accordance with what was expected, it should be emphasized that, overall, all XAI models performed closely to RF, with DLIME and AL-DLIME obtaining the closest results to the black box model.

In regards to the values obtained for Jaccard's distance, through Figure 5.1 we may see the results regarding a random instance from the BCD. The results of Jaccard's distance are represented through confusion matrices, from which we may view the variety of values through their color representation, depicted in the color bars on the right side of each matrix. Therefore, DLIME and AL-DLIME, whose matrices (Figures 5.1a and 5.1b, respectively) are solid colors depicting the value 0, demonstrate the utmost stability. In contrast, LIME's confusion matrix (Figure 5.1c) depicts a wide variety of different values, which translates into less stability. Further examples may be found in Figures B.1, B.3, and B.5.

5.2 Non-Standardized Datasets

For the second experiment, the process of standardization was omitted, thus resulting in the exclusion of the single and incremental deletion metrics. The remaining process, as explained in section 4.6, was much the same. Table 5.5 provides the results of accuracy and F1-score obtained for the ML and XAI models, while Table 5.6 displays the results of faithfulness of the XAI models. Finally, Figure 5.2 shows the confusion matrices generated from the Jaccard's distance scores

of the XAI models on a random instance of the BCD.

Table 5.5: Accuracy and F1-Score results for all models on the non-standardized datasets. Best results for accuracy and F1-Score for each dataset are highlighted through bold text.

	Accuracy				F1-Score			
Model	**BCD**	**OCD**	**PCD**	**HDD**	**BCD**	**OCD**	**PCD**	**HDD**
RF	**0.991**	0.767	0.703	**0.870**	**0.989**	0.762	0.707	**0.897**
DT	0.956	0.698	0.585	0.804	0.943	0.629	0.589	0.830
LIME	0.982	0.767	0.644	0.864	0.976	0.737	0.644	0.893
DLIME	0.982	0.767	**0.712**	**0.870**	0.977	0.762	**0.716**	**0.897**
AL-DLIME	0.982	**0.791**	0.703	**0.870**	0.977	**0.780**	0.707	**0.897**

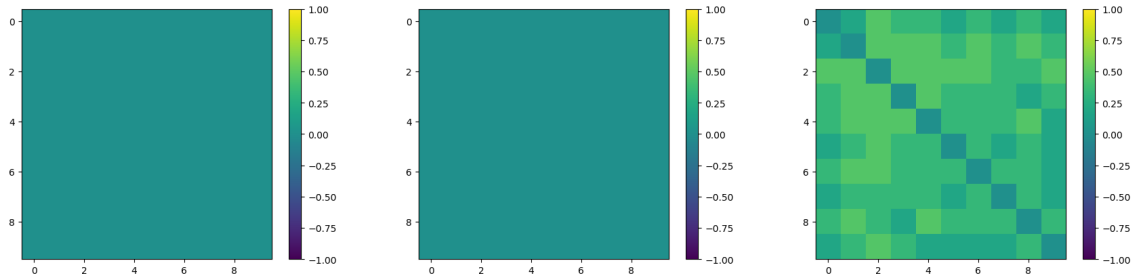
From Table 5.5, we may find the results of accuracy and F1-Score obtained for the ML and XAI models in this experiment. Once again, AL-DLIME demonstrated a strong performance, although it did not dominate as in the previous experiment. In fact, the performance of DLIME and AL-DLIME was similar, with both models managing once again to outperform RF. DLIME, however, performed in the most similar manner to RF, with the biggest discrepancy in terms of accuracy and F1-score being 0.011. Comparatively, LIME was the least faithful to RF in regards to these metrics, with the biggest discrepancy between both models' values being 0.063. Beyond this, it is worth noting that the datasets with which these models performed most similarly to RF were the OCD and HDD, in regards to LIME and DLIME, and the PCD and HDD for AL-DLIME.

Additionally, the results show that RF outperformed DT, with particularly remarkable results in accuracy for the BCD (0.991) that is in line with many state of the art models, some of which are mentioned in section 3.1. Overall, RF performed well, with results in both accuracy and F1-score consistently above 0.700. In contrast, DT also achieved acceptable results, with the exception of those obtained through the PCD, which were notably worse.

Table 5.6: Faithfulness of the XAI models on the non-standardized datasets. Best results for each dataset are highlighted through bold text.

	Faithfulness			
Model	**BCD**	**OCD**	**PCD**	**HDD**
LIME	**0.800**	**0.600**	**0.833**	**1.000**
DLIME	0.344	0.272	0.667	0.929
AL-DLIME	0.379	0.237	0.667	0.955

Regarding faithfulness, which may be viewed in Table 5.6, LIME once again demonstrated the strongest performance. In fact, LIME achieved the overall highest faithfulness score (1.000) among all XAI models, which was attained through the HDD. While the remaining models performed similarly, AL-DLIME outperformed DLIME marginally across all datasets. Moreover, it is worth noting that all models obtained their highest scores with the HDD, while their lowest scores were obtained with the OCD. This is most certainly due to the difference between the amount of features present in either dataset: while the OCD contains



(a) DLIME (b) AL-DLIME (c) LIME

Figure 5.2: Results for Jaccard's distance across ten iterations on a single, random instace from the non-standardized BCD, presented in confusion matrices.

the largest amount at 47 features, the HDD only contains 11, of which 10 were selected for classification. Following this logic, it is unsurprising that the second highest scores were obtained with the PCD, which only contains 8 features, 6 of which were used at any point in time for classification. Despite this, there is a considerable difference between the results of faithfulness from the PCD between LIME, and both DLIME and AL-DLIME, which, in addition to LIME's overall high scores, suggests that LIME is a better candidate when considering the aspect of faithfulness to the underlying black box model.

In spite of the high scores of faithfulness, in combination to the close performance to RF of the XAI models regarding accuracy and F1-score, through the HDD, it is necessary to highlight the effect of the dataset's small size of features. Instead, it is perhaps more productive to consider the relation between the scores attained through the PCD: all models performed closely to RF, in terms of accuracy and F1-score, while at the same time obtaining their lowest overall scores for faithfulness.

Finally, DLIME and AL-DLIME once again outperform LIME in regards to the stability of their explanations, as shown by Figure 5.2. As demonstrated by the lighter colors, moreover, we may conclude that in this instance, LIME demonstrated even further instability than the previous example with the standardized data. Of course, it is important to stress once again that these are only random, singular instances from one of the four selected datasets. However, the stability of DLIME and AL-DLIME in comparison to LIME remains undeniable. Further examples may be found in Figures B.2, B.4, and B.6.

5.3 Comparative Analysis

In both experiments, RF outperformed DT by a significant enough margin to justify its use over the transparent model, especially when one considers the scores from the standardized OCD, and the non-standardized PCD. Beyond this, the scores themselves were satisfactory for the area of application in question. Moreover, it is noteworthy to mention that there was a noticeable improvement in results with the use of the non-standardized datasets. In fact, as aforementioned, RF

managed an impressive score of 0.991 in accuracy with the BCD, which is on par with many state of the art models mentioned in 3.2. This dataset is very widely used, and an important benchmark for training and testing many ML algorithms; thus, attaining such a score is not only testament to an optimized model, but also a well constructed dataset.

One of the main arguments for XAI is the perceived superiority of black box models over white box models. The results of the experiments conducted in this thesis in regards to RF and DT initially appear to support this claim. However, it was observed on several occasions that both DLIME and AL-DLIME achieved higher accuracy and F1-score than RF in both experiments, albeit with small margins. Future research could investigate a wider range of both black box and white box models to reach a consensus on this issue. Nonetheless, given the results of this work, it may be more advantageous to use a simpler, more explainable model such as ridge linear regression, which is the surrogate model used in the selected XAI models, instead of a more complicated model such as RF.

Regarding the performance of the XAI models, the results showed that DLIME and AL-DLIME consistently outperformed LIME in relation to accuracy and F1-score across both experiments. Moreover, on several instances, AL-DLIME performed slightly better than its predecessor, though the difference was negligible. When viewing these results from the perspective of the scores for faithfulness, however, a trend was noted. The scores of accuracy and F1-score that are the closest to those of RF correlate with the XAI model's lowest results in faithfulness. As explained before, the fact that 10 out of 11 features were used for the classification of the HDD skewed the results of faithfulness for this dataset, and thus should not be considered. In addition to this, it is important to note that LIME performed significantly better than the remaining XAI models in terms of faithfulness. As a result, one must question what is most important in the selection of a XAI model. Though DLIME and AL-DLIME perform most closely to RF, in terms of faithfulness to the black box, they are lacking. On the contrary, LIME offers a generally faithful representation of what the selected black box model deems as important, however, there is a larger discrepancy when considering accuracy and F1-score. A possible explanation for LIME's greater faithfulness to the black box model may be due to its method of selecting instances that are used to train the surrogate model. By performing a random selection (as opposed to selecting the most informative instances, in the case of AL-DLIME), it is possible that LIME rids itself of any bias towards the data, thus approximating a more faithful representation.

Given the origin of XAI is based on the value of trust, we would propose that faithfulness should be viewed above accuracy and F1-score; additionally, the discrepancy noted never surpassed 0.250 across both experiments, and therefore may still be considered as acceptable. However, the fact remains that the stability of DLIME and AL-DLIME is undeniable in comparison to the explanations generated through LIME.

5.4 Summary

In this chapter, the results from the thesis were presented, followed by a discussion surrounding them, as well as a comparative analysis between both experiments.

5.4.1 Standardized Datasets

In terms of accuracy and F1-score, AL-DLIME performed the best. Both AL-DLIME and its predecessor performed very closely to RF, even outperforming the ML model on select instances. As expected, RF outperformed DT.

LIME was the best XAI model in terms of faithfulness, though AL-DLIME performed the best on the HDD.

In terms of the single and incremental deletion metrics, not as much perturbation was registered as was expected. LIME registered the largest perturbations in regards to the single deletion metric, while the performance of the XAI models was more balanced on the incremental deletion metric.

Finally, AL-DLIME and DLIME clearly outperformed LIME in terms of stability.

5.4.2 Non-Standardized Datasets

AL-DLIME and DLIME performed similarly in terms of accuracy and F1-score, both in relation to each other, as well as to RF. LIME once again performed the worst out of all XAI models. However, in terms of faithfulness, LIME outperformed AL-DLIME and DLIME across all datasets.

In this experiment, the stability of AL-DLIME and DLIME was once again superior to that of LIME.

5.4.3 Comparative Analysis

There was an improvement in terms of accuracy, F1-score, and faithfulness in the second experiment. RF outperformed DT across both experiments, however, the XAI models were also able to outperform RF, in spite of those improvements being slight. AL-DLIME was especially robust in regards to these metrics.

LIME sucessively outperformed AL-DLIME and DLIME in regards to faithfulness. This may be due to its non-deterministic nature of randomly perturbing data points reducing the risk of bias towards data.

A trend between the results of accuracy and F1-score, and faithfulness: the closer the XAI models performed to RF in terms of classifications, the lower their results of faithfulness. This may enforce the idea of the explainability-accuracy tradeoff.

Chapter 6

Conclusion

The present thesis served, first and foremost, to present a novel XAI model based on the DLIME framework. This was achieved through the use of AL, applying pool-based sampling and an uncertainty sampling query strategy in order to select the most informative examples. Beyond this, I aimed to provide a comprehensive comparison between the proposed model, AL-DLIME, and DLIME and LIME, across four datasets related to medicine. Thus, several metrics were selected, including faithfulness to the black box model from Ribeiro et al. [14], Jaccard's distance to measure stability of the explanations proposed by Zafar and Khan [15], single and incremental deletion, and, finally, accuracy and F1-score. Due to the nature of the white box model used in the XAI models, ridge linear regression, it was necessary to perform two experiments: the first which involved standardized versions of the selected datasets, and the second, non-standardized data.

The proposed model obtained satisfactory results in terms of accuracy and F1-Score, managing to outperform RF on several instances in both experiments. DLIME performed similarly to AL-DLIME in regards to accuracy and F1-Score, though AL-DLIME managed to outperform its predecessor ever so slightly. However, despite their strong performance related to classifications, AL-DLIME and DLIME's scores of faithfulness were lacking on both the BCD and OCD when compared to LIME through either experiment, with results as low as 0.219. This may be due to their deterministic nature – in the case of AL-DLIME, for example, by using uncertainty sampling for the selection of instances to use to train the surrogate model, bias may be introduced into the model and, thus, interfere with its ability to faithfully approximate the underlying black box model. LIME, on the other hand, randomly selects instances, which may reduce the risk of introducing bias.

Furthermore, it is important to reiterate the finding made in Section 5.3: regarding the XAI models, the higher the scores for accuracy and F1-score, the lower the scores for faithfulness.

As for the single and incremental deletion metrics, there was not as much perturbation as expected, including the scores from RF. LIME, overall, obtained the best results for these metrics, in the sense that the model showed the most perturba-

tion from the removal of features that were deemed the most important by the RF model. However, despite this, it is necessary to stress the fact that RF itself did not present much perturbation. Therefore, as the ultimate goal is to gauge how faithful the XAI models are to approximating their underlying black box model's behaviour, it is important to note that AL-DLIME and DLIME performed the most similarly to RF regarding these metrics.

Finally, AL-DLIME and DLIME both received perfect scores of stability, as demonstrated through the Jaccard's distance scores, while LIME's explanations proved a considerable degree of instability, especially in the case of the second experiment.

Overall, the results are satisfying, as well as exciting, due to the plethora of possibilities for future research. In particular, the aspect of AL in the proposed model allows for many different avenues of research, including different sampling scenarios and query strategies.

Chapter 7

Future Work

It is important to stress the lack of human-based metrics, which is especially important in this area of application. While monetary and time-related constraints impeded the inclusion of such metrics in the current work, a questionnaire was prepared in the event that an available medical professional could answer. This questionnaire may be found in Figures C.1 to C.12, and remains as a potential avenue of research for the future. Beyond this, the application of other XAI-related evaluation metrics could be another point of interest, as it is an ever-expanding topic for this field. In fact, as stated previously, experts are yet to agree upon a "gold set" of metrics to evaluate XAI models. Finally, it is important to acknowledge the lack of other relevant metrics, such as specificity and sensibility, that are widely used to evaluate classification models applied to the medical field. While the focus of this thesis was the comparison of XAI models through metrics that may accurately evaluate, for example, the quality of an explanation, it is also necessary to consider other relevant metrics.

The implementation of other black box models, such as SVMs, or neural networks, in order to provide a more comprehensive comparison between the accuracy and F1-Score of AL-DLIME and other models may also be of interest going forward.

In terms of the proposed model, there are several possible variations that remain to be explored. Future research could implement different sampling scenarios, as well as query strategies. Furthermore, it is important to acknowledge the possibility of the selected query strategy being unable to determine sufficiently uncertain instances. In other words, instances that belong within the determined interval of 47% to 53%. For this case, two possible avenues may be explored: 1) the interval of uncertainty may be widened, which, in the case of the same issue occurring once more, leads to 2) the AL-DLIME algorithm proceeds to function as the DLIME algorithm, utilizing AHC and kNN to select samples.

Lastly, another point of interest that was identified was that DLIME and, by extension, AL-DLIME could support image-based datasets in the future. Many datasets used in the medical field are image-based, and thus I view this goal as being vital for the continued use of AL-DLIME and DLIME in such domains.

References

[1] S. J. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Pearson Education, third ed., 2010.

[2] B. C. Stahl, *Ethical Issues of AI. In: Artificial Intelligence for a Better Future*, pp. 35–53. Cham, Switzerland: Springer Cham, first ed., 2021.

[3] V. C. Müller, "Ethics of Artificial Intelligence and Robotics," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Summer 2021 ed., 2021.

[4] E. Parliament, D.-G. for Parliamentary Research Services, J. Fox-Skelly, E. Bird, N. Jenner, A. Winfield, E. Weitkamp, and R. Larbey, *The ethics of artificial intelligence : issues and initiatives*, p. 13. European Parliament, 2020.

[5] M. Smith and S. Miller, "The ethical application of biometric facial recognition technology," *AI and Society*, vol. 37, pp. 167—-175, 2022.

[6] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, 2018.

[7] S. Gerke, T. Minssen, and G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare," *Artificial Intelligence in Healthcare*, pp. 295—-336, 2020.

[8] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," 2019.

[9] H. Singh, G. D. Schiff, M. L. Graber, I. Onakpoya, and M. J. Thompson, "The global burden of diagnostic errors in primary care," *BMJ Quality and Safety*, vol. 26, pp. 484–494, 2017.

[10] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, and K. D. Mandl, "Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency," *npj Digital Medicine*, vol. 3, 2020.

[11] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med Inform Decis Mak*, vol. 20, 2020.

[12] R. R. Fletcher, A. Nakeshimana, and O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," *Front. Artif. Intell.*, vol. 3, 2021.

[13] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLoS Med*, vol. 15, 2018.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," 2016.

[15] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," 2021.

[16] M. R. Zafar and N. M. Khan, "Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," 2019.

[17] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *CoRR*, vol. abs/1806.07538, 2018.

[18] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.

[19] S. Hooker, D. Erhan, P. Kindermans, and B. Kim, "Evaluating feature importance estimates," *CoRR*, vol. abs/1806.10758, 2018.

[20] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, USA: Springer, second ed., 2009.

[22] S. Holm and L. Macedo, "The accuracy and faithfulness of AL-DLIME - Active Learning-based Deterministic Local Interpretable Model-Agnostic Explanations: a comparison with LIME and DLIME in medicine." Proceedings of the 1st World Conference on XAI, July 2023, (Forthcoming).

[23] K. Capek, *R.U.R. (Rossum's Universal Robots)*. Garden City, New York: Doubleday, Page and Company, 1923.

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

[25] S. J. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, p. 16. Upper Saddle River, New Jersey: Pearson Education, third ed., 2010.

[26] A. Turing, "Computing machinery and intelligence," *Mind*, pp. 433–460, 1950.

[27] C. H. Hoffman, "Is AI intelligent? An assessment of artificial intelligence, 70 years after turing," *Technology in Society*, vol. 68, 2022.

[28] H. of Lords Select Committee on Artificial Intelligence, "AI in the UK: ready, willing and able?," *??*, p. 13, 2018.

[29] A. Kaplan and M. Haenlein, "Siri, Siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence," *Business Horizons*, pp. 15–25, 2019.

[30] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, pp. 685–695, 2021.

[31] K. Kersting, "Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines," *Frontiers in Big Data*, vol. 1, 2018.

[32] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[33] S. J. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, p. 698. Upper Saddle River, New Jersey: Pearson Education, third ed., 2010.

[34] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature biotechnology*, vol. 26, pp. 1011–1013, 2009.

[35] L. Rokach and O. Maimon, *Decision Trees. In: Data Mining and Knowledge Discovery Handbook*, pp. 165–192. Boston, MA: Springer New York, first ed., 2009.

[36] J. M. Luna, E. D. Gennatas, L. H. Ungar, E. Eaton, E. S. Diffenderfer, S. T. Jensen, C. B. Simone, J. H. Friedman, T. D. Solberg, and G. Valdes, "Building more accurate decision trees with the additive tree," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, pp. 19887–19893, 2019.

[37] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, pp. 3–29, 2020.

[38] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments torecent advancements," *Systems Science and Control Engineering*, vol. 2, pp. 602—-609, 2014.

[39] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers, N. Urquhart, and S. Wells, "Trusting intelligent machines: Deepening trust within socio-technical systems," *IEEE Technology and Society Magazine*, vol. 4, pp. 76—-83, 2018.

[40] M. Ryan, "In ai we trust: Ethics, artificial intelligence, and reliability," *Science and Engineering Ethics*, vol. 26, pp. 2749—-2767, 2020.

[41] M. Sutrop, "Should we trust artificial intelligence?," *Trames. Journal of the Humanities and Social Sciences*, vol. 23, p. 499, 2019.

[42] M. R. Carrillo, "Artificial intelligence: From ethics to law," *Telecommunications Policy*, vol. 44, p. 101937, 2020.

[43] C. Prunkl, "Human autonomy in the age of artificial intelligence," *Nature Machine Intelligence*, vol. 4, pp. 99–101, 2022.

[44] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *CoRR*, vol. abs/2006.00093, 2020.

[45] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[46] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *CoRR*, vol. abs/1806.00069, 2018.

[47] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, pp. 1–45, 2021.

[48] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[49] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[50] J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the need for explainable artificial intelligence (xAI)," pp. 1284–1293, 2020.

[51] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017.

[52] B. Herman, "The promise and peril of human evaluation for model interpretability," 2017.

[53] T. W. Kim, "Explainable artificial intelligence (xai), the goodness criteria and the grasp-ability test," 2018.

[54] Z. C. Lipton, "The mythos of model interpretability," *ACM Queue*, vol. 16, 2018.

[55] A. Rai, "Explainable ai: from black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137–141, 2020.

[56] A. Asteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115—-118, 2017.

[57] A. K. Rehme, L. J. Volz, D.-L. Feis, I. Bomilcar-Focke, T. Liebig, S. B. Eickhoff, G. R. Fink, and C. Grefkes, "Identifying Neuroimaging Markers of Motor Disability in Acute Stroke by Machine Learning Techniques," *Cerebral Cortex*, vol. 25, no. 9, pp. 3046–3056, 2014.

[58] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," *International Journal of Applied Information Systems*, vol. 3, pp. 25–30, 2012.

[59] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[60] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A random forest based predictor for medical data classification using feature ranking," *Informatics in Medicine Unlocked*, vol. 15, p. 100180, 2019.

[61] European Public Health Aliance (EPHA), the European Chronic Disease Alliance (ECDA), and the NCD Alliance, "Towards an EU strategic framework for the prevention of non-communicable diseases (NCDs)," 2019.

[62] C. P. Wild, E. Weiderpass, and B. W. Stewart, *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon, France: International Agency for Research on Cancer, 2020.

[63] World Health Organization, *WHO report on cancer: setting priorities, investing wisely and providing care for all*. Geneva, Switzerland: World Health Organization, 2020.

[64] K. S. Reddy, "Prevention and control of non-communicable diseases," *Oxford Textbook of Global Public Health*, 2015.

[65] World Health Organization, "Action plan for the prevention and control of noncommunicable diseases in the who european region," 2016.

[66] National Institutes of Health (US), "Understanding cancer," 2007.

[67] G. M. Cooper, "The development and causes of cancer," 2000.

[68] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, 2000.

[69] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA*, vol. 71, 2021.

[70] World Health Organization, *Global Atlas on Cardiovascular Disease Prevention and Control*. Geneva: World Health Organization, 2011.

[71] M. Thiriet, "Cardiovascular disease: An introduction," *Vasculopathies*, vol. 8, pp. 1–90, 2019.

[72] J. Scott, "Pathophysiology and biochemistry of cardiovascular disease," *Current Opinion in Genetics Development*, vol. 14, no. 3, pp. 271–279, 2004.

[73] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, and N. C. Barengo, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020.

[74] S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, S. Cheng, F. N. Delling, M. S. V. Elkind, K. R. Evenson, J. F. Ferguson, D. K. Gupta, S. S. Khan, B. M. Kissela, K. L. Knutson, C. D. Lee, T. T. Lewis, and J. Liu, "Heart disease and stroke statistics-2021 update: A report from the american heart association," *Circulation*, vol. 143, no. 8, 2021.

[75] M. Lu, Z. Fan, B. Xu, L. Chen, X. Zheng, J. Li, T. Znati, Q. Mi, and J. Jiang, "Using machine learning to predict ovarian cancer," *International Journal of Medical Informatics*, vol. 141, 2020.

[76] Q. S. Setiawan, Z. Rustam, S. Hartini, V. V. P. Wibowo, and J. E. Aurelia, "Comparing decision tree and logistic regression for pancreatic cancer classification," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 623–627, 2020.

[77] A. Osmanović, L. Abdel-Ilah, A. Hodžić, J. Kevric, and A. Fojnica, "Ovary cancer detection using decision tree classifiers based on historical data of ovary cancer patients," in *CMBEBIH 2017* (A. Badnjevic, ed.), (Singapore), pp. 503–510, Springer Singapore, 2017.

[78] E. Kawakami, J. Tabata, N. Yanaihara, T. Ishikawa, K. Koseki, Y. Iida, M. Saito, H. Komazaki, J. S. Shapiro, C. Goto, Y. Akiyama, R. Saito, M. Saito, H. Takano, K. Yamada, and A. Okamoto, "Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers," *Clinical Cancer Research*, vol. 25, no. 10, pp. 3006–3015, 2019.

[79] H. CAU, I. J, I. R, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors (Basel)*, vol. 22, 2022.

[80] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized bi-rads features and machine learning methods," *Ultrasound in Medicine Biology*, vol. 42, no. 4, pp. 980–988, 2016.

[81] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pp. 1–8, 2019.

[82] Z. Rustam, F. Zhafarina, G. Saragih, and S. Hartini, "Pancreatic cancer classification using logistic regression and random forest," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, p. 476, 06 2021.

[83] G. L. Simegn, W. B. Gebeyehu, and M. Z. Degu, "Computer-aided decision support system for diagnosis of heart diseases," *Research Reports in Clinical Cardiology*, vol. 13, pp. 39–54, 2022.

[84] M. M. Ahamad, S. Aktar, M. J. Uddin, T. Rahman, S. A. Alyami, S. Al-Ashhab, H. F. Akhdar, A. Azad, and M. A. Moni, "Early-stage detection of ovarian cancer based on clinical data using machine learning approaches," *Journal of personalized medicine*, vol. 12, pp. 39–54, 2022.

[85] R. Massafra, A. Latorre, A. Fanizzi, R. Bellotti, V. Didonna, F. Giotta, D. La Forgia, A. Nardone, M. Pastena, C. M. Ressa, L. Rinaldi, A. O. M. Russo, P. Tamborra, S. Tangaro, A. Zito, and V. Lorusso, "A clinical decision support system for predicting invasive breast cancer recurrence: Preliminary results," *Frontiers in Oncology*, vol. 11, 2021.

[86] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 4765—4774, Curran Associates, Inc., 2017.

[87] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[88] A. Saini and R. Prasad, "Locally interpretable model agnostic explanations using gaussian processes," *CoRR*, vol. abs/2108.06907, 2021.

[89] X. Zhao, X. Huang, V. Robu, and D. Flynn, "Baylime: Bayesian local interpretable model-agnostic explanations," *CoRR*, vol. abs/2012.03058, 2020.

[90] D. Garreau and U. von Luxburg, "Explaining the explainer: A first theoretical analysis of lime," *CoRR*, vol. abs/2001.03447, 2020.

[91] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 2018.

[92] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv*, 2017.

[93] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[94] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.

[95] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *ACM Computing Surveys*, 2023.

[96] M. Lu, Z. Fan, B. Xu, L. Chen, X. Zheng, J. Li, T. Znati, Q. Mi, and J. Jiang, "Using machine learning to predict ovarian cancer," *International Journal of Medical Informatics*, vol. 141, pp. 104–195, 2020.

[97] S. Debernardi, H. O'Brien, A. S. Algahmdi, N. Malats, G. D. Stewart, M. Plješa-Ercegovac, E. Costello, W. Greenhalf, A. Saad, R. Roberts, A. Ney, S. P. Pereira, H. M. Kocher, S. Duffy, O. Blyuss, and T. Crnogorac-Jurcevic, "A combination of urinary biomarker panel and pancrisk score for earlier detection of pancreatic cancer: A case–control study," *PLOS Medicine*, vol. 17, pp. 1–23, 2020.

[98] M. Gaillochet, C. Desrosiers, and H. Lombaert, "Active learning for medical image segmentation with stochastic batches," 2023.

[99] M. Kholghi, L. Sitbon, G. Zuccon, and A. Nguyen, "Active learning: a step towards automating medical concept extraction," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 289–296, 2015.

[100] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[101] J. Brownlee, *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-to-end*. Jason Brownlee, 2016.

[102] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: A big comparison for NAS," *CoRR*, vol. abs/1912.06059, 2019.

[103] W. Guo, S. Huang, Y. Tao, X. Xing, and L. Lin, "Explaining deep learning models - A bayesian non-parametric approach," *CoRR*, vol. abs/1811.03422, 2018.

[104] X. Han, B. Wallace, and Y. Tsvetkov, "Explaining black box predictions and unveiling data artifacts through influence functions," pp. 5553–5563, 01 2020.

Appendices

Appendix A

Datasets

Table A.1: All features pertaining to the BCD, including a brief description of them.

Id	Name	Type	Description
1	radius_mean	Linear	Mean radius of lobes
2	texture_mean	Linear	Mean of surface texture
3	perimeter_mean	Linear	Mean outer perimeter of lobes
4	area_mean	Linear	Mean area of lobes
5	smoothness_mean	Linear	Mean of smoothness levels
6	compactness_mean	Linear	Mean of compactness
7	concavity_mean	Linear	Mean of concavity
8	concave points_mean	Linear	Mean of concave points
9	symmetry_mean	Linear	Mean of symmetry
10	fractal_dimension_mean	Linear	Mean of fractal dimension
11	radius_se	Linear	Standard error of radius of lobes
12	texture_se	Linear	Standard error of surface texture
13	perimeter_se	Linear	Standard error of outer perimeter of lobes
14	area_se	Linear	Standard error of area of lobes
15	smoothness_se	Linear	Standard error of smoothness levels
16	compactness_se	Linear	Standard error of compactness
17	concavity_se	Linear	Standard error of concavity
18	concave points_se	Linear	Standard error of concave points
19	symmetry_se	Linear	Standard error of symmetry
20	fractal_dimension_se	Linear	Standard error of fractal dimension
21	radius_worst	Linear	Worst of radius of lobes
22	texture_worst	Linear	Worst of surface texture
23	perimeter_worst	Linear	Worst of outer perimeter of lobes
24	area_worst	Linear	Worst of area of lobes
25	smoothness_worst	Linear	Worst of smoothness levels
26	compactness_worst	Linear	Worst of compactness
27	concavity_worst	Linear	Worst of concavity
28	concave points_worst	Linear	Worst of concave points
29	symmetry_worst	Linear	Worst of symmetry
30	fractal_dimension_worst	Linear	Worst of fractal dimension

Table A.2: All features pertaining to the OCD, including a brief description of them.

Id	Name	Type	Description
1	MPV	Linear	Mean platelet volume
2	BASO#	Linear	Basophil cell count
3	PHOS	Linear	Phosphorus
4	GLU	Linear	Glucose
5	CA72-4	Linear	Carbohydrate antigen 72-4
6	K	Linear	Kalium
7	AST	Linear	Aspartate aminotransferase
8	BASO%	Linear	Basophil cell ratio
9	Mg	Linear	Magnesium
10	CL	Linear	Chlorine
11	CEA	Linear	Carcinoembryonic antigen
12	EO#	Linear	Eosinophil count
13	CA19-9	Linear	Carbohydrate antigen 19-9
14	ALB	Linear	Albumin
15	IBIL	Linear	Indirect bilirubin
16	GGT	Linear	Gama glutamyltransferasey
17	MCH	Linear	Mean corpuscular hemoglobin
18	GLO	Linear	Globulin
19	ALT	Linear	Alanine aminotransferase
20	DBIL	Linear	Direct bilirubin
21	RDW	Linear	Red blood cell distribution width
22	PDW	Linear	Platelet distribution width
23	CREA	Linear	Creatinine
24	AFP	Linear	Alpha-fetoprotein
25	HGB	Linear	Hemoglobin
26	Na	Linear	Natrium
27	HE4	Linear	Human epididymis protein 4
28	LYM#	Linear	Lymphocyte count
29	CA125	Linear	Carbohydrate antigen 125
30	BUN	Linear	Blood urea nitrogen
31	LYM%	Linear	Lymphocyte ratio
32	Ca	Linear	Calcium
33	AG	Linear	Anion gap
34	MONO#	Linear	Mononuclear cell count
35	PLT	Linear	Platelet count
36	NEU	Linear	Neutrophil ratio
37	EO%	Linear	Eosinophil ratio
38	TP	Linear	Total protein
39	UA	Linear	Urie acid
40	RBC	Linear	Red blood cell count

Table A.2: Continued.

Id	Name	Type	Description
41	PCT	Linear	Thrombocytocrit
42	CO2CP	Linear	Carbon dioxide-combining power
43	TBIL	Linear	Total bilirubin
44	HCT	Linear	Hematocrit
45	MONO%	Linear	Monocyte ratio
46	MCV	Linear	Mean corpuscular volume
47	ALP	Linear	Alkaline phosphatase
48	Age	Linear	Age of the patient
49	Menopause	Categorical	Whether the patient has had menopause (0) or not (1)

Table A.3: All features pertaining to the PCD, including a brief description of them.

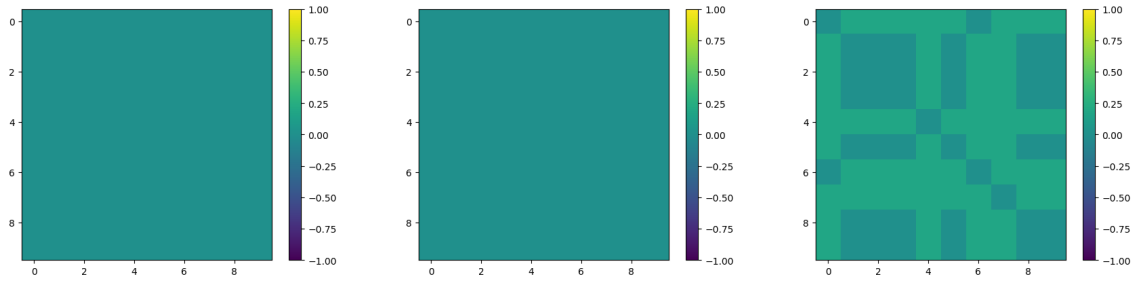
Id	Name	Type	Description
1	Age	Linear	Age of the patient
2	Sex	Categorical	Sex of the patient (M = Male, F = Female)
3	Plasma_CA19_19	Linear	Blood plasma levels of the CA19-19 monoclonal antibody
4	Creatinine	Linear	Urinary biomarker
5	CA72-4	Linear	Carbohydrate antigen 72-4
6	LYVE1	Linear	Urinary leveles of the protein lymphatic vessel endothelieal hyaluronan
7	REG1B	Linear	Urinary levels of the protein regenerating family member 1 beta
8	TFF1	Linear	Urinary levels of trefoil factor
9	REG1A	Linear	Urinary levels of the protein regenerating family member 1 alpha
10	patient_cohort	Categorical	Cohort from which the patient belonged to (Cohort1, Cohort2)
11	sample_origin	Categorical	Origin of the sample (BPTB: Barts Pancreas Tissue Bank, ESP: Spanish National Cancer Research Centre)
12	benign_sample_diagnosis	Categorical	The diagnosis for those with a benign, non-cancerous diagnosis (Pancreatitis, Other)

Table A.4: All features pertaining to the HDD, including a brief description of them.

Id	Name	Type	Description
1	Age	Linear	Age of the patient
2	Sex	Categorical	Sex of the patient (M = Male, F = Female)
3	ChestPainType	Categorical	Type of chest pain the patient exhibits (ASY, NAP, or Other)
4	RestingBP	Linear	Resting blood pressure
5	Cholesterol	Linear	Serum cholesterol
6	FastingBS	Linear	Fasting blood sugar
7	RestingECG	Categorical	Resting electrocardiogram results (Normal, LVH, or Other)
8	MaxHR	Linear	Maximum heart rate achieved
9	ExerciseAngina	Categorical	Whether or not exercise induced angina was achieved (Y = Yes, N = No)
10	Oldpeak	Linear	Value of the peak exercise ST segment
11	ST-Slope	Categorical	The slope of the peak exercise ST segment (Flat, Up, or Other)

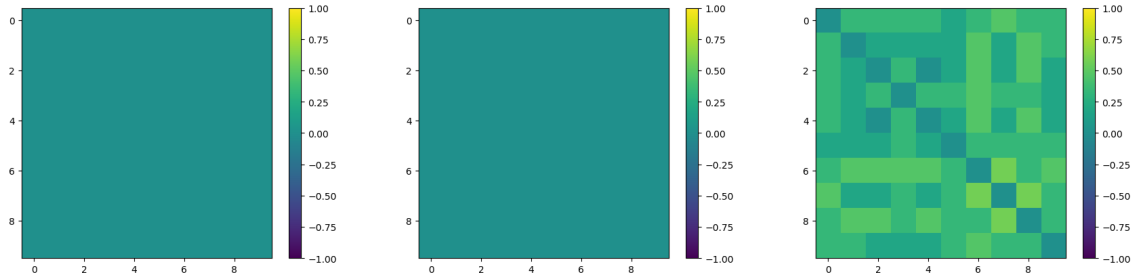
Appendix B

Results



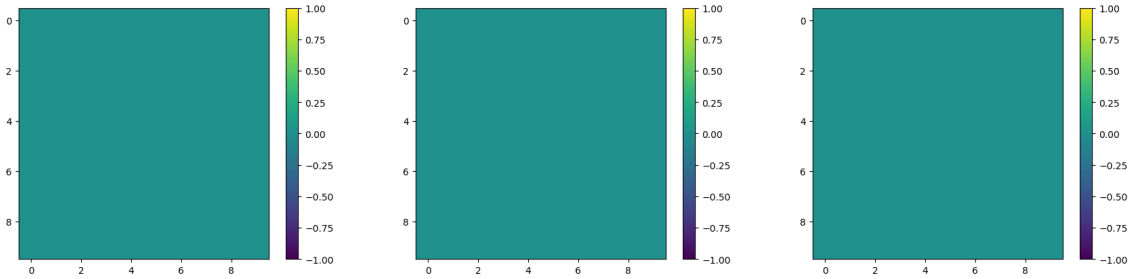
(a) DLIME (b) AL-DLIME (c) LIME

Figure B.1: Results for Jaccard's distance across ten iterations on a single, random instace from the standardized OCD, presented in confusion matrices.



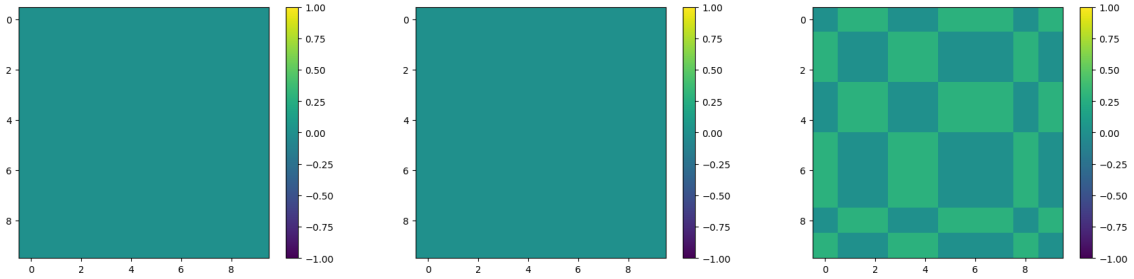
(a) DLIME (b) AL-DLIME (c) LIME

Figure B.2: Results for Jaccard's distance across ten iterations on a single, random instace from the non-standardized OCD, presented in confusion matrices.



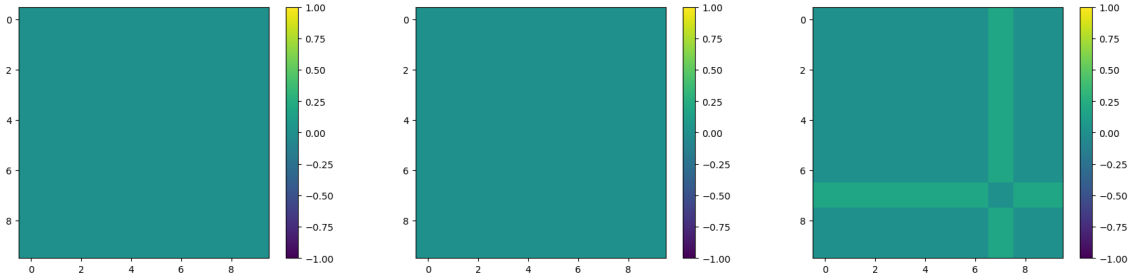
(a) DLIME (b) AL-DLIME (c) LIME

Figure B.3: Results for Jaccard's distance across ten iterations on a single, random instace from the standardized PCD, presented in confusion matrices.



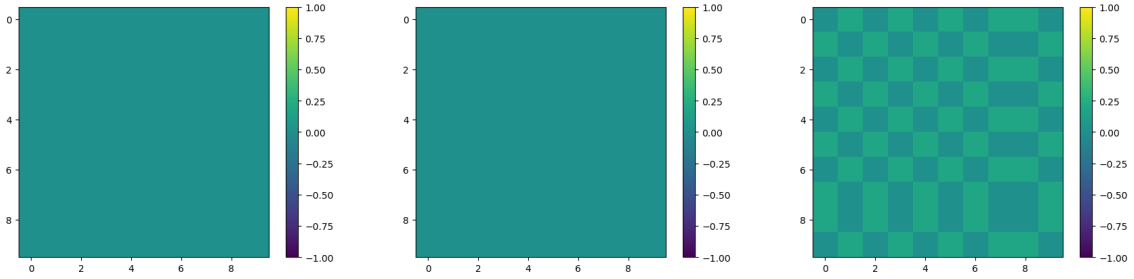
(a) DLIME (b) AL-DLIME (c) LIME

Figure B.4: Results for Jaccard's distance across ten iterations on a single, random instace from the non-standardized PCD, presented in confusion matrices.



(a) DLIME (b) AL-DLIME (c) LIME

Figure B.5: Results for Jaccard's distance across ten iterations on a single, random instace from the standardized HDD, presented in confusion matrices.

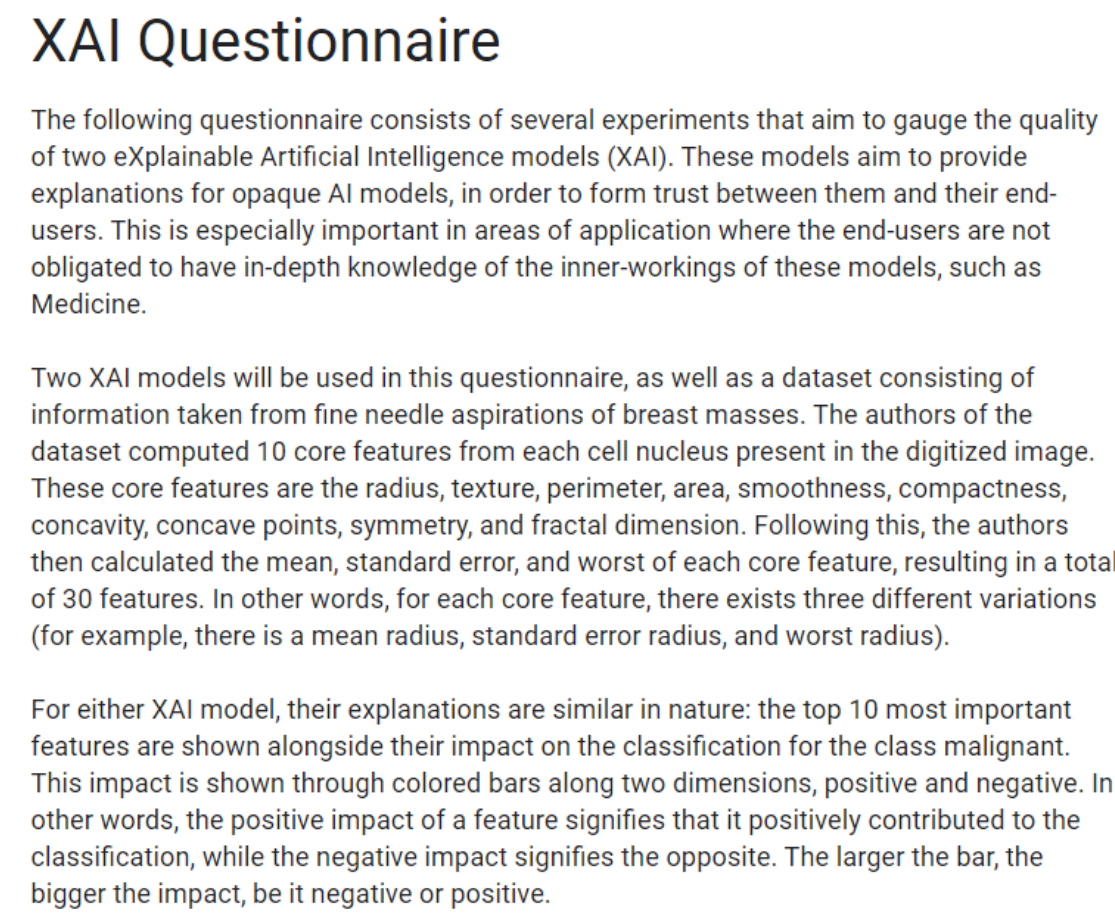


(a) DLIME (b) AL-DLIME (c) LIME

Figure B.6: Results for Jaccard's distance across ten iterations on a single, random instace from the non-standardized HDD, presented in confusion matrices.

Appendix C

Future Work



XAI Questionnaire

The following questionnaire consists of several experiments that aim to gauge the quality of two eXplainable Artificial Intelligence models (XAI). These models aim to provide explanations for opaque AI models, in order to form trust between them and their end-users. This is especially important in areas of application where the end-users are not obligated to have in-depth knowledge of the inner-workings of these models, such as Medicine.

Two XAI models will be used in this questionnaire, as well as a dataset consisting of information taken from fine needle aspirations of breast masses. The authors of the dataset computed 10 core features from each cell nucleus present in the digitized image. These core features are the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Following this, the authors then calculated the mean, standard error, and worst of each core feature, resulting in a total of 30 features. In other words, for each core feature, there exists three different variations (for example, there is a mean radius, standard error radius, and worst radius).

For either XAI model, their explanations are similar in nature: the top 10 most important features are shown alongside their impact on the classification for the class malignant. This impact is shown through colored bars along two dimensions, positive and negative. In other words, the positive impact of a feature signifies that it positively contributed to the classification, while the negative impact signifies the opposite. The larger the bar, the bigger the impact, be it negative or positive.

Figure C.1: Introduction to the questionnaire. This section has the objective of explaining the core concepts of the experiment so that the medical professionals filling out the questionnaire have some context regarding the following questions.

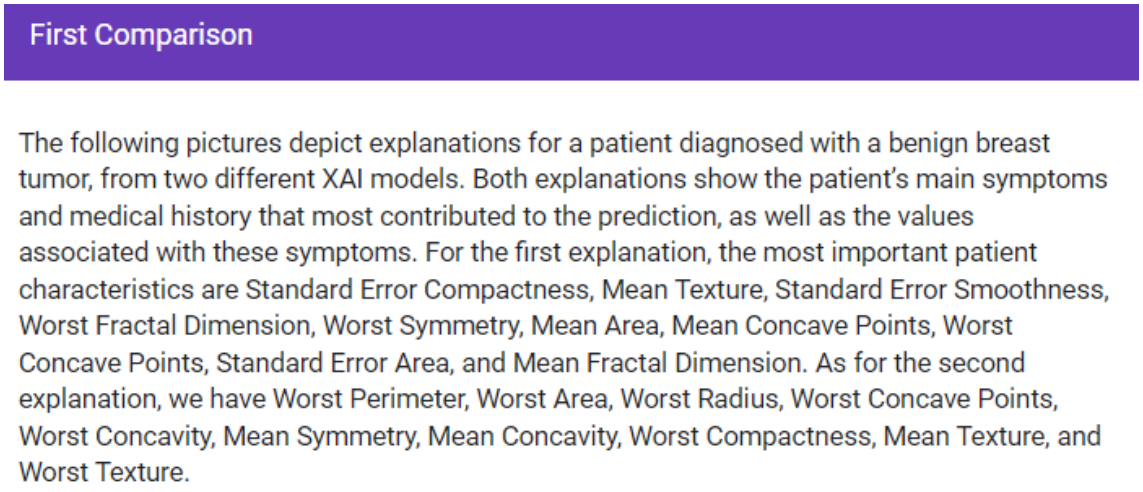
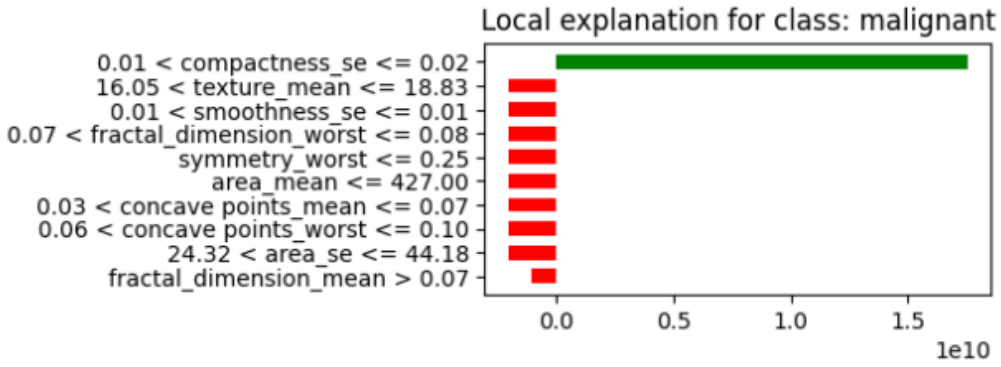
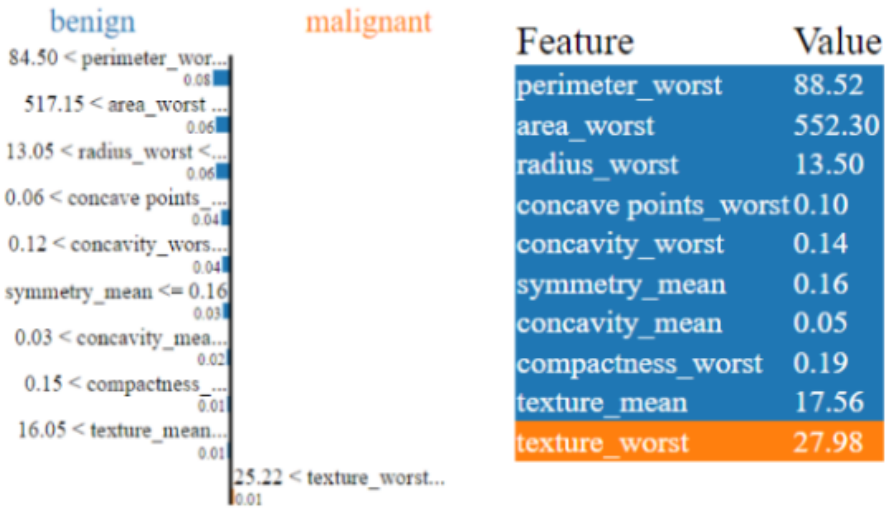


Figure C.2: Introduction to the first section. The goal of this section is to collect a general first impression from the medical professionals regarding their preference between LIME and DLIME, given some further context.

Between the two explanations presented, which one do you prefer? *





◯ Top Explanation (Model 1)

◯ Bottom Explanation (Model 2)

Please explain your choice *

A sua resposta

Figure C.3: The first and second questions from the first section. The first aims to collect a preference from the medical professionals, with an explanation for that choice required in the second question.

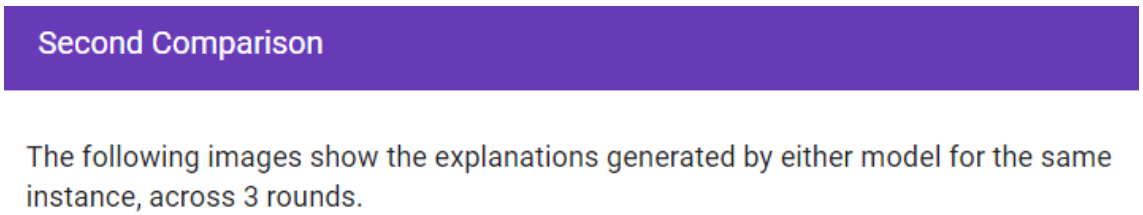
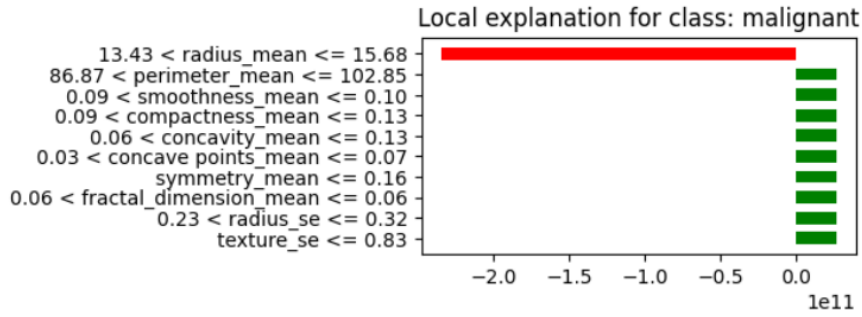
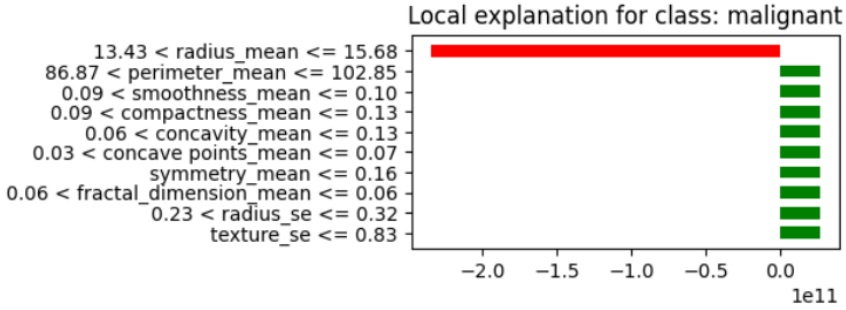


Figure C.4: Introduction to the second section. The goal of this section is to collect a more informed decision from the medical professionals, through a demonstration of the stability of explanations from DLIME and LIME.

Round 1, Model 1



Round 2, Model 1



Round 3, Model 1

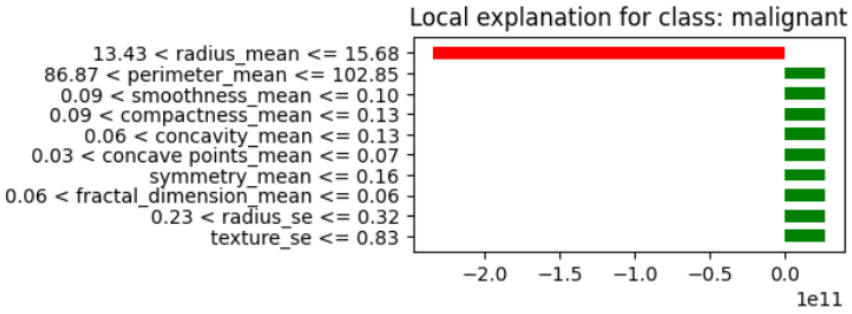
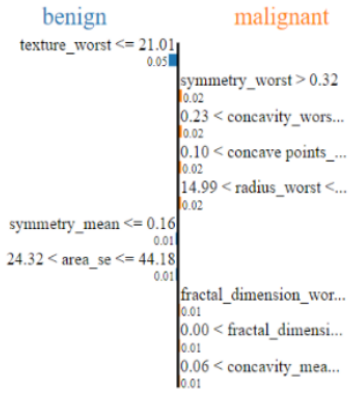


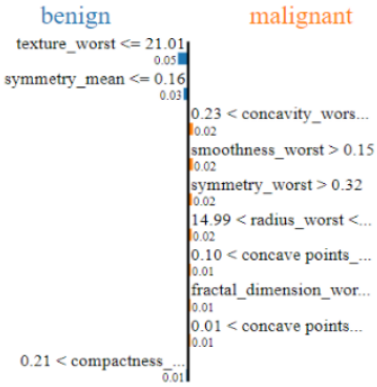
Figure C.5: Examples of DLIME's explanations regarding the same instance across three different rounds, intended to demonstrate DLIME's stability.

Round 1, Model 2



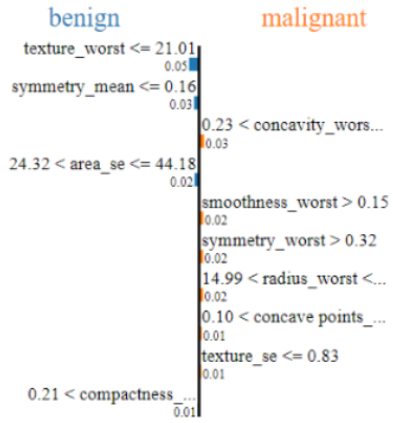
Feature	Value
texture_worst	20.24
symmetry_worst	0.34
concavity_worst	0.33
concave points_worst	0.13
radius_worst	17.77
symmetry_mean	0.16
area_se	26.44
fractal_dimension_worst	0.10
fractal_dimension_se	0.00
concavity_mean	0.08

Round 2, Model 2



Feature	Value
texture_worst	20.24
symmetry_mean	0.16
concavity_worst	0.33
smoothness_worst	0.15
symmetry_worst	0.34
radius_worst	17.77
concave points_worst	0.13
fractal_dimension_worst	0.10
concave points_se	0.01
compactness_worst	0.33

Round 3, Model 2



Feature	Value
texture_worst	20.24
symmetry_mean	0.16
concavity_worst	0.33
area_se	26.44
smoothness_worst	0.15
symmetry_worst	0.34
radius_worst	17.77
concave points_worst	0.13
texture_se	0.36
compactness_worst	0.33

Figure C.6: Examples of LIME's explanations regarding the same instance across three different rounds, intended to demonstrate the lack of LIME's stability.

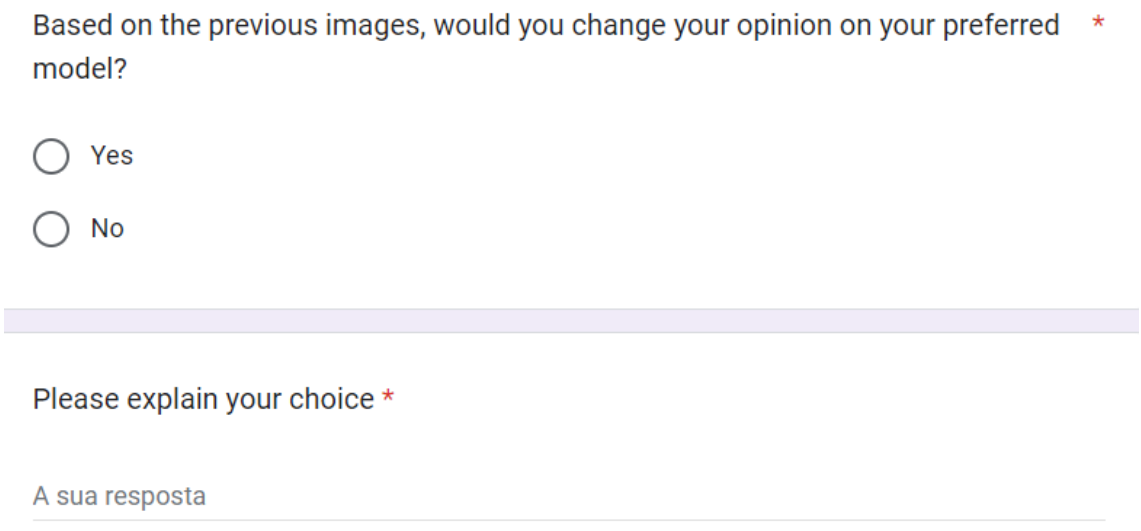


Figure C.7: The first and second question from the second section. Intended to gauge whether or not the previous demonstrations of the models' stability would change the medical professionals' previously established opinion. They are then requested to provide an explation regarding their choice.

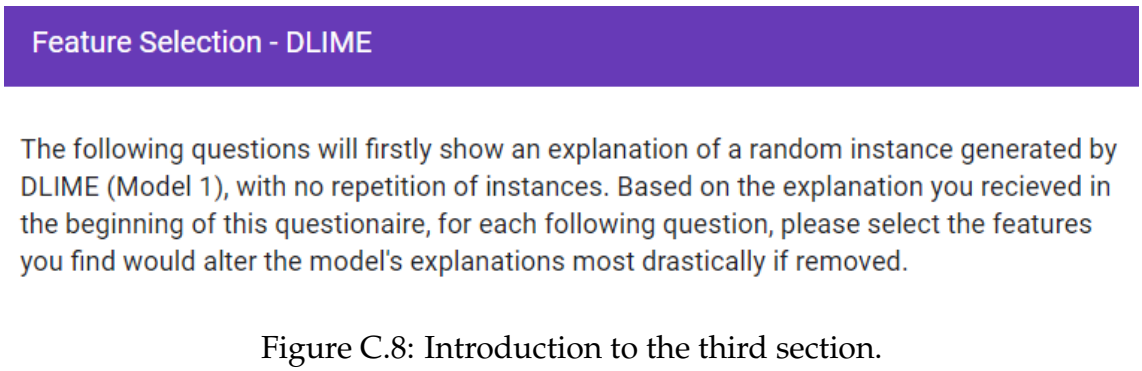
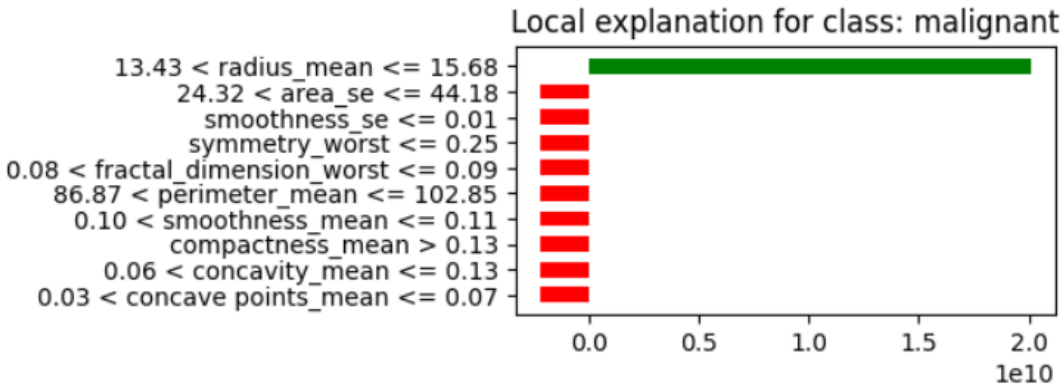


Figure C.8: Introduction to the third section.

Most vital features for instance 1 *



- [ ] Mean Perimeter
- [ ] Worst Fractal Dimension
- [ ] Worst Symmetry
- [ ] Mean Compactness
- [ ] Standard Error Smoothness
- [ ] Mean Radius
- [ ] Mean Concavity
- [ ] Mean Concave Points
- [ ] Standard Error Area
- [ ] Mean Smoothness

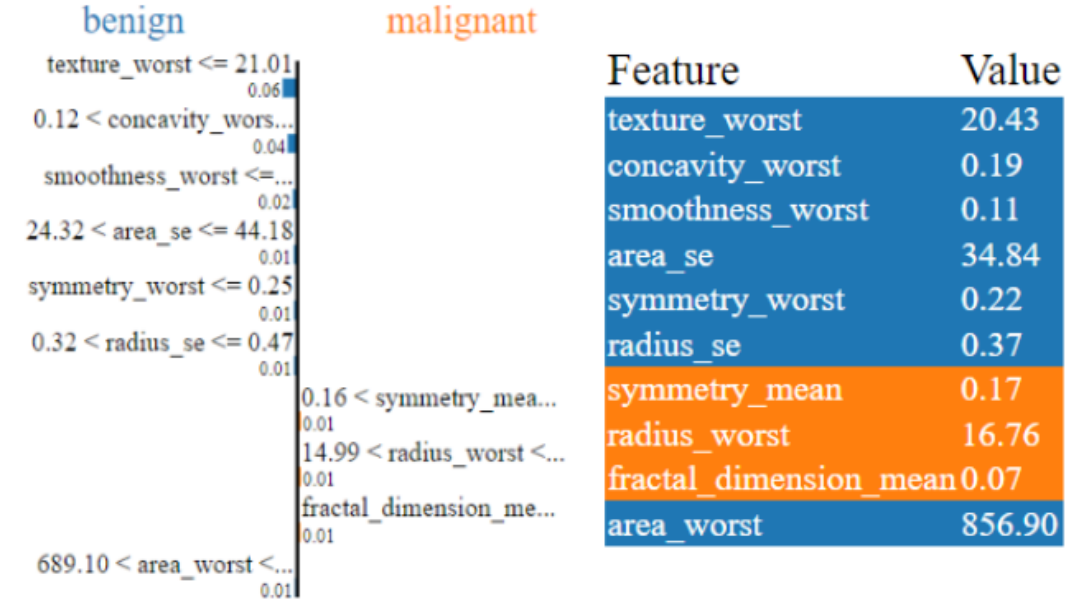
Figure C.9: An example of one of the three questions from the third section. The objective of these questions is to gauge how well the medical professionals taking this questionnaire have come to understand how DLIME functions, while simultaneously evaluating DLIME's ability to present information in an understandable manner. The second and third questions are presented in the same manner, with the exception of the initial figure, which depict explanations from DLIME of different instances.

Feature Selection - LIME

The following questions are of the same kind as the previous section, this time with LIME (Model 2). Please proceed in the same way.

Figure C.10: Introduction to the fourth section.

Most vital features for instance 1 *



- [ ] Worst Radius
- [ ] Worst Texture
- [ ] Worst Concavity
- [ ] Worst Smoothness
- [ ] Mean Fractal Dimension
- [ ] Worst Symmetry
- [ ] Mean Symmetry
- [ ] Standard Error Radius
- [ ] Standard Error Area
- [ ] Worst Area

Figure C.11: An example of one of the three questions from the fourth section. The objective of these questions is to gauge how well the medical professionals taking this questionnaire have come to understand how LIME functions, while simultaneously evaluating LIME's ability to present information in an understandable manner. The second and third questions are presented in the same manner, with the exception of the initial figure, which depict explanations from LIME of different instances.

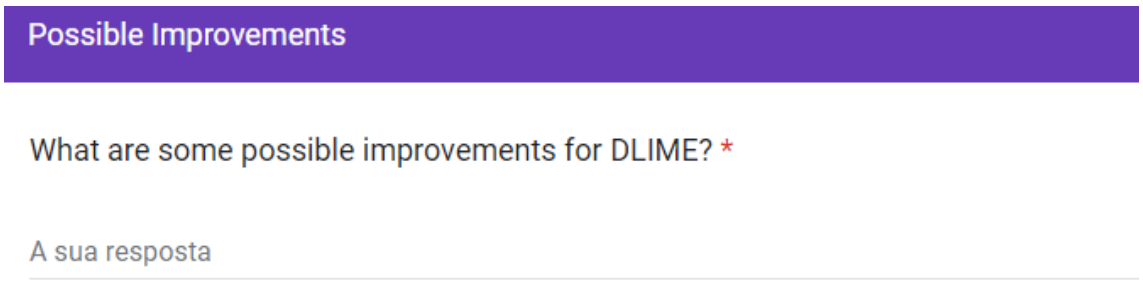


Figure C.12: The fifth and final section of the questionnaire, which aims to collect any and all suggestions the medical professionals have on how to improve DLIME and, by extension, AL-DLIME.