Universidade de Aveiro
2023

Beatriz Tavares
da Costa

Explainable AI em aplicações médicas

Explainable AI in medical applications

Universidade de Aveiro
2023

Beatriz Tavares da Costa

Explainable AI em aplicações médicas

Explainable AI in medical applications

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Robótica e Sistemas Inteligentes, realizada sob a orientação científica do Doutor Pétia Georgieva, Professor associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

Dedico este trabalho aos meus avós por terem sido o meu primeiro exemplo do que é lutar sempre pelos nossos objetivos e nunca desistir.

o júri / the jury

presidente / president

Professor Doutor Vítor Manuel Ferreira dos Santos
Professor Associado C/ Agregação, Universidade de Aveiro

vogais / examiners committee

Professora Doutora Catarina Helena Branco Simões Silva
Professora Auxiliar, Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Professora Doutora Pétia Georgieva Georgieva
Professor Associado C/ Agregação, Universidade de Aveiro

agradecimentos / acknowledgements

À minha orientadora, Professora Doutora Pétia Georgieva, aos meus pais e família, namorado e amigos,

Gostaria de expressar a minha mais profunda gratidão a todos por serem parte fundamental da minha jornada académica e pessoal durante a realização desta dissertação.

À minha Orientadora: Agradeço por toda a paciência mais principalmente por todo o conhecimento que me foi transmitindo ao longo do meu percurso de mestrado. Desde a disponibilidade a toda a hora, comunicação incrível até à ajuda nos momentos mais complicados da escrita da tese, tenho só a agradecer por me ter cultivado esta paixão por este tema e por ser uma figura que eu considero um exemplo a nível profissional. Sem a sua orientação, este trabalho não teria alcançado o mesmo nível de excelência.

Aos meus Pais: Acho que não preciso dizer muito, a não ser um obrigada por me apoiarem em todas as fases desta jornada. Obrigada por acreditarem em mim e pelo amor constante que me deram. Nunca esquecerei toda a dedicação e sacrifícios que fizeram na vossa vida para que eu pudesse chegar até aqui. Devo-vos o mundo.

Ao meu Namorado: O apoio constante, carinho, compreensão e incentivo foram o meu refúgio durante os momentos mais desafiadores desta jornada. A tua presença trouxe equilíbrio à minha vida e tornou cada obstáculo mais fácil de superar. É fácil viver na tua companhia.

Aos meus Amigos: O que seria da minha vida sem vocês. A minha segunda família. As gargalhadas partilhadas, as conversas motivadoras e o apoio mútuo tornaram este caminho mais significativo e memorável.

À Universidade de Aveiro: Obrigada por proporcionar um ambiente académico estimulante e inspirador. Os recursos que disponibilizam, o corpo docente e a comunidade académica contribuíram imensamente para o meu crescimento pessoal e profissional.

Esta tese é o resultado de um esforço coletivo, e todos vocês desempenharam papéis cruciais nesta minha jornada. Espero que continuemos a partilhar momentos de alegria, sucesso e aprendizagem.

Com profunda gratidão.

Palavras Chave

inteligência artificial, machine learning, aplicações médicas, revisão sistemática de literatura

Resumo

A Inteligência Artificial (IA) tem vindo a revolucionar o sector da saúde a vários níveis, através da automatização de tarefas, melhorando as previsões e sendo capaz de analisar grandes quantidades de dados. Uma das suas maiores aplicações tem sido na análise de diagnósticos médicos, a fim de detetar padrões e prever potenciais problemas de saúde numa fase inicial. No entanto, existe um desafio: sendo "modelos de caixa negra", muitas vezes não apresentam transparência nos seus resultados, o que levanta preocupações relativamente à compreensão dos processos de tomada de decisão do modelo. Para resolver este problema, foi criada a *Explainable AI* (XAI). Esta pode oferecer explicações claras e compreensíveis para as acções da IA, através da geração de explicações visuais. Esta dissertação centra-se na utilização de métodos XAI para aumentar a transparência, a equidade e a segurança num ambiente de classificação de imagens médicas, mais especificamente num modelo personalizado para a deteção de doenças pulmonares. A primeira fase do projeto envolve uma revisão sistemática da literatura para analisar a investigação existente sobre XAI no ramo da saúde. A revisão centra-se em artigos publicados entre 2016 e 2022, obtidos a partir de várias bases de dados. Os critérios de seleção incluem a relevância da XAI nos cuidados de saúde e os artigos elegíveis finais foram submetidos a um exame minucioso e a uma análise estatística. A partir desta pesquisa sistemática, foram selecionados três métodos XAI diferentes para serem aplicados no nosso modelo. Na fase de implementação, o GRAD-CAM, o LIME e o RISE foram integrados em duas abordagens de modelos: um modelo pré-treinado e um modelo totalmente treinado para a classificação de imagens de raios-X do tórax. Duas das técnicas XAI forneceram explicações visuais para previsões individuais, melhorando a interpretabilidade. A dissertação é concluída com a avaliação do desempenho do modelo e das técnicas XAI. O modelo pré-treinado obteve um elevado desempenho na deteção de doenças pulmonares. Duas das técnicas XAI tiveram resultados semelhantes, aumentando a confiança nos padrões do modelo e no processo de tomada de decisão. Este modelo tem como objetivo melhorar os cuidados e os resultados dos doentes. Existe uma longa jornada para o desenvolvimento de modelos de IA fiáveis e interpretáveis nos cuidados de saúde e isto deve ser mantido como uma prioridade a fim de melhorar a prestação de cuidados de saúde.

Keywords explainable AI, artificial intelligence, machine learning, healthcare, medical applications, systematic literature review

Abstract

Artificial Intelligence (AI) has been revolutionizing the healthcare industry by automating tasks, improving predictions, and being able to analyze large amounts of data. One of the biggest applications has been in terms of analyzing medical diagnoses in order to detect patterns and predict potential health problems early on. However, there's a challenge that comes to this: being "black-box models", they often lack transparency in their results, and this raises concerns regarding understanding the model's decision-making processes. To address this, Explainable Artificial Intelligence (XAI) was created. It can offer clear and understandable explanations for AI actions, through visual explanations. This dissertation focuses on utilising XAI methods to enhance transparency, fairness, and safety in a medical image classification, more specifically in a customised model for detecting pulmonary diseases. The first phase of the project involves a systematic literature review to analyse existing research on XAI in healthcare applications. The review focuses on articles published between 2016 and 2022, obtained from various databases. The selection criteria includes relevance to XAI in healthcare, and the final eligible articles underwent thorough scrutiny and statistical analysis. From this systematic search three different XAI methods were chosen to be applied in our model. In the implementation phase, GRAD-CAM, LIME, and RISE were integrated into two model approaches: a pre-trained and fully-trained model for chest X-ray image classification. Two of the XAI techniques provided visual explanations for individual predictions, improving interpretability. The dissertation is concluded by evaluating the performance of both the model and the XAI techniques. The pre-trained model achieved high performance in detecting pulmonary diseases. Two of the XAI techniques had similar results, enhancing the trust in the model patterns and decision-making process. This model aims to improve patient care and outcomes. The ongoing journey towards developing trustworthy and interpretable AI models in healthcare must be kept as a priority when enhancing healthcare delivery.

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
AUC	Area Under the ROC Curve
CAM	Class Activation Mapping
CEM	Contrastive Explanation Method
CNN	Convolutional Neural Network
CXRs	Chest X-Rays
DARPA	Defense Advanced Research Projects Agency
DenseNet	Dense Convolutional Network
DL	Deep Learning
EHR	Eletronic Health Records
GAP	Global Average Pooling
GPUs	Graphics Processing Units
Grad-CAM	Gradient-weighted Class Activation Mapping
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
NIH	National Institutes of Health
NLP	Natural Language Processing
PDPs	Partial Dependence Plots
PRISMA	Preferred Reporting Items on Systematic Reviews and Meta-analysis
ReLU	Rectified Linear Unit
RISE	Randomized Input Sampling for Explanation
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
RQs	Research Questions
SHAP	SHapley Additive exPlanations
TCN	Temporal Convolutional Network
TPUs	Tensor Processing Units
XAI	Explainable Artificial Intelligence

CHAPTER 1

Introduction

This chapter delivers the motivation and objectives behind the design of the subsequent work plan and gives some context for the problem description.

> *"The only way to do great work is to love what you do."*
>
> *Steve Jobs,* 2005

1.1 Context and Motivation

Artificial Intelligence (AI) has been helping the healthcare field through task automation, providing more accurate predictions and being able to analyse large amounts of data. These tasks can be applied in a variety of healthcare applications, including diagnosis and patient monitoring. AI can also be used to analyse Eletronic Health Records (EHR) and other data sources to identify patterns and predict earlier-stage potential health problems. Overall, the use of AI in healthcare has the potential to improve the efficiency and effectiveness of the healthcare system, as well as to enhance patient care.

An AI model is often referred to as having a "black-box" nature since it is difficult or impossible to understand how it got a certain prediction. This is caused by the model layers being invisible and not being able to explain how it arrived at a particular output to its final stakeholder.

One of the current challenges of using AI models is that they can be complex and may involve many layers of analysis and processing. The training data of the model usually has thousands of images containing millions of pixels, making it not immediately apparent for the human eye to perceive the relationships between the data and the outputs.

This lack of transparency can become a concern in critical situations or systems, such as in legal or medical contexts. Also, for researchers and developers trying to improve the performance of an AI model, it can be too difficult to understand where errors or biases come from.

The concept of Explainable Artificial Intelligence (XAI) was created to tackle these problems and therefore ensure trust, transparency, and ethics in various fields.

Although it has no formal definition, according to the Defense Advanced Research Projects Agency (DARPA) [1], XAI is a type of AI that can provide clear, understandable explanations for its actions and decisions. The goal of XAI is to make it easier for humans to understand and trust the actions of AI systems, particularly in complex or high-stakes situations. DARPA has funded

research in this area to develop technologies that can help improve the transparency, accountability, and interpretability of AI systems.

XAI allows for a better understanding of how AI algorithms reach their decisions, and can even be used to optimise and de-bias these algorithms. This makes it a valuable tool in fields such as healthcare, where it can promote fairness and safety in decision making. Additionally, XAI can be applied to a wide range of Machine Learning (ML) models, making the evolution process of current approaches easier.

1.2 Problem Description and Objectives

This dissertation aims at the development of algorithms and models to make the medical image classification process understandable to medical professionals. The primary objectives for this dissertation are as follows:

1. Critically analyse the state of the art, with a particular emphasis on understanding the importance of the integration and evaluation of XAI within medical applications;
2. Perform dataset pre-processing as well as data augmentation techniques;
3. Implement a medical image classification system using a pre-trained DenseNet model and adapt it to our dataset;
4. Perform experimental evaluation of the effectiveness and performance of the medical image classification system using Explainable AI.

1.3 Document Structure

After consideration of the objectives depicted, the remainder of this dissertation is divided into five chapters, described as follows. Chapter 2 outlines some fundamental concepts related to the problem's paradigm and context. Chapter 3 is dedicated to a systematic review of the literature and state-of-the-art solutions. Following the document, Chapter 4 describes the proposed approach. Chapter 5 is dedicated to the experimental evaluation of the proposed methods, including the datasets and dataset treatment used to test the proposed architecture. Lastly, Chapter 6 concludes this work and addresses future research directions.

CHAPTER 2

Concepts

This chapter gives an overview of some of the key concepts, AI models and explainability techniques that were chosen to be applied in the implementation phase of the project.

> *Learning never exhausts the mind.*
>
> *Leonardo da Vinci,* Unknown

2.1 Definitions

2.1.1 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence XAI is referred often to as a set of techniques and methodologies that aim for a more transparent and understandable decision-making process in terms of AI models. [2].

XAI is unfolding the black-box process and is turning out to be essential in providing insights into how AI systems arrive at their predictions or decisions. This enables all types of stakeholders (users, medical staff, and programmers) to comprehend and trust the AI-driven outcomes. It employs a wide range of methods, from unveiling the features that influence most of the predictions to generating interpretable explanations for AI predictions.

2.1.2 Interpretability vs Explainability

When talking about the concepts of Interpretability and Explainability we have to note that they are similar in certain ways but turn out to be distinct concepts in the field of AI and ML.

On one hand, interpretability is described as the understatement of the way AI models work internally. It focuses on how the model processes and manipulates data. It also involves the visualization of the model's internal components. It is considered an advantage to use when you need to understand why a model made a specific prediction.

On the other hand, explainability's focus is on providing understandable explanations for AI decisions to human stakeholders, even if they are non-technical users. It is often used when we want to answer the "why" behind a model's output. Explainability is critical for building trust in AI systems, particularly in critical applications.

Overall, interpretability unveils the model's mechanics, while explainability goes further by generating accessible explanations for AI predictions.

The choice between applying both of these concepts depends on whether the need is to either understand the model's inner workings or provide clear, user-friendly explanations.

2.1.3 Typical Challenges in XAI

XAI grapples with several common challenges. These include the complexity of explaining intricate AI models, the trade-off between model performance and explainability, and securing consistency in explanations. Also, XAI faces difficulties in creating user-friendly explanations. It is often difficult to find the right balance between comprehensiveness and clarity, and scaling methods for large datasets and real-time applications.

Robustness against adversarial attacks and noisy data, as well as defining effective evaluation criteria, remain ongoing challenges. Ensuring the use of model-agnostic XAI, addressing bias and fairness in explanations, and the adaptation to dynamic or temporal data present complications. It is a priority to pay attention to legal and ethical considerations, especially in terms of privacy and data protection.

Developing interactive XAI systems and optimizing human-computer interaction for clear communication of AI explanations is multidisciplinary. Bridging the gap between AI experts and end-users, along with educating druggies about AI limitations, is a vital aspect of XAI's development. These challenges punctuate the need for nonstop exploration and invention in XAI to establish transparent, reliable, and secure AI systems across different applications.

2.1.4 Explanation Evaluation

A crucial part of Explainable Artificial Intelligence XAI, which measures the effectiveness and usefulness of AI system explanations, is explanation evaluation. Assessing human-centric elements like clarity and usefulness, measuring explanation fidelity and comprehensibility, and taking into account the impact on system performance are important evaluation aspects. It is crucial to ensure fairness, legality, and relevance to particular domains. It is also crucial to evaluate how well an explanation can be applied to various AI models. Users and stakeholders alike benefit from the ongoing development of evaluation methodologies in XAI because it increases transparency and trust in AI systems.

2.2 DenseNet

Dense Convolutional Network (DenseNet), originally proposed by [3], is referred to *dense* due to its dense connectivity pattern. The authors focused on solving the problem of connectivity patterns between layers, ensuring maximum information and gradient flow. Their solution is to connect every layer directly with each other, requiring fewer parameters than a traditional Convolutional Neural Network (CNN).

In a traditional CNN, for L layers, there are L direct connections - one between each layer and its subsequent layer. For L layers, there are $L(L+1)/2$ direct connections. For each layer, the feature maps of all the preceding layers are used as inputs, and its feature maps are used as input for each subsequent layer.

A DenseNet architecture is divided into multiple dense blocks, as observed in Figure 2.1.

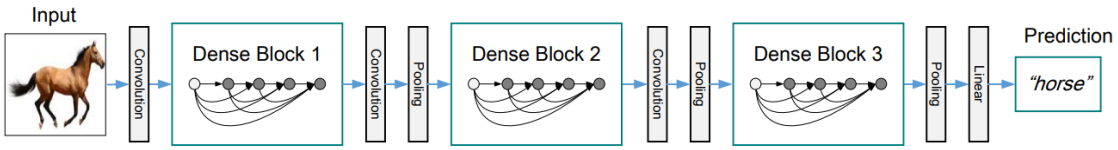

Figure 2.1: DenseNet model architecture. Extracted from [3].

DenseNets can be used to classify images of pulmonary diseases in a medical imaging dataset. In such a scenario, the dense net architecture can be used to learn features from the images and make predictions about the presence of different types of pulmonary diseases. We will discuss now a specific approach of using DenseNets for the interpretation of chest X-rays.

2.2.1 CheXNet

Chest X-Rays (CXRs) are one of the most commonly performed diagnostic imaging procedures in the world, playing a crucial role in the detection and monitoring of various pulmonary and cardiovascular conditions. However, the interpretation of CXRs is often a complex and time-consuming task, requiring the expertise of trained radiologists. The demand for timely and accurate diagnoses, coupled with the shortage of radiologists in many regions, has fueled the need for automated solutions. ChexNet addresses this need by harnessing the power of deep learning to analyse chest X-rays swiftly and accurately.

Deep learning, a subset of machine learning, has demonstrated remarkable success in various domains, including computer vision and natural language processing. CNN, a class of deep learning models, have been particularly effective in image analysis tasks. Their ability to automatically learn hierarchical features from raw data makes them well-suited for medical image analysis.

ChexNet, introduced in 2017 by researchers at Stanford University [4], was designed specifically for the automated interpretation of chest X-rays. This concept uses a DenseNet-121 architecture [3] as its backbone to analyse CXRs and provide insights into the presence or absence of various thoracic abnormalities, including pneumonia, cardiomegaly, and pneumothorax, among others.

ChexNet was designed with interpretability in mind. It can provide heatmaps highlighting the regions of the X-ray image that contributed to its decision, allowing healthcare professionals to understand the model's reasoning and increasing their confidence in its recommendations. As observed in Figure 2.2, the model takes an image as input and outputs the pathology's probability, alongside a heatmap that highlights the area most indicative of this pulmonary disease.

In summary, the ChexNet concept represents a transformative approach to chest X-ray analysis, showcasing the potential of deep learning in revolutionizing medical imaging and diagnosis. This subsection has provided an overview of ChexNet's background, development, key features, clinical applications, and ethical implications, setting the stage for a deeper exploration of its impact in subsequent sections of this thesis.

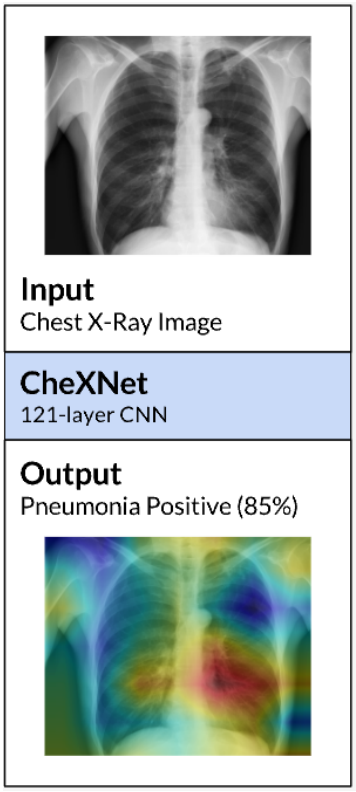

Figure 2.2: Example of application of the CheXNet model. Extracted from [4].

2.3 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) technique was first proposed by [5] in 2019 and aims at producing visual explanations for decisions from a wide variety of CNN-based models, to make them more transparent and explainable.

It is considered a generalization of original Class Activation Mapping (CAM), originally proposed in [6], taking out its limitations. CAM requires the application of Global Average Pooling (GAP), making it only applicable to a particular kind of CNN architectures that make feature maps directly precede softmax layers, as observed in Figure 2.3.

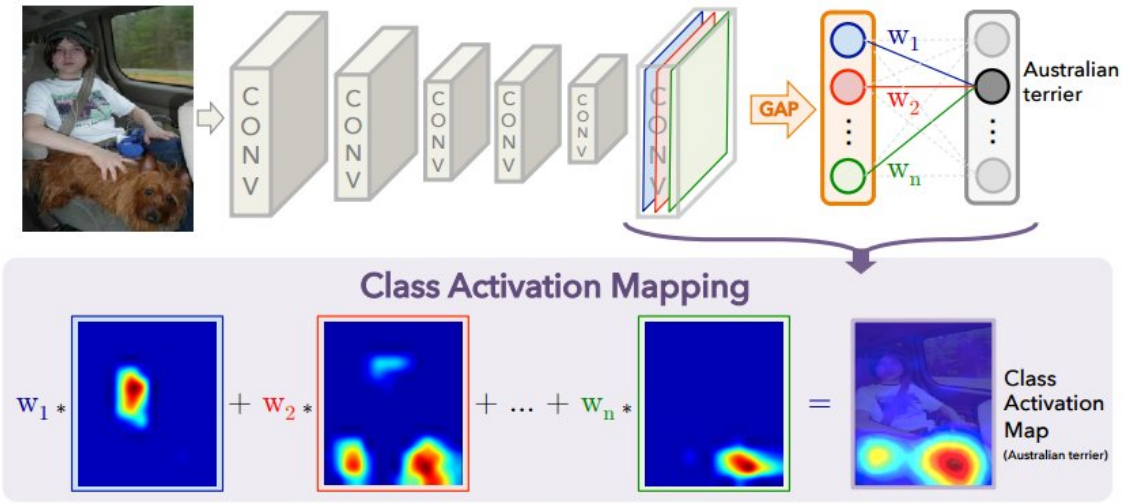

Figure 2.3: Overview of CAM algorithm proposed and extracted from [6].

Grad-CAM method extracts gradients from a CNN final convolutional layer and uses this to highlight regions most responsible for the predicted probability that the image belongs to a predefined class. Figure 2.4 represents how the algorithm works.

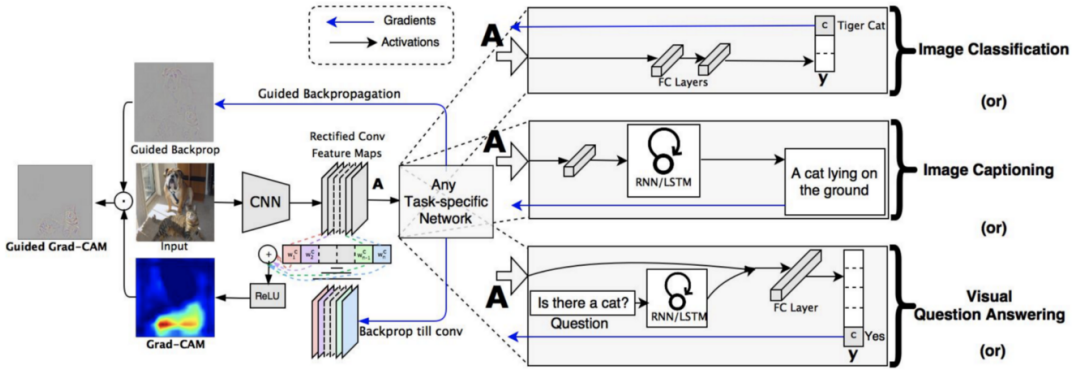

Figure 2.4: Grad-CAM overview. Extracted from [5].

Given two inputs - an image and a class of interest (e.g., 'cat') - the image is forward propagated through the CNN and obtains a raw score for the category through task-specific computations. The gradients are then set to zero for all classes except the desired class (cat), which is set to 1. Then, this signal is backpropagated to the feature maps of interest, which are combined to compute the Grad-CAM localization (blue heatmap). This blue heatmap represents where the model has to look to make the decision. Finally, using guided backpropagation, they pointwise multiply the heatmap to get Guided Grad-CAM visualizations.

In the case of medical imaging, the attention map reflects which parts of the image are affecting the model's predictions most. As observed from Figure 2.5, when looking through COVID-19 X-Ray images from a dataset available in [7], the network focused on the ground glass opacity, which is considered the most prevalent clinically observed pathology for COVID-induced pneumonia. On the other hand, for the Pneumonia cases, the highlights are from typical lung inflammation indicative of pneumonia. In normal cases, no highlighted regions are observed.

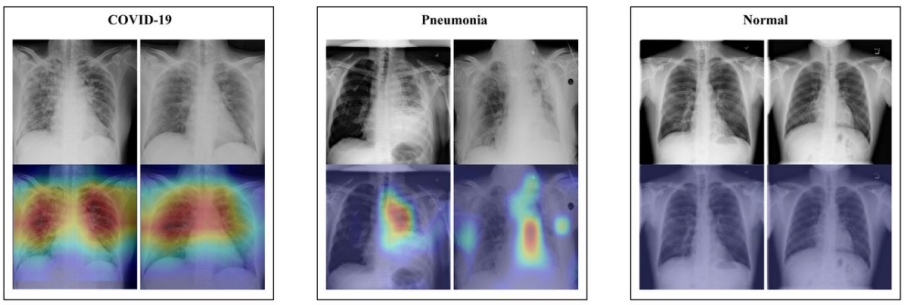

Figure 2.5: Visual explanations obtained by application of Grad-CAM method on different disease classes. Extracted from [7].

This information can be valuable to guide medical experts in the search for confirmation of the suspected diagnosis. In addition, given that the algorithm is accurate, its repeated use in different patients will increase a physician's trust in the algorithm's predictions, helping to make it more explainable.

2.4 LIME

Local Interpretable Model-agnostic Explanations (LIME), originally proposed by Ribeiro et al. in [8], is one of the most used XAI techniques. It can be applied to a variety of data since explanations that are generated by LIME are based on interpretable components. These components might differ from the input features of the original model since they are representations of the underlying data which are understandable to humans. For example, an interpretable component can be a contiguous region of an image, as presented in Figure 2.6.

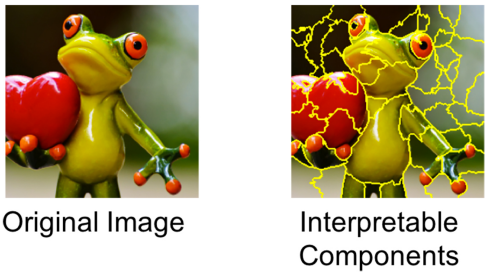

Figure 2.6: LIME determined interpretable components by pixel segmentation. Extracted from [8].

In the context of image data, LIME works by doing input data permutation (i.e. producing a set of images that are comparable to our input picture by turning on and off parts of the image's super-pixels).

The cosine distance between each sample and the original image will be computed. The greater the resemblance between a synthetic image and the original image, the greater the weight and importance of the sample.

The class of each synthetic image is then predicted and used to estimate the importance of different regions of the original image to the classifier's prediction.

This importance is often represented as a heatmap or by presenting the isolation of the most important features in an image from the unimportant features, as observed in Figure 2.7.

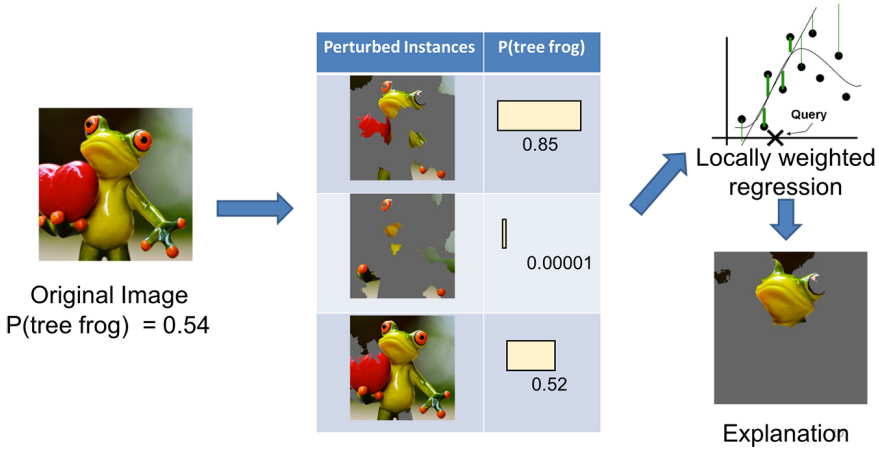

Figure 2.7: LIME overview. Extracted from [8].

2.5 RISE

Randomized Input Sampling for Explanation (RISE), introduced as an XAI technique firstly introduced by Petsiuk et al. in their paper [9], is another vastly-used method. It can provide insights

into black-box machine learning models, particularly for image data. The interpretations it generates are built on interpretable components so it can be used on various data types. These components may differ from the original input features of the model since they represent distinct regions of an image, which are understandable to humans. These interpretable components in the context of rise are segments of the image.

The process behind RISE's when working with image data involves the creation of synthetic data by randomly masking parts of the input image. These masks determine which regions of the image are "hidden" or "masked" in each synthetic image.

The degree of similarity between each of these synthetic images and the original image is calculated, often using a distance metric like cosine distance. The more a synthetic image resembles the original, the higher its weight and importance in the explanation process.

The black-box model is then used to make predictions on these synthetic images, allowing us to estimate the importance of different regions in the original image about the model's predictions.

This important information is frequently represented in the form of a heatmap, as in the methods presented above. The heatmap showcases which regions of the image have the most significant influence on the model's decision, highlighting these areas in the original image.

As a reference, you can see an overview of RISE and a visual representation of how RISE works, in Figure 2.8.

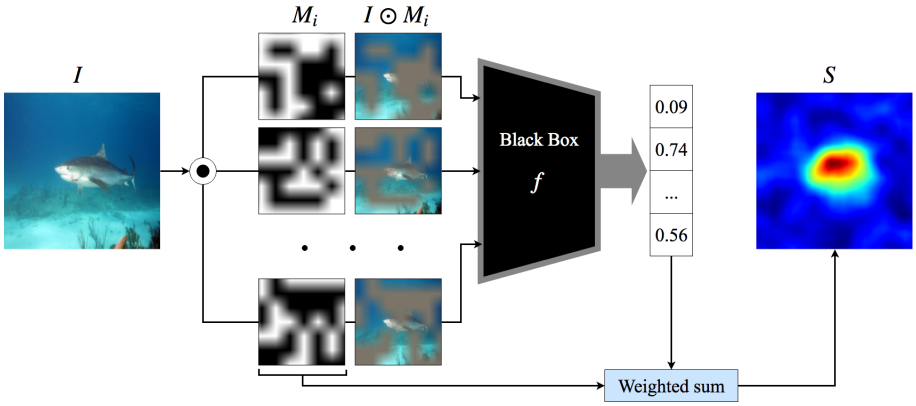

Figure 2.8: RISE overview. Extracted from [9].

CHAPTER 3

State of the Art

This chapter aims to carry on a systematic review of the literature published that shows improvements in using XAI techniques in the healthcare field. Specifically, it works towards the identification of the purpose and context of the use of interpretability and explainability in ML in medical applications, considering the main findings and thus reviewing and discussing the applied methodological strategies. For a comprehensive review of the existing literature on the topics described, there must be an introduction of recurrent concepts to all strategies reviewed, with an exploration of the purpose of the existence of the tools mentioned and their advantages, opposing with a depiction of the challenges that come with the creation of explanations for certain models.
The aim is then shifted towards state-of-the-art explainability methods. There are some types of explainability, such as intrinsic methods (that can restrict the complexity of the model before the training), or post-hoc methods (applied after the model training). These methods can then also fall into some other categories, model-agnostic and model-specific (limited to specific classes of the model). Model-agnostic methods can be employed in any kind of ML model, and this category can then be divided into global (explain the whole model behaviour) and local (explain an individual prediction) explanations.
The chapter concludes with a succinct mention of other explainability methods, followed by a compendium of the techniques reviewed and a final discussion of the reviewed literature.

3.1 Purpose of the Review

This chapter aims to systematically review the literature that employed Explainable AI to enhance the understanding of the reasons behind a decision made by a specific model given a scenario, catching on to the findings that are of particular relevance and reviewing the methodological approaches, stoking knowledge about the subject and equipping them with knowledge of the strategies used today.

A systematic literature review of IEEExplore and ScienceDirect indexed databases has been followed through. Given the low search results, other search tools were used to broaden the analysis range of my review, such as the Google Scholar search engine. Given the selected papers to review, there was scrutinized matter and statistical analysis.

The present chapter is an analysis of the current state-of-the-art solutions for applying Explainable AI in medical applications to gain trust and eliminate doubts when in a diagnosis of a certain condition. The collection of articles identified for screening was published from 2016 to 2022 and focused on the

use of XAI in the healthcare field. After screening and obtaining the final eligible articles, a study was conducted on the following research questions.

- **RQ1**: How can I compare current XAI methods?
- **RQ2**: What is the importance and utility of XAI in healthcare applications?
- **RQ3**: What are the challenges that come with the application of XAI models?
- **RQ4**: How can I evaluate the current XAI methods and which are the perfect fit for each case?

Although there have been an increasing number of studies on this content in recent years, they are still fairly small, remaining largely unknown to the scholar community, hence this chapter aims to systematically review the literature that used XAI to create a better understanding of an AI diagnose on a certain condition, catching on the findings that are of particular relevance and reviewing the methodological approaches. A systematic review of the literature has benefits for all parties involved in academic research, by stoking cognizance about the subject and equipping them with knowledge of the strategies used nowadays, easing a better understanding of the strengths and weaknesses behind the use of XAI in the healthcare field.

3.2 Methodology of Research

According to the findings of Tranfield et al. [10], narrative reviews show evident weaknesses compared to other forms of review. Given this, the present study adopted an "evidence-informed, systematic literature review approach" [11], following the five-step process proposed by Denyer and Tranfield [12] that includes a pilot search as the first phase to gain more awareness of the current literature, and then collect the criteria that help in the selection of the article. The subsequent steps come from applying the criteria to the databases chosen for research. For a general overview of the sections presented in this document, the adopted strategy is depicted below.

- First Phase: **Pilot search** - Formulate the Research Questions (RQs) for the study and get an overview of the current literature.
- Second Phase: **Location of studies** - Selection of the search engine(s) and databases as well as the keywords and search strings that offer relevant articles to review.
- Third Phase: **Study selection and evaluation** - Definition of the inclusion and exclusion criteria for the articles.
- Fourth Phase: **Analysis and synthesis** - Description of a set of characteristics to answer the research questions, including the results of this study in the form of tabulations and statistics.
- Fifth Phase: **Discussion** - Phase where the focus is to answer the RQs.

3.2.1 Article Screening

The literature gathering of three databases, EBSCOhost, ScienceDirect and IEEExplore, was done on 28 September 2022. A ‘snowball’ search was also carried out to identify additional studies by searching the reference lists of previously considered eligible studies and using Google Scholar to widen and improve the research range, on 30 September 2022. These sources were selected due to being among the largest and most popular abstract and citation databases of peer-reviewed publications and widely used for literature reviews [13][14]. In total, 883 articles were in the pilot search.

The search strategy development process started with the identification of candidate search terms by analysing the titles, abstracts and subject indexing of the database records. A draft search strategy was then developed using those terms and additional search terms were identified from the results of that strategy. The search strings on the title, abstract and keywords had to follow the pattern: ("explainability" OR "explainable ai" OR "explainable") AND ("healthcare" OR "medical applications")

AND ("artificial intelligence" OR "machine learning "). As per the identification and limits criteria, the strategy was limited to English language studies, in electronic format and published in the last six years. This period was chosen because not only DARPA launched the XAI initiative in 2016 but also in terms of literature available on this specific topic when it is considered the appearance and usage of XAI in machine learning algorithms.

To filter the already collected articles, a decision was made on complementary parameters to ensure accuracy and credibility. Duplicate records were treated and in some cases considered a continuation of the study. The number of citations and peer reviews would be taken as evidence of trust. In subsection 3.2.2 is described the chosen final criteria/characteristics the chosen articles must follow.

The selection process of the articles to be reviewed in the present document is presented in Figure 3.1, based on the Preferred Reporting Items on Systematic Reviews and Meta-analysis (PRISMA) statement [15], which guides reporting for systematic reviews, helping to identify, select, appraise, and synthesise studies.

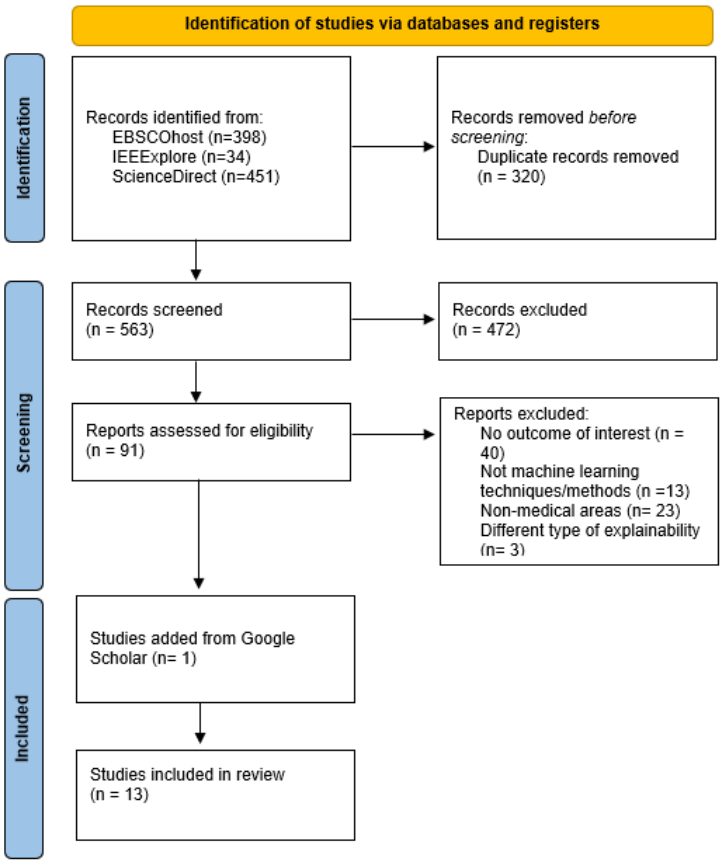

Figure 3.1: PRISMA flow diagram for the selection process.

3.2.2 Data Auditing

Following the step of selecting the articles to be reviewed, it is important to group a set of characteristics to answer the formulated RQs. This phase is vital to ensure that the information from every article is well distributed between categories to make a correct comparison and get the desired answers.

These chosen characteristics are as follows:

- The type of the study (i.e. literature review, proposed method, case study);

- The respective purpose of the study;
- The XAI technique(s) used;
- The outcomes and primary findings;
- Limitations of the study;
- Novelty of the study;
- Number of citations.

In summary, it was found 883 records in database searching. After duplicate removal, the screening process included 563 records, from which were assessed for eligibility 91 full-text documents. After reviewing for specific criteria of exclusion, 13 final papers were included in the review, each cited in Table 3.2. Later, a search was conducted of documents that cited any of the initially included studies as well as the references. However, no extra articles that fulfilled the inclusion criteria were found in these searches.

79 studies were excluded from the review, given a list of exclusion topics. The summarization of the exclusion criteria of the following publication types is listed below.

- Articles not containing a peer-review process;
- Methods developed only for enhancing model transparency with feature selection and better data visualization but not directly focused on explanation.
- Methods developed for other areas of interest rather than medical applications, medical diagnosis or in a healthcare environment;
- Studies using or describing explainability in different contexts other than AI and Computer Science, such as Psychology;

3.3 Analysis and Findings

After collecting the relevant papers for this research, it begins the fourth phase of the process as stated in subsection 3.2.1, the data analysis and synthesis. While the focus of the analysis is to break down each study into its relevant parts and describe relationships and connections between the collection, synthesis aims to identify the associations between parts of different studies, based on the author of [10]. This part of the study is represented through the following subsections. Figure 3.2 presents the categories of the selected articles while Figure 3.3 shows the year of publication of those articles.

3.3.1 Characterization of Selected Articles

After a thorough analysis of all the selected articles, given the complexity of the problem, there was an attempt to categorise the included studies along four dimensions, listed as follows:

- Reviews on explainability methods - This includes literature and/or systematic reviews of methods that include proposals and/or tests of solutions that apply to the explainability of data and knowledge-driven models;
- Development and evaluation of methods - This category includes articles that propose novelty methods that are devoted to the enhancement of the data explainability/knowledge-driven models;
- Evaluation of methods - It includes articles that focus on the results of scientific studies and that are dedicated to the performance evaluation of those said methods for explainability.
- Case studies - Includes all research with a detailed examination of a particular case (or cases) that explores the use of explainability within a real-world context.

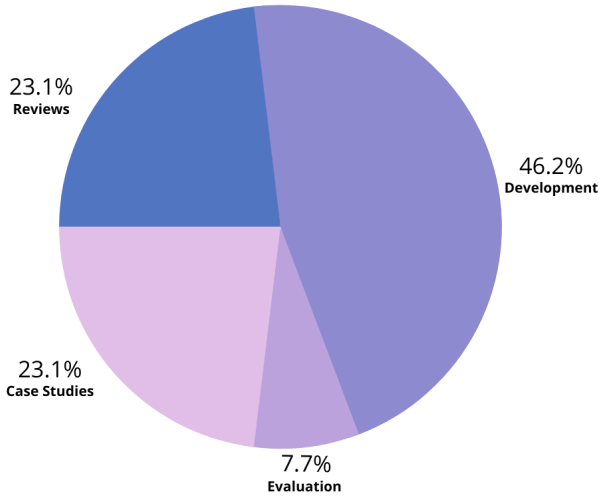

Figure 3.2: Categories pie chart of selected articles.

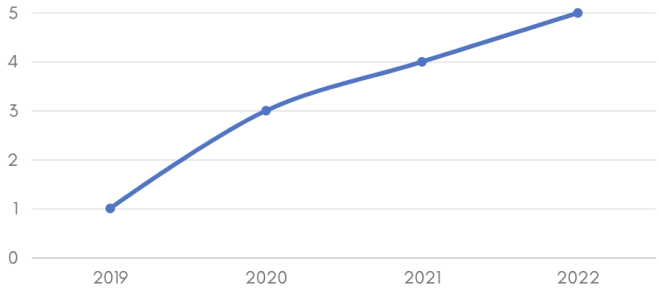

Figure 3.3: Years of publication line chart of selected articles.

As observed in Table 3.1, I developed a 'Characterization of the selected articles' that comprise summaries of the reviewed studies, given the year of publication, academic journal and citations.

Table 3.1: Characterization of the selected articles.

Year of Publication	Articles	Academic Journal	Google Scholar Citations
2019	[16]	Artificial Intelligence in Medicine	201
2020	[17]	IEEE Access	30
	[18]	Nature Communications	150
	[19]	arXiv	35
2021	[20]	IEEE Access	4
	[21]	Cancers 2021	8
	[22]	Sensors 2021	9
	[23]	The Lancet Digital Health	168
2022	[24]	IEEE Journal of Translational Engineering in Health and Medicine	22
	[25]	Procedia Computer Science	0
	[26]	Computer Vision in Co-clinical Medical Imaging for Precision Medicine	4
	[27]	IEEE Access	2
	[28]	IEEE Reviews in Biomedical Engineering	6

3.3.2 Co-occurrence Analysis

Co-occurrence analysis is the counting of paired data within a collection unit. This is a useful tool in this type of research, since there is an interest in analysing networks of, for example, documents, keywords, authors or journals [29]. Figure 3.4 visualizes the co-occurrences of the key vocabularies of XAI/AI concepts, XAI methodologies, and types of conditions diagnosed from my reviewed papers. The word clusters are represented by different colours, with the bubble size denoting the number of

publications, while the connection width reflects the frequency of co-occurrence. Due to the length of the full-text documents, it was only considered the title and abstracts of the document in this analysis.

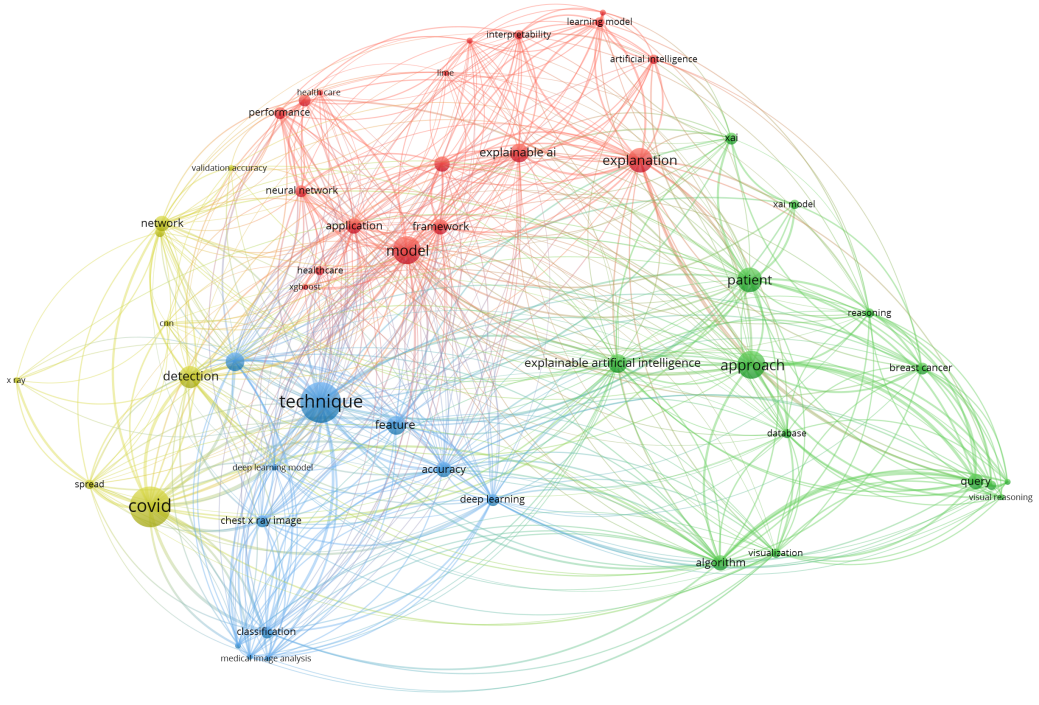

Figure 3.4: Co-occurrence network of the commonly used words in reviewed studies.

The current study provides an overview of the applications of XAI techniques in medical diagnosis applications. I focused on image classification approaches.

In Table 3.2 I present a summary table of the reviewed methods. Each method is described according to the methodology/taxonomy and key contributions of the XAI model.

Table 3.2: Summary of included articles by authors, year, methods, techniques or taxonomies and key contributions.

Author	Title	Year	Method/Taxonomy	Key Contributions
Lamy et al. [16]	Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach	2019	CBR	Proposes a quantitative and qualitative method with a visual interface and applies it to breast cancer management datasets, with a user study
Ibrahim et al. [17]	Explainable Prediction of Acute Myocardial Infarction Using Machine Learning and Shapley Values	2020	SHAP, XGBoost	Presents a detection framework of acute myocardial infarction with two deep learning models, a CNN and an RNN, and uses a decision-tree-based model (XGBoost) with Shapley values for feature relevance identification
Lauritsen et al.[18]	Explainable artificial intelligence model to predict acute critical illness from electronic health records	2020	LRP, DTD	Present an explainable AI system for the prediction of acute critical illness using EHRs
Dave et al. [19]	Explainable AI meets Healthcare: A Study on Heart Disease Dataset	2020	LIME, SHAP, CEM, Feature-Based Techniques, Example-Based Techniques	Reviews different interpretability techniques in healthcare with examples based on a heart disease dataset
Ren et al. [20]	Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models With Multisource Data	2021	Grad-CAMs	Proposes an approach to combine neural networks with an explainable model (Bayesian Network)
Chakraborty et al. [21]	Explainable Artificial Intelligence Reveals Novel Insight into Tumor Microenvironment Conditions Linked with Better Prognosis in Patients with Breast Cancer	2021	SHAP, XGBoost	Develops XAI models to establish and investigate the data-driven relationship between critical tumour microenvironment features
Sousa, Vellasco and Silva [22]	Explainable Artificial Intelligence for Bias Detection in COVID CT-Scan Classifiers	2021	Grad-CAMs, LIME, RISE, Squaregrid, direct Gradient approaches (Vanilla, Smooth, Integrated)	Reviews state-of-the-art classifications, with the application of several XAI techniques for comparison and bias evaluation purposes
Ghassemi, Oakden-Rayner and Beam [23]	The false hope of current approaches to explainable artificial intelligence in health car	2021	Saliency Maps, LIME, SHAP, LRP, Grad-CAMs, direct Gradient approaches (Vanilla, Smooth, Integrated), GuidedBP	Presents the state-of-the-art approaches and limitations of explainable AI in healthcare
Mondal et al. [24]	xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography	2022	Saliency Maps, Gradient Attention Rollout	Proposes the use of vision transformers with explainable features for COVID-19 screening, that focuses on meaningful regions in the images
Vishwarupe et al. [25]	Explainable AI and Interpretable Machine Learning: A Case Study in Perspective	2022	ELI5 XAI,LIME,SHAP, PDP	Presents a case study based on diabetes detection using ELI5 XAI toolkit and other XAI techniques
Khan et al. [26]	COVID-19 Classification from Chest X-Ray Images: A Framework of Deep Explain	2022	Grad-CAMs	Proposes a feature optimization algorithm with Grad-CAM-based visualization for an improved feature selection process
Saraswat et al. [27]	Explainable AI for Healthcare 5.0: Opportunities and Challenges	2022	Grad-CAMs, Dimension Reduction, Feature Extraction, Knowledge Refining, Proxy Representation, Attention Mechanism	Reviews methods based on taxonomy and proposes a new solution architecture for classification of COVID-19 patients in the healthcare 5.0 environment
Giusti et al. [28]	Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review	2022	Data Augmentation, Outcome Prediction, Unsupervised Clustering, Image Segmentation	Reviews XAI methods focused on the evaluation of results and present a practice guideline and potential solutions to some of the present challenges

The research question **RQ1** is discussed in this section as I present some of the current strategies to differentiate and compare XAI methods.

According to Christoph Molnar [30], there is a wide variety of criteria to classify methods for ML interpretability.

With the amount of XAI methods growing in quantity, there is an increased need for a taxonomy of methods. The right taxonomy helps to compare state-of-the-art methods and therefore to select the right method based on traits required by a specific use-case context. That is how different explainability taxonomies appear. XAI methods can be described and compared using the different explainability taxonomies currently available.

One example appears from the authors Deepti Saraswat et al. in [27] that propose a contribution for a solution taxonomy of XAI, with a particular use-case taxonomy for healthcare applications, integrating different AI techniques with XAI.

Machine learning intrinsic methods provide explanations directly as part of the model's output, as they are built into the model itself. A model's internal workings and decision-making process can be explained using these methods during the training phase.

In contrast, post-hoc methods provide explanations based on the output of a model after it has been trained. Users can use these methods to understand the factors that contributed to the model's predictions and to provide a more detailed explanation of the model's decisions.

Explainability methods can also be classified according to their type of explanation. For example, some methods may provide global explanations that describe the overall behaviour of a model. Conversely, others may provide local explanations that focus on specific predictions or decisions made by the model. Several methods provide quantitative or textual explanations, while others offer visual explanations, such as heat maps or decision trees.

3.4.1 Intrisic versus Post-hoc Methods

In terms of definitions, Intrisic methods define XAI modules that are applied during the training of the model. On the other hand, Post-hoc methods comprise the XAI methods that are applied after the model was trained. In essence, the key difference lies in the timing and focus of analysis.

A Bayesian network is a type of probabilistic graphical model that represents the probabilistic dependencies between a set of variables. It consists of a directed acyclic graph (DAG) with nodes representing the variables and edges representing the probabilistic dependencies between the variables. Bayesian networks can be considered an intrinsic explanation method, as they provide an explicit representation of the probabilistic dependencies between the variables and can be used to understand the decision-making process of the model as it is being trained, as stated by H. Ren et al. in [20].

The process of post-hoc explainability, also known as Post-model explainability, shifts the focus to an attempt to dissect the procedure behind the model's decision-making. One of the most used forms of this method is heat maps. As stated by M. Ghassemi et al. in [23], "Heat maps (or saliency maps) highlight how much each region of the image contributed to a given decision" and therefore give an overview of the limitations behind the use of post-hoc explanations, as there is a possibility that the hottest parts of the map may contain both useful and non-useful information for the medical experts and not reveal exactly what the model considered useful.

The authors M. Ghassemi et al. of [23] also state that when using post-hoc explanations, in reality, there is an addition of another source of error, since "not only can the model be right or wrong, but so can the explanation".

3.4.2 Results of the Interpretation Methods

The various interpretation methods can be broadly classified based on the type of explanations they provide.[30]

- Feature summary statistic: Provide statistics for each feature (feature importance, feature interaction). SHapley Additive exPlanations (SHAP) values, as stated by Lundberg and Lee in [31], are an example of this, where feature interaction metrics are given.
- Feature summary visualization: Most of the feature summary statistics can also be visualized and present advantages to data on a table (partial dependence of a feature). Partial Dependence Plots (PDPs), as shown by Varad Vishwarupe et al. in [25], show how a feature impacts the model's predictions while keeping other features constant.
- Model internals: Another way to classify intrinsically interpretable models (learned weights). In neural networks, interpreting the importance of individual neurons or connections can offer insights into how the model makes decisions.
- Data point: Includes all methods that return data points to make a model interpretable. Counterfactual explanations, also considered model-agnostic, can also be categorised into this.
- Intrinsically interpretable model: The model itself is interpreted by looking at internal model parameters or feature summary statistics. Decision trees or linear models are often considered intrinsically interpretable due to their transparent decision-making processes.

3.4.3 Model-Specific versus Model-Agnostic Methods

As specified by Christoph Molnar [30], another important criterion to take into consideration when classifying explainability methods is the "agnosticity criterion", i.e., a method can be model-specific or model-agnostic.

Model-specific interpretation tools are designed to work with specific classes of machine learning models, such as linear models or neural networks. These tools are typically used to interpret the internal workings of the model and how it processes data. In contrast, model-agnostic interpretation tools can be used with any machine learning model and are applied after the model has been trained. These tools usually work by analyzing the input and output of the model and do not have access to the internal structure or parameters of the model. Therefore, model-agnostic interpretation tools are less powerful than model-specific interpretation tools but can be more widely applicable.

A type of model-specific method is the Layer-wise Relevance Propagation (LRP). LRP, as stated by Bach, Sebastian et al. in [32] belongs to the category of model-specific interpretability methods, as it is a technique that is specifically designed for explaining the predictions made by artificial neural networks, like deep-learning models. Authors Simon Meyer Lauritsen et al. in [18] apply LRP in a Temporal Convolutional Network (TCN) model instead of Recurrent Neural Network (RNN) with attention. They use LRP analysis to discover new and unknown correlations. The LRP decomposes an explanation into simpler local updates and "ensures that the total back-propagated relevance amounts to the extent to which the illness of interest is detected by the function."

Two of the most used model-agnostic methods are LIME, originally proposed by Ribeiro et al. in [8], and SHAP, proposed by Lundberg and Lee in [31].

Ghassemi et al. [23] expose the problems and concerns behind some model-agnostic methods such as LIME and SHAP, defending that they are "generic and not specific to images" despite being the routine application on health-care data, even in data from electronic health-care records. The main difference between these two methods resides in the ability of SHAP being able to explain both globally and locally, while LIME is only capable of local explanations as stated by Devam Dave et al. [19].

An example of the use of SHAP values in this environment is presented in [18], as they adopt it to visualize explanations of global parameter importance and local explanation summary. In this adaptation of SHAP, about the global parameter importance, the higher the confidence about a decision, the higher the probability and larger relevance scores. But when the model has a contrary response, the output will display low probability and relevance scores. On the other hand, the local summary distribution allows the healthcare stakeholders to get an overview of the expectations from the model.

Other authors have applied LIME and SHAP values to their models to compare behaviour with other types of state-of-art methods. Authors in Lauritsen et al. [18], Dave et al., [19] and Lundberg and Lee [31] conclude that applying a similar technique called Kernel Shapley, introduced by Lundberg and Lee [31]. This presents a quicker return of results and is mathematically more efficient.

Anchors, another type of model-agnostic approach, generate a local region that provides a more accurate representation of the data that is being explained, also proposed by the authors of LIME by M. Ribeiro et al. in [33]. This approach is also a better solution than the two methods described above, generalizing better than LIME and having a smaller computation cost than SHAP. The disadvantage of anchors is that they may provide a set of rules that are difficult to interpret, especially if they include a large number of feature predicates. In addition, if the anchors are too specific, the coverage area for explaining observations may be significantly reduced.[19]

Counterfactual explanations, as seen by S. Wachter et al. in [34], are a model-agnostic XAI technique that works by identifying what changes I can make to the input to get a certain result. This method can also be applied to classification datasets with more than two target classes, but may not perform as well as they do on binary classification datasets. A faster and more accurate evolution of this technique is counterfactuals guided by prototypes. This makes the search process significantly faster by directing the counterfactual to the prototype of a particular class, later reviewed in [19].

3.4.4 Global versus Local Methods

Explainability methods can also be classified according to their type of explanation. For example, some methods may provide global explanations that describe the overall behaviour of a model. On the contrary, others may provide local explanations that focus on specific predictions or decisions made by the model. Several methods provide quantitative or textual explanations, while others offer visual explanations, such as heat maps or decision trees.

A local interpretation is an explanation that focuses on the decision made by the model for a specific input or a small subset of inputs. Local interpretations are useful for understanding how the model uses specific features of the input data to make a prediction and can help debug and analyse the performance of the model.

A global interpretation is an explanation that captures the overall behaviour of the model across all inputs. Global interpretations can provide a more comprehensive understanding of how the model works and how it makes decisions but may be less helpful for understanding the specific decision-making process for individual inputs.

When talking about the difference between global and local explainability methods, Christoph Molnar in [30] asks two important questions that lead to the discussion.

In terms of global interpretability, "How does the trained model make predictions?" is what lets me understand how this level of interpretability is difficult to achieve. The relevant features, how are they connected and interact with each other, therefore all about understanding how the model makes decisions. In summary, global interpretability helps "to understand the distribution of your target outcome based on the features." [30].

Authors Ghassemi et al. [23] defend that global descriptions of the functionality of a model are more realistic than using local explanations as a way to produce and justify model predictions.

On local interpretability, the question that is asked is "Why did the model make a certain prediction for an instance?".

Locally, these explanations can be more accurate than global explanations as the way predictions of the model have a chance of only depending in a linear or monotonic way on certain features.

Works like Devam Dave et al. [19] reviewed various types of methods that produce local explanations, such as Contrastive Explanation Method (CEM). CEM is a technique for providing explanations for classification models by identifying the features that are most important for maintaining the original prediction class and the features that are necessary to differentiate the prediction class from the nearest different class. CEM is the first method to provide explanations for both the features that should be minimally present and the features that should be necessarily absent to maintain the original prediction class. While counterfactual explanations identify the changes that should be made to the input features to produce a predefined output, CEM does the opposite and identifies the features that should be present to maintain the original prediction class.

Integrated Gradients is a local explainability method that is used to explain individual predictions by identifying the specific contributions of the input features to the final prediction, showing which features had positive and negative attributions. It is a widely used method that is often used to identify where a machine learning model is making mistakes so that improvements can be made to increase its accuracy. This method can be useful for debugging and can also make the model more transparent for users. [19] [23]

3.4.5 Comparison

The taxonomy that dedicates to the division of methods as global or local considers relevant criteria to take into account when categorising XAI methods but it fails to distinguish methods like Gradients from Shapley Values in works like Ghassemi et al. [23], since the first method is local and model-specific, while the second is also local, but model-agnostic. With this, I can conclude that this taxonomy alone is not enough to well categorise XAI methods.

Another example that a singular taxonomy cannot fit all models is the case of PDPs [25], which are model-agnostic and global, while Counterfactual Explanations (Devam Dave et al.) [19] are also model-agnostic, but local, showing that the agnosticity criterion that I described before is also not suitable to classify all explainability methods (Rio-Torto et al.) [35].

In terms of the taxonomy based on the results of the interpretation, detailed in Subsection 3.4.2, this divides methods according to their output, which I can consider be a broad classification since it fails to separate certain explainability methods like PDPs and Shapley Values in works like Varad Vishwarupe et al. [25]. Although they are global and local methods, respectively, this taxonomy considers them both to fall into the same category, as feature summary methods.

3.5 XAI Applications and Challenges in Healthcare

This section focuses on answering the research questions **RQ2** and **RQ3** as it presents the applications and challenges in integrating XAI in healthcare. In the healthcare industry, AI has been widely adopted to improve analytics and prediction models, identify anomalies, and detect diagnostic patterns. AI is used for tasks such as image classification, segmentation, and disease prediction. However, AI decisions in healthcare are critical and require explainability to ensure transparency and traceability in clinical output. XAI techniques, such as Bayesian teaching (H Ren et al.) [20] and

saliency maps (Marzyeh Ghassemi et al.)[23], (Christoph Molnar)[24], can provide insight into how the AI model arrived at its prediction and highlight important features that contribute to accurate predictions in the medical domain. XAI is particularly useful for deep learning models used in imaging analysis, such as tumour segmentation, where data collection, labelling, and augmentation are critical [16] [21] [27]. Works like L. Ibrahim et al. [17] and Devam Dave et al. [19] reveal and reassert the importance of implementing more XAI solutions to systems that detect Cardiovascular diseases, considered the number one cause of death globally.

In terms of global pandemics, COVID-19 fast detection has been a priority in recent years and led to most of the state-of-the-art AI solutions available nowadays. Authors of Giuste et al. [28] discuss the XAI utility during COVID-19 through a systematic review of literature employing AI for COVID-19 detection and risk assessment. Publications like Sousa et al. [22], Mondal et al. [24], Khan et al. [26] take on the feature importance and relevance when it comes to multiclass classification such as COVID-19, viral pneumonia, lung opacity, and normal images, given the similarity among each image being very high, and leading to a chance of misleading the correct classification accuracy.

On another topic, integrating EHR in AI has been used to collect information from patient data and reduce the administrative workload of healthcare workers. Authors of [18] present a contribution to this by proposing an AI model with visual explanations for predicting acute critical illness from EHR. This helps clinicians focus on which relevant EHR data the prediction is supported by.

After concluding the discussion of the relevance and applications of explainability, I have to address the challenges that come with this.

In Ghassemi et al. [23], the authors defend that despite the current explanation modules being unable to provide a reliable evaluation of AI models, the focus should shift to advocating for robust and unbiased validation of these systems. They consider that explainability methods should be a tool for developers of the systems (to evolve them and reduce their bias) rather than serve the final stakeholders (the medical community). The main challenge that is pointed out is the rare testing of the performance of explanations, which leads to an added source of error.

Saraswat et al. reveal in [27] that one of the main current challenges for XAI resides in the human-machine interaction in a way that is necessary to design better feedback mechanisms into the explainability modules. Combining different types of social and human behaviour studies will lead to an enhancement of this interaction as it incorporates ethics, transparency and morality. In terms of data availability, sharing, and security, there will be a need to create more robust XAI solutions as the number of collaborators will increase exponentially in a medical environment, and it is a priority that the privacy of the data that are handled remains.

Authors Giuste et al. in [28] present some of the common challenges XAI techniques face and potential solutions for them. One of them is the limited availability or imbalance of data, like in the first stage of the pandemic of COVID-19, where there was much more data for normal samples rather than COVID-19-positive samples. The solution to that challenge may reside in data augmentation of the minority classes. The lack of annotated images and redundant data is another typical challenge that leads to weaker models, and it can be solved using feature extraction, which transforms raw data from the initial input into numerical features that can be more manageable.

The creation of properties to correctly evaluate XAI methods is also taken into consideration as one of the challenges, because it is not clear yet how to formalise how they could be calculated, and works like Christoph Molnar [30] and Giuste et al. [28] defend this need to study further into this. The next section will present some of the current approaches to this challenge.

3.5.1 Evaluation of XAI Methods

The research question **RQ4** is discussed in this section as I present some of the current evaluation measurements for XAI methods. After being discussed what are some of the current XAI approaches, it is essential to talk about how to measure and evaluate those methods, to be able to choose which is the better fit for each case. One of the first attempts at creating an evaluation measurement was in 2017 when Doshi-Velez and Kim proposed three main levels for the evaluation of interpretability, enumerated below. [36]

1. Application-level evaluation (real task): Focus on evaluating the interpretability of an AI system in a specific application. For example, by conducting user studies to assess how well users can understand and use the AI system in a real-world setting.
2. Human-level evaluation (simple task): It is a simpler version of the application-level evaluation, which focuses on evaluating the interpretability of an AI system from the perspective of human users. This might involve conducting user studies to assess how well people can use the AI system and understand how it works.
3. Function-level evaluation (proxy task): Focus on evaluating the interpretability of an AI system in terms of its function or purpose. These approaches might require analyzing the different inputs and relationships between the input data and the output decisions of the AI system.

The authors Giuste et al. of [28] propose a guideline based on the key criteria of Doshi-Velez and Kim [36] for evaluating XAI methods from two perspectives, model behaviour, and human understanding. They consider that in terms of how to evaluate the generation of explanations and the model behaviour itself, the accuracy, completeness, and robustness of the model have to be taken into consideration. In terms of how to evaluate explanation representations, the important characteristics to consider are clarity, generalizability and simplicity. The described guideline properties are defended by Molnar in his works [30], as these can be used to judge how good an explanation method or explanation is.

3.6 Conclusion

I can now consider, after studying the benefits of applying explainability techniques that they provide insights into how the different features of a model contribute to its outcomes. These approaches help me to understand the decision-making process of a black box model and explain its behaviour. [19]

This systematic review also demonstrates the need for the involvement of medical stakeholders in the development process of any type of ML, Deep Learning (DL) or XAI methods for the medical domains.

After a comprehensive analysis of Table 3.2, I can observe that the majority of analysed explainability methods are post-hoc methods and that there is a clear evolution of the adoption of some of the techniques, like Grad-CAM and LIME. As a result of the analysis of the articles presented in the table, I will apply these two XAI techniques in the context of my practical work and compare them to the baselines through XAI evaluation techniques.

In terms of explainability taxonomies, there is a variety of taxonomies available to describe the reviewed methods and there is not a "one fits all" solution taxonomy when it comes to differentiating the state-of-the-art techniques available. In the present paper, four taxonomies were reviewed, where the differences between the methods were described and presented with examples for each of the subcategories.

I suggest using a combination of two taxonomies to better classify the methods reviewed. The first taxonomy separates methods into those specific to a particular model and those not related to a specific

model (Subsection 3.4.3). The latter group is further divided into global and local methods (Subsection 3.4.4). The second taxonomy focuses on the results of the interpretation, which differentiates how the outputs can classify a method (Subsection 3.4.2).

The same applies to the evaluation of methods, I consider that there is still a long way to go in terms of creating a correct guideline so it is a priority to carefully consider what are the goals and requirements of the XAI system when selecting and evaluating these types of methods.

I suggest using the framework established by Doshi-Velez and Kim in [36], as it has been widely cited in the research community and has influenced the development of other evaluation frameworks for XAI. The main advantage is that it provides a clear and structured way of comparing different XAI methods, applied in several empirical studies to evaluate their effectiveness.

To summarize, the rapid use of AI solutions in critical systems like healthcare decision-making applications creates the need for trustworthy and transparent explanations of how AI-based decisions are reached. Explanation modules can help clinicians and developers increase the confidence of a model in a certain decision by elucidating previously hidden patterns in the data and also reducing potential biases. I find that the application of XAI in AI-based solutions is an essential step in understanding black-box decision-making for the healthcare industry and overall improving patient care quality. The present systematic review highlights through different taxonomies the current state-of-the-art XAI approaches in healthcare applications and their potential challenges.

CHAPTER 4

Implementation

In this chapter, we delve into the practical implementation of our medical image classification system. We will provide a detailed account of the methods, tools, and technologies used to build and deploy the system. As we navigate through the implementation process, we aim to demystify the intricate workings of our AI solution, offering insights and practical guidance.
In this chapter, we address data collection and pre-processing, the foundation upon which our AI system is built. We then delve into the architecture of our deep learning model, outlining the choices made and the rationale behind them. Throughout the implementation, we encountered and tackled various challenges, such as class imbalance, overfitting, and the computational demands of training complex models. The main focus is on integrating the XAI techniques into our baseline model and trying to elevate the understanding of the decisions it makes. Some of the methods applied include LIME, Grad-CAM, and RISE.

> *“The best programs are written so that computing machines can perform them quickly and so that human beings can understand them clearly.”*
>
> *Donald Knuth,* Selected Papers on Computer Science[37]*, 2003*

4.1 Data Collection and Preprocessing

The quality and integrity of the data used to train and evaluate models are at the heart of any machine learning research. This section goes into the critical process of data collection and pre-processing, providing a detailed overview of how the chest X-ray dataset was obtained, cleaned, and prepared for analysis. The careful management of data, including ethical issues and approaches for dealing with class imbalances, is the foundation of our research, ensuring the dependability and robustness of our machine learning models.

4.1.1 Data Source Selection

The chest X-ray dataset used in this research was obtained from a publicly available source. It is the ChestX-ray14, the largest publicly available chest X-ray dataset. This dataset comprises X-ray images and associated metadata, including labels for various pathological conditions. It was elaborated to be used with the CheXNet model [4]. The dataset, released by the National Institutes of Health (NIH),

contains in total 112,120 frontal view X-ray images of 30,805 unique patients, which are annotated with up to 14 different thoracic pathology labels using Natural Language Processing (NLP) methods in radiology reports. The original paper [4] obtained a test set of 420 frontal chest X-rays in which annotations were obtained by practising radiologists at Stanford University. These practitioners were asked to label the 14 pathologies. In terms of ground truth, they evaluated the performance of an individual radiologist by using the majority vote of the other 3 radiologists.

4.1.2 Dataset Description and Splitting

The dataset consists of a total of 415 prior chest radiograph images from the data source mentioned in 4.1.1. These images were stored in a specific format (JPEG). The **ImageDataGenerator** is responsible for the dataset's splitting and is detailed in Table 4.1.

Dataset Split	Number of Images
Training Set	226
Validation Set	65
Test Set	124

Table 4.1: Dataset composition.

4.1.3 Data Preprocessing

Data pre-processing plays a fundamental role in ensuring the integrity of data, serving as an essential step in the realm of building operational data analysis. Given the inherent intricacies of building operations and inherent data quality issues, data pre-processing encompasses a range of methods aimed at improving the raw data's quality. These methods encompass tasks like the identification and removal of outliers, as well as the imputation of missing values [38].

Before analysis, the raw dataset was subjected to several pre-processing tasks, presented in Figure 4.1 which we will describe in more detail in the following subsections.

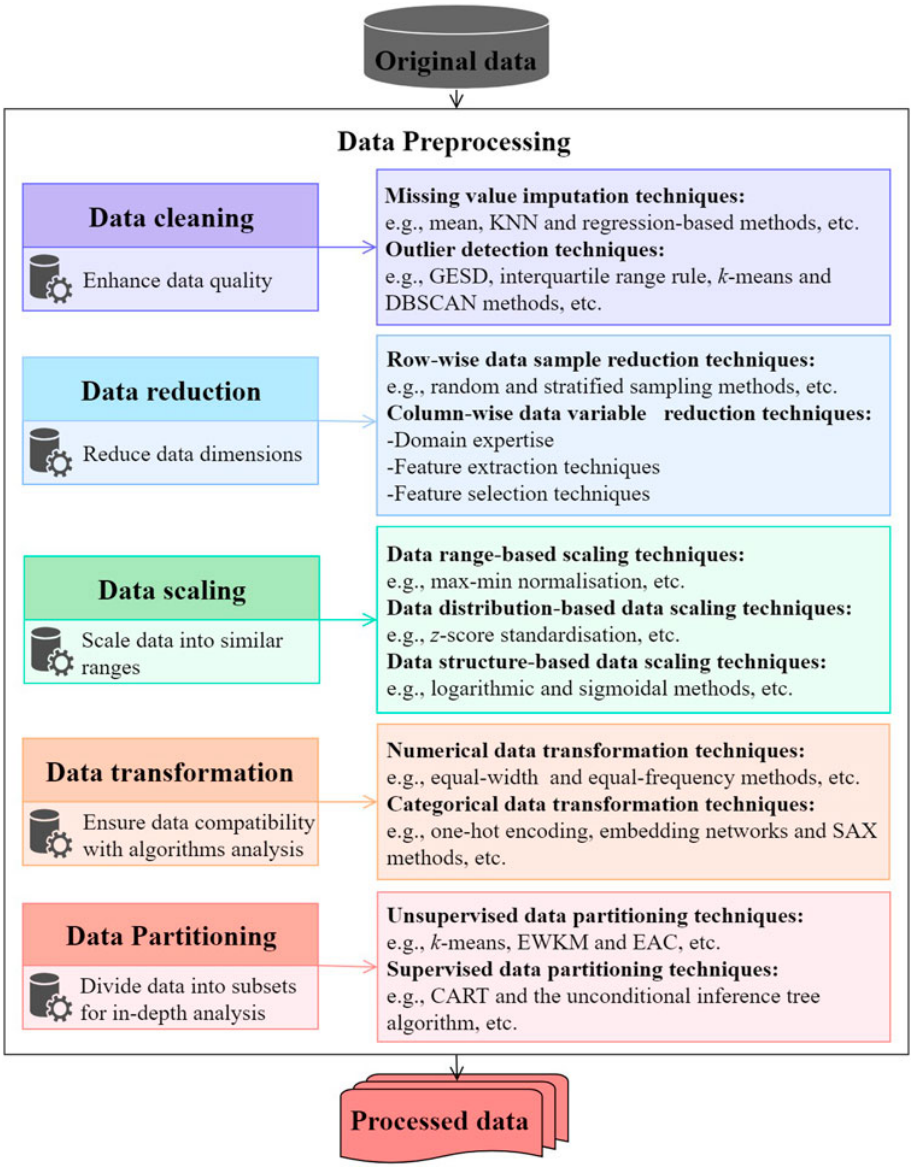

Figure 4.1: Typical tasks for data pre-processing when performing data analysis. Extracted from [38].

Data Exploration

In the beginning, the dataset was explored to understand its structure and characteristics. Descriptive statistics and visualizations were used to gain insights into the dataset's distribution of labels and the presence of data imbalances.

Data Label Analysis

The dataset contained 14 different pathological conditions, and the distribution of positive (1) and negative (0) labels for each condition was analysed. Class imbalance was observed when plotting the class distribution, as seen in Figure 4.2, indicating that the dataset is not balanced.

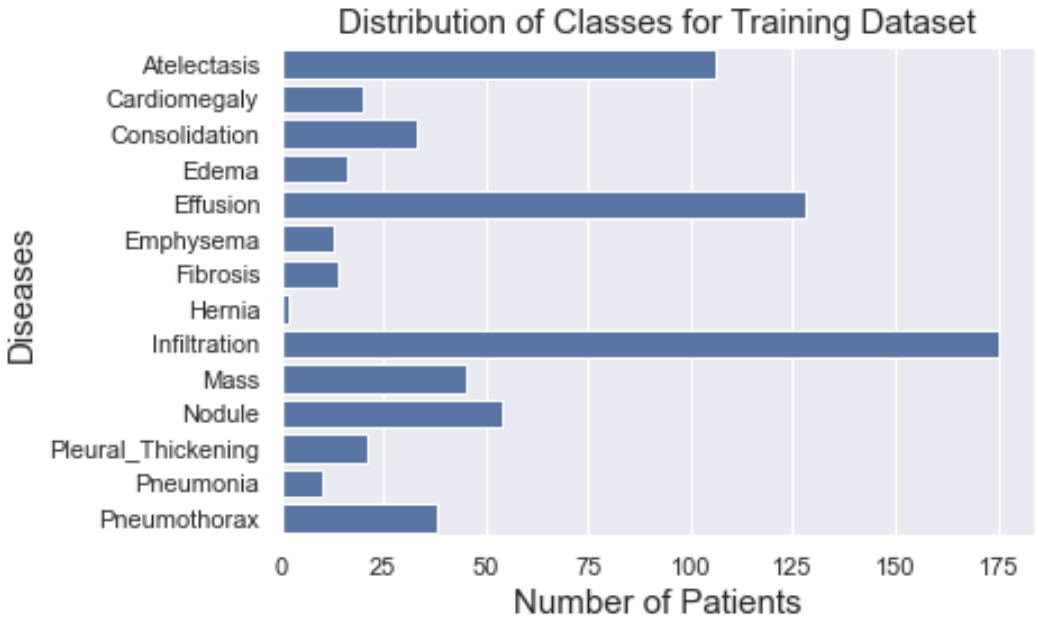

Figure 4.2: Distribution of classes for the training dataset.

Patient ID Analysis

Patient IDs were checked to determine if there were repeated data for certain patients. The goal was to ensure that patients with multiple records did not appear in both the training and test sets to prevent data leakage.

Data Visualization

Sample images from the dataset were visualized to provide an overview of the data, as seen in Figure 4.3.

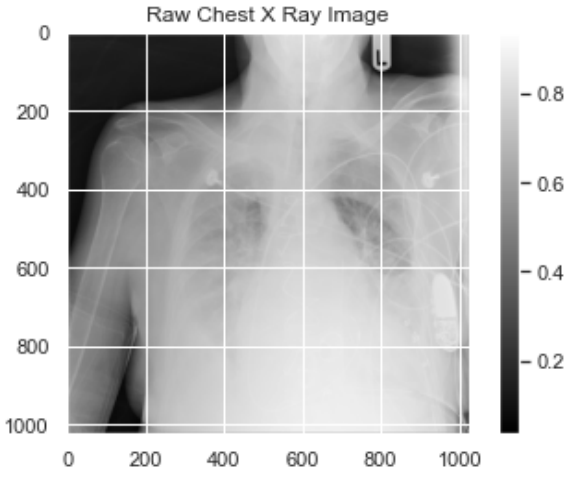

Figure 4.3: Details of the image contents.

Image dimensions, pixel value ranges, and pixel intensity distributions (Figure 4.4) were explored and analysed, based on certain types of characteristics of medical images. [39]

- **Image Dimensions**:

The images in the dataset had a size of 1024 pixels in width and 1024 pixels in height. This indicates the physical size or resolution of each image. The images were grayscale, meaning they had only one colour channel, which is typical for medical images.

- **Pixel Value Range**:

  Pixel values in images typically represent the intensity or brightness of each pixel. In this dataset, the pixel values ranged from a minimum value of 0.0353 to a maximum value of 0.9373. This range represents the variation in brightness or intensity across the image. In medical images like X-rays, pixel values often represent the level of X-ray attenuation by different tissues. Lower values may indicate less dense or more transparent areas, while higher values may indicate denser or more opaque areas.

- **Mean Pixel Value**:

  The mean pixel value for the images was 0.6060. This value represents the average brightness or intensity across all pixels in the image. It indicates the overall brightness level of the images. In the context of X-ray images, it could suggest the average X-ray attenuation across the image.

- **Standard Deviation**:

  The standard deviation of pixel values was 0.2223. This statistic measures the degree of variation or spread in pixel values within the image. A higher standard deviation indicates greater variability in pixel intensities, while a lower standard deviation suggests more uniform pixel values.

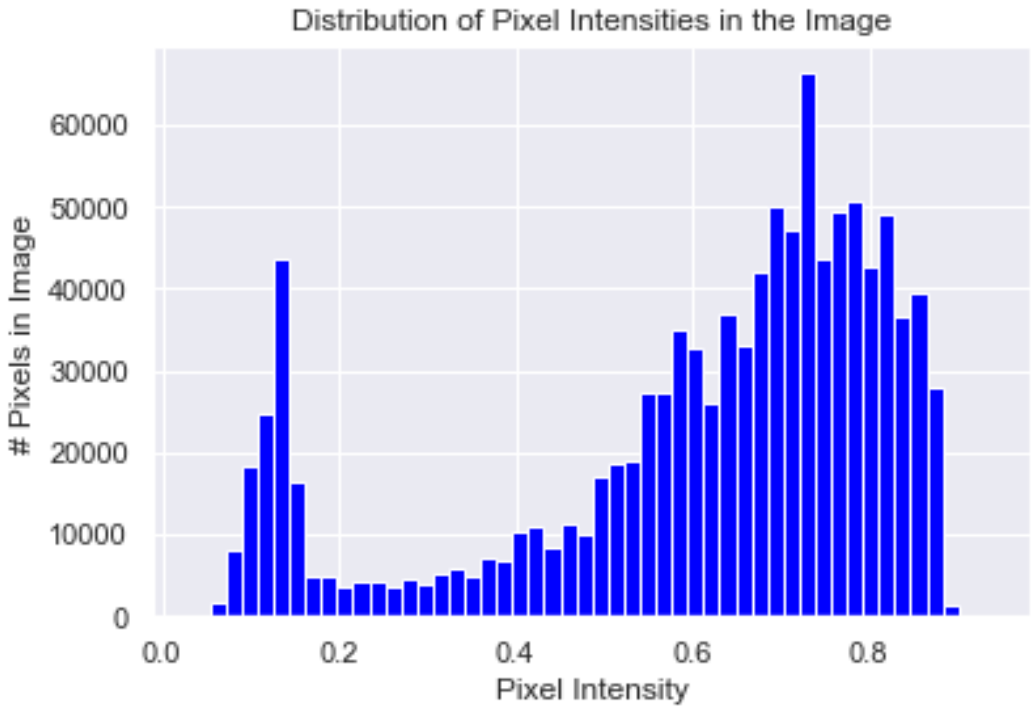

Figure 4.4: Distribution of pixel intensities in the image.

Data Scaling

Data scaling (also known as data normalization) is a critical data pre-processing step in deep learning and machine learning tasks. It aims to ensure that the input data features have consistent scales and distributions. In the context of the project, it was essential for the following reasons:

1. **Convergence**:

Standardizing the data helps the neural network converge more quickly during training. When the input features have consistent scales and are centred around zero, it can speed up the optimization process, also allowing the model to be faster.

2. **Gradient Descent**:

   During the training process, many optimization algorithms, such as gradient descent, can be sensitive to the scale of input features. Standardized data can mitigate issues related to uneven feature scales, preventing slow convergence or divergence during training.

3. **Model Stability**:

   Data scaling can lead to better model stability. Neural networks often benefit from having inputs that are centred around zero, as this can help prevent activations from becoming too large or too small, which could lead to issues like vanishing or exploding gradients.

4. **Generalization**:

   Standardizing the data can improve the generalization of the model. When input data has a consistent scale and distribution, the model is more likely to perform well on unseen data that follows a similar pattern.

Therefore, applying sample-wise centring and standardization to pixel values fell under the category of data pre-processing, specifically data scaling, and was a crucial step to ensure that our neural network performs effectively during training and inference.

The **ImageDataGenerator** from the Keras library [40] was employed to perform data scaling and other pre-processing tasks on the X-ray images. We will now discuss the procedure behind the data scaling task using this tool.

- **Sample-Wise Centering and Standardization**:

  The key aspect of data scaling in this project was the sample-wise centring and standardization of pixel values. For each image in the training dataset, the generator subtracted the mean pixel value and divided by the standard deviation. This operation was performed individually for each image, making it sample-wise.

- **Mean and Standard Deviation**:

  After the mean pixel value and the standard deviation were computed for each image in the previous task 4.1.3, the generator subtracted this mean value from each pixel in that image. Similarly, the generator divided the pixel values by the standard deviation.

- **Normalization**:

  The result of this operation was that, after data scaling, all the pixel values in each image had a consistent scale and distribution. The data scaling process effectively normalized the pixel values. The result of the normalization can be seen in Figure 4.5

- **Scale and Distribution**:

  By normalizing the pixel values, as seen in Figure 4.6, we ensured that they were centered around zero and had a standard deviation of one. This meant that the pixel values of the images had a similar scale and distribution, which is crucial for training a deep learning model effectively.

In terms of benefits, data scaling in this manner had several benefits, including faster convergence during training, improved model stability, and better generalization to unseen data. It helped the CNN model process the X-ray images more efficiently and effectively.

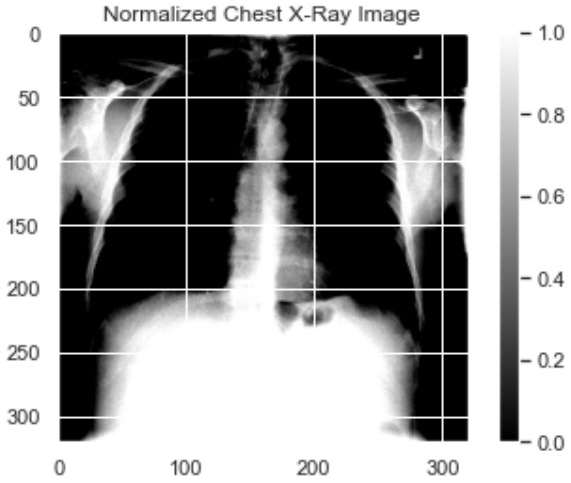

Figure 4.5: Normalized chest x-ray image.

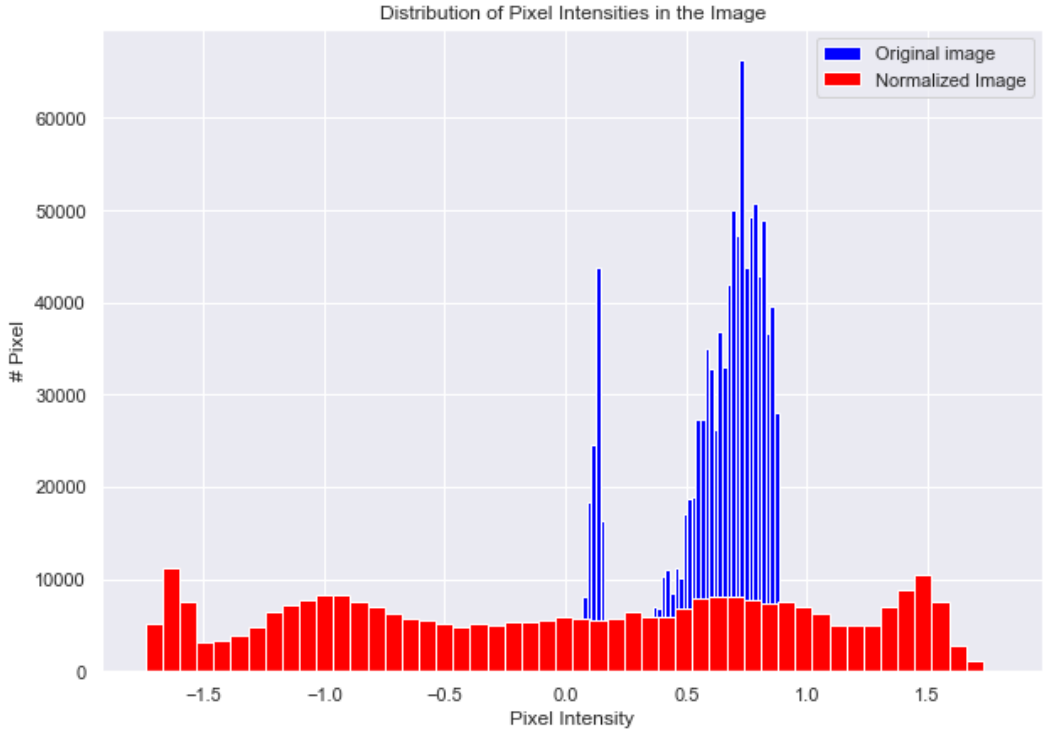

Figure 4.6: Comparison between the distribution of pixels in the original and normalized image.

Data Partitioning

Data partitioning was then performed to create distinct subsets for training and validation. The removal of overlapping patients from the validation set can be considered a form of data partitioning, ensuring that no patients appear in both the training and validation sets.

The task described here, known as "Patient Overlap and Data Leakage" is an essential step in the data pre-processing pipeline for medical image analysis, particularly when dealing with datasets containing medical images and patient-related information. Let's break down this task and its significance:

- **Patient Overlap**:

In medical datasets, each patient may have multiple associated images or records. Patient overlap refers to situations where the same patient's data appears in multiple subsets of the dataset, such as training, validation, and test sets.

- **Data Leakage**:
  Data leakage occurs when information from the validation or test set inadvertently influences the training process. In the context of medical image analysis, data leakage can lead to overly optimistic model performance estimates, as the model might inadvertently learn patterns specific to the validation or test data.
- **Significance**:
  Identifying and addressing patient overlap and data leakage is crucial for maintaining the integrity of the machine learning experiment and ensuring that the model generalizes well to new, unseen data. If patient overlap exists and is not handled correctly, the model may inadvertently learn to recognize specific patients rather than generalizing from the medical conditions present in the data.

The task involved comparing the patient IDs present in the training set with those in the validation and test sets. Overlapping patient IDs were identified, indicating that the same patients were present in multiple datasets.

To prevent data leakage, we removed the patient data associated with overlapping patients from the training set. This ensured that the model did not learn from the same patients whose data was used for validation or testing.

The decision to remove overlapping patients from the training set (and not the validation set) is crucial for maintaining the data's representativeness and integrity. The validation set is typically used to assess how well the model generalizes to unseen data. Removing overlapping patients from the validation set might make it less representative of real-world scenarios, where the model needs to make predictions for patients it has not seen during training. By only removing overlapping patients from the training set, we ensure that the model's training data is not contaminated by data that it will later evaluate in the validation or test sets.

In summary, the task of identifying patient overlap and removing overlapping patients from the training set is a critical step to maintain the integrity and fairness of your machine learning experiment when working with medical image datasets. It helps to ensure that the model learns to generalize from a diverse and unbiased set of patient data.

Data Augmentation

Besides being used for data scaling tasks, the **ImageDataGenerator** [40] was also used for data augmentation, which is a crucial aspect of training a deep learning model for medical image classification.

By applying random transformations such as rotation, width and height shifts, horizontal flips, zooming, and brightness adjustments to chest X-ray images, the generator creates variations of the original images.

This is essential because it simulates different positions, orientations, and lighting conditions that can occur during the imaging process, therefore increasing the diversity of the training dataset, and making the model more robust and better at generalizing to unseen chest X-ray images with varying characteristics.

Random horizontal flipping of X-ray images was performed using the **ImageDataGenerator** [40]. This means that during training, some X-ray images are horizontally flipped with a certain probability.

Horizontal flipping helps the model learn to recognize features from different orientations, enhancing its ability to identify patterns in X-ray images regardless of whether they are taken from the left or right side.

The images were then resized to a uniform size of 320x320 pixels. This resizing ensures that all input images have consistent dimensions. It is important because deep learning models, particularly CNNs, typically require input data to have the same dimensions. By resizing the images, we make sure that the model can process them consistently, regardless of their original dimensions.

As we saw in 4.1.3, the original image is grayscale. The **ImageDataGenerator** [40] also converted the single-channel X-ray images (gray-scale) to a 3-channel format. Many pre-trained deep learning models, including the one used in our project, require 3-channel inputs. This was done by repeating the values in the image across all three colour channels (e.g., R, G, B). This ensures compatibility with models designed for colour images.

After each epoch of training, the **ImageDataGenerator** [40] shuffled the input data. Shuffling helps prevent the model from memorizing the order of the training samples and ensures that it learns to generalize from the data more effectively.

4.1.4 Handling Class Frequencies Imbalance

The task described here involved calculating class frequencies for each label in the dataset. Specifically, it aimed to determine the proportion of positive and negative examples for each class (i.e., pathological condition) within the dataset and check if the contributions were equal.

Computing class frequencies is essential to understand the distribution of classes and the degree of class imbalance. These class frequencies are later used to address class imbalance issues, such as assigning appropriate weights to classes or customizing loss functions, which we will discuss.

As we can see in Figure 4.7, the contributions of positive cases were significantly lower than those of the negative ones. However, we want the contributions to be equal.

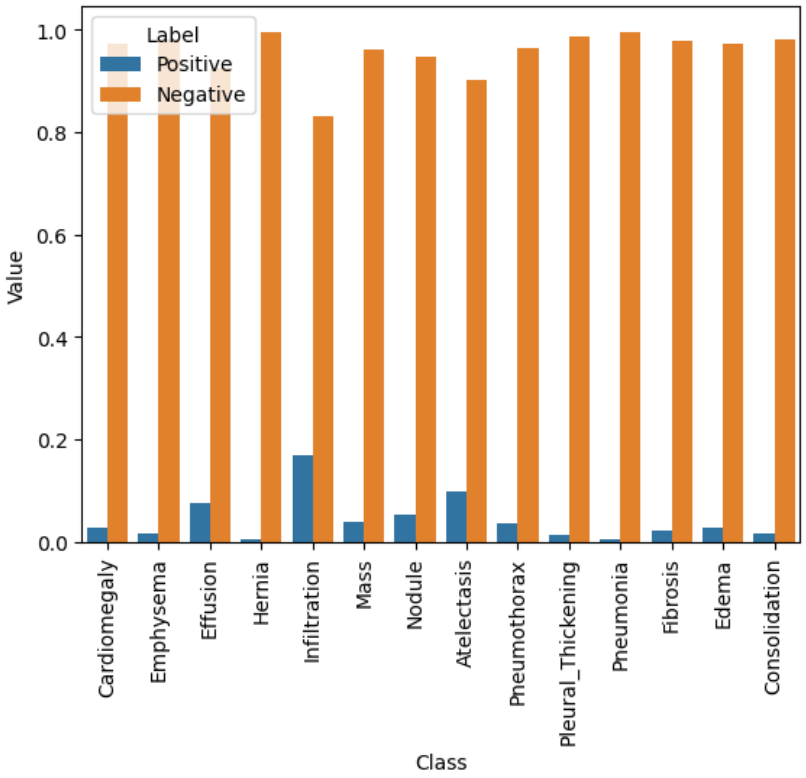

Figure 4.7: Contribution of positive and negative labels before balance.

There are consequences of class imbalance on model training and performance. Specifically, when using a standard cross-entropy loss function on a highly imbalanced dataset, it could lead the model to prioritize the majority class (i.e., negative cases) since they contribute more to the loss.

To balance these contributions, we had to multiply each frequency from each class ***freqp*** (positive) and ***freqn*** (negative) by a class-specific weight factor, ***wpos*** and ***wneg***, so that the overall contribution of each class is the same, as shown in Equation 4.1.

$$\begin{aligned} wpos \times freqp &= wneg \times freqn, \\ wpos &= freqneg \\ wneg &= freqpos \end{aligned} \tag{4.1}$$

This way, we balanced the contribution of positive and negative labels, as shown in Figure 4.8.

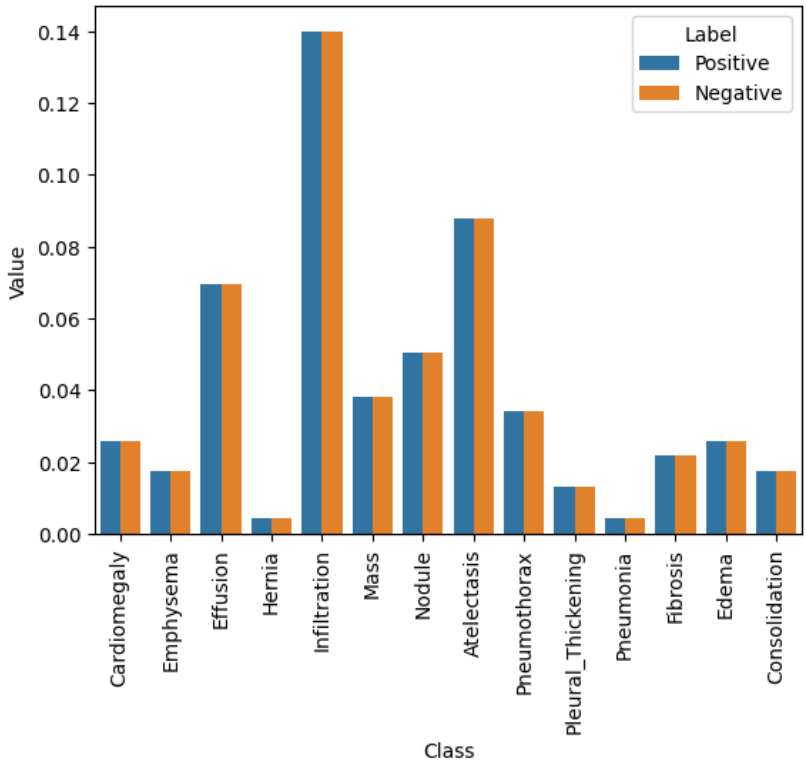

Figure 4.8: Contribution of positive and negative labels after balance.

After computing the weights, the final weighted loss for each training case is presented in the next formula.

$$\mathcal{L}^{w}_{cross-entropy}(x) = -(w_p y \log(f(x)) + w_n(1-y)\log(1-f(x))).$$

The formula above computes the average weighted loss for all training examples (batch data). For the multiclass loss, the average loss for each individual class is computed. The small value, ϵ, is added to the predicted values before taking their logs. This is done to avoid a numerical error that would otherwise occur if the predicted value happens to be zero.

4.2 Model Architecture

The pulmonary diseases detection model used in this project is based on a CNN architecture because they are a popular choice for image classification tasks due to their ability to automatically learn hierarchical features from image data.

We will now break down the architecture and discuss the thinking behind its design.

4.2.1 Choice of the Model

In terms of choosing the model to apply, the decision was made in an early stage of the project, when collecting the advantages of a CNN architecture for this type of image classification task. Being able to adapt and interpret models was also a factor in deciding what model to use. We ended up choosing a CNN architecture because it can capture local patterns and hierarchical features in images. The model architecture consisted of the following components:

Base Model - DenseNet121

The base of the model is the DenseNet121 architecture [3], specified in Chapter 3, Section 2.2. The base is pre-trained on a large dataset and serves as a feature extractor. It has multiple convolutional layers, each followed by a batch normalization layer and a Rectified Linear Unit (ReLU) activation function. These layers are densely connected, meaning each layer receives inputs from all previous layers. This architecture promotes feature reuse and gradient flow, leading to better training and convergence. The pre-trained weights of DenseNet121 are used, leveraging transfer learning to extract meaningful features from chest X-ray images.

GAP Layer

After the convolutional layers of DenseNet121, a Global Average Pooling 2D layer was added. GAP reduces the spatial dimensions of the feature maps and outputs a fixed-size vector for each feature map. GAP was used to obtain a compact representation of the learned features. It simplified the model and reduced the number of parameters, which can help prevent overfitting, one of the encountered challenges.

Dense Layer with Sigmoid Activation

Following the GAP layer, a dense layer with sigmoid activation was added. This dense layer is designed for multi-label classification, where each class (pathological condition) can be present or absent independently. The sigmoid activation function assigned a probability score between 0 and 1 to each class, indicating the likelihood of the presence of that condition.

Compilation

The model was then compiled using the Adam optimizer, a widely used optimizer for deep learning tasks. Adam adapts learning rates for each parameter, which can lead to faster convergence. The custom loss function was specified using the ***get_weighted_loss*** function. This loss function takes into account class weights to handle class imbalance, which we saw as one of the challenges and that was treated in a previous stage of the implementation, the Data Preprocessing task, in Subsection 4.1.4.

4.3 Pre-trained Model

The base model, DenseNet121 [3], was initialized with pre-trained weights. These weights were obtained from a model that was trained on a large dataset, in our case the ChexNet [4], as discussed in Chapter 3, Subsection 2.2.1, which contains a wide variety of chest X-ray images. Using pre-trained weights allowed the model to capture general image features and patterns, which in our case were for the same purpose, and were then fine-tuned.

In summary, the choice of the model architecture, including the use of DenseNet121 as a feature extractor and the addition of GAP and sigmoid activation layers, is well-suited for the task of detecting pulmonary diseases in chest X-ray images. The utilization of pre-trained weights facilitates feature extraction and transfer learning, enabling the model to learn relevant patterns from medical images effectively.

4.4 Model Training

Aside from the pre-trained model, we tried to fully train the model, using the training dataset previously preprocessed in an earlier stage. We used the validation data for monitoring the model

performance. Training typically involves multiple epochs, where each epoch represents one pass through the entire training dataset. The model learns to predict the presence or absence of each pathological condition in chest radiographs. The challenges encountered with this training are explained in the next section.

4.5 Challenges Encountered

Deep learning is revolutionising the field of medical image analysis and diagnosis. It can offer many advantages to improve healthcare outcomes, such as its ability to incorporate XAI techniques. However, its advantages in the medical domain also come with some challenges.

In this section, we expose the challenges encountered when employing deep learning models to our dataset and explore the strategies applied to overcome them. These challenges included class imbalance, overfitting, and the considerable training time required for complex models. We will now discuss each of these challenges in detail and highlight the solutions employed to address them.

- **Class Imbalance:**
  - **Issue:** Medical datasets often suffer from class imbalance, where certain classes have significantly fewer samples compared to others. For example, in a dataset to diagnose rare diseases, the number of positive cases may be very small compared to the number of negative cases.
  - **Solution:** To address this challenge, we assigned class-specific weights during training. These weights gave more importance to the underrepresented classes, ensuring that the model paid sufficient attention to them. This helped in preventing the model from biased predictions towards the majority class.
- **Overfitting:**
  - **Issue:** Overfitting occurs when a deep learning model performs exceptionally well on the training data but poorly on unseen data. This can happen when the model memorises the training data instead of learning general patterns.
  - **Solution:** Several techniques were employed to combat overfitting. We applied data augmentation, for instance, creating new samples and helping the model generalize better. Regularization techniques like L1 or L2 regularization were also applied to penalize complex model weights. Additionally, the use of GAP layers in the neural network architecture helped reduce spatial dimensions before the final classification layer, which acted as a form of regularisation and feature extraction. As discussed in the subsections above, we applied all of these solutions to prevent overfitting of the data, and the measures were successful.
- **Training Time:**
  - **Issue:** Training deep neural networks, especially on large medical datasets, can be time-consuming and computationally intensive. Training a deep model from scratch can take a significant amount of time and resources.
  - **Solution:** There were several strategies that we tried to mitigate the long training times:
    - **Mini-Batch Training:** Training on smaller mini-batches of data rather than the entire dataset at once can expedite training and enable efficient use of hardware resources. This was our first approach, and even with the smallest batch, the training times were very time-consuming.
    - **Hardware Accelerators:** Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) can significantly speed up training by parallelising computations. These

are optimised for deep learning tasks. We tried using hardware accelerators by purchasing a Google Colab subscription, but the execution times were not viable and we could not reach considerable results.

* **Transfer Learning:** Instead of training a model from scratch, transfer learning involves using pre-trained models, such as the one that we used in an early stage of the project, described in Subsection 4.3, and fine-tuning them based on our dataset. This dramatically reduced training time, as the model had already learnt useful features. This ended up being the final approach. We applied transfer learning in the experimental evaluation explored in the next chapter, Chapter 5.

4.6 Integration of Explainable AI Techniques

In this section, we present the incorporation of XAI techniques into our model for chest X-ray image classification. These strategies provided useful insights into the model's decision-making process, assisting us in understanding why and how the model makes its predictions. We chose to employ Grad-CAM, LIME and RISE as part of our XAI techniques. The choice was made based on the State of the Art study presented in Chapter 3.

4.6.1 Grad-CAM

Grad-CAM [7] is a technique that shows the parts of an image on which the model concentrates when generating a certain prediction. It provides insight into the importance of different parts of the image for a specific class.

We will now enumerate the process behind generating Grad-CAM explanations.

1. We selected a class of interest and obtained the class score gradient concerning the feature maps in a specific layer of the model.
2. The weighted combination of these gradient-based feature maps was calculated by the Grad-CAM, showing regions that contributed favourably to the class score.
3. The heatmap that results illustrates which portions of the image were important in the model's choice for that class.

4.6.2 LIME

LIME [8] is a method for producing local explanations for individual predictions. It works by triggering the input image to change and then observing how the model's predictions change. LIME generates surrogate models that locally approximate the behaviour of the original model. In our scenario, we use LIME to explain why the model predicted a specific chest radiograph image.

We will now enumerate the process behind generating LIME explanations.

1. We loaded and pre-processed the input image.
2. We developed a surrogate model that replicates the behaviour of the original model.
3. LIME repeatedly perturbed the input image and observed the model predictions.
4. It found the image's most influential regions by examining altered images and their accompanying predictions.
5. Finally, LIME generated an explanation that highlighted these influential regions, to help us understand which parts of the chest X-ray image contributed to the model's decision.

4.6.3 RISE

RISE [9] is a technique that provides insight into the model's decision boundaries by generating a set of masks and observing their impact on predictions. It helps us understand the robustness and stability of the model's predictions.

We will now enumerate the process we applied to generate RISE explanations.

1. We created a set of randomised masks, each with different patterns of pixel values (0 or 1).
2. These masks were applied to the input image, effectively hiding or highlighting different parts of the image.
3. We observed how the model's predictions varied when different masks were applied.
4. By aggregating the predictions across multiple masks, RISE generated an explanation highlighting regions of the image that strongly influence the model's decisions.

We encountered some challenges when trying to apply RISE that ended up making not possible to retrieve satisfactory results and heatmaps. We will enumerate some of the causes of this obstacle.

1. **Computationally Intensive:** RISE can be computationally expensive. Since we had 14 classes, we had to generate a large number of synthetic images by masking and evaluating them, which required significant computational resources and time.
2. **Sensitivity to Masking Parameters:** The quality of RISE explanations depends a lot on the choice of masking parameters, such as the number and size of masks. This makes it time-consuming and provides us with a lot of misleading or less informative explanations.

By integrating these XAI techniques into our model, we gained a deeper understanding of its behaviour and can better trust its predictions in chest X-ray image analysis.

CHAPTER 5

Experimental Evaluation

In this chapter, we explore the experimental evaluation phase. Our primary objectives are to rigorously assess the performance and effectiveness of the model and proposed methods. Our aim was to validate our hypotheses, examine the results, and draw meaningful conclusions from the data gathered during this research.
The chapter is structured in different sections, and described as follows. We will first detail the experimental setup, including the evaluation metrics used to assess the model's performance and the interpretation of results gained by the XAI techniques applied. We will present some of the results for each class while interpreting them. At the end, we discuss the implications of the results in terms of ethical considerations and bias analysis. We also address some challenges encountered during the evaluation process.

> *“ In science, there are no shortcuts to truth. ”*
>
> *Karl Popper,* The Logic of Scientific Discovery[41]*, 1959*

5.1 Evaluation and Validation of the Model

In this section, we'll present the outcomes achieved by fine-tuning the initially employed pre-trained model. After fine-tuning this model, we employed the earlier described XAI techniques. We'll discuss the specific methods chosen to evaluate and validate our model.

5.1.1 Evaluation Metrics

To assess the performance of the model, various evaluation metrics were used for the multilabel classification tasks. In the next subsections, we will present a comprehensive review of the used evaluation metrics for the data splitting made on our dataset. We will enumerate the results and insights gained with those results for the training data, validation data and testing data.

5.1.2 Model Results on Training Data

We present the results of the model in the training data in Table 5.1.

Metric	Value
Train Loss	0.552
Binary Accuracy	0.782
False Negatives (images)	39

Table 5.1: Model evaluation results on training data.

Discussion of Model Results on Training Data

Based on the evaluation metrics, the model's performance on the training data in detecting pulmonary diseases provided us with valuable insights into its effectiveness.

- Loss: The training loss is 0.5521, which means that, on average, the model's predictions deviated from the actual values by this amount during the training process. A lower training loss suggests a better fit of the model to the training data.
- Binary accuracy: The training accuracy is 78.22%, indicating that the model correctly predicted approximately 78.22% of the samples in the training dataset. It is a measure of how well the model performed on the data it was trained on.
- False Negatives: The model had 39 false negatives during training. These are instances that were incorrectly predicted as negative when they were actually positive.

5.1.3 Model Results on Validation Data

We present the model results on Validation Data in Table 5.2.

Metric	Value
Validation Loss	0.793
Binary Accuracy	0.8
False Negatives (images)	13

Table 5.2: Model evaluation results on validation data.

Discussion of Model Results on Validation Data

In terms of the model's performance on the validation data in detecting pulmonary diseases, we will now discuss and provide some insights into its effectiveness.

- Loss: The validation loss is 0.7928, which is higher than the training loss. This suggests that the model's performance is slightly worse on data it hasn't seen during training. A reasonable degree of overfitting may occur if the gap between training and validation loss is significant.
- Binary Accuracy: Validation precision is 80%, indicating that the model correctly predicted approximately 80% of the samples in the validation dataset. This is generally a good sign, as the model is performing well on unseen data.
- False Negatives: The model had 13 false negatives in the validation dataset. The decrease in false negatives from training to validation is positive, indicating that the model improved in identifying true positive cases in the validation phase.

5.1.4 Model Results on Test Data

We present the model results on Test Data in Table 5.3.

Metric	Value
Test Loss	1.387
Binary Accuracy	0.710
False Negatives (images)	75

Table 5.3: Model evaluation results on test data.

Discussion of Model Results on Testing Data

Based on the evaluation metrics, the model's performance on the test data in detecting pulmonary diseases provided us with valuable insights into its effectiveness. Here is a summary of the performance:

- Loss: The test loss is 1.3874, which is notably higher than both the training and validation losses. This suggests that the model's performance on the test data is not as good as on the training and validation data. It might indicate that the model is overfitting the training data or that the test data is significantly different.
- Binary Accuracy: The test accuracy is 71.03%, which is lower than the validation accuracy. This means that the model correctly predicted approximately 71.03% of the samples in the test dataset. It's slightly lower than the validation accuracy, indicating a potential drop in performance when moving from the validation to the test set.
- False Negatives: The model had 75 false negatives in the test dataset. This is a significant increase from the validation phase, and it suggests that the model is missing a substantial number of true positive cases in the test set. We will observe this in action when evaluating the visual explanations provided by the employed XAI methods.

Another way to evaluate the model's ability to generalize to unseen data is to plot the AUC-ROC scores. Receiver Operating Characteristic Curve (ROC) and Area Under the ROC Curve (AUC) are both used to evaluate the performance of classification models. They provide a way to assess how well a model can distinguish between two classes, typically a positive class and a negative class. In terms of multilabel classification, it can be used for each label to assess the model's ability to distinguish between positive and negative labels. The results of the AUC-ROC scores and curves are presented next, in Figure 5.1.

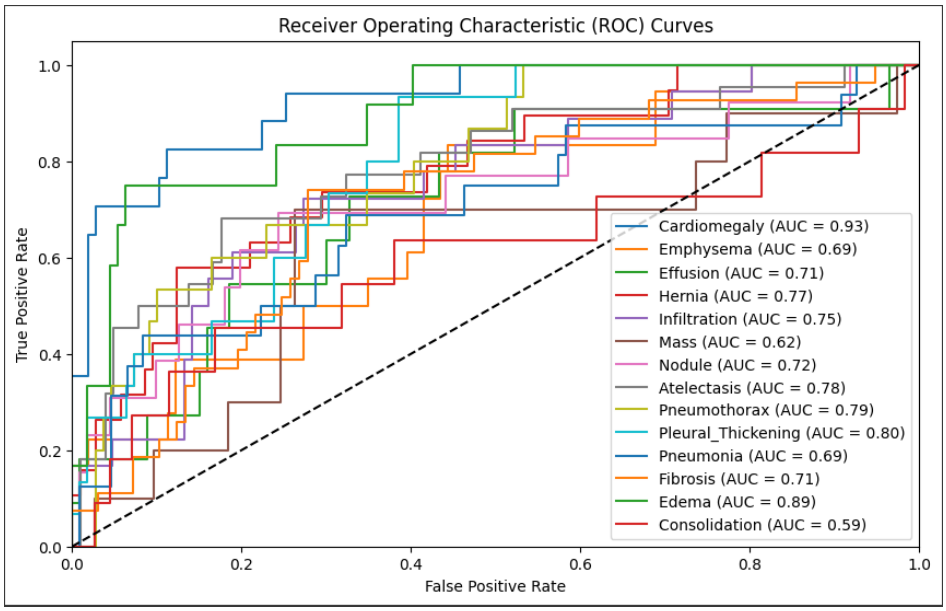

Figure 5.1: AUC-ROC scores and curves in the testing data for each class of our dataset.

By reviewing the plotted scores and curves, we can conclude that the class that has a better AUC score (0.93) is Cardiomegaly, which is the closest to being considered a perfect classifier for that specific class. A label with a high AUC score indicates that the model can effectively distinguish between that label's positive and negative instances. The worst AUC score is considered to be for the Consolidation class, which has an AUC of 0.59. A label with an AUC score close to 0.5 suggests that our model had a weaker ability to separate the positive and negative instances for that label, possibly due to the class imbalance.

As referred to in Subsection 4.1.4, the class imbalance was a challenge, and the contribution of the positive labels was really low compared to the negative ones. Despite balancing the dataset, we can still consider that the model would benefit from more images contributing to the positive labels. Overall, the model shows promise in detecting pulmonary diseases but may require further refinement to minimize false negatives, which can have significant clinical implications.

5.1.5 Error Analysis

To assess the model's performance and the effectiveness of XAI techniques on a broader scale, we computed and visualized the training and validation errors per class. This analysis helps us understand which pathologies are challenging for the model and where XAI techniques might be particularly useful.

Train Error

Similarly, we calculated the error percentage for each class in the training dataset, as observed in Figure 5.2. This analysis provides insights into whether there are specific pathologies that are consistently challenging for the model during training.

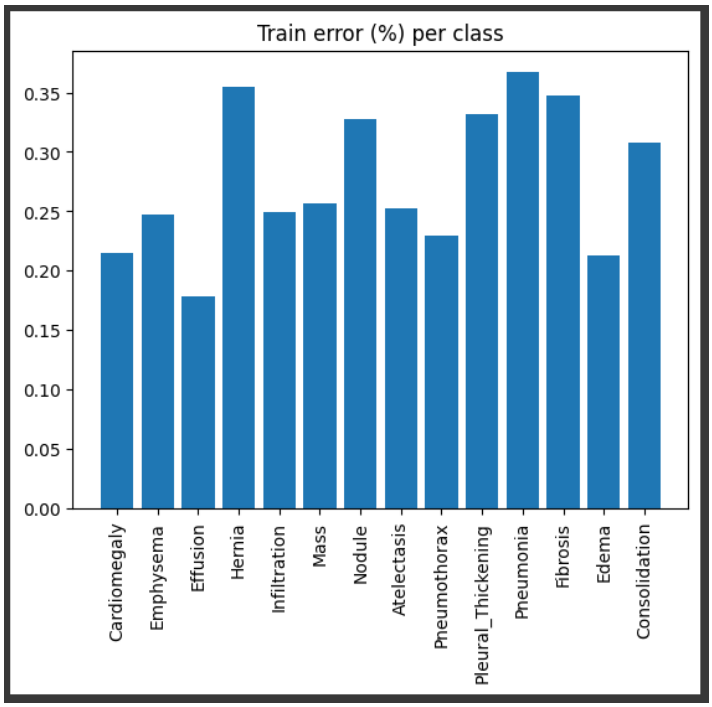

Figure 5.2: Training error for all classes.

Validation Error

We calculated the error percentage for each class in the validation dataset, as we can see in Figure 5.3. The error represents the discrepancy between the predicted class probabilities and the ground

truth labels. The bar chart shows the validation error for each pathology, helping us identify where the model struggles the most.

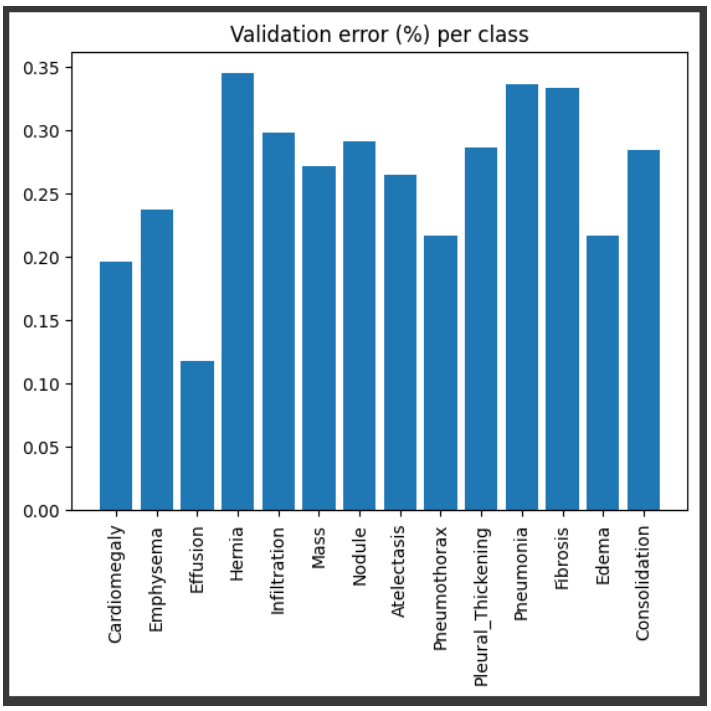

Figure 5.3: Validation error for all classes.

Test Error

After testing, we calculated the error percentage for each class in the testing dataset. As we can observe in the bar chart in Figure 5.4, the conditions with a higher loss were also the ones that could not provide good visual explanations, as described in Subsection 5.2.1.

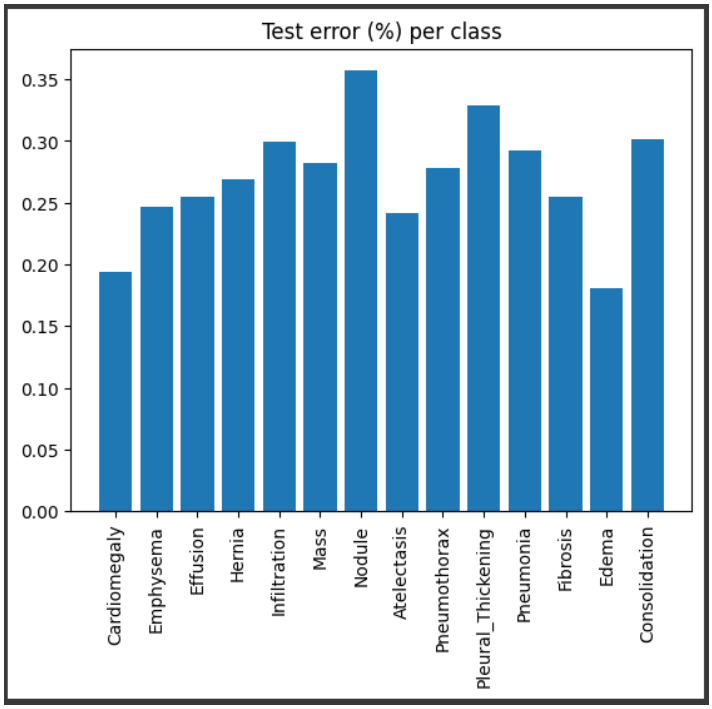

Figure 5.4: Test error for all classes.

In conclusion, the error analysis provided a wider view of the model's performance on different pathologies, aiding in the identification of areas where the model would benefit from further refinement or additional data.

5.2 Explainable AI Evaluation

The integration of XAI techniques into the model was crucial for understanding its decision-making process and ensuring the reliability of predictions in the medical domain. In this section, we will evaluate the effectiveness of the employed XAI techniques, including Grad-CAM and saliency maps, in providing interpretable insights into the model's predictions.

5.2.1 Grad-CAM and LIME Analysis

Let's evaluate Grad-CAM's and LIME's performance on several specific cases:

Case 1: Cardiomegaly Prediction

- True Labels: ['Cardiomegaly']
- Predicted Label: Cardiomegaly
- Class Probability: 0.9997

In Figure 5.5 we can observe the visual explanations of each XAI method for the considered primary pathology.

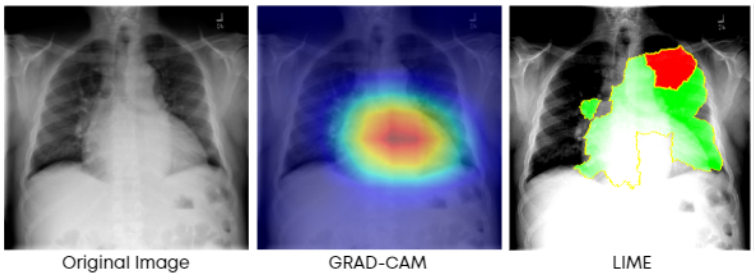

Figure 5.5: Visual explanations for case 1 (Cardiomegaly).

Case 2: Emphysema Prediction

- True Labels: ['Emphysema']
- Predicted Label: Emphysema
- Class Probability: 0.9983

In Figure 5.6 we can observe the visual explanations of each XAI method for the considered primary pathology.

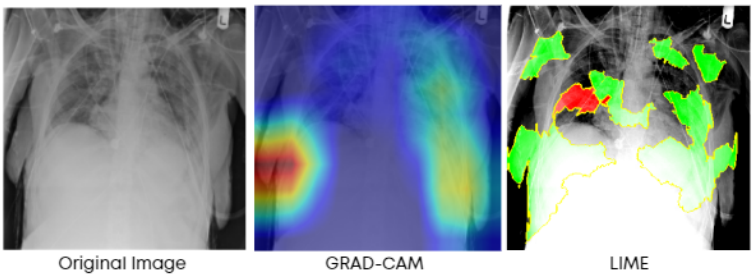

Figure 5.6: Visual explanations for case 2 (Emphysema).

Case 3: Effusion and Fibrosis Prediction

In this case, we have multiple true labels, but the model predicts Effusion as the primary pathology. We applied Grad-CAM and LIME, helping us understand the model's decision process.

- True Labels: ['Effusion', 'Fibrosis']
- Predicted Label: Effusion
- Effusion Class Probability: 0.9846
- Fibrosis Class Probability: 0.3596

In Figure 5.7 we can observe the visual explanations of each XAI method for the considered primary pathology.

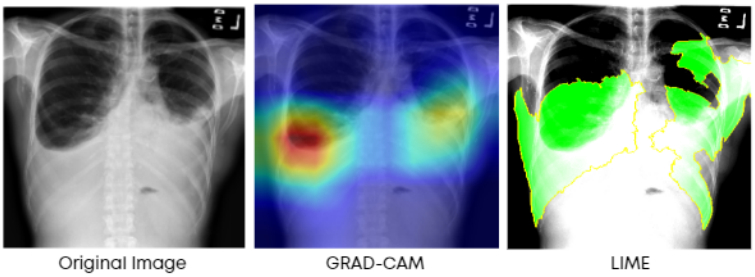

Figure 5.7: Visual explanations for case 3 (Effusion).

Case 4: Edema, Pneumoniax and Infiltration Prediction

In this case, we have multiple true labels, but the model predicts Edema as the primary pathology. We applied Grad-CAM and LIME, helping us understand the model's decision process.

- True Labels: ['Infiltration', 'Pneumonia', 'Edema']
- Predicted Label: Edema
- Edema Class Probability: 0.9784
- Pneumonia Class Probability: 0.8574
- Infiltration Class Probability: 0.8138

In Figure 5.8 we can observe the visual explanations of each XAI method for the considered primary pathology.

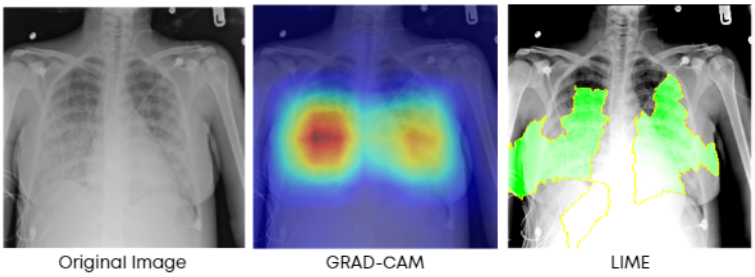

Figure 5.8: Visual explanations for case 4 (Edema).

Case 5: Nodule Prediction

- True Labels: ['Nodule']
- Predicted Label: Nodule
- Class Probability: 0.9696

In Figure 5.9 we can observe the visual explanations of each XAI method for the considered primary pathology.

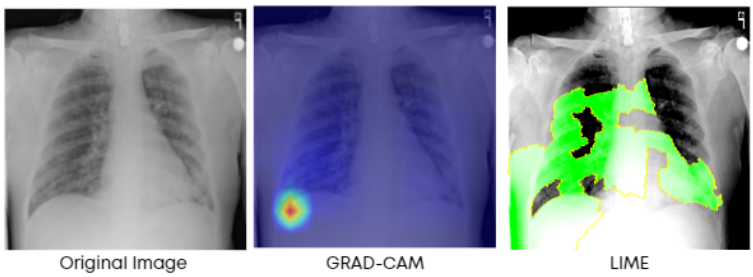

Figure 5.9: Visual explanations for case 5 (Nodule).

Case 6: Emphysema, Atelectasis, and Pneumothorax Prediction

- True Labels: ['Emphysema', 'Atelectasis', 'Pneumothorax']
- Predicted Label: Atelectasis
- Atelectasis Class Probability: 0.9566
- Emphysema Class Probability: 0.2046
- Pneumothorax Class Probability: 0.2883

In Figure 5.10 we can observe the visual explanations of each XAI method for the considered primary pathology.

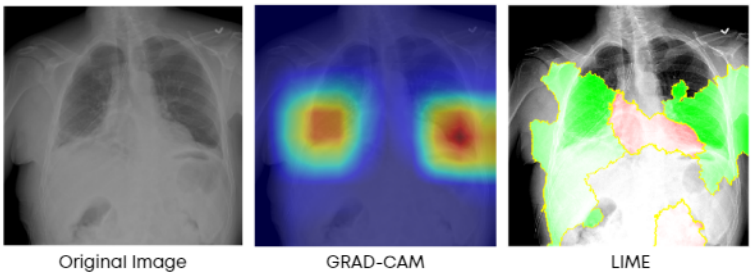

Figure 5.10: Visual explanations for case 6 (Atelectasis).

Case 7: Hernia Prediction

- True Labels: ['Hernia']
- Predicted Label: Hernia
- Class Probability: 0.9533

In Figure 5.11 we can observe the visual explanations of each XAI method for the considered primary pathology.

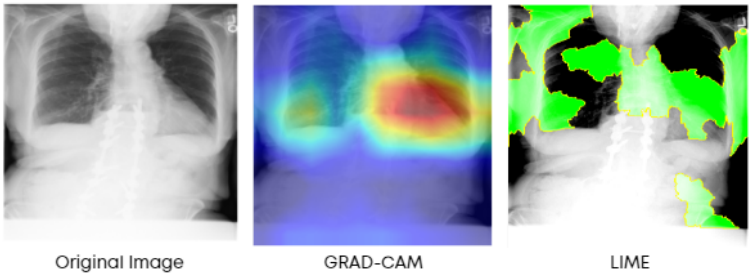

Figure 5.11: Visual explanations for case 7 (Hernia).

Case 8: Fibrosis Prediction

- True Labels: ['Fibrosis']
- Predicted Label: Fibrosis
- Class Probability: 0.9495

In Figure 5.12 we can observe the visual explanations of each XAI method for the considered primary pathology.

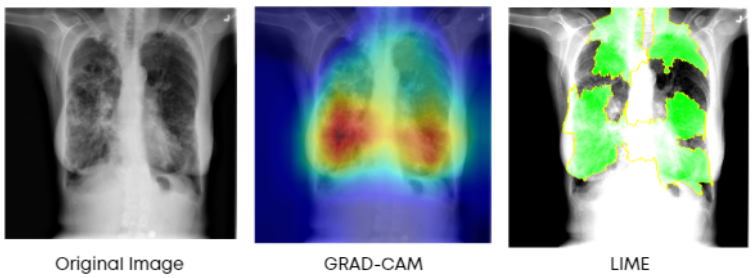

Figure 5.12: Visual explanations for case 8 (Fibrosis).

Case 9: Pneumothorax and Pleural Thickening Prediction

- True Labels: ['Pneumothorax', 'Pleural_Thickening']

- Predicted Label: Pneumothorax
- Pneumothorax Class Probability: 0.9241
- Pleural Thickening Class Probability: 0.8511

In Figure 5.13 we can observe the visual explanations of each XAI method for the considered primary pathology.

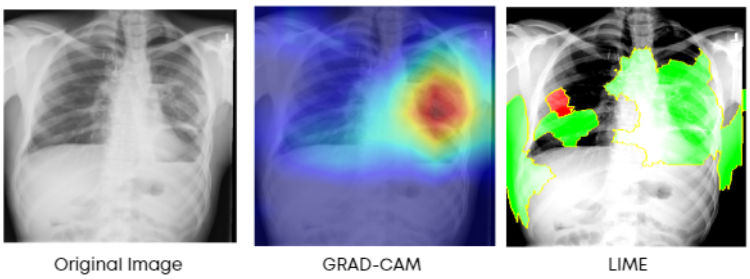

Figure 5.13: Visual explanations for case 9 (Pneumothorax).

Case 10: Effusion, Infiltration, Mass, and Nodule Prediction

- True Labels: ['Effusion', 'Infiltration', 'Mass', 'Nodule']
- Predicted Label: Mass
- Mass Class Probability: 0.8879
- Effusion Class Probability: 0.5369
- Infiltration Class Probability: 0.3381
- Nodule Class Probability: 0.8279

In Figure 5.14 we can observe the visual explanations of each XAI method for the considered primary pathology.

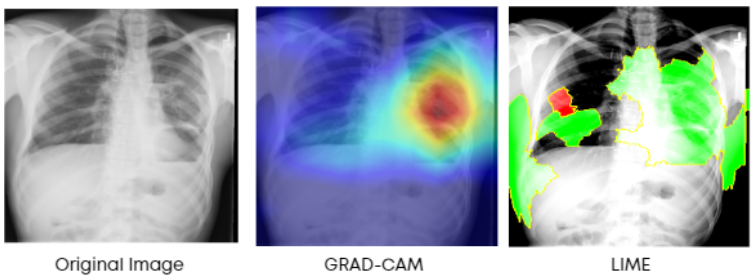

Figure 5.14: Visual explanations for case 10 (Mass).

Pleural Thickening, Infiltration, Pneumonia and Consolidation Cases

These four pulmonary pathologies presented certain challenges, leading to less consistent visual results for individual cases. When we examined the test loss bar charts for each class, it became evident that these specific classes had notably higher testing losses. This could be attributed to the fact that these pathologies typically occur in conjunction with other conditions rather than in isolation, making them more challenging to distinguish from the primary pathologies, when on unseen data.

For cases where these pathologies co-occur with others, our analysis of the plotted visual explanations revealed that they were seldom identified as the primary pathology. Consequently, due to these factors, we encountered difficulties in providing coherent and relevant visual explanations for these four classes.

We can see the demonstration of these conclusions for the infiltration class in Subsections 5.2.1 and 5.2.1. On the other hand, as an example of the pleural thickening class, we can observe the Subsection 5.2.1.

In the case of Pneumonia cases, we also verified that some of the few images that were on the testing set were not properly displayed. An example of this is Figure 5.15. Besides that, it is the class with less positive label images, as we can see in Figure 4.7. These two aspects contribute to the consequence of not being able to portray satisfactory visual explanations for the class.

Figure 5.15: Example of a poorly displayed pneumonia radiograph image.

5.2.2 Insights Gained from XAI Techniques

The quality and effectiveness of explanations generated by XAI techniques can be assessed through various means. The application of XAI techniques has yielded valuable insights and measures into the model's behaviour and decision-making:

1. **Localization of Features**: Techniques like Grad-CAM and LIME highlighted regions in the X-ray images that the model deems important for its predictions. These regions corresponded to areas of interest in disease detection.
2. **Comparison**: By having different XAI techniques, we were able to compare them to identify which one provided more useful insights and if they provided similar results, which can help when evaluating the consistency of the model.
3. **Robustness**: We checked if the explanations remained consistent across different images and datasets, as robust explanations are more trustworthy.
4. **Feedback Loop**: By incorporating XAI insights into the model's development, it was possible to create a feedback loop for continuous model improvement and enhanced accuracy in disease detection.

In summary, the applied techniques provided valuable insights that contribute to both model refinement and we consider that the visual explanations can help healthcare professionals in their decision-making processes.

5.3 Discussion of Interpretability Results

The overall performance and effectiveness of the model had satisfactory results. We can conclude that if we could have made a full training of the model, the results would be a lot better than they appear using a pre-trained model. Despite that, the results are still positive in terms of interpretability and trustworthiness, facilitated by the integrated XAI techniques, which is the focus of our study. Despite the results not being ground-breaking, we can still conclude that using these XAI techniques was a positive measure in terms of understanding the model and its results. In conclusion, integrating XAI techniques like Grad-CAM and LIME into the model's evaluation process enhanced our understanding of its decision-making process. These methods allowed us to visualize and interpret individual predictions,

making it easier for potential experts to validate model predictions and identify areas where the model may require further improvement.

CHAPTER 6

Conclusion and Future Work

This chapter concludes the dissertation by summarizing the key findings, achievements, and contributions. Additionally, it outlines potential avenues for future research and development in the field of Explainable AI in medical image classification.

> *“ Every new beginning comes from some other beginning's end. ”*
>
> *Lucius Annaeus Seneca,* Unknown

6.1 Conclusion

In this project, our objective was to address a critical challenge posed by the use of AI in the domain of medical image classification, which is the inherent "black-box" nature of AI models, that often raises concerns about their transparency, interpretability, and ethical implications.

To demystify and close the gap between AI's decision-making processes and human understanding, we proposed and implemented a medical image classification system that not only provides accurate predictions to a certain variety pulmonary diseases but also offers explanations for those predictions. This was achieved by exploring and employing Explainable AI methods.

Before implementation of this system, we had to do a thorough investigation of the state of the art in both medical image classification and XAI, to get to know the latest trends in this field and to compare the available methods that would better fit our model and case study.

Throughout the implementation and experimentation phases, the system exhibited promising performance, effectively categorising medical images and providing valuable visual insights into the decision-making process. This achievement aligned with the initial goal of enhancing the efficiency and effectiveness of healthcare systems while maintaining transparency and accountability.

6.2 Future Work

While the results obtained were satisfactory, there is still room for further research and development. We will now enumerate some of the potential measures to enhance this solution.

1. **Enhanced XAI Techniques**: With future research comes new and advanced XAI techniques and methodologies that can provide even more comprehensive and detailed explanations for

AI model predictions. An example of this is producing human-readable explanations through natural language processing.

2. **Visual Framework**: Employing a visual framework for this system is likely the next step in mind. This can help stakeholders analyse a single image by uploading it and analysing the visual explanation in real time.
3. **Real-world Integration**: Collaboration with healthcare professionals and institutions is essential to ensure seamless integration and practical utility.
4. **Human-AI Collaboration**: In terms of ground truth, having annotations and bounding boxes created by healthcare practitioners and making them evaluate the performance of the system will likely help to make the system more trustworthy and also fine-tune our model to better fit these accommodations. Having different radiologists label the pathologies in real time and checking the model results will be a great evaluation measure.
5. **Ethical Considerations**: We have to consider this a priority and keep continued attention to ethical considerations, particularly in terms of data privacy, fairness, and bias mitigation.
6. **Multi-modal Data Integration**: Exploring the integration of multiple data modalities, such as employing EHR and patient history, alongside our medical images, can enhance the overall diagnostic capabilities of the AI system.

In conclusion, this project has illuminated the path toward achieving transparency and interpretability in AI-driven medical image classification. We consider that the priority is to narrow the differences between AI's decision-making processes and human understanding, especially in terms of healthcare systems.

As the field of Explainable AI continues to evolve, we hope that the findings and insights presented here will contribute to the ongoing transformation of healthcare systems for the betterment of society.

References

[1] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "Darpa's explainable ai (xai) program: A retrospective," *Applied AI Letters*, vol. 2, no. 4, e61, 2021. DOI: https://doi.org/10.1002/ail2.61. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018. DOI: 10.1109/access.2018.2870052. [Online]. Available: https://doi.org/10.1109/access.2018.2870052.

[3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2016. DOI: 10.48550/ARXIV.1608.06993. [Online]. Available: https://arxiv.org/abs/1608.06993.

[4] P. Rajpurkar *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017. [Online]. Available: https://arxiv.org/abs/1711.05225.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019. DOI: 10.1007/s11263-019-01228-7. [Online]. Available: https://doi.org/10.1007%2Fs11263-019-01228-7.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Learning deep features for discriminative localization*, 2015. DOI: 10.48550/ARXIV.1512.04150. [Online]. Available: https://arxiv.org/abs/1512.04150.

[7] S. Basu, S. Mitra, and N. Saha, "Deep learning for screening covid-19 using chest x-ray images," *medRxiv*, 2020. DOI: 10.1101/2020.05.04.20090423. eprint: https://www.medrxiv.org/content/early/2020/05/08/2020.05.04.20090423.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/05/08/2020.05.04.20090423.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, *"why should i trust you?": Explaining the predictions of any classifier*, 2016. DOI: 10.48550/ARXIV.1602.04938. [Online]. Available: https://arxiv.org/abs/1602.04938.

[9] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," *CoRR*, vol. abs/1806.07421, 2018. arXiv: 1806.07421. [Online]. Available: http://arxiv.org/abs/1806.07421.

[10] D. Tranfield, D. Denyer, and P. Smart, "Towards a methodology for developing evidence-informed management knowledge by means of systematic review," *British journal of management*, vol. 14, no. 3, pp. 207–222, 2003.

[11] R. Toorajipour, V. Sohrabpour, A. Nazarpour, P. Oghazi, and M. Fischl, "Artificial intelligence in supply chain management: A systematic literature review," *Journal of Business Research*, vol. 122, pp. 502–517, 2020, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2020.09.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014829632030583X.

[12] D. Denyer and D. Tranfield, "Producing a systematic review," in *The Sage handbook of organizational research methods*, D. A. Buchanan and A. Bryman, Eds., Sage Publications Ltd., 2009, pp. 671–689.

[13] M. Gusenbauer and N. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed and 26 other resources [open access]," *Research Synthesis Methods*, vol. 11, pp. 181–217, Mar. 2020. DOI: 10.1002/jrsm.1378.

[14] C. Wohlin, M. Kalinowski, K. Romero Felizardo, and E. Mendes, "Successful combination of database search and snowballing for identification of primary studies in systematic literature studies," *Information and Software Technology*, vol. 147, p. 106 908, 2022, ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2022.106908. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584922000659.

[15] M. J. Page *et al.*, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021. DOI: 10.1136/bmj.n71. eprint: https://www.bmj.com/content/372/bmj.n71.full.pdf. [Online]. Available: https://www.bmj.com/content/372/bmj.n71.

[16] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, 2019,

ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2019.01.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718304846.

[17] L. Ibrahim, M. Mesinovic, K. .-. Yang, and M. A. Eid, "Explainable prediction of acute myocardial infarction using machine learning and shapley values," *IEEE Access*, vol. 8, pp. 210 410–210 417, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3040166.

[18] S. M. Lauritsen *et al.*, "Explainable artificial intelligence model to predict acute critical illness from electronic health records," *Nature communications*, vol. 11, pp. 1–11, 1 2020.

[19] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable ai meets healthcare: A study on heart disease dataset," 2020. DOI: 10.48550/ARXIV.2011.03195. [Online]. Available: https://arxiv.org/abs/2011.03195.

[20] H. Ren *et al.*, "Interpretable pneumonia detection by combining deep learning and explainable models with multisource data," *IEEE Access*, vol. 9, pp. 95 872–95 883, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3090215.

[21] D. Chakraborty *et al.*, "Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer," *Cancers*, vol. 13, p. 3450, 14 2021, ISSN: 20726694. DOI: 10.3390/cancers13143450. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=a9h&AN=151564794&site=ehost-live&scope=site.

[22] I. de Sousa, M. M. B. R. Vellasco, and E. da Silva, "Explainable artificial intelligence for bias detection in covid ct-scan classifiers," *Sensors (14248220)*, vol. 21, p. 5657, 16 2021, ISSN: 14248220. DOI: 10.3390/s21165657. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=a9h&AN=152146211&site=ehost-live&scope=site.

[23] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, e745–e750, 11 2021, ISSN: 2589-7500. DOI: https://doi.org/10.1016/S2589-7500(21)00208-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589750021002089.

[24] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "Xvitcos: Explainable vision transformer based covid-19 screening using radiography," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–10, 2022, ISSN: 2168-2372. DOI: 10.1109/JTEHM.2021.3134096.

[25] V. Vishwarupe, P. M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, and V. Pawar, "Explainable ai and interpretable machine learning: A case study in perspective," *Procedia Computer Science*, vol. 204, pp. 869–876, 2022, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.08.105. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922008432.

[26] M. A. Khan *et al.*, "Covid-19 classification from chest x-ray images: A framework of deep explainable artificial intelligence," *Computational Intelligence & Neuroscience*, pp. 1–14, 2022, ISSN: 16875265. DOI: 10.1155/2022/4254631. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=a9h&AN=157989308&site=ehost-live&scope=site.

[27] D. Saraswat *et al.*, "Explainable ai for healthcare 5.0: Opportunities and challenges," *IEEE Access*, vol. 10, pp. 84 486–84 517, 2022. DOI: 10.1109/ACCESS.2022.3197671.

[28] F. Giuste *et al.*, "Explainable artificial intelligence methods in combating pandemics: A systematic review," *IEEE Reviews in Biomedical Engineering*, pp. 1–17, 2022. DOI: 10.1109/RBME.2022.3185953.

[29] M. Sedighi, "Application of word co-occurrence analysis method in mapping of the scientific fields (case study: The field of informetrics)," *Library Review*, vol. 65, pp. 52–64, Feb. 2016. DOI: 10.1108/LR-07-2015-0075.

[30] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, 2nd ed. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[32] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, Jul. 2015. DOI: 10.1371/journal.pone.0130140. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. DOI: 10.1609/aaai.v32i1.11491. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11491.

[34] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," 2017. DOI: 10.48550/ARXIV.1711.00399. [Online]. Available: https://arxiv.org/abs/1711.00399.

[35] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, "Towards a joint approach to produce decisions and explanations using cnns," in *Iberian Conference on Pattern Recognition and Image Analysis*, Springer, 2019, pp. 3–15.

[36] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017. DOI: 10.48550/ARXIV.1702.08608. [Online]. Available: https://arxiv.org/abs/1702.08608.

[37] D. E. Knuth, *Selected Papers on Computer Science*. Stanford, California, USA: Stanford University, 2003.

[38] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, 2021, ISSN: 2296-598X. DOI: 10.3389/fenrg.2021.652801. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801.

[39] *Diagnostic Radiology Physics* (Non-serial Publications). Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY, 2014, ISBN: 978-92-0-131010-1. [Online]. Available: https://www.iaea.org/publications/8841/diagnostic-radiology-physics.

[40] TensorFlow Authors. "ImageDataGenerator." [Online; accessed 30-October-2023]. (Accessed 2023), [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator.

[41] K. Popper, *The Logic of Scientific Discovery*. Basic Books, 1959.