INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




EVALUATION OF EXPLAINIBILITY AI (XAI) TECHNIQUES FOR MITIGATING ETHICAL AND LEGAL CHALLENGES

RAFAEL PORCIDONIO FERNANDES ESQUIÇATO
julho de 2025

P.PORTO

EVALUATION OF EXPLAINIBILITY AI (XAI) TECHNIQUES FOR MITIGATING ETHICAL AND LEGAL CHALLENGES

Rafael Porcidonio Fernandes Esquiçato

Student nº: 1230375

Thesis submitted for the degree of
Master in Artificial Intelligence Engineering

Supervisor: Maria Goreti Carvalho Marreiros, Full Professor at the School of Engineering – Polytechnic Institute of Porto

Examination committee:

President:

Luís Manuel Silva Conceição, Adjunct Professor at the School of Engineering – Polytechnic Institute of Porto

External Examiner: Dalila Alves Durães, Assistant Professor at the University of Minho

Porto, June 2025

Dedication

To my father, **Laércio**, for doing everything he could to support and help me during all the steps of my career, since the rides to the University a long time ago until helping to organize things in my birth city while I'm living abroad.

To my mother, **Isabel**, for instilling in me from an early age the value of education, the importance of continuous learning and also the importance of being ambitious, believing that with good preparation it's possible to achieve things.

And to my wife, **Ana Cláudia**, for her patience, encouragement, and unwavering support during the many moments of uncertainty and indecision throughout this journey. Doing this master program after a 10-year academic break wasn't easy but her support was fundamental to make me believe that would be possible and to keep me on the track and achieve this important milestone.

I love each of you and know how you three cheers on me and are happy with achievement!

Abstract

The integration of Artificial Intelligence (AI) into healthcare systems raises significant ethical and legal concerns. This study investigates how Explainable AI (XAI) techniques can enhance the transparency and trustworthiness of medical image classification systems. Through a systematic literature review of 860 papers and experiments using COVID-19 radiography and skin lesion datasets, the research identifies and evaluates XAI methods such as Grad-CAM, SHAP, and ABELE. These methods were assessed for their ability to clarify decision-making processes, improve model accountability, and support regulatory compliance. The study proposes an explainability module that combines different techniques to provide human-readable explanations, aiming to bridge the gap between AI predictions and clinical trust. Findings indicate that XAI not only addresses transparency and bias issues but can also improve diagnostic performance and decision support in critical applications.

Keywords: Explainable AI, Healthcare, Image Classification, Deep Learning, Medical Ethics, Transparency

Resumo

A aplicação de Inteligência Artificial (IA) em sistemas médicos, especialmente em diagnóstico por imagem, levanta importantes questões éticas, legais e de confiança devido à natureza opaca dos modelos de aprendizagem profunda. Este trabalho investiga como técnicas de Inteligência Artificial Explicável (XAI) podem mitigar esses desafios, promovendo maior transparência e interpretabilidade. A pesquisa realiza uma revisão sistemática da literatura, analisando 860 artigos recentes, e identifica como a falta de explicações compreensíveis afeta diretamente a adoção desses sistemas na prática clínica.

Além da revisão, foram desenvolvidos e treinados dois modelos baseados em redes neurais convolucionais (CNNs), utilizando conjuntos de dados reais: um de radiografias torácicas para detecção de COVID-19 e outro de lesões de pele. Sobre esses modelos, aplicaram-se três técnicas de XAI – Grad-CAM, SHAP e ABELE – com o objetivo de explicar as decisões do sistema e identificar quais regiões ou atributos foram mais relevantes para cada predição.

Os resultados mostram que as técnicas de XAI melhoram a confiança dos usuários ao fornecerem explicações visuais ou estatísticas compreensíveis, auxiliando na detecção de viés, na melhoria da performance do modelo e na conformidade com regulamentos como o GDPR. Além disso, a explicabilidade pode ser integrada como um módulo funcional nos sistemas de diagnóstico, ampliando sua utilidade clínica.

Conclui-se que XAI desempenha um papel fundamental na construção de sistemas de IA mais éticos, transparentes e confiáveis no contexto da saúde. Este trabalho contribui ao propor uma abordagem prática que combina diferentes métodos explicativos, com potencial para impactar positivamente a adoção de IA em ambientes médicos reais.

Palavras-chave: Inteligência Artificial Explicável, Saúde, Classificação de Imagens, Deep Learning, Ética, Transparência.

Acknowledgement

I would like to thank the School of Engineering of the Polytechnic Institute of Porto for providing the academic environment and resources that supported the development of this work.

I am grateful to all the professors who contributed throughout the program with their knowledge and guidance. I would like to express my particular appreciation to Professor Goreti Marreiros, my supervisor, for her valuable support and supervision during this project.

I also acknowledge the support of Professor Carlos Ramos, the program director, whose assistance was especially important during the enrolment process.

Finally, I would like to thank my parents — my father, Laércio, and my mother, Isabel — for their continued support throughout my academic and professional path. I am also especially grateful to my wife, Ana Cláudia, for her constant support throughout this stage of my studies.

Summary

List of Images

List of Tables

List of Code Blocks

Abbreviations

AAE	Adversarial Auto-Encoder
ABELE	Adversarial Black Box Explainer Generating Latent Exemplars
AI	Artificial Intelligence
CAM	Class Activation Mapping
CBR	Case-Based Reasoning
CNN	Convolutional Neural Network
DL	Deep Learning
DSR	Design Science Research
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
KNN	K-nearest neighbours
ISIC	The International Skin Imaging Collaboration
LLM	Large Language Model
LRP	Layer-wise Relevance Propagation
LIME	Local Interpretable Model-agnostic Explanation
LORE	Local Rule-Based Explanations
ML	Machine Learning
MRI	Magnetic Resource Imaging
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations

1 Introduction

This chapter will introduce the explainability concept, how the existing literature defines it and the importance of this concept to bring transparency to Artificial Intelligence (AI) systems. Moreover, it will be presented the justification behind the decision of researching this topic in the medical area and the goals for this work in terms of research and experimentation.

1.1 Explanibility concept

Explainability is related to the concept of interpretability, which is the measure of how understandable something is for a human. It can be applied for many areas, such as math formula, a piece of code and a machine learning model. According to (Molnar, 2024), the measurement of interpretability or explainability of a machine learning model is directed connected to how easy is for a human comprehend its output according to the data set used to train that model. Machine learning models use a mix of statical methods to sort out a prediction when fed with new data, if it's hard for a human comprehend how some outputs are being generated, we can say that this model is less explainable.

It's important to refer that (Molnar, 2024) and (Dunn et al., 2021) use explainability and interpretability as synonyms but it's not consensual. (Amazon Web Services, n.d.), for example, differs both concepts stating that Interpretability is related to observe the internal mechanisms of a machine learning and interpretate the model's weights and features that determined a given model output. The same article also highlights that implementing those kinds of observations usually decreases the model performance. They also state that Explainability is connected to making a given output human understandable, for example, using methods to correlate instances in the dataset or specific features to a given output. This work will keep the focus on Explainability discussing the current state for methods that help humans to understand

output of machine learning models including deep-leaning and generative artificial intelligence (or GenAI).

(Hauptman et al., 2024) reinforces that explainability information extracted from any model is just valuable if understandable by humans. It reinforces the user-centric character of XAI, what means that any method must produce relevant information readable and interpretable by a human. They also bring the concept of level of explainability, low-level information refers to basic information about the decision, usually mentioning the algorithm behinds that while high-level information refers to more detailed explanations about the entire process, including the decision logic.

(Adadi & Berrada, 2018) also corroborates the connection between explainability and human readable justifications in machine learning models. According to the paper, XAI is a research field that aims to make AI systems results more understandable to humans. They also provide a historical context about the term that was mentioned for the first time by (Lent et al., 2004) in a study about a training system developed by U.S. Army. Anyway, they also reinforce that the need for comprehendible reasons behind some decisions was also a concern for expert system in the 1970's. Rule based systems were the first artificial intelligence commercial programs and the explainability module providing the "why's" behind any system output brought confidence to the systems adoption. The term has gained notability decades after this period according to modern artificial intelligence models became more popular. XAI methods have their importance related to the fact that as more understandable AI models output more trustworthy, they are.

(Adadi & Berrada, 2018) correlates XAI to the concept of Responsible Artificial Intelligence that has three main pillars: accountability (AI systems must justify their decisions to users, partners or other systems), responsibility (AI systems must be able to answer for a specific output and identify possible errors), and transparency (AI system must be inspectable having their processes and internals reproduceable).

Visibility about the reasons behind machine learning models is also relevant to bring more foundations to the knowledge acquired according (Molnar, 2024). Predict the likelihood of a client churn is important but more than it, is to understand the reasons behind an eminent churn. Scientifically, knowledge is only created with the proper justifications and openness about the learning and experimentation process, what means that a new machine model can just be considered dependable if their process and mechanisms are transparent.

Ethical aspects such as bias detection and potentially reduction are also another point that XAI helps to address. Gaining some understanding on how a model generalizes, which concepts are more relevant, which instances are more influential in the output can help to identify unexpected bias and plan actions to mitigate then: review the data set, rethink the architecture, etc. This aspect is highlighted by (Adadi & Berrada, 2018), and (Hauptman et al., 2024).

1.2 Health-care systems need more than performance metrics

According to (Molnar, 2024), explainability can be more or less important according to the criticality of the decision made by the system. While an e-commerce recommendation system can be reliable if provides meaningful recommendations even without deeper explanations, the same thing cannot be said for medical systems for example. (Longo et al., 2024) corroborates this affirmation stating the usage of AI in medical system increases the need for explanations, it offers support for doctors and clinicians to improve the decision-making process in healthcare, communicating diagnosis, selecting treatments. It's clear that any decision in that field can impact seriously the life of a patient, a wrong diagnosis (false-positive or false-negative) will be really impacting for anyone. It urges the need of more than just a system output for medical systems.

Yet about health-care systems, (Talia, 2022) presents two cases of class lawsuits being filled against health insurance providers in the United States due to high error rate for a machine learning algorithm that calculates the required number of rehabilitation sessions according to the patient condition. The system, that has been used by both providers to deny rehabilitation payments to the users impacting their daily live, lacks transparency not disclosing details behind the decisions that could help doctors and patients to understand better some conclusions.

The scenario for health-care systems using AI is that only the normal performance metrics (accuracy, recall, etc) are not enough to make people trust them to make decisions that will impact the course of a medical treatment. As described in the last two paragraphs, medical system makes decisions that can affect directly people lives. To be considered sufficiently reliable to be adopted a system must somehow produce some evidence that help people to understand how some decisions were made.

This special need (not exclusive for health care) justifies why, in this work, medical area will be used to experiment and explore how XAI could be a possible solution for this problem and accelerate the adoption of AI in medical systems. Being more specific, diagnosis per image systems will be the target in terms of experimentation and assessment.

1.3 Objectives

This work aims to analyse how XAI techniques have been applied into medical systems to mitigate the trustworthy and ethical issues in their adoption. The focus will be machine learning models based on images commonly used for diagnosis such as image classification and generative AI models. The hypothesis is that XAI can be integrated to medical systems to support and disclose details of the decision-making process improving the trustworthiness of the system. In summary, below are enlisted the main hypothesis and goals for this work:

Hypothesis: Can XAI help to mitigate legal, ethical and trustworthiness issues in medical imaging system?

1. Identify how the available literature addresses the problem
2. Identify potential relevant XAI methods to be experimented according to the goal 1
3. Identify potential medical data sets/DL models could be used to experiment the XAI selected in the goal 2
4. Experiment the selected XAI methods using the selected data sets
5. Produce an assessment about how each of the methods could or not improve solve mentioned issues and also be integrated to real systems

1.4 Methodology

This work will follow the Design Science Research (DSR) methodology to better organize the steps and leverage the produced outcomes. According to (vom Brocke et al., 2020), "Design Science Research is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts. Simply stated, DSR seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiate". The DSR process can be defined as sequencing of steps presented in the following paragraphs.

Activity 1 - problem identification and motivation: this is the initial activity where the problem space is defined along with theorical justification and relevance corroborated by the existing literature. In the chapter 1 Introduction, the explainability concept is described along with the basis to justify the decision of investing time writing about this concept and what kind of problems it might solve.

Activity 2 - define the objectives for a solution: after the problem definition, the next step is to set the goals for the study that could support a solution to address the problem identified in the initial stage. In the chapter 1 Introduction > 1.3 Objectives, the study goals were presented to explore XAI concept and methods.

Activity 3 - design and development: one or more artifacts are designed and crafted to meet the goals described in the last activity and solve or address problems identified so far. For this item, the intention is to identify a couple of relevant XAI methods, experiment them and produce a comprehensive comparison with focus on how they could leverage the trustworthiness in medical system using AI. Chapters 3, Methods and Materials, and 4, Implementation, are the places where this artifact will be designed and elaborated.

Activity 4 - demonstration: this activity demonstrates the usage of the artifact to solve problems defined in the beginning of the study. Chapter 4 Implementation will present the XAI methods exploration and their outcome for two machine models.

Activity 5 - evaluation: the evaluation measures how well the artifact supports a solution, the Chapter 5 Conclusion will present an evaluation on each method do solve the transparency problems related to AI systems in medical care.

Activity 6 – communication: here all aspects of the problem and the designed artifact are communicated to the relevant stakeholders.

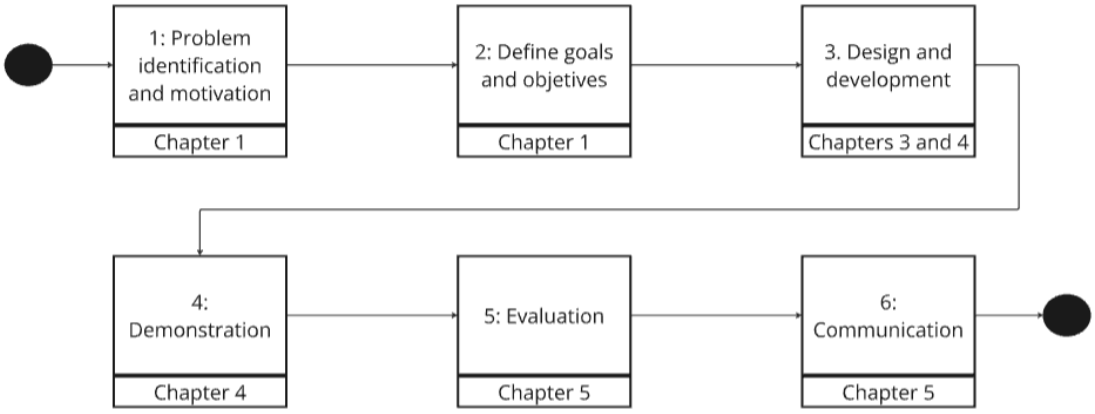

Figure 1 DSR methodology workflow

1.5 Contributions

This section aims to present the academic contributions that this study might have. This work contains a systematic literature review answering research questions about how the lack of transparency affects the usage of AI in medical image systems, the XAI techniques in use to try to face these challenges and how XAI has been integrated into medical system. During the systematic review process, 860 papers from 4 renowned and relevant sources were analysed to provide a literature panorama about how XAI is impacting healthcare systems.

The experimentation of some relevant XAI methods (according to the systematic review) to explain two common deep-learning models containing technical details about the implementations along some assessment about their effectiveness to improve the trustworthiness of AI medical solutions providing human readable explanations is another contribution.

The proposition of an explanation module to be integrated in image health-care diagnosis systems is another contribution to the community looking for options to improve the transparency of medical systems.

Moreover, there is the intention of publishing potentially two articles in scientific publications: one first paper with the systematic review produced as the theoretical part of this study and another one to expose the experimentation process and results.

1.6 Document Strucuture

The first chapter – Introduction – presents the explainability concept, defines the problem, goals, methodology and contributions.

The second chapter – State-of-the-art – presents a systematic review of the literature about the impact of XAI in healthcare.

The third chapter – Methods and materials – presents the XAI methods experimented, the justification, Deep-Learning models and datasets used.

The fourth chapter – Implementation – presents the experimentation process of XAI methods to explain and bring light to the decisions produced by the ML model. Technical details and evidence about each XAI method are presented.

The fifth chapter – Conclusion – presents the evaluation of each XAI method on how they can be used to produce effective explanations. An assessment about the methods effectiveness is presented and a comparison between them. Future works and possible improvements are also presented.

2 State-of-the-art

This section aims to explore the existing literature looking for answers to our research questions. The methodology is presented introducing the questions and data sources, the chapter continues with the papers selection and finishes with a discussion providing answers to each question according to the explored papers.

2.1 Methodology

The next pages will describe the steps chosen to conduct this systematic review. Starting by the questions and data sources and wrapping-up with the papers selection workflow.

2.1.1 Research questions

This research aims to investigate the influence of XAI in the healthcare industry with focus in solutions using diagnosis by image. As stated before, the adoption of AI in critical use-cases as healthcare demands more than just the system output or normal accuracy and performance metrics, details about the decision-making process are welcomed to make the system more reliable for clinicians and patients. The selected questions aim to capture the panorama of how XAI has been used in healthcare systems to enhance the trustworthiness of their decisions and address legal and ethical issues caused by their adoption. The first question “RQ1: what are the most recurrent trustworthy challenges identified?” aims to identify which issues are commonly cited in the examined papers setting the focus on how it affects the AI systems adoption. The second question “RQ2: how can XAI leverage trustworthy for image classification use cases?” aims to gather which XAI methods have been used to address transparency issues in healthcare image systems based on classifiers. The third question “RQ3: are there use-cases using Gen-AI being applied to leverage reliability of image system models?” aims to focus on how Gen-AI techniques have been contributed to enhance the performance and reliability of medical

imaging systems. The fourth question "RQ4:how has XAI been integrated to the current medical systems?" aims to investigate how XAI methods have been (or might be) integrated to health care systems beyond academical methods analysis.

Table 1 Research questions

	Research Question
RQ1	what are the most recurrent trustworthy challenges identified?
RQ2	how can XAI leverage trustworthy for image classification use cases?
RQ3	are there use-cases using Gen-AI being applied to leverage reliability of image system models?
RQ4	how has XAI been integrated to the current medical systems?

2.1.2 Definition of search strategy

To answer the research questions, the literature available for medical imaging systems was investigated. The process started with the definition of relevant sources to this topic, definition of search terms, definition of inclusion and exclusion criteria, papers selection and, in the end, data extraction.

2.1.3 Definition of sources

This research has considered 4 electronic sources (Table 2) considering works published in journal articles, papers and conference proceedings.

Table 2 Data sources

Identifier	Database	URL
ED1	PubMed	https://www.ncbi.nlm.nih.gov/pubmed/
ED2	Science Direct	https://www.sciencedirect.com/
ED3	IEEE Explore	https://ieeexplore.ieee.org/Xplore/home.jsp
ED4	Web Of Science	https://www.webofscience.com/wos/woscc/basic-search

2.1.4 Inclusion and exclusion criteria

Inclusion and exclusion criteria can be found in the Table 3 and Table 4.

Table 3 Inclusion criteria

Identifier	Criteria
IC1	Work in the medical (or healthcare) field
IC2	Work mentioning XAI methods
IC3	Work mentioning experimentation
IC4	Work applying ML image models

Table 4 Exclusion criteria

Database	URL
EC1	Work not written in English
EC2	Work not written in 2023, 2024 and 2025.
EC3	Work is a systematic review
EC4	Work besides journal articles, papers and conference proceedings.
EC5	Work duplicated in other sources

2.1.5 Definition of search queries

The search queries selected to be used in the 4 databases can be found in the Table 5 and Table 6.

Table 5 Search queries for image classification problems

Scope	Query
XAI	("explainable artificial intelligence" OR "xai") AND
Image Classification	("image analysis" OR "image classification") AND
Healthcare	("medical" OR "healthcare" OR "health-care")

Table 6 Search queries for Gen-AI use-cases

Scope	Query
XAI	("explainable artificial intelligence" OR "xai") AND
Gen-AI	("text-to-image" OR "synthetic image" OR "image generation") AND
Healthcare	("medical" OR "healthcare" OR "health-care")

2.1.6 Papers selection

The initial step was querying in the data sources listed in the Table 2 using the search terms identified in Table 5 and Table 6. The data sources allow to filter by year and result type allowing to reduce the initial set of papers according to EC2 and EC4.

To organize the papers triage and selection, the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) methodology was applied according to the Figure 2 to select the studies to be analysed. This methodology states four stages: Identification, Screening, Eligibility and Inclusion. During identification phase 860 articles have been found. In the following, during the Screening phase, 91 articles have been removed due to duplications (according EC5) and 574 articles have been removed after papers' abstracts analysis considering the inclusion and exclusion criteria (Table 3 and Table 4). During the Eligibility phase, 125 papers have been excluded after full-text analysis considering the same criteria from the last phase. In the end, 70 articles have been selected and considered to answer the research question.

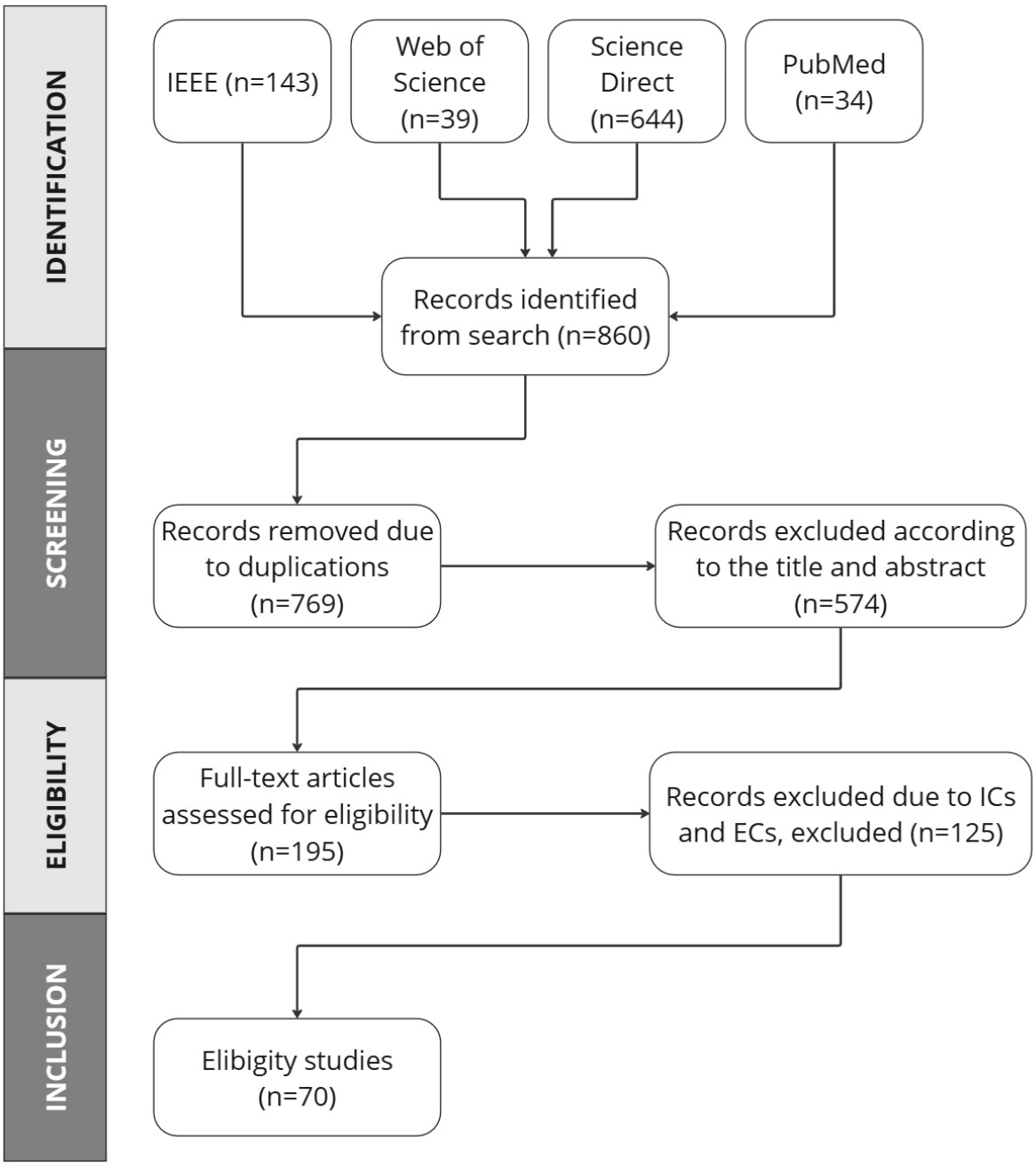

Figure 2 Papers selection process using PRISMA

2.2 Results

2.2.1 RQ 1: What are the most recurrent trustworthy challenges identified?

After the papers analysis, it was possible to list the most common challenges to be overcome while adopting AI solutions based in the healthcare area. As stated before, these open questions

affect the trustworthiness of any inference produced by AI system. The literature provides an overview of how ethical and legal concerns affect the application of AI in healthcare.

Transparency and lack of explainability was referred by several papers as a recurrent obstacle according to the research. (Alkhalaf et al., 2023) state that decision-making by AI methods is fundamentally a black-box procedure making it problematic for doctors to determine the validity of their choices. (Wickstrøm et al., 2023) refer that CBIR (Content based image retrieval) systems also suffer from an intrinsic lack of explainability that may have detrimental effects in a clinical setting, as long deep learning-based systems are known to mix several factors and artifacts to make their predictions. The need of validation and the lack of proper tools in cancer diagnosis tools is also mentioned in (Burgos et al., 2024). (Akay et al., 2023) also corroborate the need of proper explanations to the clinical staff in their stroke detection model. This can be clearly defined as the most recurrent challenge according to the research, many other papers refer the same, the need of explainability to bring transparency and improve trustworthiness of patients and doctors in the systems output, it was also reported by (Lamba & Rani, 2024), (Bouabdallah et al., 2024), (Singh et al., 2023), (Shojaei et al., 2023), (Song et al., 2023), (T. Mahmud et al., 2024), (Mahamud et al., 2024), (Srinivasu et al., 2024), (F. Mahmud et al., 2023), (T R et al., 2024), (Schweizer et al., 2023), (M. Wang et al., 2023), (Oliveira et al., 2024), (Deepanshi et al., 2023), (Pisarcik et al., 2024), (Akbar et al., 2024), (Heng & Abdul-Kadir, 2023), (Thiruvenkadam et al., 2024), (Kim et al., 2024), (R. Rahman et al., 2023), (Akhlaq et al., 2024), (Lo et al., 2024), (Thapar & Tiwari, 2024), (Mukhtorov et al., 2023), (Bardozzo et al., 2024), (Z. Wang et al., 2025), (Islam et al., 2023), (Rahimiaghdam & Alemdar, 2024), (Shaheema et al., 2024), (Alami et al., 2024), (Pereira et al., 2024), (Wester Trejo et al., 2024), (Zahoor et al., 2023), (Barua et al., 2023) (Harikumar et al., 2024), (Ghnemat et al., 2023), (Singhal et al., 2024), (Mu et al., 2024), (Laguna et al., 2023), (Hroub et al., 2024), (Hussain et al., 2024), (Carloni & Colantonio, 2024), (Grillo et al., 2024), (L. Wang et al., 2023) , (Chaddad et al., 2024), (Grillo et al., 2024), (Vairetti et al., 2024), (Alomar et al., 2023), (Jiang et al., 2024), (Huang et al., 2024), (Ullah et al., 2024), (Rahim, El-Sappagh, et al., 2023), (Saeed et al., 2024), (Flores-Araiza et al., 2024), (Latha et al., 2024), (Parola et al., 2024) (Rahim, Abuhmed, et al., 2023), (Biswas et al., 2024), (Nikolić et al., 2024), (Hassan et al., 2024), (Ellis et al., 2023), (Nazir et al., 2024), (Tanone et al., 2025) and (Metta, Beretta, Guidotti, Yin, et al., 2024).

Bias mitigation was also mentioned by several papers. (Lamba & Rani, 2024) highlighted the need for techniques, such as XAI, that could help to identify bias and error sources. (Metta, Beretta, Guidotti, Yin, et al., 2024) also corroborate the need of tools to detect and mitigate bias. (Veetil et al., 2024) focused on the bias and data leakage identification in MRI classification models highlighting the importance of these topics in medical AI solutions. (Shojaei et al., 2023) also highlight the need of tools to combat to detect bias and model overfitting. It was also referred by (Patel et al., 2024), (Singhal et al., 2024), (Laguna et al., 2023), (Hussain et al., 2024), (Chaddad et al., 2024), (Huang et al., 2024), (R. Rahman et al., 2023), (Hassan et al., 2024), (Nikolić et al., 2024), (Latha et al., 2024), (Nazir et al., 2024) and (Pisarcik et al., 2024).

(Wickstrøm et al., 2023) refer in their study about liver diseases prediction that CBIR systems requires labelled data that can be costly and time consuming. The same challenge, the need of

trustable and sufficient data to train a ML model is also referred by (M. Wang et al., 2023). The problem with imbalanced datasets is also referred by (Ukwuoma et al., 2023). (Bardozzo et al., 2024) also refer the need of quality data (in their case, images) to achieve good results in a diagnosis per image system. It's also referred by (Ghnemat et al., 2023) as obstacle to achieving good results in their COVID-19 detection using lungs images model.

Data privacy is mentioned in (Metta, Beretta, Guidotti, Yin, et al., 2024) claiming that medical image data is sensitive and requires safe manipulation and handling. Data protection is also referred as concern by (T R et al., 2024), according to the paper it requires robust data anonymization, secure storage and controlled access. (Mu et al., 2024) mention the same challenge applying Federation Learning and Blockchain to mitigate risks. It's also referred by (Oliveira et al., 2024), (Laguna et al., 2023), (De Aguiar et al., 2024), (Chaddad et al., 2024), (Nazir et al., 2024) and (Singhal et al., 2024).

Compliance with regulation and standards is mentioned in (Metta, Beretta, Guidotti, Yin, et al., 2024) stating that medical AI systems must comply with government and industry standards what can be costly and labour-intensive. (T R et al., 2024) mention the need of compliance with privacy laws like GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act. GDPR and other regulations are also mentioned by (Oliveira et al., 2024) and (Nikolić et al., 2024).

Table 7 Challenges summary

Challenge	**Number of papers**
Transparency and trustworthiness	67 (95,71%)
Bias detection and mitigation	17 (24,29%)
Privacy in the data handling	9 (12,86%)
Quality data	4 (5,71%)
Compliance	4 (5,71%)

2.2.2 RQ 2: How can XAI leverage trustworthy for image classification use cases?

The analysed papers present a list of methods and techniques in the field of XAI that can be used to provide some feasible explanations and justifications for healthcare image classification models. The research provides a rich panorama of how those methods has been applied for diagnosis of many diseases on top of many deep learning models.

Grad-CAM (Gradient-weighted Class Activation Mapping) method and variations are the top cited method according to the literature review. (Burgos et al., 2024) use Grad-CAM (Gradient-weighted Class Activation Mapping) in their study about deep learning models for cancer diagnosis. The method is applied on top of a CNN (Convolutional Neural Network) model used to classify exam images (such as mammograms) as possible positive or negative cancer diagnosis. The last layer of a CNN model is used to derive semantic information that is used by the XAI algorithm to create an activation map. (Bouabdallah et al., 2024) bring a different perspective in the usage of Grad-CAM with CNN, they proposed a thorax lesion diagnosis model where Grad-CAM is used on top of a first classification model (healthy or unhealthy) to identify

areas in unhealthy radiographies where is more likely to see specific micro-lesions. That second classification supported by Grad-CAM is, in the end, used to catalogue micro-lesion images.

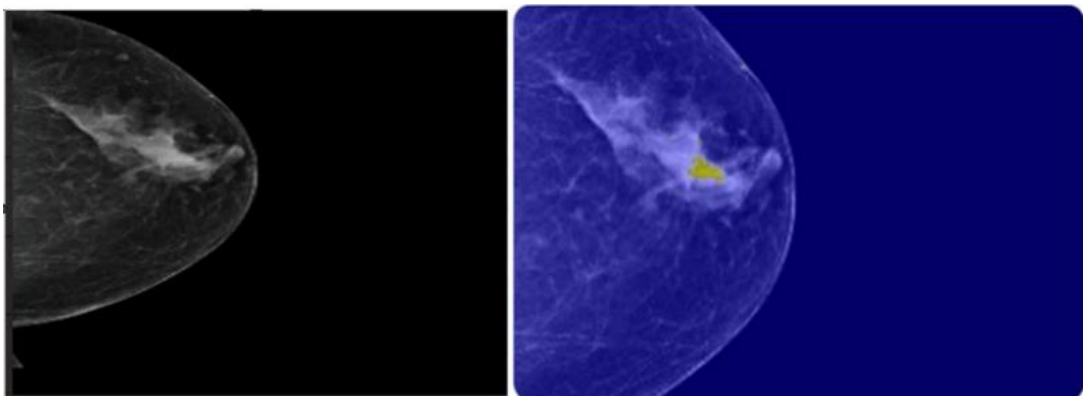

Figure 3 Left: mammogram, right: the same mammogram with Grad-CAM (Burgos et al., 2024).

(Song et al., 2023) also invest in the same architecture, using Grad-CAM to explain a CNN model but the difference is that they propose a new explainer model trained on top of the predictions and the heat maps to identify key features achieving better saliency maps. One of the reasons, according to the authors, is that their Explainer model is trained with inputs not only from the last layer (as the Grad-CAM) but also from the intermediate layers, in the Figure 4 it's possible to check the explainer architecture.

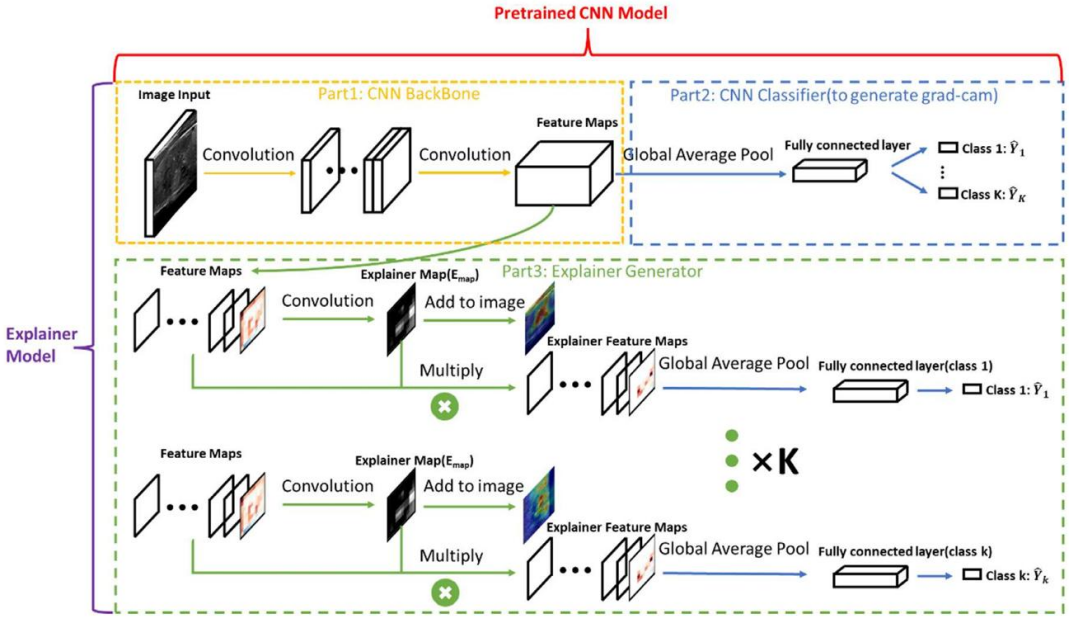

Figure 4 Explainer CNN model trained on top of Grad-CAM heatmaps (Song et al., 2023).

Grad-CAM or some variation are also cited in many more papers analysed (Lamba & Rani, 2024), (Veetil et al., 2024), (T. Mahmud et al., 2024), (Patel et al., 2024), (T R et al., 2024), (Schweizer et al., 2023), (M. Wang et al., 2023), (Ukwuoma et al., 2023), (Thiruvenkadam et al., 2024), (R.

Rahman et al., 2023), (Lo et al., 2024) , (Z. Wang et al., 2025), (Islam et al., 2023), (Shaheema et al., 2024), (Alami et al., 2024), (Wester Trejo et al., 2024), (Zahoor et al., 2023), (Carloni & Colantonio, 2024), (Flores-Araiza et al., 2024), (Jiang et al., 2024), (Rahim, El-Sappagh, et al., 2023), (Alomar et al., 2023), (De Aguiar et al., 2024), (Nikolić et al., 2024), (Latha et al., 2024), (Rahim, Abuhmed, et al., 2023), (Biswas et al., 2024), (Nazir et al., 2024) and (Shojaei et al., 2023).

(Hroub et al., 2024) compare five vision-based explainability algorithms (GradCAM, GradCAM++, EigenGradCAM, AblationCAM, and Random CAM) in a lung diagnosis model analysis to provide local explanation on interest areas. Grad-CAM++ algorithm is applied to a CNN model in (Akay et al., 2023) in their stroke diagnosis image system using SmoothGrad to produce saliency maps, (F. Mahmud et al., 2023) also present a model using the same composition (CNN+SmoothGrad). Grad-CAM++ plus CNN is also the choice for (Deepanshi et al., 2023). (Thapar & Tiwari, 2024) have applied Score-CAM and SmoothGrad as XAI mechanisms to explain their CNN model in their study about cancer lesion classification. (Mukhtorov et al., 2023) explored a list of Grad-CAM variations, such as Grad-CAM++, Layer-CAM, Hires-CAM and XGrad-CAM to explain their endoscopic image classification model. (Akhlaq et al., 2024) refer the usage of a YOLO object detection model in study of bone anomalies but using a variation of Grad-CAM, EigenCAM, to explain the model. According to the paper, EigenCAM generates heatmaps for each model layer to highlight the areas the more active areas. (Chaddad et al., 2024) also applie multiple CAM methods variations (such as XGrad-CAM, Layer-CAM and Grad-CAM++).

(Alkhalaf et al., 2023) present a cancer classification model built using RNN (Recurrent Neural Network) and applying LIME (Local interpretable model-agnostic explanation) as XAI method. The authors state that this method provides a clear explanation for black-box classifiers. The algorithm works by perturbing the input dataset (removing specific words, hiding part of the image, and adding random noise) and seeing how prediction changes. One advantage of this method is that it's model agnostic working on top of variations in the dataset and analysis of the correspondent variation in the output. LIME is also the choice in (Kim et al., 2024) at their ADPKD (Autosomal Dominant Polycystic Kidney Disease) model study using DCNN (Diffusion-CNN). CNN and Lime is also the choice of (R. Rahman et al., 2023), (Islam et al., 2023), (Rahimiaghdam & Alemdar, 2024), (Pereira et al., 2024), (Zahoor et al., 2023), (Alami et al., 2024), (Harikumar et al., 2024), (Ghnemat et al., 2023), (Alomar et al., 2023), (Laguna et al., 2023), (De Aguiar et al., 2024), (Ullah et al., 2024), (Nikolić et al., 2024), (Hassan et al., 2024), (Nazir et al., 2024), (Biswas et al., 2024) and (Lo et al., 2024).

(Singh et al., 2023) present a different approach, they propose a framework using PyRadiomics (a Python package to extract features from radiographies images) and make usage of a random forest classifier to predict the diagnosis. They use SHAP (SHapley Additive exPlanations) method to find out which features was more important in each prediction, helping to explain the system output. The usage of SHAP on top of a CNN model is also referred by (Srinivasu et al., 2024) in their study about diabetes diagnosis. It was also the choice in (Lo et al., 2024), (Alami et al., 2024), (Nazir et al., 2024), (De Aguiar et al., 2024), (Biswas et al., 2024), (Zahoor et al., 2023), (Tanone et al., 2025) and (Islam et al., 2023).

(Mahamud et al., 2024) present a mixed approach where they make a mix of multiple explainability models: SHAP, LIME, Grad-CAM and Grad-CAM++, to provide explanations on how their lung diseases classifier inferences. According to the authors, SHAP and LIME provided insights on the decision-making process while the Grad-CAM and Grad-CAM++ was used to present how specific areas in the radiographies activated the classifier.

(Hussain et al., 2024) refer the usage of seven different interpretation techniques (Saliency, XRAI, Integrated Gradients, Smooth Gradients, Smooth Integrated Gradients, Grad-CAM, Smoothgrad-CAM) to explain their CNN model. The results are that Saliency Maps were able to match closer the experts annotations.

(Pereira et al., 2024) benchmark multiple XAI methods to evaluate how they could provide explanations for two different CNN chest X-Ray classifiers (DenseNet121 and ResNet50). GradCAM, Grad CAM++, EigenGrad-CAM, Saliency maps, LRP and DeepLift were applied and compared with radiologists' annotations. According to the paper, Grad-CAM++ and Saliency methods offer the most accurate explanations and that the effectiveness of visual explanations is found to vary based on the model and corresponding input size. Grad-CAM and Saliency maps are also the choice in (Barua et al., 2023).

Grad-CAM and Guided Grad-CAM are referred by (Vairetti et al., 2024) in their study about retina images analysis. According to the paper, Grad-CAM offers a coarse visual justification, highlighting regions where retinal damage may be located, such as the central zone of the macula or fovea centralis while Guided Grad-CAM provides a fine-grained visualization of the different retinal layers. They also mention this two-step analysis is equivalent to the diagnostic process of an ophthalmologist: first examining the area of interest, then inspecting the retinal layers.

(Oliveira et al., 2024) present a novel where multiple XAI methods are utilized to benchmark the performance of many pre-trained MR image classification models: Integrated Gradients, Gradient SHAP, LRP, DeepLIFT, Saliency Maps, Deconvolution, and Guided Backpropagation. (Pisarcik et al., 2024) also utilized Integrated Gradients and Saliency Maps as explainers to CNN models. LRP is also referred by (Saeed et al., 2024). (Ellis et al., 2023) refer the usage of LRP and Saliency to provide explanations in their neuroimaging system.

(Akbar et al., 2024) utilized Attention Mechanism method to visualize interest areas in their COVID-19 detection through radio-X images. This method allows to visualize graphically which areas were more important in the model analysis. They also use charts to visualize the weight per pixel (Figure 5). (L. Wang et al., 2023) also uses an Attention Mechanism (Deformable Attention) to overcome limitations of Grad-CAM methods, according to the paper, those methods fail these methods do not adequately capture fine-grained information regarding object boundaries, and as external extensions, they do not offer an intrinsic understanding of the decision-making process.

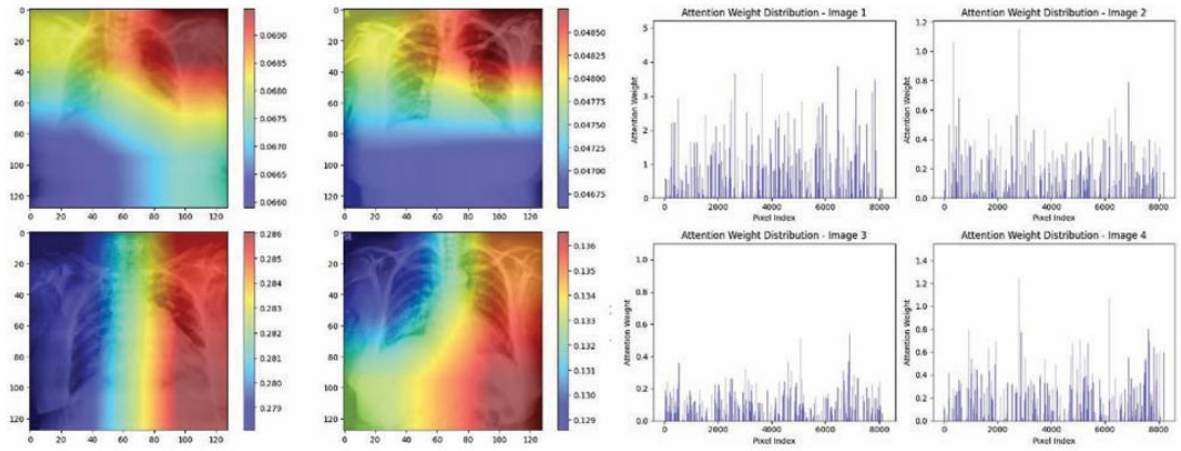

Figure 5 Attention Mechanism in place to visualize regions with more relevance to the model (Akbar et al., 2024)

Occlusion sensitivity is used by (Heng & Abdul-Kadir, 2023) along with other two methods (GCAM and LIME) in their study about hair disease detection through image analysis. As other mechanisms already mentioned here, it relies on how the model reacts to small perturbations in the input images. In the Figure 6, it's possible to visualize the sensitivity map for a given input and prediction (alopecia).

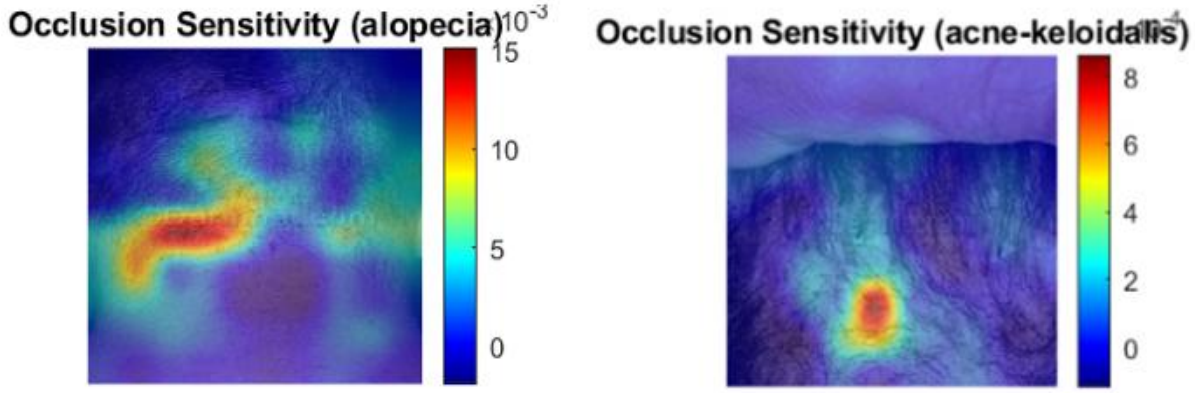

Figure 6 Occlusion sensitivity charts for two input images (Heng & Abdul-Kadir, 2023)

(Wickstrøm et al., 2023) make usage of RELAX, another occlusion-based method, to provide explanations in their CBIR deep-learning system of liver images. It's an explainability framework that provides input feature importance in relation to a vector representation, as opposed to a classification or similarity score. The core idea of RELAX is to evaluate how the representation of an image changes as parts of the image are removed using a mask. They apply this method on top of their deep-learning model (CNN ResNet50 architecture) that uses Simple Siamese (SimSiam) to represent learning, a core feature for a CBIR system that needs to represent knowledge from a given image and find similar instances in a database.

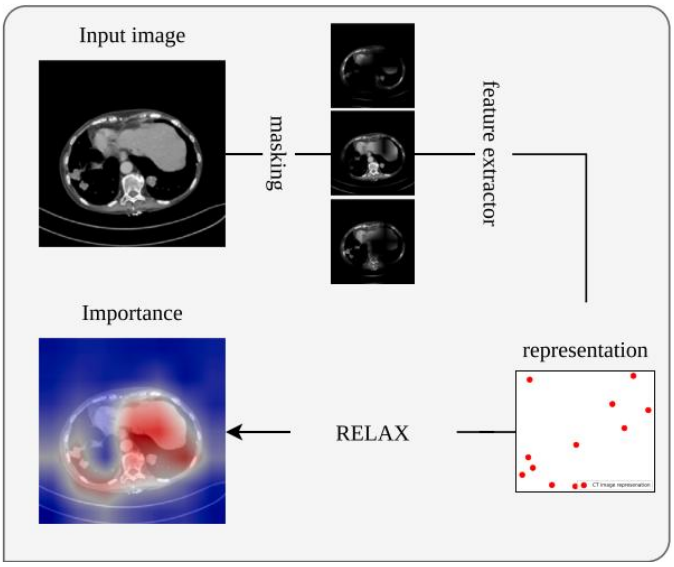

Figure 7 Representation of RELAX workflow in the CBIR system (Wickstrøm et al., 2023).

(Singhal et al., 2024) present a novel about how XAI can improve not just the trustworthiness of healthcare models but also their performance and accuracy once they can help to identify and visualize key features and patterns that can assist in the model optimization. They propose a explainability framework using many techniques categorized by them in end-to-end explainable evaluation (SHAP, Lime and Integrated Gradients), rule-based explanations (Sequential Covering Rule-based Learning algorithms, Association Rule Mining, or Decision Trees using open-source libraries such as EBM – Explainable Boosting Machine or Rule Fit) and user-adaptative explanations (integration of specialists feedback to improve the explanations using Reinforcement Learning Techniques).

(Mu et al., 2024)refer the usage of Counterfactual Reasoning to explain their distributed model inferring causes and effects and generating visual explanations in CNN. Through causal reasoning, AI model can understand how different factors interact with each other, enabling more accurate predictions or decisions.

(Parola et al., 2024) propose a Case-Based Reasoning (CBR) methodology to provide explanations in their cancer identification model. They utilize a k nearest neighbour (kNN) algorithm to find similar cases in the data set and use that similarity explanation. According to the paper, doctors feel more confident in the system output checking other cases that the same system has identified as similar. The similarity is calculated according to the instance features.

Table 8 XAI methods summary

XAI method	Number of papers
GCAM	42 (60%)
Lime	21 (30%)
SHAP	14 (20%)
GCAM++	10 (14,29%)
SaliencyMaps	6 (8,57%)
LRP	4 (5,71%)
Integrated Gradients	4 (5,71%)
EigenCAM	3 (4,29%)
Attention	2 (2,86%)
LayerCAM	2 (2,86%)
XGrad-CAM	2 (2,86%)
DeepLift	2 (2,86%)
Guided Backpropagation	2 (2,86%)
CBR KNN	1 (1,43%)
Guided GCAM	1 (1,43%)
ScoreCAM	1 (1,43%)
HiresCAM	1 (1,43%)
RandomCAM	1 (1,43%)
AblationCAM	1 (1,43%)
Relax	1 (1,43%)
Deconvolution	1 (1,43%)
Smooth Integrated Gradients	1 (1,43%)
Smooth Gradients	1 (1,43%)
Occlusion sensitivity	1 (1,43%)
User-Based (RL)	1 (1,43%)
Rule-Based	1 (1,43%)
Counterfactual Analysis	1 (1,43%)

2.2.3 RQ 3: Are there use-cases using Gen-AI being applied to leverage reliability of image system models?

Diagnosis systems using image models are normally implemented using deep learning classifiers. As discussed in the research question 1, CNN and RNN are used to infer the likelihood of specific disease or diagnosis. According to this literature review, with the advent of Gen-AI (Generative Artificial Intelligence Gen-AI), this scenario starts to change with those techniques being applied to medical solutions.

(Metta, Beretta, Guidotti, Yin, et al., 2024) proposes the usage of generative artificial intelligence techniques to overcome some blind spots of the XAI methods. They refer saliency maps can be fragmented and challenging to be interpret in an urgent medical situation while LIME and SHAP lacks reliance in image models. Their propose is the usage of ABELE (Adversarial Black Box Explainer Generating Latent Exemplars), a local, model-agnostic model tailored to image classifiers. ABELE explainer works by providing exemplar and counter-exemplar images

to be classified by the model alongside an input image, a saliency map is generated to highlight decision making areas. Their proposal advances in the usage of generative AI techniques as long they make usage of AAE (adversarial auto-encoder) to generate similar (exemplar) or different (counter-exemplar) instances that are used to fulfil the explainer and provide explanations through similarity to the final user.

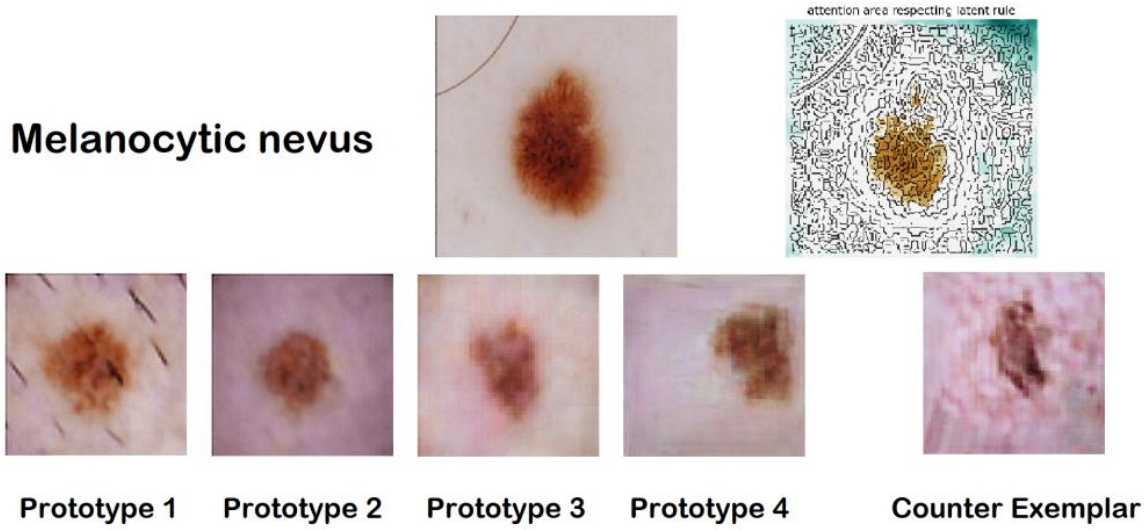

Figure 8 ABELE explainer example with an input image and their exemplar and counter-exemplar (Metta, Beretta, Guidotti, Yin, et al., 2024)

(Bardozzo et al., 2024) refer the usage of GAN (Generative Adversarial Network) to enhance the quality of microscopic images. The paper states this model was able to improve the quality of kidneys samples achieving high-resolution images that allows the visualization of microscopic structures. To prove the effectiveness of the model, besides the specialists' opinions, they use activation maps of the Grad-CAM family to unhide features importance for the model in real and generated images.

(Grillo et al., 2024) mention the usage of ChatGPT-4 (generative model developed by Open AI) to provide textual explanations to the final user. They have used the system API to train an ChatGPT agent to identify tumours in MRI images using labelled data and leverage the text generation capability from the same to provide explanations about the reasoning behind a classification (Figure 9).

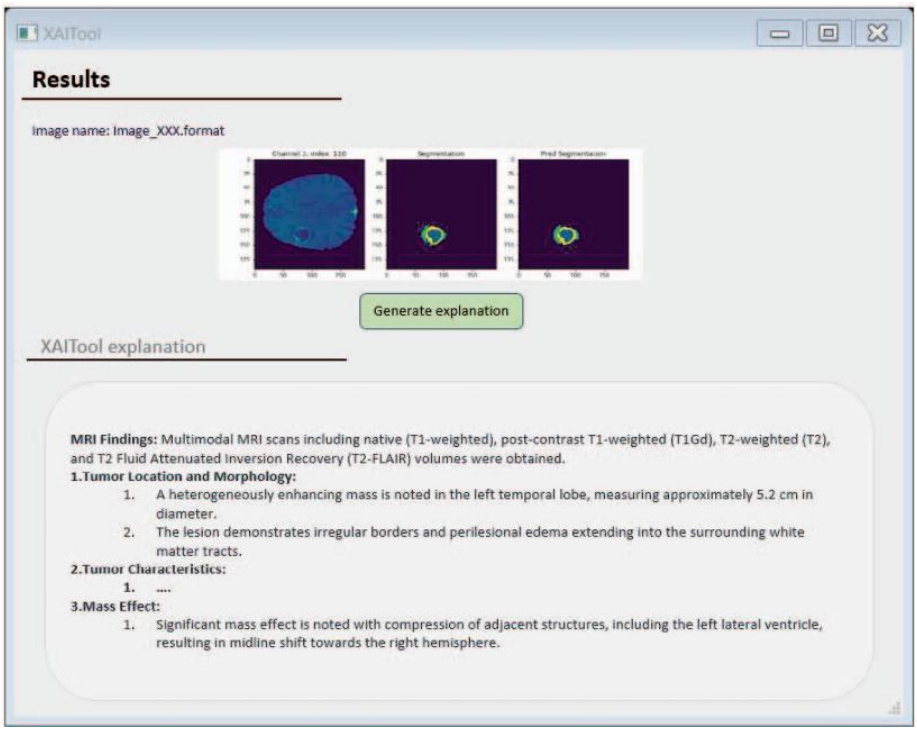

Figure 9 XAI tool using ChatGPT-4 (Grillo et al., 2024)

2.2.4 RQ 4: How has XAI been integrated to the current medical systems?

This literature review presents several use-cases of how XAI can increase the confidence of medical systems, bringing clarity to the black-box pattern that many ML models present. However, multiple articles keep their focus in theorical frameworks only applying XAI methods to explain how their model generalizes while this research question try to address how XAI has been (or might be) applied to real medical systems.

(Lamba & Rani, 2024) present a novel about a neurorehabilitation program using deep learning models applying Grad-CAM as explainer. The paper presents a fictional patient, Alex, taking part of this program. According to the paper, Grad-CAM not only improves the BCI and Neuron-Electronics system's interpretability, but it also gives Alex and the medical team valuable insights into the rehabilitation process. This openness in the decision-making encourages a collaborative approach between the multiple professionals involved. The paper refers the creation of the heatmaps in real time to bring insights about specific brain regions to the doctors and physicians. The system workflow can be visualized in the Figure 10.

(Bouabdallah et al., 2024) propose a classification model to identify and catalogue thorax lesions that leverages Grad-CAM as part of the final system, not only as the explainability mechanism. They built a pipeline to identify thorax lesions that uses Grad-CAM to localize areas in the X-ray images where there is significant likelihood of a lesion exists. The process starts with a CNN model that classifies X-ray images as healthy and unhealthy, after that, Grad-CAM saliency maps are used to crop parts of the images where the lesions can be identified. It's

possible to see the entire process in the Figure 11. The usage of XAI techniques to improve the model performance helping specialists to identify key features is also referred by (Singhal et al., 2024).

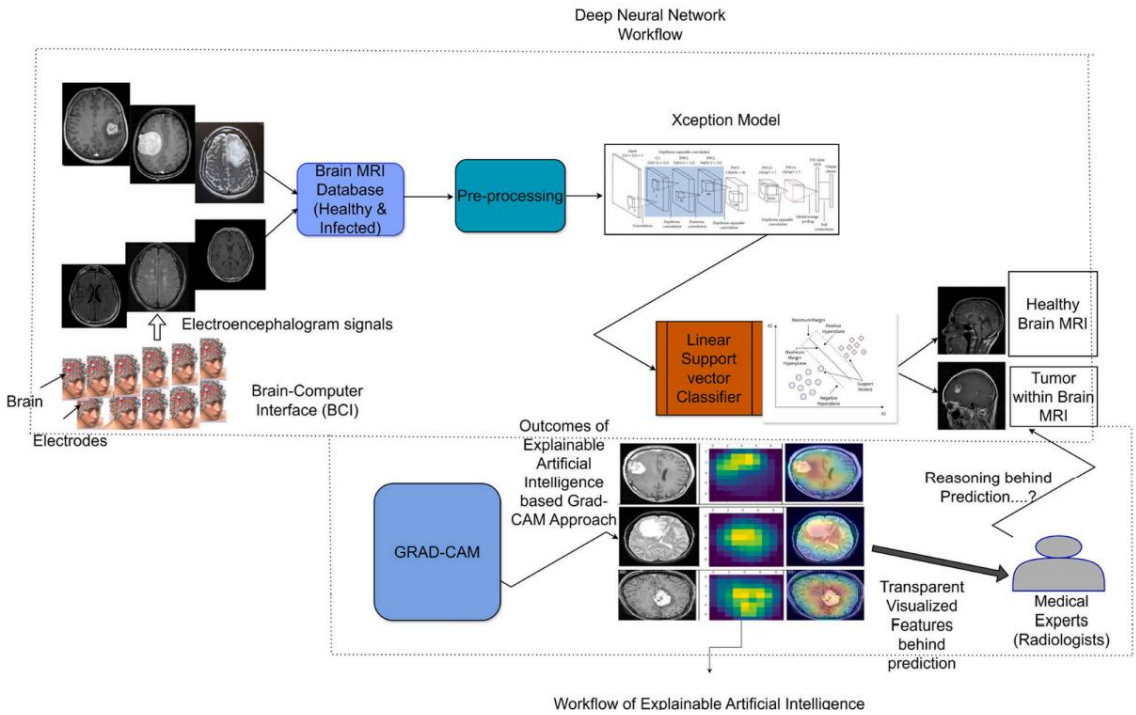

Figure 10 Grad-CAM heat maps being generated in real-time to produce valuable insights to the medical team (Lamba & Rani, 2024)

(Metta, Beretta, Guidotti, Yin, et al., 2024) proposes an explainability module using ABELE explainer that relies on exemplar and counter-exemplar synthetic images to provide how a give image was classified. It could be integrated for a diagnosis system justifying the black-box model classification. The Figure 12 presents a screenshot of the proposed explainer module.

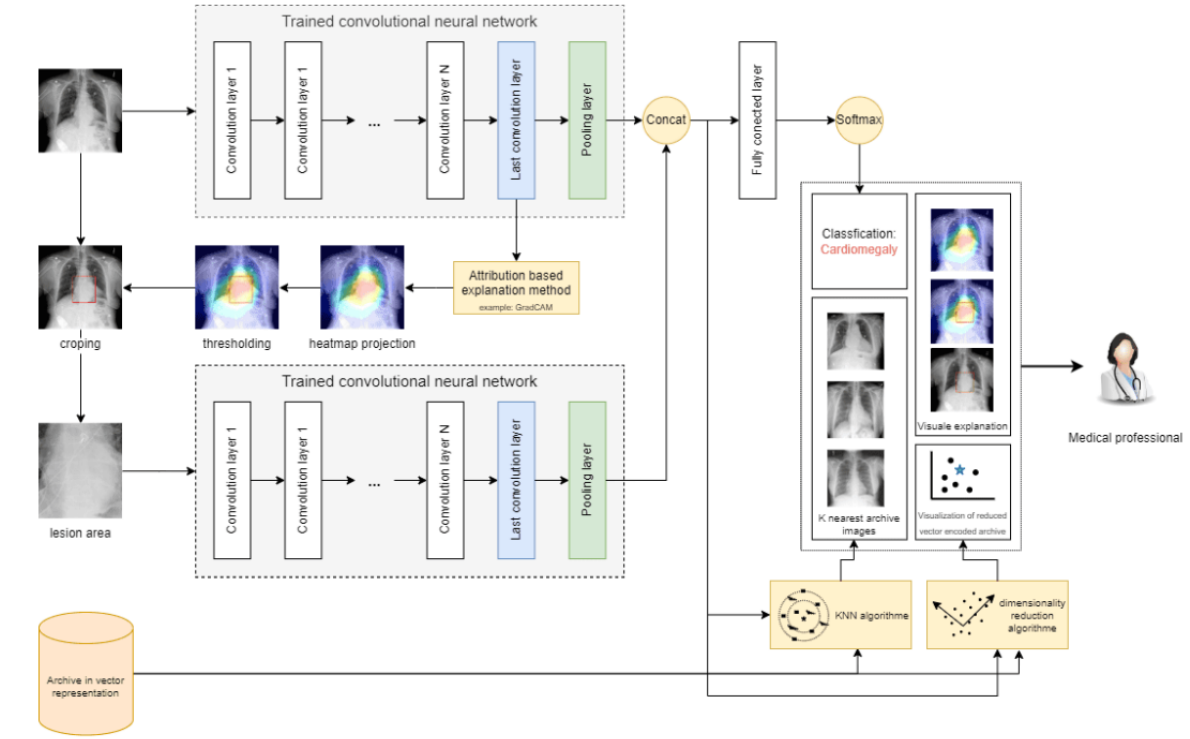

Figure 11 Overview of thorax diagnosis system using Grad-CAM (Bouabdallah et al., 2024)

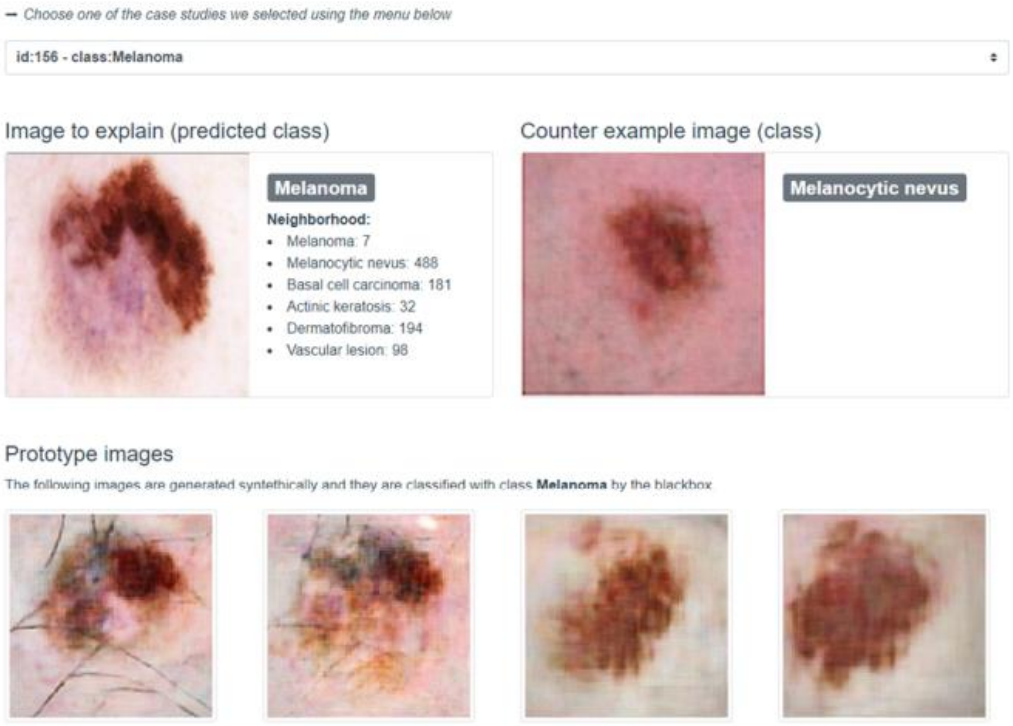

Figure 12 Example of how a real image classified as melanoma is explained by synthetic images generated by ABELE (Metta, Beretta, Guidotti, Yin, et al., 2024)

(Thiruvenkadam et al., 2024) refer the need of an Explanation Interface to enhance the trustworthiness in clinical image systems by the clinical staff and patients to overcome the black-box problem of ML models. The Figure 13 presents how this workflow with explainability would lead to an increase in the system reliability.

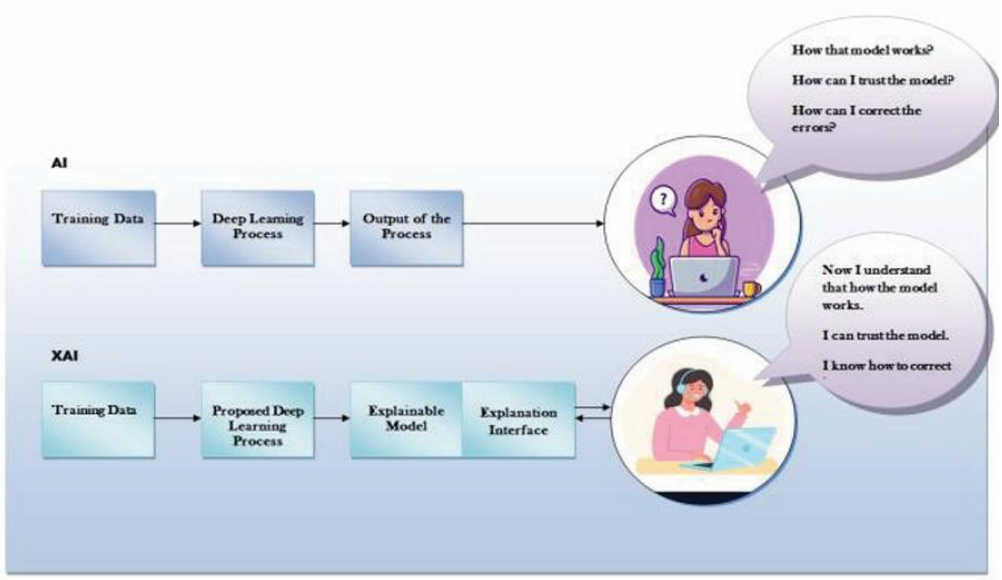

Figure 13 Explainable interface addressing reliability issues in black-box models (Thiruvenkadam et al., 2024)

(Lo et al., 2024) speak about the integration of XAI into a vascular wound diagnosis system. To improve the confidence by the medical staff, the system presents importance heatmaps to help the users to understand which region was more relevant in the classification. The system also presents SHAP charts to provide a better understanding of how the model considered features. In the Figure 14, it's possible to see how the system interface looks like.

(*Chest Radiographic Findings in COVID-19*, n.d.) mention how the usage of heat maps produced using LIME can be integrated in a fetal analysis model based on ultrasound images. Beyond bringing more transparency to the system output, the paper refers how the heat maps can help clinicians to identity interest areas. They also mention, as future works, the integration of other XAI methods to health care analytics system.

(Hroub et al., 2024) mention the creation of lung disease diagnosis tool to support less-experienced clinicians using activation maps in x-ray analysis. Their tool has an "interpretation" module integrated with the inference model to support the model conclusions and highlight the interest areas (Figure 15).

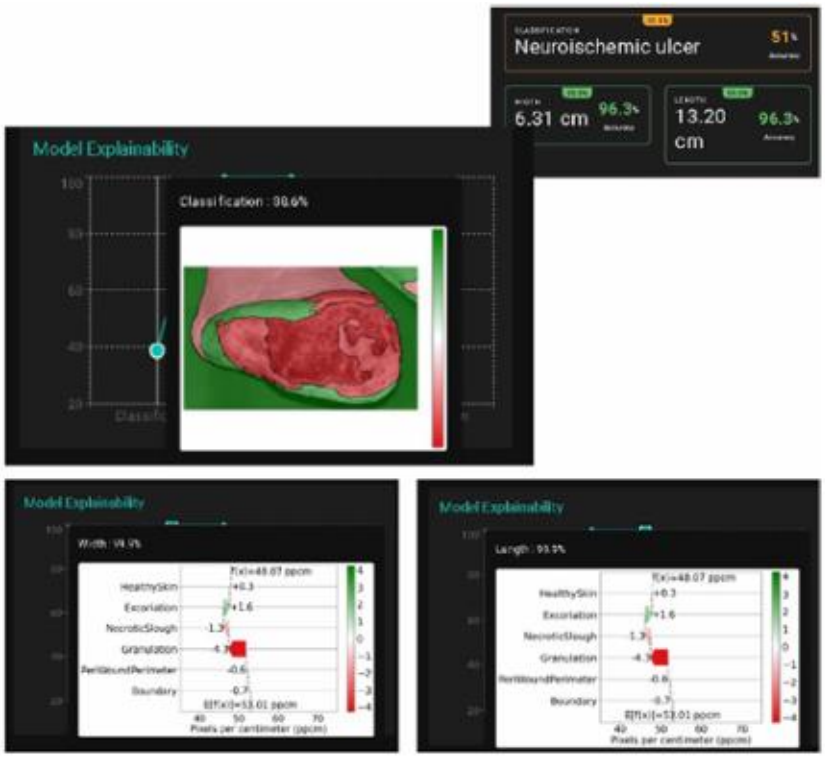

Figure 14 XAI visualizations integrated to the vascular wound classifier (Lo et al., 2024)

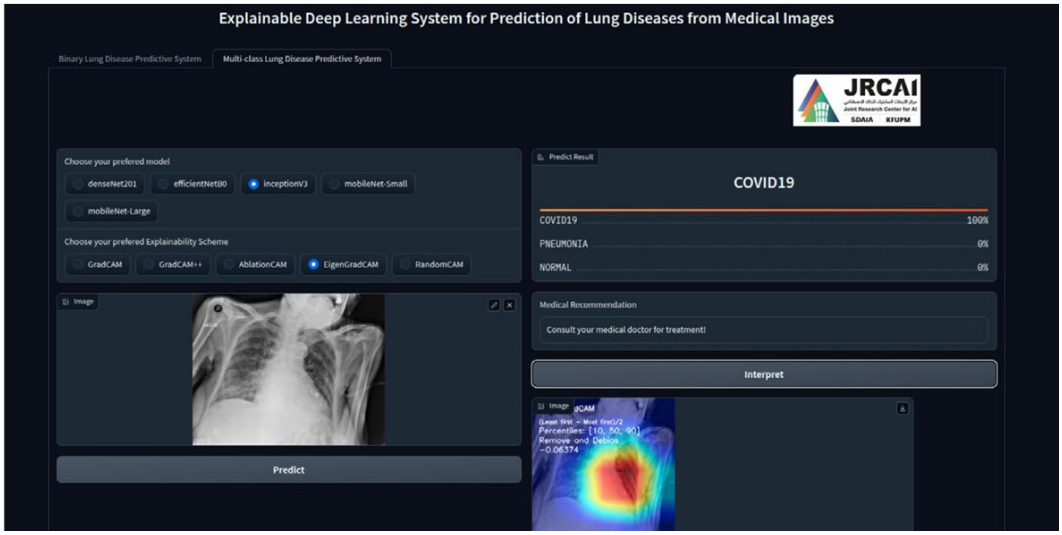

Figure 15 lung diseases analysis system with Interpretation module.

(Grillo et al., 2024) propose a MRI tumours detection system using ChatGPT-4 to classify images and generate textual explanations. The text-generation features available in LLM (Large Language Models) can be integrated to explainability models providing user-friendly explanations. The XAI tool proposed by in the article can be found in the Figure 9.

2.3 Discussion

The research has identified the importance of XAI techniques to accelerate the adoption of AI in healthcare systems. As concluded in the first research question, machine learning models are commonly identified as 'black boxes' since their decision-making process is complex and not intuitive usually based on mathematical relationships extracted from a historical dataset. This lack of transparency is referred by most of the analysed papers as an obstacle to be overcome in medical system using AI, and XAI techniques have been mentioned as key factors to achieve this improving the reliability and trustworthiness of clinicians in the system decision.

In health-care image systems (focus of this research), visual explanations providing local explanations (about individual inferences) are a clear tendency according to the analysed papers. Methods such as Grad-CAM (and many other variations), LIME and Saliency are referred to generate heat-maps presenting which areas in each classified image are more relevant in the model's decision making. Figure 3 presents an example of heat-maps using Grad-CAM. The highlighted areas in the visual explanation map are connected to the internal neural network layers nodes weight presenting how important each image pixel was to the inference.

Beyond the need for more transparency and details about the models' decision-making process, other challenges are also mentioned in the analysed papers. The necessity of bias identification and mitigation, one relevant ethical aspect, is very important to build reliable systems with the ability to generalize and predict with accuracy for a diverse group of patients and conditions. According to the research, more clarity about the decision-making process can highlight blind spots or lack of representation for some specific groups. XAI global methods that do not explain specific decisions but bring light to internal statistics and characteristics of the model. Many methods rely on the concept of perturbation where some small changes in the input are produced to check how they influence the output, it can also uncover the decision-making process helping to identify sources of errors or bias. SHAP can be used to generate statistics about how each feature is important to the model. This level of internal detail can help to identify not expected correlations or distortions. SHAP can also be used for local explanations (providing feature importance for individual predictions) as referred by (Srinivasu et al., 2024) and other analysed papers.

The same characteristics in XAI methods can also help with other recurrent challenges: legal and regulatory standards. Health care systems deal with critical decisions that can affect directly people's lives, that's why they are continuously audited and regulated (Patel et al., 2024) and (Chaddad et al., 2024). XAI methods can disclose details about why some specific decisions have been made. For example, Grad-CAM (and other similar methods) generates heat maps to highlight which areas for a classified image was more relevant to the model. It can be used as evidence to prove how the system is deciding between positive or negative diagnostics. (Metta, Beretta, Guidotti, Yin, et al., 2024) reinforces that the XAI domain in medical imaging is governed by a complex web of regulations and standards and ensuring compliance with these regulatory frameworks is both challenging and labour-intensive.

The research presented deep learning classifiers (using CNNs) as the most used method in healthcare diagnosis per image systems, but some papers also present the usage of Generative AI methods such as GAN being used to solve medical problems. Not specifically about XAI (Bardozzo et al., 2024) cite the usage of GAN to improve the quality of microscopic images improving the overall system accuracy. Gen AI is also referred to enhance explainability solutions, for example (Grillo et al., 2024) cite the generation of textual explanations using ChatGPT 4 API about MRI images classification (Figure 9). (Metta, Beretta, Guidotti, Yin, et al., 2024) refer the usage of GAN to generate counter-factual images to explain decisions using ABELE method (Figure 8).

Another important aspect to highlight is how XAI can be used not only for enhancing AI systems reliability and trustworthiness but also for enhancing their performance and accuracy. (Bouabdallah et al., 2024) and (Singhal et al., 2024) refer that details about how the model is making decisions can help scientists and engineers to tune hyper-parameters and adapt model architecture. Heat maps can, for example, disclose important or non-important areas to the model, according to the specialist's analysis, the model or the data set must be adjusted if some important features have been disregarded.

The research also provide insight into the usage of XAI in healthcare systems. Many papers keep their focus in academical studies, but some articles provide hypotheses or real examples of how XAI could be integrated into real systems. Those articles refer some sort of explanation module attached to the diagnosis system to provide heat maps, feature importance, exemplars and counter exemplars, textual explanations, etc, about a specific diagnostic by image. According to the same papers, all those tools enhance the reliability of the system by the medical staff.

Another interesting aspect mentioned by (Harikumar et al., 2024) and (Lamba & Rani, 2024) is the fact of that heat maps presenting activated regions according to the relevance into the model can also be used as tool to guide the diagnosis process. Activated areas can also be used by experienced medical professionals to evaluate specific lesions or formations accelerating decisions.

As described in the Research Methodology section, the research analysed about 800 recent papers from 4 renowned sources checking textually about 70 to answer research questions about the adoption of XAI techniques in health care imaging systems. The outcome presents clear concerns about the lack of transparency in deep learning models setting this as challenge their widely adoption in medical areas. Visual explanation maps (highlighting Grad-CAM as the most used) are referred as solutions to solve this and other challenges (with many other XAI methods). As already mentioned, local XAI methods are the preferred since they can clearly explain what, in a given image, makes the system made a specific decision (or diagnosis). Even though the majority of the papers refers normal image classifiers models as the main technology, Gen-AI is also referred by some papers to solve quality image problems but also to provide explanations. The research also provides a panorama of XAI integration in health care systems highlighting the creation of an explanation module to provide local explanations about any system diagnostic. The research was able to provide answers for all the four research questions and conclude that XAI has been used to enhance the reliability and transparency of

image model systems, addressing legal and ethical questions, making more feasible their adoption in real world critical systems.

3 Methods and Materials

This chapter provides an overview about methods and materials employed in the development of the experimentation for this work. The dataset will be presented with some key categories and preprocessing steps. It also presents the methods applied to evaluate the effectiveness of XAI in health care systems. Additionally, privacy concerns are addressed.

3.1 Datasets

In order to explore the effectiveness of XAI methods in the Experimentation chapter, some Datasets need to be selected to feed DL models training. As long image healthcare systems are the focus of this work, the datasets should be image-based containing medical exams samples (radiographies for example). Public and verified datasets were also preferred as long they were already utilized in other scientific studies. Datasets only containing non-sensitive data were also preferred.

A radiography database with chest images labelled with COVID-19 positive or negative diagnosis proposed by (Chowdhury et al., 2020) and (T. Rahman et al., 2021) was picked to train the DL model target of the explainers experiment. The sub-set of the dataset used by the DL model is composed of 13,808 chest images categorized in two classes: 3,616 positive COVID-19 cases and 10,192 negative COVID-19 samples. The dataset also contains images classified as Lung Opacity and Viral Pneumonia that were not included in the experimentation.

The dataset is composed of PNG images with 299x299 size and was created by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors. The datasets presented good quality, images have enough resolution to be used in the training and are standardized in terms of the chest positioning, so no need for major pre-processing steps, the unique step was the images resizing to 70x70 pixels.

To complement the experimentation, HAM10000 dataset (Tschandl et al., 2018) containing skin lesion images is also used to train another classification model. This is a dataset consisting of 10,015 dermatoscopic images which are released as a training set for academic machine learning purposes and are publicly available through the ISIC (The International Skin Imaging Collaboration - https://api.isic-archive.com/collections/212/) archive.

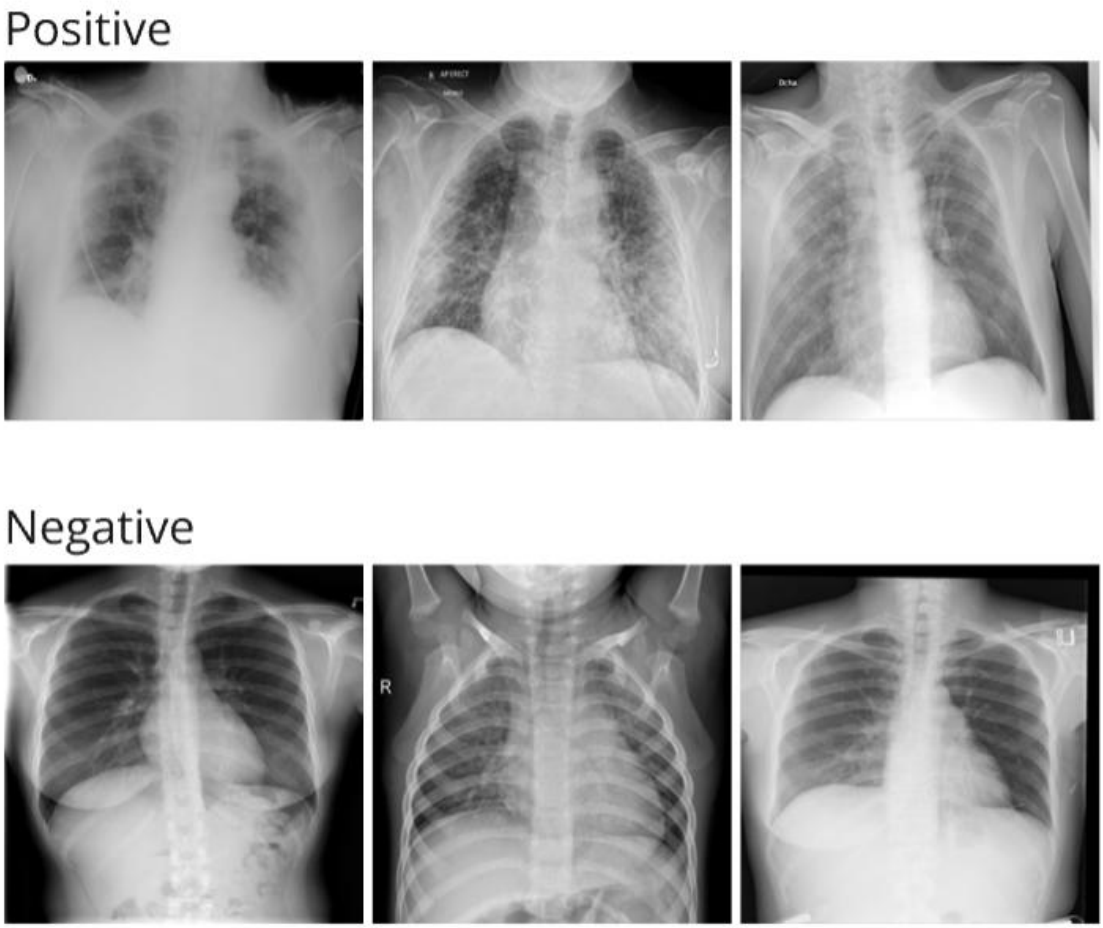

Figure 16 Samples of the radiographies in the Data Set selected to train the model

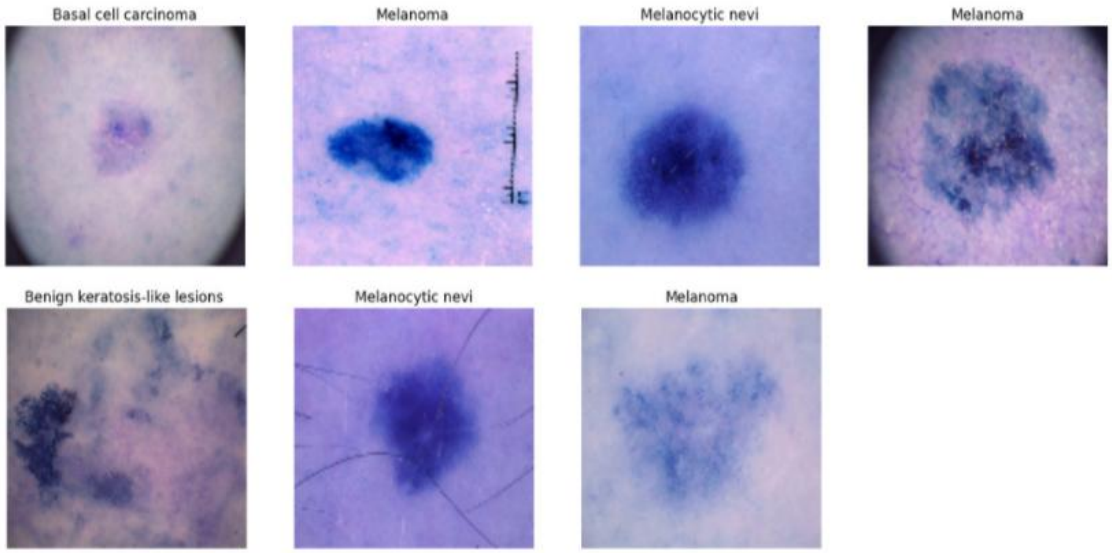

Figure 17 Samples of HAM1000 dataset

The images are classified in seven classes: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions and dermatofibroma. The dataset also contains a metadata CSV with additional information about each case: sex, age and localization (in the body, such as back, face, etc).

3.2 XAI Methods

In the systematic review more than 20 methods have been identified to explain deep learning models. In order to explore how they work, experiment and produce a comparison, three methods have been selected. In the end, this work aims to propose an Explainability module that could be integrated into any medical system, the methods selection considers how they could complement each other.

Grad-CAM is the most relevant according to the systematic review referred by 60% of the papers and, more than it, other CAM family methods such as Grad-CAM++ and ScoreCAM are also mentioned by many other articles. This is a local method that highlights the more important pixels in a given image after an inference. It's done according to the weights in the last convolution layer in the model. The choose of this method is justified by their relevance mainly.

ABELE is referred in the systematic review as a method using Gen-AI techniques to provide counterfactual images to explain the output of a DL method. It's one of the only methods that does not produce only heatmaps, so it could be a relevant complement for a Explainability module using Grad-CAM.

SHAP is also relevant (mentioned by 20% of the papers) and can produce feature importance values not only to the predicted class but also for all the other classes, this ability is the

justification for its selection. These feature importance values can complement the explanation providing global and local statistics about how the model consider each feature.

3.3 COVID-19 Model

In the experimentation chapter, XAI will be applied to DL models to explore their capacity to disclose decision-making details about the classification process. The first chosen model is a DL model built to predict the presence of COVID-19 in chest's radiographies images. The selected model is built on top of a CNN architecture using ADAM optimizer. One of the goals for this work is the production of a complement to this model in the form of an Explainability module providing more details about the decision-making process of the system what should bring more transparency and reliability from any clinicians adopting the system. The training was done using the dataset described in the Data Sets section achieving satisfied results after 47 training epochs.

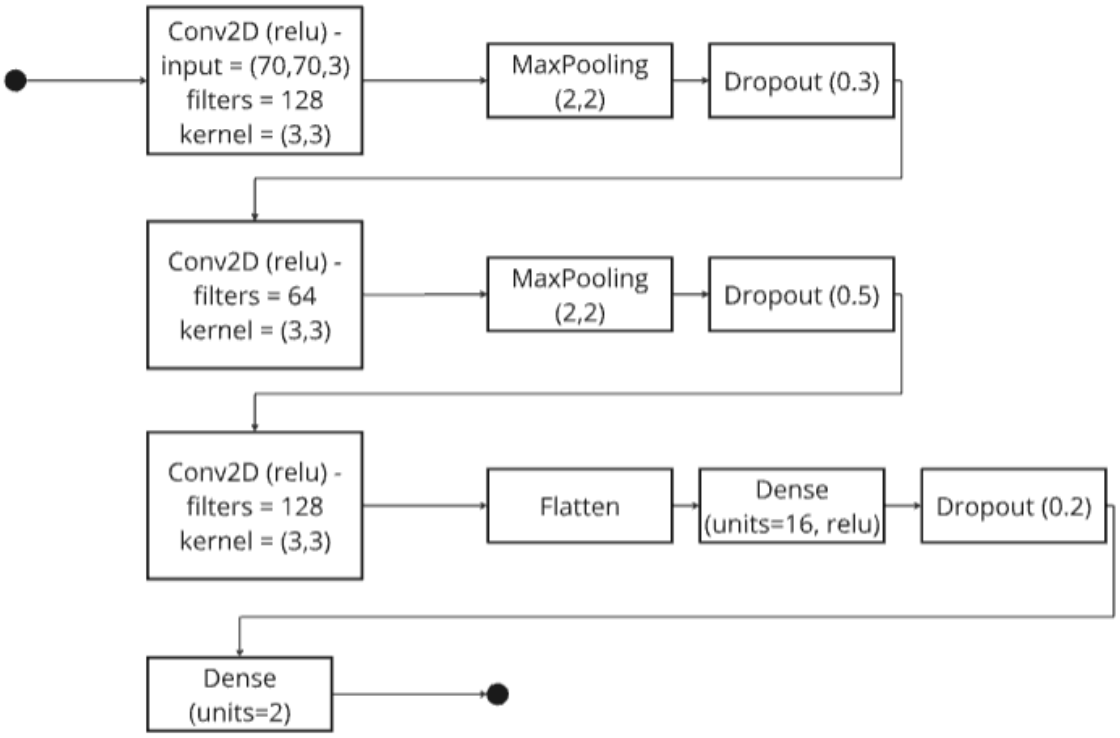

Figure 18 COVID-19 prediction CNN model architecture

The train has followed a train & test split approach considering 20% of the samples to evaluate the system performance. The model achieved the following performance metrics considering the test data:

Accuracy on Test Data: 0.9627
Precision on Test Data: 0.9359
F1 Score Test Data: 0.9257

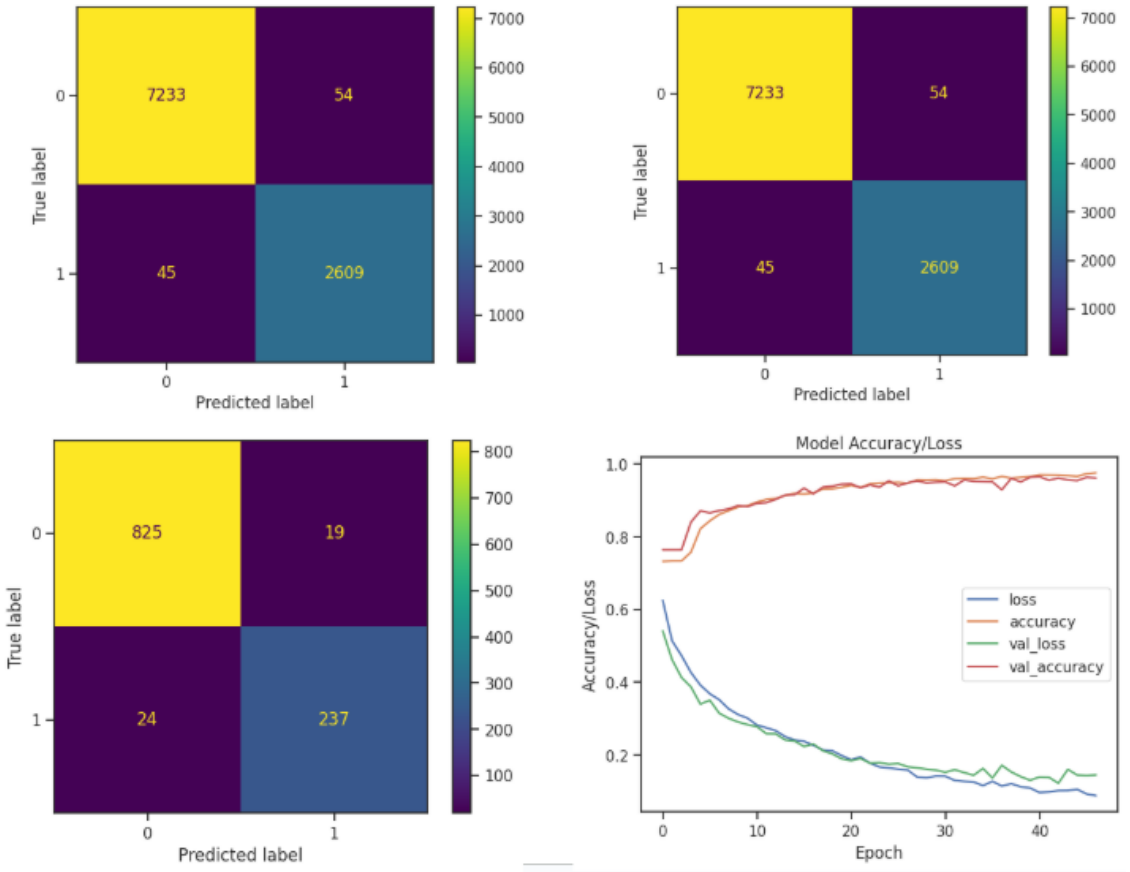

Figure 19 Model accuracy evolution during the training epochs

3.4 Skin Lesion Classifier Model

To complement the experimentation chapter, another model was selected: a DL classification model to classify skin lesion. As the COVID-19 example, it's a CNN model that receives individual skin lesion images (coloured and resized to 100x100 pixels) and calculates the likelihood for each of the seven classes: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions and dermatofibroma.

The model was trained using HAM 1000 skin lesion image dataset and a train split random method was used to segregate the samples into 33% to test and the remaining to train. After 100 epochs of training, it achieved accuracy of 84%.

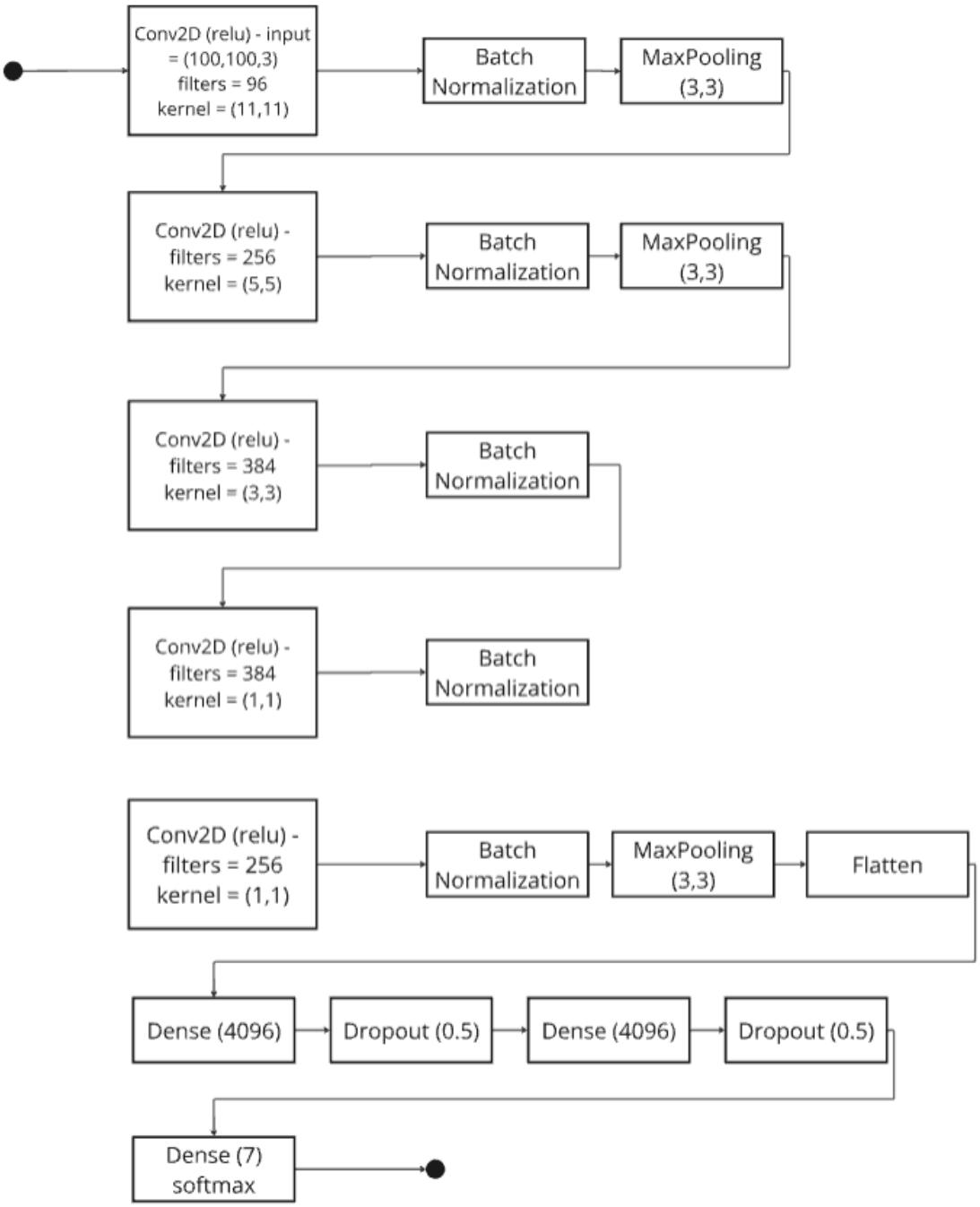

Figure 20 Skin lesion classifier model architecture

3.5 Ethics and Privacy

Privacy and confidentiality are two relevant aspects in any medical system. Artificial intelligence systems rely on historical data to be trained, so producing a good dataset to train a machine learning model in healthcare can be a challenge.

For this work, the Dataset described in the section 3.1 Dataset will be used to train a Convolutional Neural Network model (presented in the section 3.3 Model) that will, in the end, be used to explore XAI methods.

The dataset proposed by (Chowdhury et al., 2020) and (T. Rahman et al., 2021) is composed of PNG images with 299x299 size and was created by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors. The dataset does not have any kind of personal identification that could be used to track any specific x-ray image to a real person. It mitigates any kind of data breach issue that could expose sensitive data trackable to people.

According to the authors, the dataset is composed of a couple of other public datasets. Public repositories usually consider the user consent to share their medical images, but it's important to refer that this topic is not explicitly expressed in the paper that presents the Dataset.

As stated before, the HAM10000 dataset is public and maintained by ISIC to be used in ML models related to skin cancer studies. The dataset does not contain any kind of personal identifier or information that could be used to correlated samples with real people.

The dataset contains images in the PNG format of micro lesions in many sizes and also a CSV with metadata about each sample (age, sex and localization in the body), none of them can be used to identify people.

4 Implementation

This chapter covers the implementation details of the XAI methods selected to explain the decisions produced by COVID-19 CNN and Skin lesion models. Specific pre-processing steps needed by each method, details of the implementation made and some evidence of the outcome for each algorithm will be disclosed in the next sections.

4.1 Grad-CAM

According to the systematic review conducted in this work, Gradient-weighted Class Activation Mapping (Grad-CAM) is the most mentioned method used to explain DL models in medical systems. Grad-CAM is a local method once explain a specific inference and is model specific, only works with CNN.

One possible outcome of the method is a heatmap that can present which pixels was more relevant in each inference. This outcome is precisely one of the reasons behind it widely adoption. The possibility of checking visually how a model analysed an image and what it considered more relevant can make easier for a clinician understand the decision and trust on it according to (Burgos et al., 2024).

For a given CNN model, Grad-CAM works analysing the gradients in the last convolution layer for a given target class (target of the explanation). This is the driver to conclude that specific positions in this layer output (representing pixels in the image being analysed) with more weight are more relevant to the final classification. This is the input to create an activation map that can be transposed in the classified image to provide a visualization about the more important pixels.

As described in the Methods and Materials methods, this work aims to explain two CNN models: COVID-19 and Skin lesion. The first step to integrate Grad-CAM to this model was to identify

the last convolution layer. A model is generated to associate this layer output with the final output allowing to calculate the gradients between the top activated class (a.k.a. model prediction) and the last convolution layer output. A correlation map is calculated between the feature map and this model allows to conclude which features (or pixels) are more active allowing to create a final heat map.

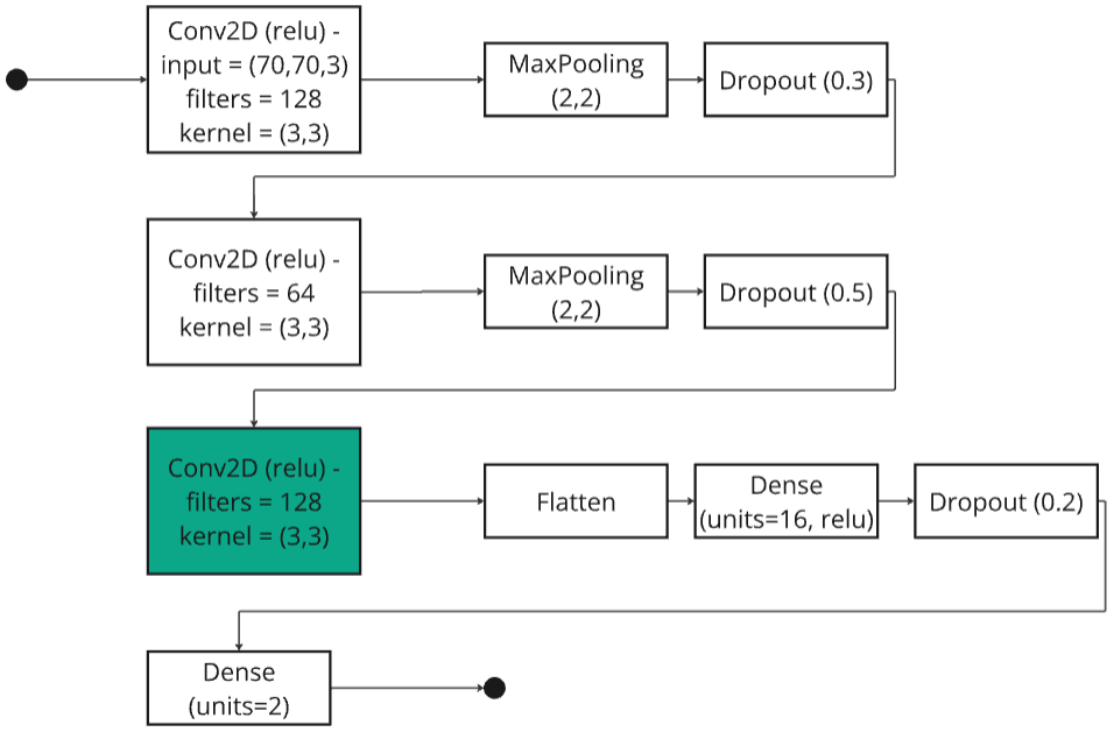

Figure 21 Last convolution layer highlighted

To experiment the method, four images from the target dataset have been selected (two from each class), predictions were created for each of them generating heatmaps to highlight which areas have been more relevant to the final output. The outcome can be analysed in the Figure 22. It's possible to see how the model considered some specific regions in the lungs to conclude the presence or not of COVID-19 aspects. Below it's possible to see the implementation using Keras:

```
def make_gradcam_heatmap(img_array, model, last_conv_layer_name, pred_index =
None):
  grad_model = tf.keras.models.Model([model.inputs],
                [model.get_layer(last_conv_layer_name).output, model.output])

  with tf.GradientTape() as tape:
    last_conv_layer_output, preds = grad_model(img_array)
    if pred_index is None:
      pred_index = tf.argmax(preds[0])
      class_channel = preds[:, pred_index]
```

```
grads = tape.gradient(class_channel, last_conv_layer_output)
pooled_grads = tf.reduce_mean(grads, axis=(0, 1, 2))

last_conv_layer_output = last_conv_layer_output[0]
heatmap = last_conv_layer_output @ pooled_grads[..., tf.newaxis]
heatmap = tf.squeeze(heatmap)
heatmap = tf.maximum(heatmap, 0) / tf.math.reduce_max(heatmap)

return heatmap.numpy()
```

Code 1: Grad-CAM implementation

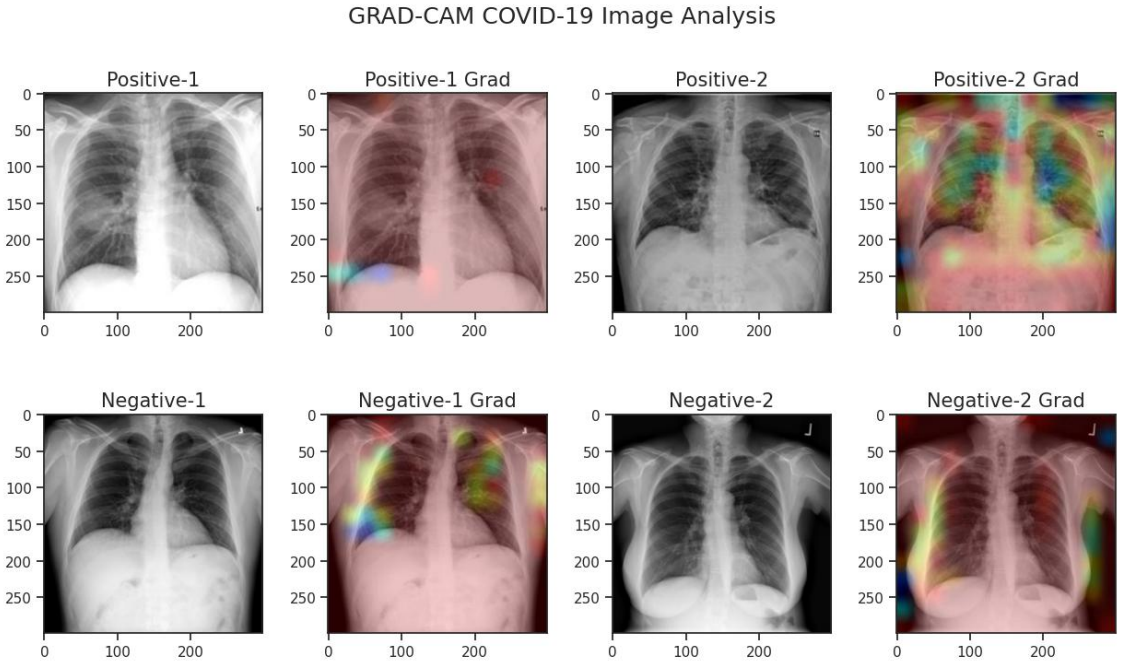

Figure 22 Heatmaps for each of the predictions

The method was able to present some details about which areas were more relevant to the model output. In order to evaluate this evidence, (*Chest Radiographic Findings in COVID-19*, n.d.) has been analysed to verify which areas medical experts consider more relevant while evaluating a lungs radiography in a potential COVID-19 diagnosis. The paper provides guidelines to help clinicians identify visual indicators in radiographs that may suggest a positive or negative COVID-19 diagnosis.

As it's possible to see in the Figure 23 the areas highlighted in some of the steps in the guide corroborate the influence regions detected by the Grad-CAM. Peripheral areas in the lungs can provide visual signs about the presence of COVID-19 according to the paper. Those areas were

also activated in the GRAD-CAM samples investigated in the Figure 22. It can give some signals that the model considered relevant the same areas a specialist would do.

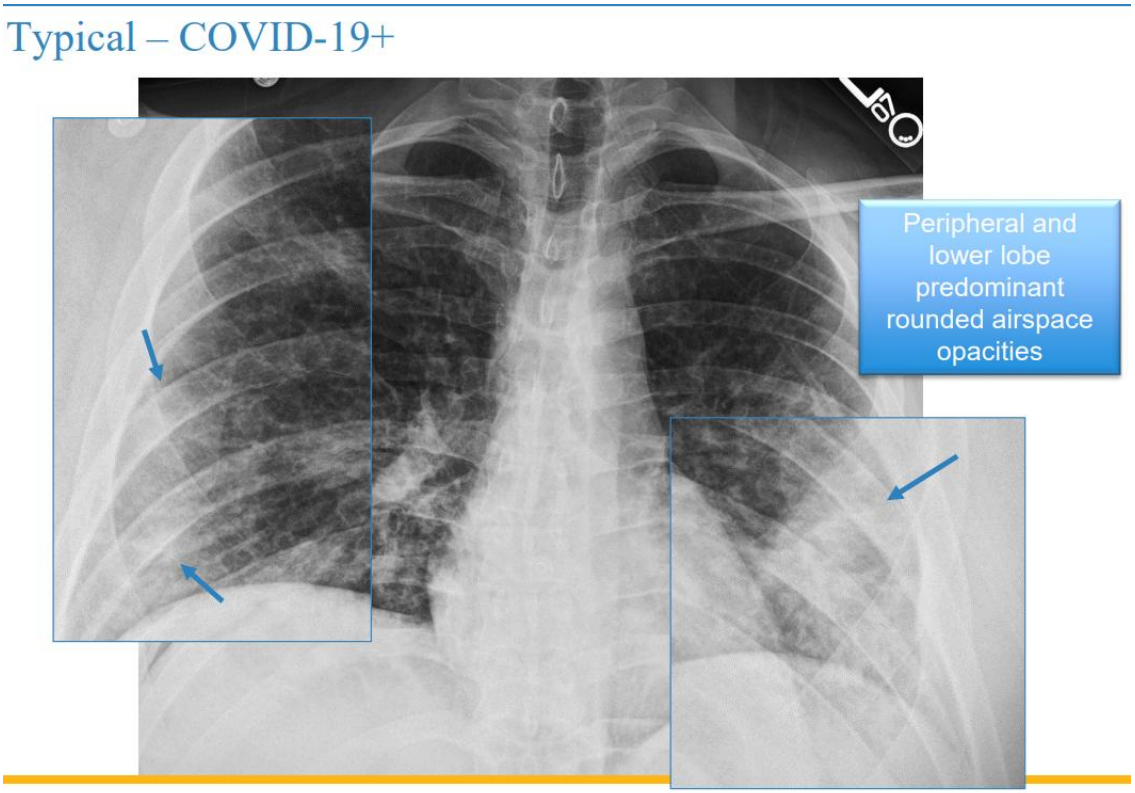

Figure 23 Lung radiography analysis according (*Chest Radiographic Findings in COVID-19*, n.d.)

The same XAI method was also applied to the skin lesion model. The first step was to identify the last convolution layer in the CNN model: the fifth one. The Figure 24 presents four samples along with their classifications and Grad-CAM heatmap. The results were satisfactory, the algorithm was able to identify and highlight specific points in the lesion area and surroundings, although in some cases such as Melanoma sample, some areas far from the lesion was marked as relevant to the model what could reduce the trustworthiness in the final classification.

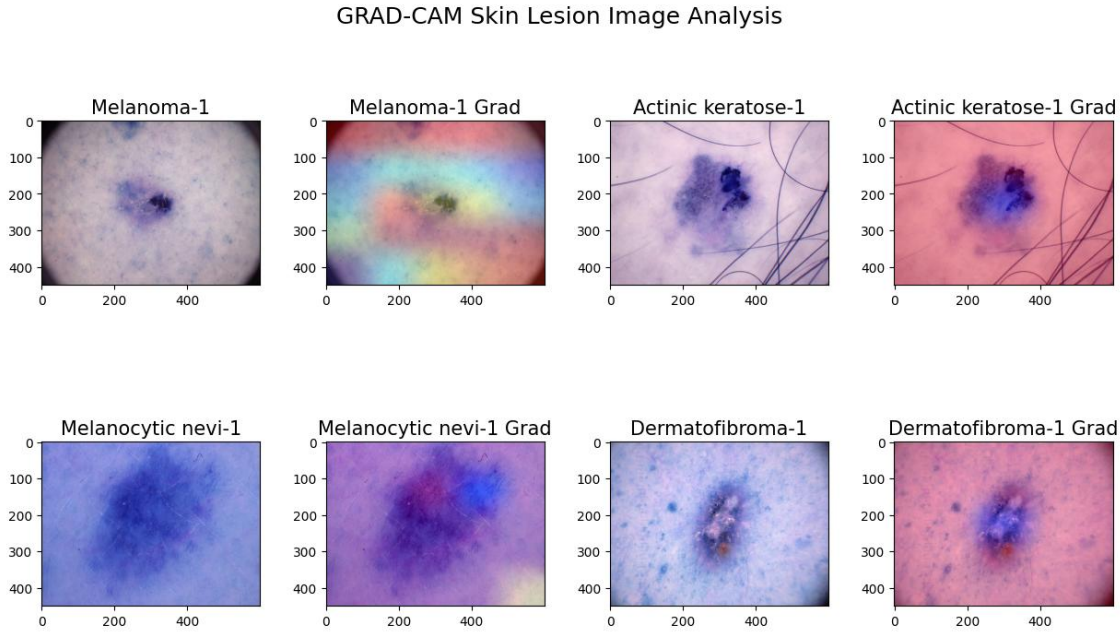

Figure 24 Heatmaps for some skin lesion samples

4.2 ABELE

ABELE, referred in the systematic review, was studied as an additional tool to provide details about the decision-making process behind the DL models target of this work. The usage of Generative AI techniques may reveal additional aspects about the model inference process with graphical and visual support as in the other methods.

According to (Gezici et al., 2024) and (Metta, Beretta, Guidotti, Yin Yuanand Gallinari, et al., 2024), ABELE - Adversarial Black Box Explainer Generating Latent Exemplars, ABELE is a model-agnostic explainer that operates locally, taking an image as input along with a black-box classifier to generate a set of exemplar and counter-exemplar images, and a saliency map. Exemplars and counter-exemplars are artificially generated images where exemplars share the same classification as the input image, while counter-exemplars are assigned a different classification. These images can be visually analysed to understand the reasoning behind the classification. The saliency map highlights the areas of the input image that contribute to a specific class and those that push it toward an alternative classification. One important concept behind the ABELE ways of working is the usage of latent space concept, it's used to compare two samples and conclude how close or not they in the problem space. To establish a neighbourhood in the latent feature space, ABELE employs an Adversarial Autoencoder (AAE). The encoder processes the input image through the AAE, extracting its latent representation based on key features. A genetic algorithm is used to generate the neighbourhood, optimizing a fitness function. In order to generate counter factual exemplars, ABELE uses LORE decision rules mechanism.

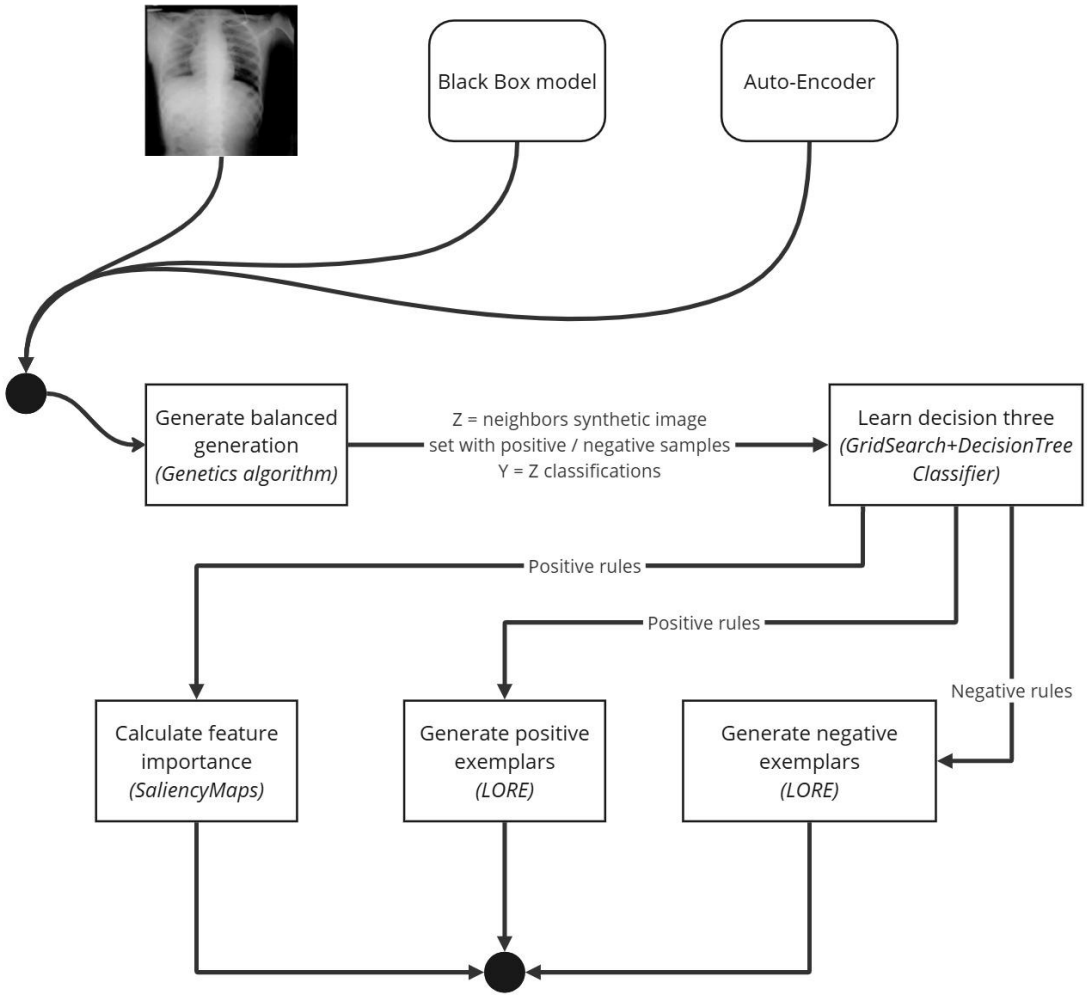

Figure 25 ABELE general flow

LORE is a robust framework designed to generate localized and interpretable explanations for machine learning models. LORE's explanations are composed of two primary elements. First, it generates a decision rule derived from the logic of the black-box model. This rule highlights the key factors influencing the model's decision, bringing light on the most relevant features and their impact. Additionally, LORE provides a set of counterfactual rules, offering alternative scenarios in which changes to the input features would result in a different outcome. By presenting actionable suggestions for modifying input variables, LORE allows users to explore "what-if" scenarios and understand how small adjustments can affect the model's predictions.

To summarize, the ABELE method generates explanations based on a given sample (local method) and a black-box model. It uses the concept of latent space to conclude proximity between instances, so the autoencoder is also relevant and fundamental to the method. Explanations are generated through three evidences:

- Feature importance saliency map (visual explanations highlighting the more relevant areas in the predicted image)
- Set of positive synthetic neighbours' images (with the same classification)

- Set of negative synthetic neighbours' images (with different classification)

To implement the ABELE model to this work, an open-source library named XAI-Lib, that implements many other XAI methods, was used https://github.com/kdd-lab/XAI-Lib. In the following sections, all the details about each implementation blocked will be described.

4.2.1 Auto-encoder

According to (Bergmann, n.d.), a latent space in machine learning algorithms is a compressed representation of samples that keep only essential features, this representation is enough to reconstruct the original data point with some level of proximity. Latent space modelling is an essential part of deep learning solutions. The concept is connected to dimensional reduction and is mandatory to make it possible to execute highly cost computation operations during machine learning training and inference reducing the complexity of data points.

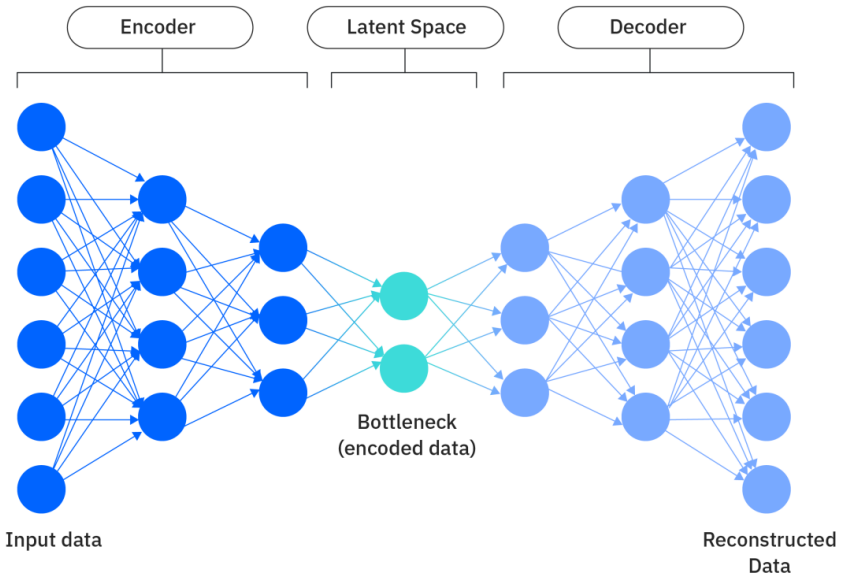

Figure 26 Latent space representation (Bergmann, n.d.)

The COVID-19 model represents an image of 70x70 pixels coloured in 3 channels as a 3-dimensional array (70x70x3) but as long ABELE extracts conclusions about the model decision making process navigating in the latent space and inferring proximity between samples (real and synthetic), an auto-encoder is needed. In the experimentation for this work, an adversarial auto-encoder was implemented.

An adversarial autoencoder is a type of encoder that integrates adversarial training to enforce a structured latent space. It consists of an encoder that compresses input data into a latent representation and a decoder that reconstructs the original input. AAEs add a discriminator,

like a Generative Adversarial Network, which intents that the learned latent representations follow a predefined prior distribution (Gaussian for example). The encoder is trained to generate latent representations that fool the discriminator, resulting in a structured latent space. This adversarial regularization makes AAEs more stable compared to variational autoencoders while still allowing control over the latent distribution.

Encoder, decoder and discriminators are implemented as neural networks. The encoder receives an image representation (70x70x3 array) and generates an essential representation in a single dimension array of 64 positions. This representation is used in several parts of the following actions during the ABELE execution. The decoder is dotted with layers that receive the latent space representation and convert it in an image in the original format (array of 70x70x3). The discriminator is used during the training to determine validity of encodings. Below the piece of code that initializes the auto-encoder model:

```
optimizer = Adam(0.0002, 0.5)
self.discriminator = self.build_discriminator()
self.discriminator.compile(loss='binary_crossentropy',optimizer=optimizer,
metrics=['accuracy'])

self.encoder = self.build_encoder()
self.decoder = self.build_decoder()

x = Input(shape=self.shape)
lx = self.encoder(x)  # latent representation (latent x)
tx = self.decoder(lx)  # reconstructed record (tilde x)
self.discriminator.trainable = False

validity = self.discriminator(lx)
self.autoencoder = Model(x, [tx, validity])
self.autoencoder.compile(loss=['mse', 'binary_crossentropy'], loss_weights=[0.999, 0.001],
optimizer=optimizer)
```

Code 2: Adversarial auto-encoder implementation

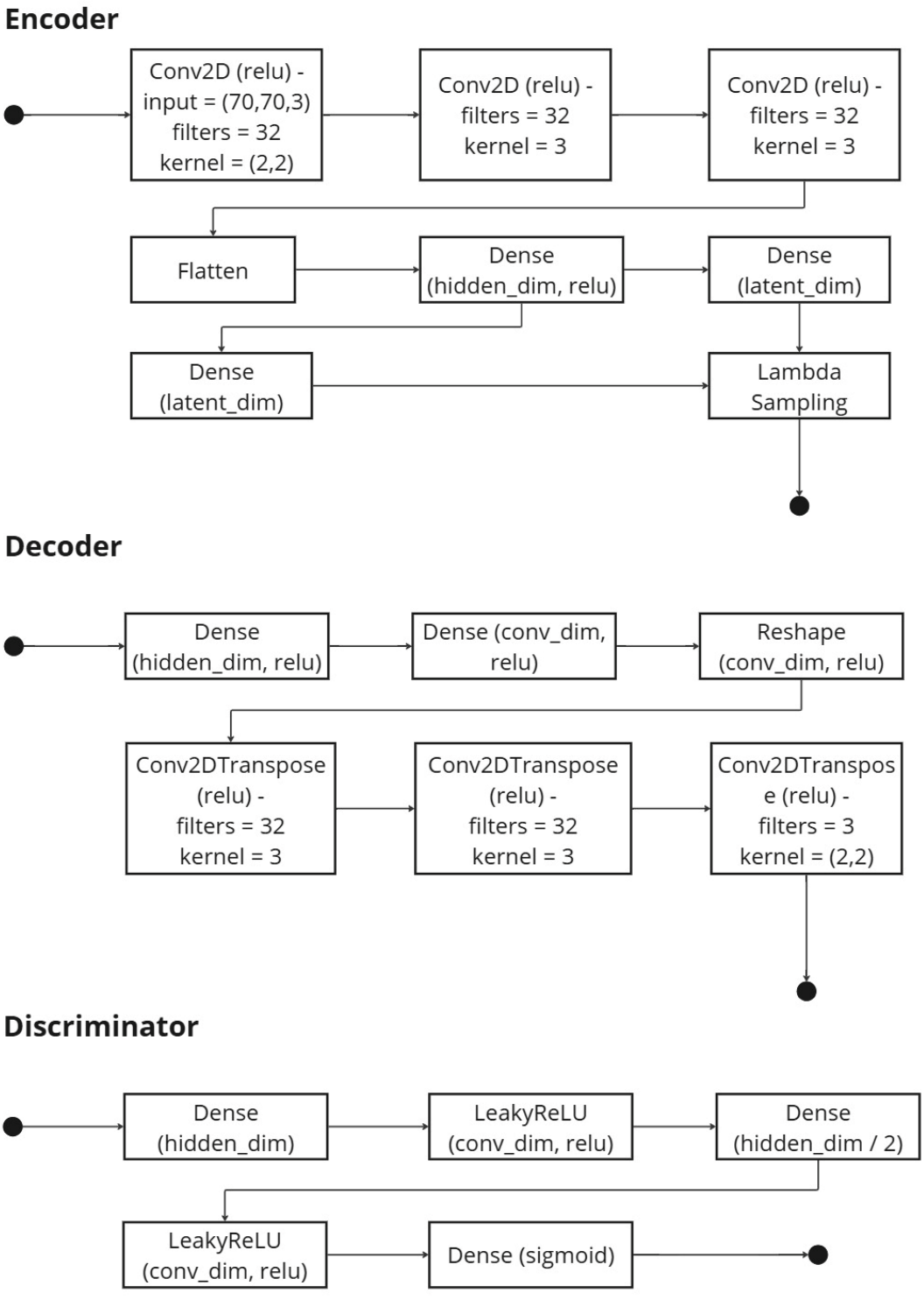

Figure 27 Encoder, decoder and discriminator neural networks architectures

During this experimentation work, some hyper parameters were adjusted to achieve the presented results:

- latent_dim: latent space dimension, basically defines the proper size of the latent space array, for this work it was defined as 64.
- hidden_dim: hidden dimension of the autoencoders layers, 1024 was selected for this work.
- num_filters: number of filters to use in the convolutional, 32 was selected for this work.
- ephocs: number of epochs in the train, 2000.

4.2.2 Balanced population generation

The following step in the ABELE process is to generate a balanced generation with samples with the same classification of the sample target of the explanation (positive) and samples with a different classification (negative). The data points are not just generated randomly, for the positive samples, feature proximity is considered to prioritize those are closer to the target sample, on the other hand, for negative data points, feature distance is prioritized. The final goal is to generate a population well-balanced (with positive and negative samples) considering the prioritization mentioned before.

To tackle this challenge, a genetics algorithm is used to start from a randomly create an initial generation and, according to the evolution process, mutate instances to generate more samples and keep the best ones according to fitness functions for positive (equal instances) and negative (not equal instances).

Both fitness functions make usage of feature similarity (or difference) calculated using the target sample encoding and a random synthetic sample. It uses Euclidean distance metric to calculate feature proximity. Bellow the code of both fitness functions:

```
def fitness_equal(self, x, x1):
    feature_similarity_score = 1.0 - cdist(x.reshape(1, -1), x1.reshape(1, -1), metric=self.metric).ravel()[0]
    feature_similarity = sigmoid(feature_similarity_score) if feature_similarity_score < 1.0 else 0.0

    y = self.bb_predict(self.autoencoder.decode(np.array([x])))
    y1 = self.bb_predict(self.autoencoder.decode(np.array([x1])))

    target_similarity_score = 1.0 - hamming(y, y1)
    target_similarity = sigmoid(target_similarity_score)
    # print(target_similarity_score, target_similarity)

    evaluation = self.alpha1 * feature_similarity + self.alpha2 * target_similarity

    return evaluation,

def fitness_notequal(self, x, x1):
```

```
    feature_similarity_score = 1.0 - cdist(x.reshape(1, -1), x1.reshape(1, -1),
metric=self.metric).ravel()[0]
    feature_similarity = sigmoid(feature_similarity_score)

    y = self.bb_predict(self.autoencoder.decode(np.array([x])))
    y1 = self.bb_predict(self.autoencoder.decode(np.array([x1])))

    target_similarity_score = 1.0 - hamming(y, y1)
    target_similarity = 1.0 - sigmoid(target_similarity_score)
    # print(target_similarity_score, target_similarity)
    evaluation = self.alpha1 * feature_similarity + self.alpha2 * target_similarity

    return evaluation
```

Code 3: Fitness functions to calculate proximity in the problem space

The final outcome of this step is a balanced population of synthetic images selected according to their proximity (positive) or distance (negative) to the target image.

4.2.3 Decision tree

Using the set of synthetic encodings selected in the last step along with their classifications, a decision tree is derived to find positive and negative rules that correlate features in the latent space, thresholds and results, for example:

r = { 35 > -0.91 } --> { class: 1 }
c = { { 35 <= -0.91 } --> { class: 0 } }

It means that if the feature 35 is greater 0,91 our result is positive. This rule set try to capture mathematically how the decision-making process behind the decision is made. The usage of decision trees to derive rules from samples and classifications is not exclusivity from ABELE, (Singhal et al., 2024) also cites the same idea in their article about cancer classification images.

4.2.4 Feature importance

This is one the outcomes of the ABELE method. Using a set of synthetic samples that are compliant with the same rules discovered in the last step and have the same classification of the target sample, a feature importance map is calculated and plotted in the target image to highlight attention areas from yellow (affecting negatively) to blue (affecting positively) (Figure 28).

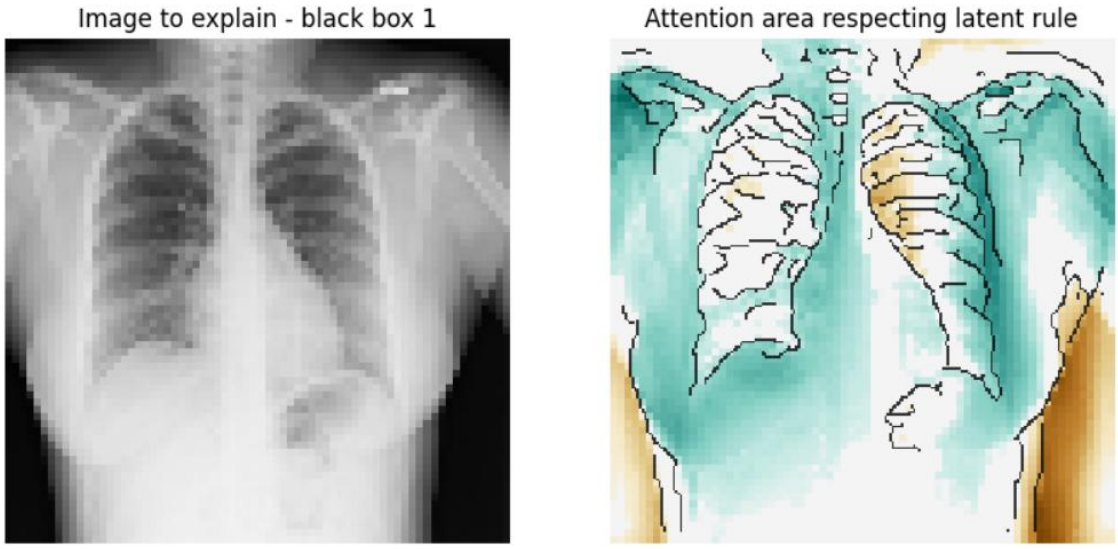

Figure 28 Example of a saliency map generated by the ABELE method

The positive hottest areas (strong tone of blue) can be corroborated by (*Chest Radiographic Findings in COVID-19*, n.d.) Figure 23. Peripheral areas in the lungs can provide visual signs about the presence of COVID-19 according to the paper. Those areas were also activated in the Saliency map.

The same method also produced saliency maps for the skin lesion classifier model, in the samples analysed, the lesion areas was properly identified along with some small spots in the skin. In the Figure 29, two saliency map produced evidence that on which areas affected the prediction positively (blue) and negatively (yellow).

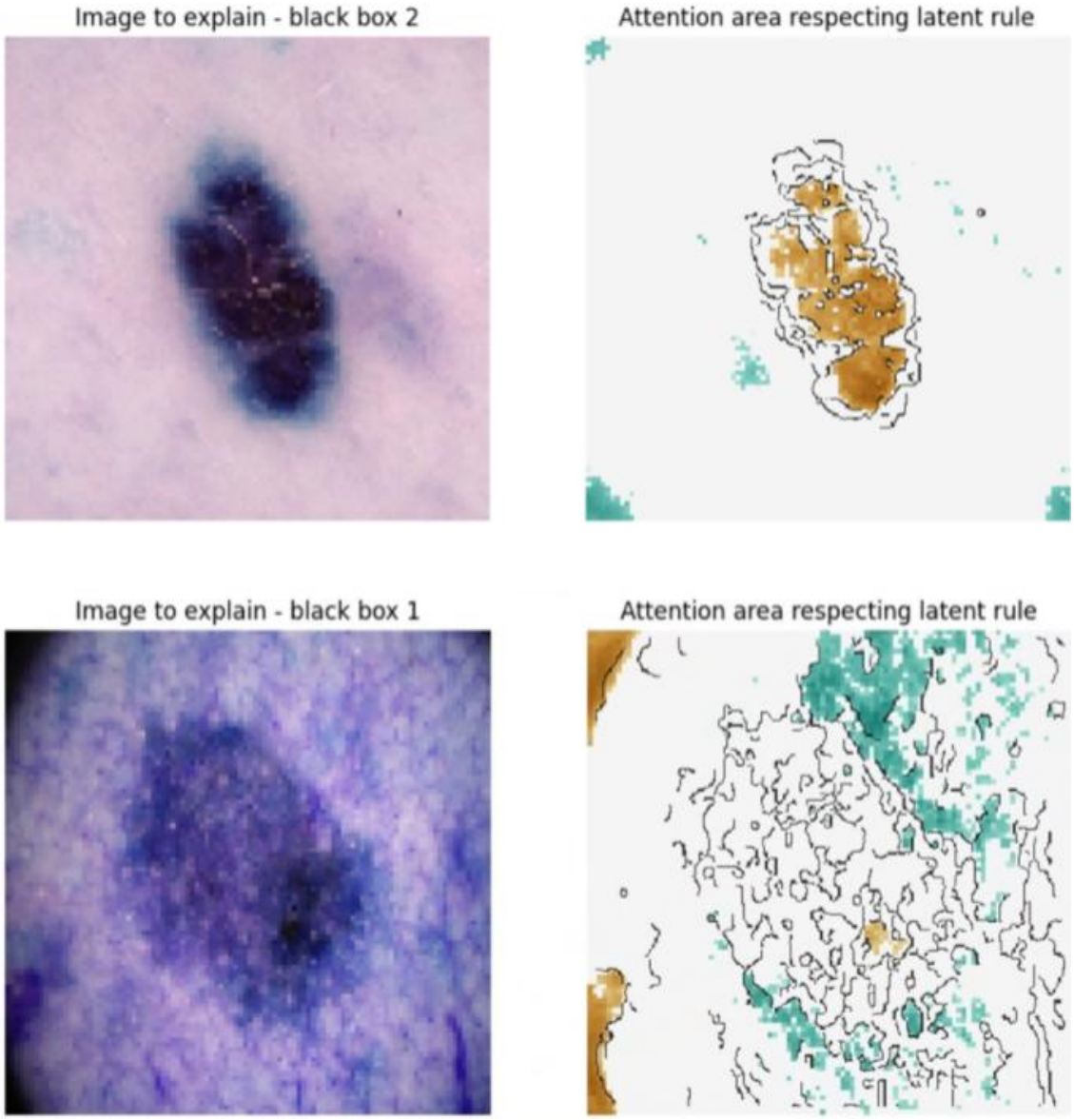

Figure 29 Saliency map created for a skin lesion sample

4.2.5 Generate positive exemplars

To make transparent the model decision-making process to the final users another outcome of ABELE is generated: set of positive exemplars or a group of synthetic images that are compliant with the same rules derived in the Decision tree step.

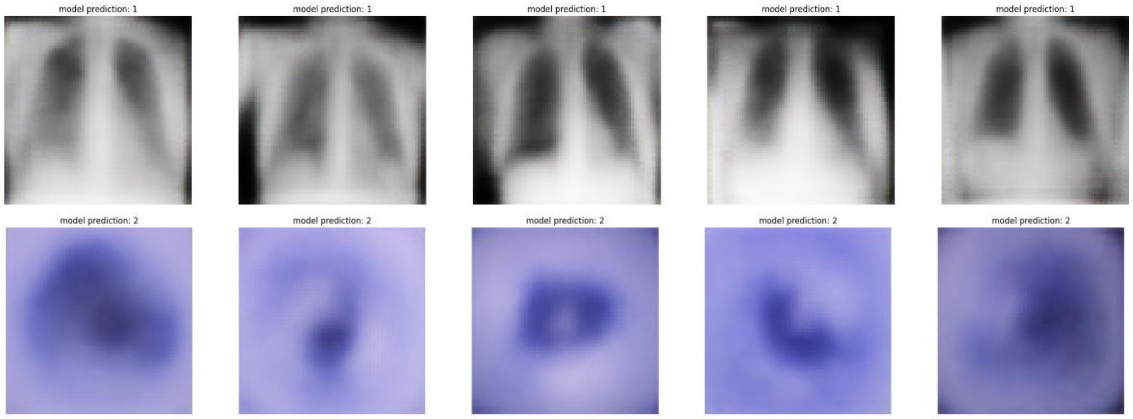

Figure 30 Set of synthetic positive examples for both models

The idea is to present a couple of examples to the final users that are classified with the same class of the target sample and more than it, have the same rules activated.

4.2.6 Generate negative exemplars

The last outcome of ABELE a set of synthetic images classified initially positively (as the target sample) and then converted into negative samples after applying small perturbations to the features values according to the rules derived in the Decision Tree step. For example, one rule says that if feature 38 is greater than 0,91 the class is 1, in this case, if this feature is set to 0,90 most likely the model outcome will change.

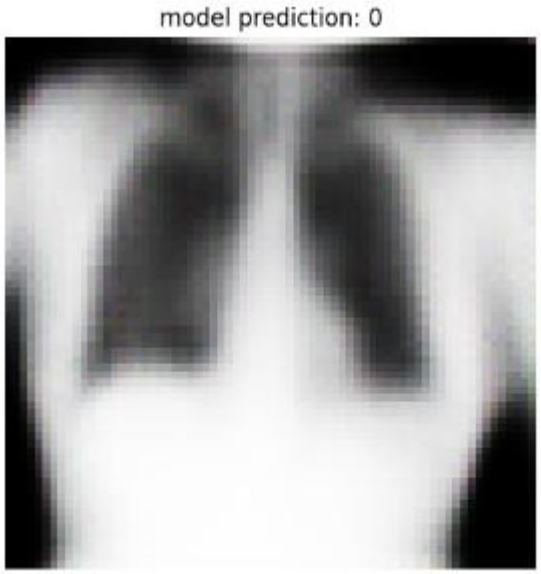

Figure 31 Synthetic negative examples

The idea is to present negative samples in the border line of the space problem as long they were generated manipulating the features that were significant to target sample prediction.

4.3 SHAP

The last XAI method explored was SHAP - SHapley Additive exPlanations. It's an agnostic method than can be used to explain predictions in any ML model. It's based in the games' theory, and it evidences how each feature contributed for a given classification. SHAP method basically works calculating Shapley values for each feature. To do so each feature value is considered a 'player' in a 'game,' while prediction value represents the "reward." (Molnar, 2024) presents an example connected to an apartment prediction's regression model: according to the model outcomes, an apartment's worth to be €300,000 while the rest of the model's predicted price for all the other apartments is €310,000. It is essential to understand how features like area, floor number, distance from a park, and banning cats led to this €10,000 difference. With the Shapley value method, this difference is allocated among the features which contributed to the Shapley value, ensuring each feature has an explanation to support the prediction. For instance to analyse the contribution for "cat-banned", predictions are calculated by incorporating and excluding the feature in conjunction with other features that are randomly filled in apartments where values are not available. The change in prediction is used as the feature's marginal contribution for the specific coalition. The entire process is done across all possible coalitions, and the average is calculated to determine the Shapley value.

The two models analysed in this work were image models, so the features are basically the image pixels. Using SHAP we can see visually how each pixel impacted the prediction positively or negatively. An interesting feature that can be explored with images classifier using SHAP is the feature importance calculation for more than one class.

It's possible to see, for example, that a given image how the model considered a given region not only to the predicted class but also how that given region contributed pushing the prediction towards other classes. Both models needed to use the GradientExplainer available in the SHAP library to generate meaningful Shapey values. The implementation is very simple using the mentioned library:

```
class_names = ["0", "1"]
explainer = shap.GradientExplainer(model, x_train)
shap_values = explainer.shap_values(x[:10])
shap.image_plot([shap_values[i] for i in range(2)], x[:10])
```

Code 4: SHAP Gradient explainer implementation

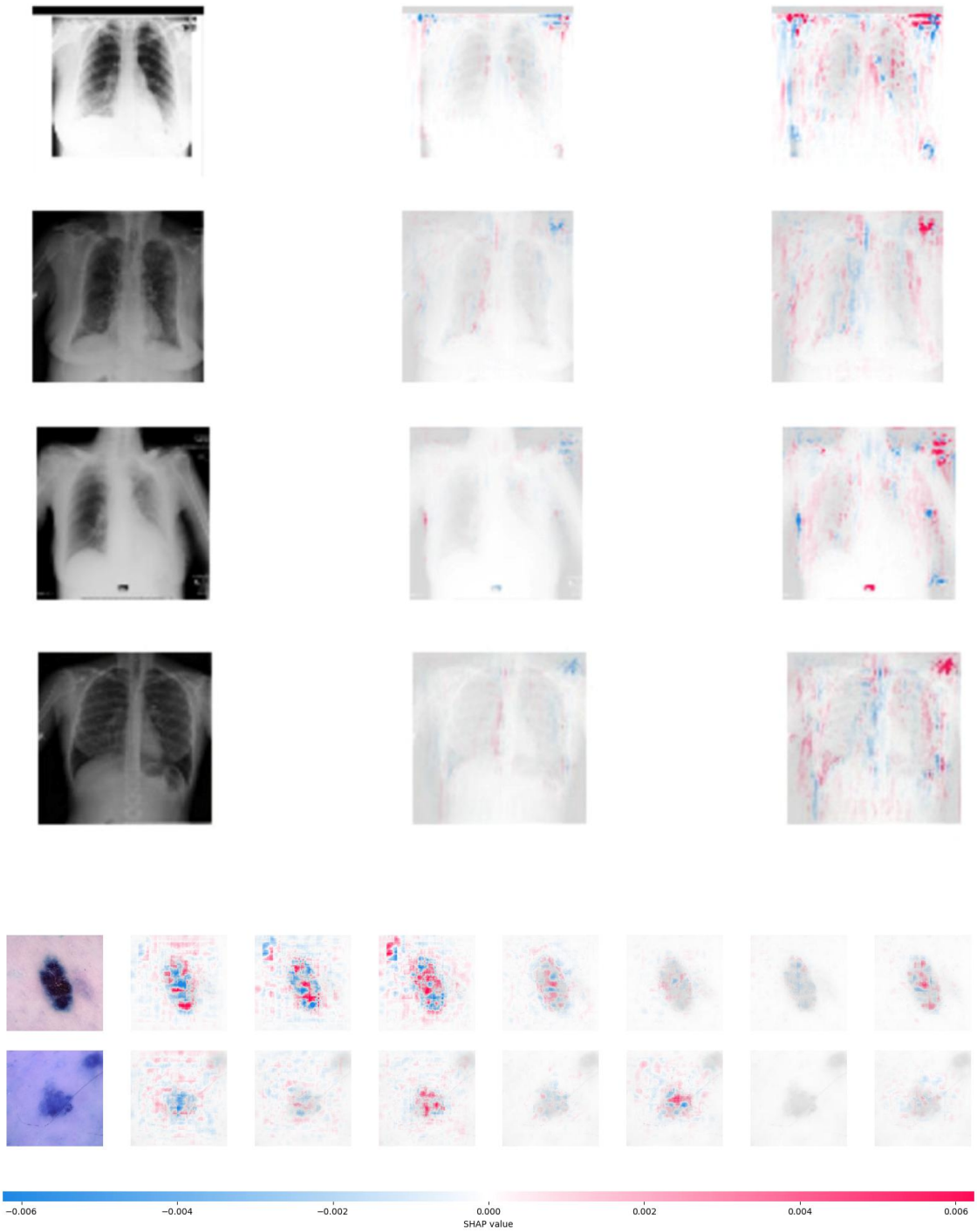

Figure 32 Shap values calculated for some Skin lesion and COVID-19 model samples

For the COVID-19 model, a binary classifier, SHAP method was applied to four positive samples, the SHAP values for the positive class can be visualized in the right image while the images in the left present the values for the same images in the negative class. It's possible to see some common areas such as the lungs perimeter in the negative and positive classes evidencing the model decision making.

For the skin lesion classifier, a multi-classifier model, it's interesting to highlight the first sample, a melona case where other two classes, melanocytic nevi and benign keratosis-like lesions, were closed to be chosen what becomes evident since the three first images have features similar importance map.

4.4 Discussion

So far, after the systematic review and the initial experimentation, it's possible to conclude that XAI such as Grad-CAM can enhance the trustworthiness of medical systems. In the experimentation, it was possible to visualize how the DL model analysed chest images, identifying correctly the lungs and which areas were crucial to the prediction (COVID-19 positive or negative).

According to (Harikumar et al., 2024) and (Lamba & Rani, 2024), visual explanations generated by Grad-CAM can not only bring more confidence and reliability to the system output but also help the technical staff to identify specific regions that they haven't checked before. As long the heat map highlights specific areas in images used to diagnosis, these regions might contain some relevant information (micro-lesion, color, etc) according to the model training (so considering the dataset) that a doctor could not note without this assistance.

When it comes to compliance or auditing questions, the heatmaps can be a tool to provide more details about the decision-making process when analysed by medical experts that assess if the areas considered relevant by the model are correct and, more important than it, if no other important region has been missed.

Speaking about ethical concerns, the heat maps can also be useful to identify bias when if the model can just predict considering as relevant some specific regions in the image disregarding other relevant areas according to medical experts (Patel et al., 2024). It may be caused by some unbalance in the dataset containing much more samples where just that region is relevant.

Even though heatmaps can provide relevant information about the decision-making process for individual inferences, it cannot bring clarity about the entire model that could made possible to check some important factors with no need of multiple individual predictions assessments. The expectation is that some global method such SHAP could help to address this issue. The expectation is the Explanation module that will be integrated to the DL model will count with the three XAI methods complementing each other to fulfil many aspects of the decision-making process in the COVID-19 model.

The Grad-CAM method was applied to the COVID-19 model in order to clarify their decision model process and bring transparency to its decisions. The experimentation could produce satisfactory evidence about how the COVID-19 make decisions. Using the state of the last convolutional layer after an inference, GRAD-CAM produces a visual heat map with the key areas in each prediction.

The visual evidence produced on top of the COVID-19 model was corroborated by (*Chest Radiographic Findings in COVID-19*, n.d.) in terms of relevant areas to visually inspect during a chest x-ray evaluation for COVID-19. Another relevant aspect is that the heat map generation is not computation expensive, what means that adding this feature to a potential explainability module would not face many technical challenges or increase deeply the cost of the solution. The heat map is generated according to the weights of the last convolution layer, no new calculations or complex routines need to be executed after the model prediction.

While experimenting GRAD-CAM with the Skin lesion model, the connection with convolution layer block size was evident. In the CNN model architecture, the last convolution layer has a block size of 5 resulting in a heat map of 5x5 transposed in the analysed image. This small heatmap is calculated using the layer gradients and after some final rounding and normalization operations a heatmap with 0 values were produced for some images, it could affect the usage in a real system since no region was highlight.

The ABELE method was applied to the COVID-19 model to explore what kind of evidence about the model decision making process it would bring. In order to explain why a given image was classified positively with COVID-19, ABELE applies Generative Artificial Intelligence techniques to generate positive exemplars (synthetic images also classified as positive with proximity in terms of key features) and negative counter factual exemplars (synthetic images classified as negative after small perturbations in key features to the sample being analysed).

In the experimentation section, the process of integrating ABELE to the COVID-19 was described. ABELE has three outcomes: set of positive exemplars, set of counter factual exemplars and a saliency map. The saliency map was able to present visually which areas most activated and attracted attention from the model, peripheral areas were activated what makes sense according to (*Chest Radiographic Findings in COVID-19*, n.d.). Saliency maps present the same benefits of Grad-CAM heatmaps in terms of compliance (producing evidence about each area were relevant for each inference) and bias mitigation (helping to identify areas overrepresented in the dataset).

Positive exemplars (or synthetic images classified with the same class of a target sample being explained compliant with the same rules that activated the class prediction) were generated to present visual support to the rules that were fundamental to the inference. For example, if a specific format of a micro lesion produced a specific model outcome, other images with the same lesion would be generated to support the model decision. For the COVID-19 model, although, the exemplars weren't so effective due to the quality of images generated after the encoding/decoding process and due to the fact of the images used to train and predict represent a medium size area (human-chest in this case) so evidence small characteristics can not be simple. (Metta, Beretta, Guidotti, Yin, et al., 2024) present a study related to skin cancer detection using micro lesion images, in this situation, the positive examples seem to be more effective.

Negative counter factual exemplars (or synthetic images not classified with the same class of a target sample explained and generated with small perturbations on key features for the

activated prediction) were also generate to disclose details about the decision-making process. For example, if a micro lesion is significant to generate a specific diagnosis the counter factual samples would be images with a slightly different lesion to make it clear how the model made the decision. According to the same points listed for positive samples, it was not so effective for COVID-19 model due to the image's characteristics.

One relevant aspect of integrating ABELE to any explainability module is that all the three outcomes rely on heavy computation processes that could be complex to execute in real-time (along with the prediction for example). To derive the decision tree, ABELE requires a balanced population of synthetic images with the same and opposite classification of a target sample (the one being explained), this population is generated using a genetics algorithm. The counter factual and positive prototypes generation also rely in some iterative methods that can take time to present results.

While experimenting ABELE in the Skin lesion model, the importance of the auto-encoder become even more evident, skin lesion images are more sensible to quality loss caused by the AAE in the synthetic samples generated, that's why only the salience map could have some real meaning in a health system, the synthetic positive and negative samples produced didn't have enough quality to be understood by the medical staff. Many attempts have been made to increase the quality of the generated images such as changing the latent space representation size, update the internal layers size, adapt the optimizer learning rate but nothing produced some visible improvement.

SHAP was also applied to both models producing heat maps for each possible class in the model (not only to the predicted class) to evidence which pixels contributed (positively or negatively) to that given classification. This feature (heat maps for all classes) can be interesting to show boundaries between two classes when more than one has relevant likelihood according to the neural network output. In terms of integration in real systems, SHAP images can be generated with reduced computational effort what would simplify its usage in a given explainability module not compromising the overall system performance.

The Skin lesion classifier is a multi-class model that brings more value to the SHAP analysis, as long inferences closed to more than one class can evidence which areas push the classification towards each of them bringing more clarity about what made the model decide by one of them.

The SHAP images evidence how the model decided between multiple classes in the model and, as Grad-CAM and ABELE Salience Maps could be beneficial to improve the confidence in the system output mitigating trustworthiness issues.

4.5 Explainability Module

One of the goals of this work was idealize how an explainability model could be integrated to any real health care system using AI to boost the confidence in the system results. According to the experimentation using COVID-19 and skin lesion classifier models, it was possible to conclude this module would use two of our the analysed methods: Grad-CAM and SHAP.

These two modules were able to produce meaningful evidence for both models while ABELE presented as the better result the saliency map, the positive and negative samples had their impact affected by poor quality of the images generated by the AAE.

While designing a module to be integrated in a real system, computational resources and mainly the time required to generate the evidence are also factors to be considered so it also contributed to the choice of Grad-CAM and SHAP that can produce evidences and artifacts in a reduced timeframe not compromising the system overall performance or requiring significant amount of extra wait for the system output.

As explained in more detail in the initial section of the Discussion chapter, Grad-CAM would evidence the high attention areas for a given diagnosis analysis providing more information about how the model predicted the output. SHAP would complement the analysis, providing more information about how the model differentiate the chosen classes from the other possible classes.

In the Figure 34 the proposed module interface is presented highlighting the analysed image in the top left corner and the classification in the center. To disclose details about the decision-making process the Grad-CAM heat map is presented in the left top corner and the SHAP images are presented in the bottom.

In this sample, the classification is Dermatofibroma, the SHAP images present how other classes were evaluated in the same image and how specific points (in red) pulled the decision out of that class. It's possible for example to see that other classes such as Benign keratosis and Basal cell carcinoma were relevant in the decision-making process.

As long Grad-CAM is not agnostic (can just be connected to CNN models) as SHAP, this module could be integrated into convolutional models. The module would disclose more details about how this given model predicted some specific diagnosis boasting the confidence in the system output and providing evidence that could be used to mitigate trustability questions. The module could also be used to produce artifacts in a regulatory audit to attest the systems well-functioning mitigating legal questions.

Grad-CAM and SHAP are also methods that can be integrated into DL models not requiring any change in the model architecture, so with no risk of impacting the model performance. This condition is not true for all XAI methods, some of them can impact the model design and performance.

Another contribution of the this module is the fact of the heatmaps can be not only to confirm diagnosis but also to support the treatment highlighting areas that should be analysed by the medical staff in the course of the treatment, this aspect was referred by (Harikumar et al., 2024) and (Lamba & Rani, 2024) as some hidden benefits of this kind of XAI method. The idea is simple, if the model paid attention in a given area it can be a hint for the medical stuff check it. In the same direction, (Hroub et al., 2024) mentioned this characteristic, high importance areas being highlighted could also be beneficial for less experienced professionals.

It's important to reiterate that this proposed explanation is scalable in the sense it does not affect the overall system response time. As mentioned before, Grad-CAM and SHAP are not compute intensive methods, so they could be integrated with no risk of making the medical staff to wait too long for the system output. The explainer module could also run both explainers in parallel (SHAP and Grad-CAM) to save even more time.

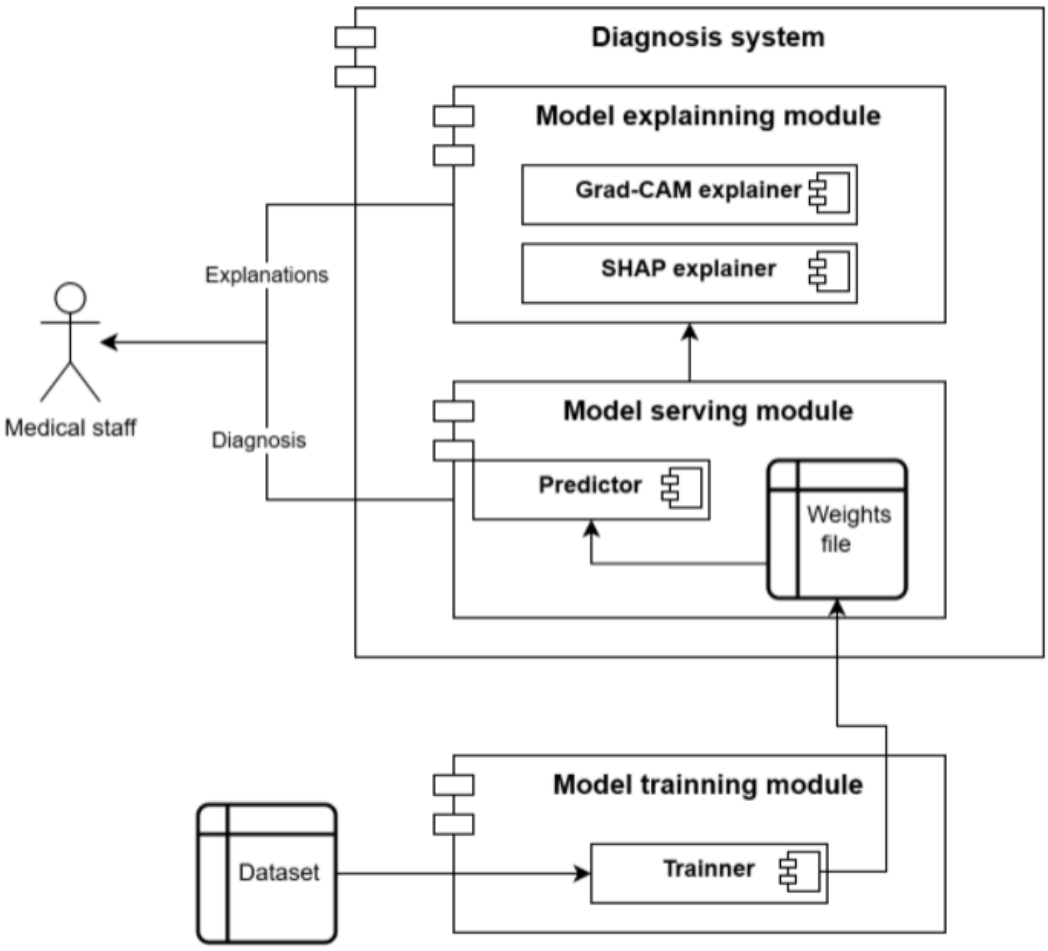

Figure 33 Explainer module integrated into a real system

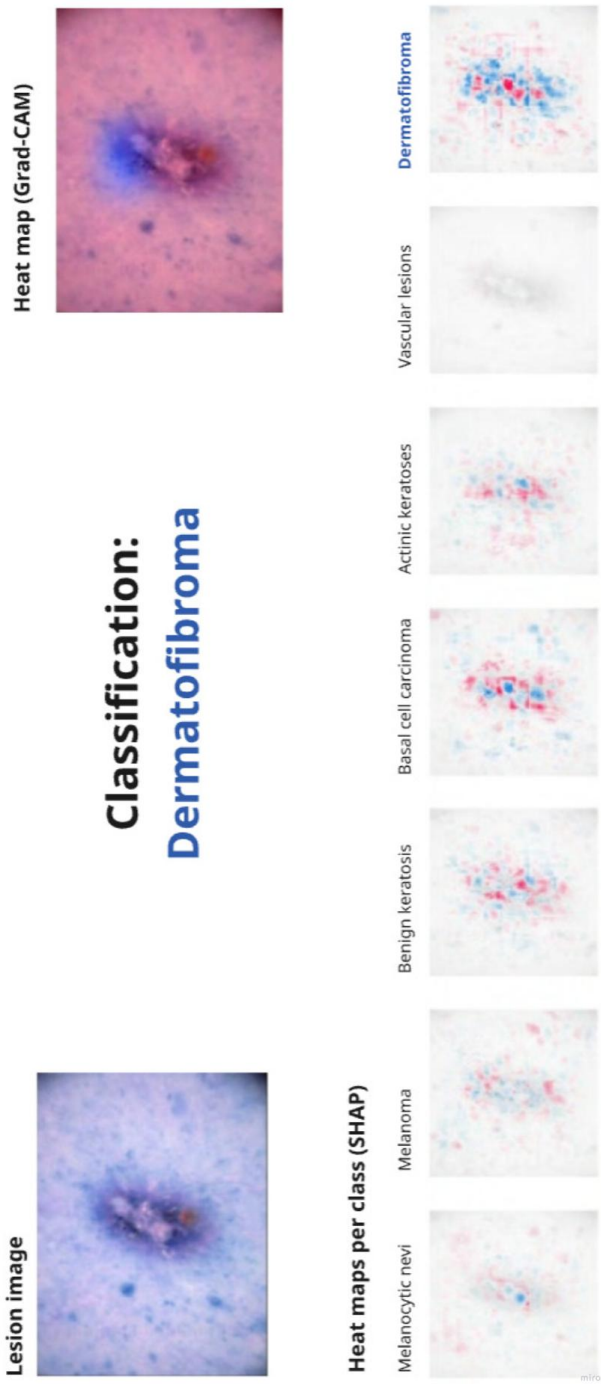

Figure 34 Proposed explainability module

4.6 Results Comparison

The systematic review presented in the second chapter of this work reviewed several works in the health care field using images to support diagnosis decisions. This section aims to compare the achievements and conclusions from this work with other academical work.

(Metta, Beretta, Guidotti, Yin, et al., 2024) presented a paper to boast the confidence in a skin classification system using ABELE XAI method as it was done by this work. They highlighted the same challenges we identified in this area: the fact of DL systems are usually black-boxes bring challenges in their adoption as long the decision-making process is obscure. They decided to apply ABELE to produce multiple evidence and artifacts related to model decision making process. As this work concluded, they also referred to the importance of the AAE being able to produce high-quality images: "*the efficacy of ABELE heavily relies on the quality of the encoder and decoder functions used; the more effective the autoencoder, the more realistic and valuable the explanations become.*". They also presented the ABELE explanations to a group of health professionals to assess their effectiveness to improve or not the trustworthiness and concluded that the confidence gain varied according to the professional level, experienced professional felt more comfortable with more detailed explanations while novice prefer less detailed evidence.

(Thapar & Tiwari, 2024) used the same dataset explored in the experimentation chapter, HAM10000, to create skin lesion classification models. Their focus was more on the classification models using multiple architectures than in XAI methods so the paper didn't present strong conclusions about the explanations. Anyway, they experimented some methods to explain the classifiers output: Faster Score-CAM and SmoothGrad, both producing some variations of heat-maps as Grad-CAM explored in this work. The same choices were made by (F. Mahmud et al., 2023) using HAM10000 to explore skin cancer classifiers and selected Score-CAM and SmoothGrad as explainers also without big focus on the explainers performance. In general, all the studies decided to use XAI to disclose more details about some DL model central focus of the pager bringing more transparency adding visual explanations on top of the normal performance metrics what corroborates the conclusion of this work that XAI methods are important and effective for health-case image systems.

(Akbar et al., 2024) presented a study to classify chest radiographies into positive or negative COVID-19 cases. As the model experimented in this work, they also built a CNN model to inference the diagnosis and used Attention mechanism to produce heatmaps to evidence areas with high or low influence in the decision. In this work, one of conclusions was that XAI methods based in heat maps can produce good evidence about image healthcare decision making process. It's also corroborates this work conclusions, that visual explanations offer a strong support for clinicians understanding how the model decided for some classification.

(Hassan et al., 2024) also presented a paper related to COVID-19 diagnosis prediction but their choice to explain their model was LIME transposing the feature importance into a heat map. (Ukwuoma et al., 2023) also presented a paper related to COVID-19 detection using ML models, it also uses lung x-ray images to train a CNN model and relies on Grad-CAM heatmaps to explain the model output exactly as experimented in this work.

The analysis of the related work presented a clear tendency in health-care systems based on image inference: the usage of heatmap like solutions to explain CNN models. As discussed earlier, visual explanations can bring insights into the model decision making process,

highlighting and lowlighting areas according to the importance to the model disclosing details to the medical staff. It's also corroborated by the Systematic Review conducted in the second chapter of this work, the question *"RQ 2: How can XAI leverage trustworthy for image classification use cases?"* was replied elucidating the XAI methods most recurrent in the literature: Grad-CAM and other heatmap solutions are the most used.

5 Conclusion

This work aimed to investigate how Explainable Artificial Intelligence (XAI) can contribute to mitigating legal, ethical, and trustworthiness issues in medical imaging systems, a field where transparency and reliability are essential due to its direct impact on human lives.

To address this goal, the research began by identifying the most recurrent challenges within medical imaging systems. Through a systematic literature review, it was found that the predominant concern is the lack of transparency in deep learning (DL) models, often described as "black boxes." This opacity reduces clinicians' ability to trust and validate the outputs. Other significant challenges include bias in data and models, data privacy, the quality of datasets, and compliance with regulatory standards.

XAI has emerged as a promising solution to address these challenges. The review revealed a wide adoption of visual explanation methods like Grad-CAM and its variants, as well as feature-based methods such as LIME and SHAP. According to the explored literature, these techniques have been used to provide mainly local interpretability, offering insight into individual predictions, for health-care system using images usually the feature importance is translated in a heat-map transposed in the input image what make simple to any human understand which areas were more relevant in a given decision. Global interpretability is also referred but this usage in the context is limited. Such transparency not only improves trust among clinical professionals but also supports compliance with legal standards. (Patel et al., 2024) refers the fact that regulatory bodies often need comprehensive documentation about decision-making process what could be achieved through XAI evidence that shows the reasoning behind the system output. (Chaddad et al., 2024) also referred how graphical evidence produced by Grad-CAM could be integrated into regulatory frameworks.

XAI methods are also recurrently mentioned to identify bias, (Alomar et al., 2023) refer that methods that reveals more important areas in a given input image producing a heatmap such as Grad-CAM can be used to determine which attributes the model is employing, this helps on

spot any possible biases or inaccuracies in the model's predictions. (Ueda et al., 2024) also corroborate this fact but refers a relevant concern: bias confirmation, when people only favour explanations that confirms their beliefs. XAI heatmaps, for example, could be mis interpretated by medical staff that just look for confirmation if specific regions were analysed that could confirm their own belief disregarding the fact that other "non-expected" areas were also relevant in the inference. (Andrade & Alves, 2024) reinforces that XAI saliency maps solutions can disclose details about the decision-making process for a given prediction in order to understand how any unethical or unexpected connections is being made.

Independent and regular audits using performance metrics and XAI explanations on top of well-distributed test datasets are necessary to address potential bias and ensure the system remains fair over the time.

The thesis identified and evaluated over 20 XAI methods, narrowing the focus to three particularly relevant techniques for experimentation: Grad-CAM, SHAP, and ABELE. These were chosen based on prevalence in literature and their complementary characteristics—covering visual, statistical, and generative explainability approaches. These methods were then applied to DL models trained on COVID-19 chest radiographs and skin lesion classification datasets. The practical implementation of these explainers showed that Grad-CAM effectively highlighted relevant image regions influencing model decisions. SHAP provided quantifiable feature importance values supporting local explainability not only in the predicted class but also evidencing how each area in the image pushed to the decision in or out every possible class. ABELE introduced a generative approach using exemplars and counter-exemplars to justify decisions, but the experimentation concluded that its explanations were not so effective due to the quality of the images produced.

Finally, the thesis proposed an explainability module that integrates these XAI methods into real-world medical systems. Such integration could enable clinicians to receive not only a diagnosis but also a justification backed by visual evidence, increasing system transparency, and fostering trust in AI-assisted healthcare. Supported by the experimentation results, the proposed module would use Grad-CAM and SHAP images to explain the system output. Both methods are complementary since Grad-CAM explains how the model choose the predicted class while SHAP also helps to explain why the system does not chose the other classes. They are also less time-heavy in terms of the amount of time required to generate the explanations, ABELE for example relies on some iterative methods such as genetic algorithms that take time to produce results.

In conclusion, XAI is not merely a technical enhancement but a critical enabler of responsible AI adoption in medical imaging. Its integration into DL-based systems is vital for aligning with ethical standards, fulfilling legal obligations, and most important ensuring clinicians and patients can trust and act on AI-generated insights.

As described in the experimentation section, the quality of the synthetic images generated during the ABELE implementation wasn't enough to produce meaningful and interpretable results. It happened due to the auto-enconder implementation that wasn't able to represent

the image in the latent space in way that could be possible to have an enough quality representation back. In order to make the ABELE output more effective, some investment would be required in the AAE, testing other models and implementations.

As described earlier, hyper-parametrization adjustments were tried (batch-size, latent space size and hidden layers size) with no success. Exploration took place using an adversarial auto-encoder architecture (AAE), so other kinds of auto-encoders such as Variational or Convolutional could be explored.

Another aspect that could be explored is some sort of qualitative feedback about the evidence produced by explainers from health professionals. As discussed in the Results comparison section, this is one common path to evaluate the quality of the explainers outcome. A general feedback survey could be conducted in a group of professionals to hear their point of view about the effectiveness of the explainers and more importantly understand how they improved or not the trustworthiness of the system output. To make it even more complete, feedback from legal professionals could be used to attest the effectiveness of the explainer's evidence as proof for auditing purposes for example.

During the experimentation process, the focus was only on the explanation of 100% image-based models, but healthcare models commonly also use patient data such as age, sex, weight, etc. as model inputs. For example, a given lung cancer detection model could rely on radiography images but also in the patient age and if smoker or not. It would create additional opportunities for explainers: analyse how each of these features influenced in a given inference (local analysis) and even how they affect the whole model (global analysis). SHAP would be interesting to calculate SHAP values in the feature level producing strong evidence about how each feature contributed.

While this work was being written, Agentic AI emerged as a hot topic in the market to orchestrate AI agents using LLMs to produce a desired output. In this field, this study only identified (Grillo et al., 2024) where ChatGPT model is used to generate textual explanations about MRI images to support image segmentation results but nothing related to agents has been mentioned. This kind of orchestration solution might be explored in the context of interpretability of machine learning (ML) models in healthcare by providing detailed, context-aware explanations for their outputs. Leveraging an AI Workflow concept, these agents can autonomously gather domain-specific knowledge from scientific articles, technical reports, and medical databases to build enriched narratives around model predictions. For instance, when an ML model identifies a high probability of skin cancer from dermatological images, an AI agent can consult recent dermatological studies, clinical guidelines, and expert analyses to produce a thorough explanation of the model's decision-making process. This would include highlighting key features such as lesion asymmetry, border irregularities, and color variations while cross-referencing these characteristics with documented medical literature. This automated knowledge synthesis not only augments the interpretability of ML outputs but also bridges the gap between technical model outputs and actionable clinical insights, thereby enhancing decision support for healthcare professionals.

Additionally, the integration of AI agents into the AI Workflow would enable a continuous learning mechanism where agents iteratively refine their explanations based on new medical findings and model updates. For example, in the case of skin cancer detection, the agent could autonomously update its reference knowledge as new research emerges, ensuring that the generated explanations remain current and clinically relevant. This dynamic adaptation is particularly valuable in healthcare, where scientific advancements are frequent and critical for accurate diagnostics and treatment planning.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Akay, E. M. Z., Rieger, J., Schöttler, R., Behland, J., Schymczyk, R., Khalil, A. A., Galinovic, I., Sobesky, J., Fiebach, J. B., Madai, V. I., Hilbert, A., & Frey, D. (2023). A deep learning analysis of stroke onset time prediction and comparison to DWI-FLAIR mismatch. *NeuroImage: Clinical*, *40*, 103544. https://doi.org/https://doi.org/10.1016/j.nicl.2023.103544

Akbar, S., Azam, H., Almutairi, S. S., Alqahtani, O., Shah, H., & Aleryani, A. (2024). Contemporary Study for Detection of COVID-19 Using Machine Learning with Explainable AI. *Computers, Materials and Continua*, *80*(1), 1075–1104. https://doi.org/https://doi.org/10.32604/cmc.2024.050913

Akhlaq, F., Ali, S., Imran, A. S., Daudpota, S. M., & Kastrati, Z. (2024). Diving Deep into Bone Anomalies on the FracAtlas Dataset Using Deep Learning and Explainable AI. *2024 International Conference on Engineering & Computing Technologies (ICECT)*, 1–6. https://doi.org/10.1109/ICECT61618.2024.10581288

Alami, A., Boumhidi, J., & Chakir, L. (2024). Explainability in CNN based Deep Learning models for medical image classification. *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1–6. https://doi.org/10.1109/ISCV60512.2024.10620149

Alkhalaf, S., Alturise, F., Bahaddad, A. A., Elnaim, B. M. E., Shabana, S., Abdel-Khalek, S., & Mansour, R. F. (2023). Adaptive Aquila Optimizer with Explainable Artificial Intelligence-Enabled Cancer Diagnosis on Medical Imaging. *Cancers (Basel)*, *15*(5).

Alomar, A., Alazzam, M., Mustafa, H., & Mustafa, A. (2023). Lung Cancer Detection Using Deep Learning and Explainable Methods. *2023 14th International Conference on Information and Communication Systems (ICICS)*, 1–4. https://doi.org/10.1109/ICICS60529.2023.10330443

Amazon Web Services. (n.d.). *Interpretability versus explainability - Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions*. Retrieved November 17, 2024, from https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html

Andrade, O. M. de, & Alves, M. A. S. (2024). Using Explainable Artificial Intelligence (XAI) to reduce opacity and address bias in algorithmic models. *Revista Thesis Juris*, *13*(1), 03–25. https://doi.org/10.5585/13.2024.26510

Bardozzo, F., Fiore, P., Valentino, M., Bianco, V., Memmolo, P., Miccio, L., Brancato, V., Smaldone, G., Gambacorta, M., Salvatore, M., Ferraro, P., & Tagliaferri, R. (2024). Enhanced tissue slide imaging in the complex domain via cross-explainable GAN for Fourier ptychographic microscopy. *Computers in Biology and Medicine*, *179*, 108861. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.108861

Barua, K., Mahmud, T., Barua, A., Sharmen, N., Basnin, N., Islam, D., Hossain, M. S., Andersson, K., & Hossain, S. (2023). Explainable AI-Based Humerus Fracture Detection and Classification from X-Ray Images. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/ICCIT60459.2023.10441124

Bergmann, D. (n.d.). *What Is Latent Space? | IBM*. Retrieved February 28, 2025, from https://www.ibm.com/think/topics/latent-space

Biswas, S., Mostafiz, R., Uddin, M. S., & Paul, B. K. (2024). XAI-FusionNet: Diabetic foot ulcer detection based on multi-scale feature fusion with explainable artificial intelligence. *Heliyon*, *10*(10), e31228. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e31228

Bouabdallah, K., Drif, A., & Kaderali, L. (2024). A Novel Black-Box Complementary Explanation Approach for Thorax Multi-Label Classification. *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 1–8. https://doi.org/10.1109/IRASET60544.2024.10548809

Burgos, D., Morshed, A., Rashid, M. D. M., & Mandala, S. (2024). A Comparison of Machine Learning Models to Deep Learning Models for Cancer Image Classification and Explainability of Classification. *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 386–390. https://doi.org/10.1109/ICoDSA62899.2024.10651790

Carloni, G., & Colantonio, S. (2024). Exploiting causality signals in medical images: A pilot study with empirical results. *Expert Systems with Applications*, *249*, 123433. https://doi.org/https://doi.org/10.1016/j.eswa.2024.123433

Chaddad, A., Hu, Y., Wu, Y., Wen, B., & Kateb, R. (2024). Generalizable and Explainable Deep Learning for Medical Image Computing: An Overview. *Current Opinion in Biomedical Engineering*, 100567. https://doi.org/https://doi.org/10.1016/j.cobme.2024.100567

Chest Radiographic Findings in COVID-19. (n.d.). UCLA Health David Geffen School of Medicine. Retrieved January 26, 2025, from https://www.uclahealth.org/sites/default/files/documents/UCLA-covid19-chest-radiographic-findings.pdf

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. Bin, Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. Al, Reaz, M. B. I., & Islam, M. T. (2020). Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access*, *8*, 132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287

De Aguiar, E. J., Traina, C., & Traina, A. J. M. (2024). RADAR-MIX: How to Uncover Adversarial Attacks in Medical Image Analysis through Explainability. *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, 436–441. https://doi.org/10.1109/CBMS61543.2024.00078

Deepanshi, Budhiraja, I., Garg, D., & Kumar, N. (2023). Choquet integral based deep learning model for COVID-19 diagnosis using eXplainable AI for NG-IoT models. *Computer Communications*, *212*, 227–238. https://doi.org/https://doi.org/10.1016/j.comcom.2023.09.032

Dunn, J., Mingardi, L., & Zhuo, Y. D. (2021). *Comparing interpretability and explainability for feature selection*. https://arxiv.org/abs/2105.05328v1

Ellis, C. A., Miller, R. L., & Calhoun, V. D. (2023). Towards greater neuroimaging classification transparency via the integration of explainability methods and confidence estimation approaches. *Informatics in Medicine Unlocked*, *37*, 101176. https://doi.org/https://doi.org/10.1016/j.imu.2023.101176

Flores-Araiza, D., Villegas-Jimenez, A., Lopez-Tiro, F., Gonzalez-Mendoza, M., Rodríguez-Guéant, R.-M., Hubert, J., Ochoa-Ruiz, G., & Daul, C. (2024). On the Link Between Model Performance and Causal Scoring of Medical Image Explanations. *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, 1–8. https://doi.org/10.1109/CBMS61543.2024.00009

Gezici, G., Metta, C., Beretta, A., Pellungrini, R., Rinzivillo, S., Pedreschi, D., & Giannotti, F. (2024). *XAI in Healthcare ⋆*. https://github.com/riccotti/LORE

Ghnemat, R., Alodibat, S., & Abu Al-Haija, Q. (2023). Explainable artificial intelligence (XAI) for deep learning based medical imaging classification. *J. Imaging*, *9*(9).

Grillo, G., Torda, T., Voena, C., Ciardiello, A., & Giagu, S. (2024). Integrating ChatGPT-4: A Novel XAI Interface for Enhanced Clinician Understanding of MRI Image Segmentation Results. *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, 320–325. https://doi.org/10.1109/CBMS61543.2024.00060

Harikumar, A., Surendran, S., & Gargi, S. (2024). Explainable AI in Deep Learning Based Classification of Fetal Ultrasound Image Planes. *Procedia Computer Science*, *233*, 1023–1033. https://doi.org/https://doi.org/10.1016/j.procs.2024.03.291

Hassan, M. M., AlQahtani, S. A., AlRakhami, M. S., & Elhendi, A. Z. (2024). Transparent and Accurate COVID-19 Diagnosis: Integrating Explainable AI with Advanced Deep Learning in CT Imaging. *CMES - Computer Modeling in Engineering and Sciences*, *139*(3), 3101–3123. https://doi.org/https://doi.org/10.32604/cmes.2024.047940

Hauptman, A. I., Schelble, B. G., Duan, W., Flathmann, C., & McNeese, N. J. (2024). Understanding the influence of AI autonomy on AI explainability levels in human-AI

teams using a mixed methods approach. *Springer - Cognition, Technology and Work*, *26*(3), 435–455. https://doi.org/10.1007/s10111-024-00765-7

Heng, W. W., & Abdul-Kadir, N. A. (2023). Deep Learning and Explainable Machine Learning on Hair Disease Detection. *2023 IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 150–153. https://doi.org/10.1109/ECBIOS57802.2023.10218472

Hroub, N. A., Alsannaa, A. N., Alowaifeer, M., Alfarraj, M., & Okafor, E. (2024). Explainable deep learning diagnostic system for prediction of lung disease from medical images. *Computers in Biology and Medicine*, *170*, 108012. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.108012

Huang, S., Wang, L., Liao, J., & Liu, L. (2024). Multi-attentional causal intervention networks for medical image diagnosis. *Knowledge-Based Systems*, *299*, 111993. https://doi.org/https://doi.org/10.1016/j.knosys.2024.111993

Hussain, E., Mahanta, L. B., Borbora, K. A., Borah, H., & Choudhury, S. S. (2024). Exploring explainable artificial intelligence techniques for evaluating cervical intraepithelial neoplasia (CIN) diagnosis using colposcopy images. *Expert Systems with Applications*, *249*, 123579. https://doi.org/https://doi.org/10.1016/j.eswa.2024.123579

Islam, M. K., Rahman, M. M., Ali, M. S., Mahim, S. M., & Miah, M. S. (2023). Enhancing lung abnormalities detection and classification using a Deep Convolutional Neural Network and GRU with explainable AI: A promising approach for accurate diagnosis. *Machine Learning with Applications*, *14*, 100492. https://doi.org/https://doi.org/10.1016/j.mlwa.2023.100492

Jiang, P., Liu, J., Feng, J., Chen, H., Chen, Y., Li, C., Pang, B., & Cao, D. (2024). Interpretable detector for cervical cytology using self-attention and cell origin group guidance. *Engineering Applications of Artificial Intelligence*, *134*, 108661. https://doi.org/https://doi.org/10.1016/j.engappai.2024.108661

Kim, Y., Bu, S., Tao, C., Bae, K. T., Steinman, T., Wei, J., Czarnecki, P., Pedrosa, I., Braun, W., Nurko, S., Remer, E., Chapman, A., Martin, D., Rahbari-Oskoui, F., Mittal, P., Torres, V., Hogan, M. C., El-Zoghby, Z., Harris, P., … Wendler, D. (2024). Deep Learning–Based Automated Imaging Classification of ADPKD. *Kidney International Reports*, *9*(6), 1802–1809. https://doi.org/https://doi.org/10.1016/j.ekir.2024.04.002

Laguna, S., Heidenreich, J. N., Sun, J., Cetin, N., Al-Hazwani, I., Schlegel, U., Cheng, F., & El-Assady, M. (2023). ExpLIMEable: A Visual Analytics Approach for Exploring LIME. *2023 Workshop on Visual Analytics in Healthcare (VAHC)*, 27–33. https://doi.org/10.1109/VAHC60858.2023.00011

Lamba, K., & Rani, S. (2024). A novel approach of brain-computer interfacing (BCI) and Grad-CAM based explainable artificial intelligence: Use case scenario for smart healthcare.

Journal of Neuroscience Methods, *408*, 110159. https://doi.org/https://doi.org/10.1016/j.jneumeth.2024.110159

Latha, M., Kumar, P. S., Chandrika, R. R., Mahesh, T. R., Kumar, V. V., & Guluwadi, S. (2024). Revolutionizing breast ultrasound diagnostics with EfficientNet-B7 and Explainable AI. *BMC MEDICAL IMAGING*, *24*(1). https://doi.org/10.1186/s12880-024-01404-3

Lo, Z. J., Mak, M. H. W., Liang Shanying and Chan, Y. M., Goh, C. C., Lai, T., Tan, A., Thng, P., Rodriguez, J., & Weyde Tillman and Smit, S. (2024). Development of an explainable artificial intelligence model for Asian vascular wound images. *Int. Wound J.*, *21*(4), e14565.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. Del, Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301. https://doi.org/10.1016/J.INFFUS.2024.102301

Mahamud, E., Fahad, N., Assaduzzaman, M., Zain, S. M., Goh, K. O. M., & Morol, Md. K. (2024). An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning. *Decision Analytics Journal*, *12*, 100499. https://doi.org/https://doi.org/10.1016/j.dajour.2024.100499

Mahmud, F., Mahfiz, Md. M., Kabir, Md. Z. I., & Abdullah, Y. (2023). An Interpretable Deep Learning Approach for Skin Cancer Categorization. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/ICCIT60459.2023.10508527

Mahmud, T., Barua, K., Habiba, S. U., Sharmen, N., Hossain, M. S., & Andersson, K. (2024). An explainable AI paradigm for Alzheimer's diagnosis using deep transfer learning. *Diagnostics (Basel)*, *14*(3), 345.

Metta, C., Beretta, A., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S., & Giannotti, F. (2024). Advancing dermatological diagnostics: Interpretable AI for enhanced skin lesion classification. *Diagnostics (Basel)*, *14*(7).

Metta, C., Beretta, A., Guidotti, R., Yin Yuan and Gallinari, P., Rinzivillo, S., & Giannotti, F. (2024). Advancing Dermatological Diagnostics: Interpretable AI for Enhanced Skin Lesion Classification. *DIAGNOSTICS*, *14*(7). https://doi.org/10.3390/diagnostics14070753

Molnar. (2024). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*.

Mu, J., Kadoch, M., Yuan, T., Lv, W., Liu, Q., & Li, B. (2024). Explainable federated medical image analysis through causal learning and blockchain. *IEEE J. Biomed. Health Inform.*, *28*(6), 3206–3218.

Mukhtorov, D., Rakhmonova, M., Muksimova, S., & Cho, Y.-I. (2023). Endoscopic image classification based on explainable deep learning. *Sensors (Basel)*, *23*(6).

Nazir, M. I., Akter, A., Hussen Wadud, M. A., & Uddin, M. A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI. *Heliyon*, *10*(20), e38997. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e38997

Nikolić, M., Janković, D., Stanimirović, A., & Stoimenov, L. (2024). The Integration of Explainable AI Methods for the Classification of Medical Image Data. *2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, 1–6. https://doi.org/10.1109/IcETRAN62308.2024.10645095

Oliveira, M., Wilming, R., Clark, B., Budding, C., Eitel, F., Ritter, K., & Haufe, S. (2024). Benchmarking the influence of pre-training on explanation performance in MR image classification. *Front. Artif. Intell.*, *7*, 1330919.

Parola, M., Galatolo, F. A., La Mantia, G., Cimino, M. G. C. A., Campisi, G., & Di Fede, O. (2024). Towards explainable oral cancer recognition: Screening on imperfect images via Informed Deep Learning and Case-Based Reasoning. *Computerized Medical Imaging and Graphics*, *117*, 102433. https://doi.org/https://doi.org/10.1016/j.compmedimag.2024.102433

Patel, A. N., Murugan, R., Srivastava, G., Maddikunta, P. K. R., Yenduri, G., Gadekallu, T. R., & Chengoden, R. (2024). An explainable transfer learning framework for multi-classification of lung diseases in chest X-rays. *Alexandria Engineering Journal*, *98*, 328–343. https://doi.org/https://doi.org/10.1016/j.aej.2024.04.072

Pereira, P., Rocha, J., Pedrosa, J., & Mendonça, A. M. (2024). Evaluating Visual Explainability in Chest X-Ray Pathology Detection. *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, 1116–1121. https://doi.org/10.1109/MELECON56669.2024.10608569

Pisarcik, A., Hudec, R., Hlavata, R., & Kamencay, P. (2024). Comparison of Deep Learning Explainable Approaches for Medical Image Analysis. *2024 International Symposium ELMAR*, 331–334. https://doi.org/10.1109/ELMAR62909.2024.10694137

Rahim, N., Abuhmed, T., Mirjalili, S., El-Sappagh, S., & Muhammad, K. (2023). Time-series visual explainability for Alzheimer's disease progression detection for smart healthcare. *Alexandria Engineering Journal*, *82*, 484–502. https://doi.org/https://doi.org/10.1016/j.aej.2023.09.050

Rahim, N., El-Sappagh, S., Ali, S., Muhammad, K., Del Ser, J., & Abuhmed, T. (2023). Prediction of Alzheimer's progression based on multimodal Deep-Learning-based fusion and visual

Explainability of time-series data. *Information Fusion*, *92*, 363–388. https://doi.org/https://doi.org/10.1016/j.inffus.2022.11.028

Rahimiaghdam, S., & Alemdar, H. (2024). Evaluating the quality of visual explanations on chest X-ray images for thorax diseases classification. *NEURAL COMPUTING & APPLICATIONS*. https://doi.org/10.1007/s00521-024-09587-0

Rahman, R., Alam, Md. G. R., Reza, Md. T., Huq, A., Jeon, G., Uddin, Md. Z., & Hassan, M. M. (2023). Demystifying evidential Dempster Shafer-based CNN architecture for fetal plane detection from 2D ultrasound images leveraging fuzzy-contrast enhancement and explainable AI. *Ultrasonics*, *132*, 107017. https://doi.org/https://doi.org/10.1016/j.ultras.2023.107017

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S. Bin, Islam, M. T., Al Maadeed, S., Zughaier, S. M., Khan, M. S., & Chowdhury, M. E. H. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, *132*, 104319. https://doi.org/10.1016/J.COMPBIOMED.2021.104319

Saeed, T., Khan, M. A., Hamza, A., Shabaz, M., Khan, W. Z., Alhayan, F., Jamel, L., & Baili, J. (2024). Neuro-XAI: Explainable deep learning framework based on deeplabV3+ and bayesian optimization for segmentation and classification of brain tumor in MRI scans. *Journal of Neuroscience Methods*, *410*, 110247. https://doi.org/https://doi.org/10.1016/j.jneumeth.2024.110247

Schweizer, L., Seegerer, P., Kim, H.-Y., Saitenmacher, R., Muench, A., Barnick, L., Osterloh, A., Dittmayer, C., Jödicke, R., Pehl, D., Reinhardt, A., Ruprecht Klemens and Stenzel, W., Wefers, A. K., Harter Patrick N and Schüller, U., Heppner, F. L., Alber, M., Müller, K.-R., & Klauschen, F. (2023). Analysing cerebrospinal fluid with explainable deep learning: From diagnostics to insights. *Neuropathol. Appl. Neurobiol.*, *49*(1), e12866.

Shaheema, S. B., K., S. D., & Muppalaneni, N. B. (2024). Explainability based Panoptic brain tumor segmentation using a hybrid PA-NET with GCNN-ResNet50. *Biomedical Signal Processing and Control*, *94*, 106334. https://doi.org/https://doi.org/10.1016/j.bspc.2024.106334

Shojaei, S., Saniee Abadeh, M., & Momeni, Z. (2023). An evolutionary explainable deep learning approach for Alzheimer's MRI classification. *Expert Systems with Applications*, *220*, 119709. https://doi.org/https://doi.org/10.1016/j.eswa.2023.119709

Singh, S., Acton, S. T., Moosa, S., & Sheybani, N. D. (2023). An Adaptive and Interpretable Framework for Biomedical Image Analysis. *2023 57th Asilomar Conference on Signals, Systems, and Computers*, 1156–1160. https://doi.org/10.1109/IEEECONF59524.2023.10476986

Singhal, A., Agrawal, K. K., Quezada, A., Aguiñaga, A. R., Jiménez, S., & Yadav, S. P. (2024). Explainable Artificial Intelligence (XAI) Model for Cancer Image Classification. *CMES - Computer Modeling in Engineering and Sciences*, *141*(1), 401–441. https://doi.org/https://doi.org/10.32604/cmes.2024.051363

Song, D., Yao, J., Jiang, Y., Shi, S., Cui, C., Wang, L., Wang, L., Wu, H., Tian, H., Ye, X., Ou, D., Li, W., Feng, N., Pan, W., Song, M., Xu, J., Xu, D., Wu, L., & Dong, F. (2023). A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM. *Computer Methods and Programs in Biomedicine*, *235*, 107527. https://doi.org/https://doi.org/10.1016/j.cmpb.2023.107527

Srinivasu, P. N., Ahmed, S., Hassaballah, M., & Almusallam, N. (2024). An explainable Artificial Intelligence software system for predicting diabetes. *Heliyon*, *10*(16), e36112. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e36112

T R, M., Gupta, M., T A, A., Kumar V, V., Geman, O., & Kumar V, D. (2024). An XAI-enhanced efficientNetB0 framework for precision brain tumor detection in MRI imaging. *Journal of Neuroscience Methods*, *410*, 110227. https://doi.org/https://doi.org/10.1016/j.jneumeth.2024.110227

Talia, D. (2022). Algorithms That Can Deny Care, and a Call for AI Explainability. *The IEEE Computer Society*, *55*(10), 82–86. https://doi.org/10.1109/MC.2022.3190786

Tanone, R., Li, L.-H., & Saifullah, S. (2025). ViT-CB: Integrating hybrid Vision Transformer and CatBoost to enhanced brain tumor detection with SHAP. *Biomedical Signal Processing and Control*, *100*, 107027. https://doi.org/https://doi.org/10.1016/j.bspc.2024.107027

Thapar, P., & Tiwari, S. (2024). Empowering Skin Cancer Diagnosis: Integrating Advanced Deep Learning Models with Explainable AI for Lesion Classification. *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, *1*, 1–6. https://doi.org/10.1109/ICEECT61758.2024.10739236

Thiruvenkadam, K., Ravindran, V., & Thiyagarajan, A. (2024). Deep Learning with XAI based Multi-Modal MRI Brain Tumor Image Analysis using Image Fusion Techniques. *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, 1–5. https://doi.org/10.1109/TQCEBT59414.2024.10545215

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, *5*, 180161. https://doi.org/10.1038/SDATA.2018.161

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/S11604-023-01474-3,

Ukwuoma, C. C., Cai, D., Heyat, M. B. Bin, Bamisile, O., Adun, H., Al-Huda, Z., & Al-antari, M. A. (2023). Deep learning framework for rapid and accurate respiratory COVID-19 prediction using chest X-ray images. *Journal of King Saud University - Computer and Information Sciences*, *35*(7), 101596. https://doi.org/https://doi.org/10.1016/j.jksuci.2023.101596

Ullah, M. S., Khan, M. A., Albarakati, H. M., Damaševičius, R., & Alsenan, S. (2024). Multimodal brain tumor segmentation and classification from MRI scans based on optimized DeepLabV3+ and interpreted networks information fusion empowered with explainable AI. *Computers in Biology and Medicine*, *182*, 109183. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.109183

Vairetti, C., Maldonado, S., & Cuitino Loreto and Urzua, C. A. (2024). Interpretable multimodal classification for age-related macular degeneration diagnosis. *PLoS One*, *19*(11), e0311811.

Veetil, I. K., Chowdary, D. E., Chowdary, P. N., Sowmya, V., & Gopalakrishnan, E. A. (2024). An analysis of data leakage and generalizability in MRI based classification of Parkinson's Disease using explainable 2D Convolutional Neural Networks. *Digital Signal Processing*, *147*, 104407. https://doi.org/https://doi.org/10.1016/j.dsp.2024.104407

vom Brocke, J., Hevner, A., & Maedche, A. (2020). *Introduction to Design Science Research*. 1–13. https://doi.org/10.1007/978-3-030-46781-4_1

Wang, L., Huang, J., Xing, X., & Yang, G. (2023). Hybrid Swin Deformable Attention U-Net for Medical Image Segmentation. *2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 1–5. https://doi.org/10.1109/SIPAIM56729.2023.10373513

Wang, M., Lin, Z., Zhou, J., Xing, L., & Zeng, P. (2023). Applications of Explainable Artificial Intelligent Algorithms to Age-related Macular Degeneration Diagnosis: A Case Study Based on CNN, Attention, and CAM Mechanism. *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, *1*, 1–5. https://doi.org/10.1109/InC457730.2023.10263077

Wang, Z., Chen, Y., Wu, Y., Xue, Y., Lin, K., Zhang, J., & Xiao, Y. (2025). Enhancing Epstein–Barr virus detection in IBD patients with XAI and clinical data integration. *Computers in Biology and Medicine*, *184*, 109465. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.109465

Wester Trejo, M. A. C., Sadeghi, M., Singh, S., Mahmoodian, N., Sharifli, S., Hruskova, Z., Tesař, V., Puéchal, X., Bruijn, J. A., Göbel, G., Bajema, I. M., & Kronbichler, A. (2024). Explainability of a Deep Learning-Based Classification Model for Antineutrophil Cytoplasmic Autoantibody–Associated Glomerulonephritis. *Kidney International Reports*. https://doi.org/https://doi.org/10.1016/j.ekir.2024.11.005

Wickstrøm, K. K., Østmo, E. A., Radiya, K., Mikalsen, K. Ø., Kampffmeyer, M. C., & Jenssen, R. (2023). A clinically motivated self-supervised approach for content-based image retrieval of CT liver images. *Computerized Medical Imaging and Graphics*, *107*, 102239. https://doi.org/https://doi.org/10.1016/j.compmedimag.2023.102239

Zahoor, K., Bawany, N. Z., & Ghani, U. (2023). Explainable AI for Healthcare: An Approach Towards Interpretable Healthcare Models. *2023 24th International Arab Conference on Information Technology (ACIT)*, 1–7. https://doi.org/10.1109/ACIT58888.2023.10453740