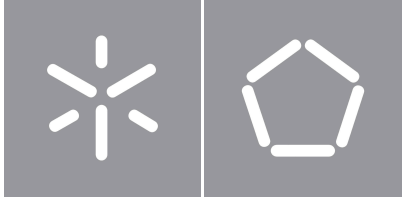


University of Minho
School of Engineering

Andreia Martins

Designing quality metrics for colonoscopy video using explainable AI techniques

december 2025



University of Minho
School of Engineering

Andreia Martins

Designing quality metrics for colonoscopy video using explainable AI techniques

Master's Dissertation in Biomedical Engineering
Medical Electronics

Dissertation supervised by
Professor Doutor Carlos Manuel Gregório Santos Lima
Professor Doutor Frederik Maes

december 2025

Copyright and Terms of Use for Third Party Work

This dissertation reports on academic work that can be used by third parties as long as the internationally accepted standards and good practices are respected concerning copyright and related rights.

This work can thereafter be used under the terms established in the license below.

Readers needing authorization conditions not provided for in the indicated licensing should contact the author through the RepositóriUM of the University of Minho.

License granted to users of this work:



CC BY-NC-ND

https://creativecommons.org/licenses/by-nc-nd/4.0/

"É preciso sair da ilha para ver a ilha. Não nos vemos se não saímos de nós."

– José Saramago

Acknowledgements

I would like to express my deepest gratitude to everyone who made this thesis possible.

First, I am profoundly grateful to my supervisors, Professor Dr. Carlos Lima from the University of Minho, for his constant approachability, and Professor Frederik Maes from KU Leuven, for providing me with such an interesting topic. I am truly thankful for their combined guidance, as well as for the opportunity to participate in the Erasmus program. I am especially grateful to my mentor, Belian Callaerts, whose careful feedback and encouragement turned difficult moments into achievable small steps.

I would also like to thank all of my colleagues, both in the laboratories in Leuven and in Braga, for welcoming me, encouraging me, and maintaining such a light-hearted spirit. I am also thankful to my Erasmus friends, especially those from hallway 0A - Blok3, the Leuven girls and all my friends who have shown me that no matter where I go, I can always find "uma casa portuguesa, com certeza".

To all of my friends, from the oldest to the newest, who have witnessed so many versions of me but have always supported me, I do not need to name each one to know that I wouldn't be here without any of you. A special mention to my girls from 3E, to my godfather and goddaughters, for all the belly laughs and for the full album of memories. Eduarda, you are persevering and an example of how to serve with a full heart.

Inês, for being my hug and my home, Hugo, for all the companionship and Bea, Rita and Raquel, thank you all. We have overcome every challenge together and I know this will always be the case throughout life. I cherish all our memories and the lessons we've learned with so much affection.

I want to express my unconditional love to my family. Everything I am, I owe to you. Thank you for being my role model of love, for being my place of comfort and for giving without measure.

To my best friend and partner in life, thank you for never stopping believing in me. You are my heart and my inspiration.

To all of you, thank you. What has carried me here is not strength, but the precious people around me on whom I can always rely.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, december 2025

Andreia Martins

Abstract

Deep Learning has emerged as a powerful tool in medical image analysis, attracting significant attention due to its potential to improve diagnostic accuracy. In the context of Colorectal Cancer - the third most commonly diagnosed cancer worldwide - early detection via colonoscopy substantially increases the odds of survival. While several approaches with Artificial Intelligence have developed an automated support system for polyp detection and characterization, they are often challenged by variations in colonoscopy image quality.

With this in mind, the primary focus of this work is to leverage explainable Artificial Intelligence to understand model decision-making and enhance model reliability, especially for clinical applications. By comparing explainability maps generated by Artificial Intelligence models with ground truth segmentation of polyps, the study assesses whether the model's focus aligns with clinically relevant regions. The underlying hypothesis is that accurate predictions are more likely when the model's attention is directed towards the polyp rather than surrounding noise.

This dissertation has four core objectives: (1) to explore different explainable Artificial Inteligence techniques, (2) to calculate metrics that evaluate the spatial overlap between explainability maps and segmentation masks, thereby identifying high- and low-quality images automatically; (3) to correlate the new metrics with the model classification performance, and (4) to investigate the potential use of the new metrics to train a model for estimating the adequacy of images for further analysis. Ultimately, the results of this research aim to enhance the reliability and transparency of Artificial Intelligence models in Colorectal Cancer detection among healthcare professionals and improve clinical outcomes.

Keywords Artificial Intelligence, Explainability, Colonoscopy videos, Colorectal polyp classification

Resumo

O Deep Learning emergiu como uma ferramenta poderosa na análise de imagens médicas, atraindo atenção significativa devido ao seu potencial para melhorar a precisão diagnóstica. No contexto do Cancro Colorretal – o terceiro tipo de cancro mais diagnosticado mundialmente – a deteção precoce através de colonoscopia aumenta substancialmente as probabilidades de sobrevivência. Embora várias abordagens baseadas em Inteligência Artificial tenham desenvolvido sistemas automatizados para a deteção e caracterização de pólipos, estas enfrentam frequentemente desafios devido às variações na qualidade das imagens de colonoscopia.

Com isto em mente, o foco principal deste trabalho é utilizar Inteligência Artificial Explicável para compreender o processo de tomada de decisão dos modelos e aumentar a sua fiabilidade, especialmente em aplicações clínicas. Ao comparar os mapas de explicabilidade com as segmentações de referência dos pólipos, pretende-se avaliar se a atenção do modelo se está a focar nas regiões clinicamente relevantes. A hipótese subjacente é que as previsões precisas são mais prováveis quando a atenção do modelo está direcionada para o pólipo em vez do ruído ao redor.

Esta dissertação tem quatro objetivos principais: (1) explorar diferentes técnicas de inteligência artificial explicável, (2) calcular métricas que avaliem a sobreposição espacial entre os mapas de explicabilidade e as máscaras de segmentação, identificando automaticamente imagens de alta e baixa qualidade, (3) correlacionar as novas métricas com o desempenho de classificação do modelo, e (4) investigar o potencial uso dessas novas métricas para treinar um modelo de regressão ou de classificação para estimar a qualidade das imagens. Em última análise, os resultados desta pesquisa visam melhorar a fiabilidade e a transparência dos modelos de inteligência artificial na deteção de cancro colorretal entre os profissionais de saúde e aprimorar os resultados clínicos.

Palavras-chave Classificação de pólipos colorretais, Explicabilidade, Inteligência Artificial, Vídeos de colonoscopia,

Contents

1 Introduction **1**

1.1 Scope . . . 1

1.2 Context . . . 1

1.3 Motivation and Aims . . . 3

1.4 Dissertation Overview . . . 4

2 Clinical Background **6**

2.1 Colon Anatomy . . . 6

2.2 Colonoscopy . . . 7

2.3 Colorectal Polyps and Tumor Development . . . 8

2.4 Polyp Classification Models . . . 9

3 State of the Art **10**

3.1 Fundamentals of Artificial Intelligence . . . 10

3.2 Classification Models . . . 11

3.2.1 Resnet50 . . . 12

3.2.2 EfficientNet . . . 13

3.3 Explainable AI . . . 16

3.3.1 Model-specific . . . 19

3.3.1.1 Saliency Maps . . . 19

3.3.1.2 CAM . . . 20

3.3.1.3 Grad-CAM . . . 22

3.3.2 Model-agnostic . . . 23

3.3.2.1 LIME . . . 23

3.3.2.2 SHAP . . . 26

3.4 Related Work on XAI in Medical Imaging . . . 29

3.5 Summary . . . 31

4 Methodology **33**

4.1 Approach Pipeline . . . 33
4.2 Preliminary Experiment . . . 35
4.3 Main Experiment . . . 36
4.3.1 Dataset Description . . . 37
4.3.2 Preprocessing, Augmentation and Crops . . . 39
4.3.3 Training Setup and Hyperparameters . . . 40
4.4 Grad-CAM . . . 41
4.5 SHAP . . . 43
4.6 Artificial Degradations . . . 45
4.7 Evaluation Metrics . . . 46
4.7.1 Accuracy . . . 48
4.7.2 Precision . . . 48
4.7.3 Recall . . . 48
4.7.4 F1 Score . . . 48
4.8 Quality Metrics . . . 48
4.8.1 Intersection over Union . . . 48
4.8.2 Score (Effective Probability) . . . 49
4.8.3 Skewness of the intensity distribution . . . 50
4.8.4 Laplacian variance (image sharpness) . . . 50
4.8.5 Centroid distance (spatial alignment) . . . 51
4.8.6 Combined Metrics . . . 51
4.9 Software and Hardware Setup . . . 52

5 Experiments and Results **54**

5.1 Architecture Choice and Training Configuration . . . 54
5.2 Grad-Cam Implementation . . . 56
5.3 SHapley Additive exPlanations (SHAP) implementation . . . 57
5.4 Preliminary Experiment Implementation . . . 58
5.5 Main Experiment Implementation . . . 64
5.6 Summary . . . 75

5.6.1 Classification model 75
5.6.2 Preliminary experiment (pipeline validation) 75
5.6.3 Polyp Experiment 75
5.6.4 XAI comparison 76
5.6.5 Robustness to image modifications and sequence effects 76
5.6.6 Composed Metric 76

6 Discussion **78**

6.1 Experiments 78
6.2 Challenges 80
6.3 Prospect for future work 81

7 Conclusions and future work **83**

7.1 Conclusions 83
7.2 Future Work 84

List of Figures

1 Colonoscopy image in white light 2

2 Structure of the human colon wall showing the main histological layers. 7

3 Different stages of colonoscopy examination. 7

4 Visual summary of the three main colorectal types of polyps. 9

5 Hierarchical relationship between AI, ML and DL. 11

6 Typical architecture of a CNN 11

7 Residual block with a skip (shortcut) connection. 12

8 Overview of the ResNet50 model architecture, showing the division into five blocks. . . . 13

9 Model size vs. ImageNet accuracy. 14

10 EfficientNet architecture using MBConv blocks. 15

11 Diagram illustrating the different purposes of Explainable Artificial Intelligence (XAI) sought by different audience profiles 16

12 Examples of XAI methods applied to different types of data. 17

13 Taxonomy of the reviewed literature for explainability techniques 18

14 Saliency maps generated using Grad-CAM, Grad-CAM++ and Smooth-Grad-CAM++, respectively 20

15 Class Activation Mapping 21

16 Overview of Grad-CAM and Guided Grad-CAM. 23

17 Steps of the Local Interpretable Model agnostic Explanations (LIME) method. 25

18 Shapley value computation process: decomposition of a model's prediction $f(x)$ into contributions from baseline ϕ_0 and individual feature contributions ϕ_1, ϕ_2, ϕ_3 28

19 SHAP visualization of image classification, highlighting regions that contribute positively (red) or negatively (blue) to the predicted class 29

20 Overall methodological workflow 33

21 Saliency thresholding and overlap strategy 35
22 Example of a cropped 256x256 region of interest from the dataset. 39
23 Development environment setup. 52
24 Example of the Grad-CAM overlay for a polyp-containing frame. The left panel of each pair of images shows the heatmap blended with the original image; the right panel shows the ground-truth segmentation mask. 57
25 Example SHAP attribution overlays on polyp images. 58
26 Dataset examples of cats and dogs. 59
27 Example of Grad-CAM for cats and dogs. 59
28 Example images of the Beagle breed: IoU vs Score = -0.492 and IoU vs distance to center = -0.151. 61
29 Example images of the Russian Blue breed: IoU vs Score = 0.452 and IoU vs distance to center = -0.677. 61
30 Original image (top-left) and 11 artificially degraded variants, including directional shifts, gamma corrections, noise, hue shifts, value shift and saturation shift applied at different intensities. 64
31 Example of a polyp annotation and its corresponding cropped image. 65
32 Original image (top-left) and seven artificially degraded variants, which include obstruction, two levels of Bright Gaussian noise, Bright Global noise, and three degrees of blur. . 67
33 Application of Grad-CAM to some of the degradations and to the original image (on the right). 67
34 Application of Grad-CAM to a frame from the sequence p046_19_R_WLI. The top-left image shows the original, with blurred variants on the top-right (blur factor 5), bottom-left (blur factor 10), and bottom-right (blur factor 15). 69
35 Application of Grad-CAM to a frame from the sequence p295_02_R_BLI. The top-left image shows the original, with blurred variants on the top-right (blur factor 5), bottom-left (blur factor 10), and bottom-right (blur factor 15). 69
36 Application of Grad-CAM to a frame from the sequence p526_03. The top-left image shows the original, with blurred variants on the top-right (blur factor 5), bottom-left (blur factor 10), and bottom-right (blur factor 15). 70
37 Application of Grad-CAM to a frame from the sequence p295_02_R_BLI. Left: original image; right: obstructed version. 74

List of Tables

1 Types of Colorectal Polyps and their Characteristics 8

2 Comparison of selected XAI techniques. 32

3 Distribution of samples in the *Cats vs Dogs* preliminary experiment. 36

4 Distribution of the main dataset. 38

5 Summary of preprocessing steps and data augmentation techniques. 39

6 Summary of the planned training setup and optimization parameters. 42

7 Summary of the implemented artificial degradations and their simulation goals — Part A. 46

8 Summary of the implemented artificial degradations and their simulation goals — Part B. 47

9 Hardware and software summary for the experimental environment. 53

10 Summary of training configuration and hyperparameter choices. 55

11 Performance comparison of the tested models in terms of accuracy, precision, recall, and F1-score. 56

12 Classification Report. 59

13 Correlation Matrix 60

14 Per-breed Pearson correlations: IoU vs Score and IoU vs Distance to Center 62

15 Abyssinian_101 — Predicted score and IoU across artificial degradation variants 63

16 Correlation matrix of SHAP implementation 65

17 Correlation matrix of Grad-Cam implementation 66

18 Correlation matrix of IoU–score correlation for each sequence and degradation level . . 68

19 Pearson Correlation by Sequence of Composed Metric vs Score 72

20 Score vs. Composed Metric — distortion images 73

21 Score vs. Composed Metric — only for originals 73

22 Correlation of the composed metric versus the true score, by sequence and modification. 74

Acronyms

ACS American Cancer Society

Adam Adaptive Moment Estimation

AI Artificial Intelligence

CAD Computer Aided Detection

CAM Class Activation Mapping

CDF Cumulative Distribution Function

CNN Convolution Neural Network

CNNs Convolution Neural Networks

CRC Colorectal Cancer

DL Deep Learning

ECIR European Cancer Inequalities Registry

GAP Global Average Pooling

GPU Graphics Processing Unit

Grad-CAM Gradient-weighted Class Activation Mapping

GT Ground-Truth

IAI Interpretable Artificial Intelligence

IoU Intersection over Union

LIME Local Interpretable Model agnostic Explanations

ML Machine Learning

SHAP SHapley Additive exPlanations

SSLs Sessile Serrated Lesions

SSPs Sessile Serrated Polyps

TSAs Traditional Serrated Adenomas

WHO World Health Organization

XAI Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Scope

This dissertation was developed within the scope of the Master's degree in Biomedical Engineering at the University of Minho. The research was carried out in collaboration with KU Leuven, where the author completed an Erasmus mobility period during the first semester of the academic year. The work was jointly supervised by both institutions, providing a multidisciplinary framework that combined expertise in biomedical engineering, medical imaging, and artificial intelligence.

1.2 Context

Colorectal Cancer (CRC) is a type of cancer that begins in the cells of the colon, part of the large intestine, or rectum [1]. It is a serious global health concern as it can cause severe harm and death [1].

This multifactorial disease ranks as the third most frequently diagnosed cancer and the second leading cause of cancer-related deaths worldwide [2]. In the late 1990s, CRC was the fourth leading cause of cancer death in both men and women. However, by 2024, according to the American Cancer Society (ACS), it is the leading cause of cancer death in men and the second leading cause in women with an estimated 2 001 140 new cancer cases in the United States [3].

The incidence of CRC correlates with age, with most cases affecting people over 50 years old [1, 2]. Geographically, CRC incidence rates were highest in Europe, Australia, and New Zealand, while mortality rates were highest in eastern Europe. The projections of the World Health Organization (WHO) estimate that by 2040 the burden of CRC will increase to 3.2 million new cases per year (an increase of 63%) and 1.6 million deaths per year (an increase of 73%) [1].

If we take a closer look at Portugal, according to the European Commission via the European Cancer Inequalities Registry (ECIR), in 2020, there were an estimated 50 deaths per 100,000 men due to

CRC, which is 1.8 times higher than the rate observed in women [4]. WHO - Global Cancer Observatory (GLOBOCAN 2020) reported that CRC in Portugal was the leading cancer type when considering both sexes combined, with an estimated incidence of approximately 10 575 new cases (15.2% of all cancers) [5].

One of the main challenges in combating CRC is that it is often asymptomatic in its initial stages, leading to delayed diagnosis when treatment options are limited [1]. Consequently, timely diagnosis, appropriate treatment, and regular follow-up care are imperative to improve survival rates and quality of life, as early-stage cancers have higher survival rates than advanced-stage cancers [1].

Medical imaging plays a central role in the early diagnosis of cancer [6]. Among the multiple techniques that can be used to visualize the colon, colonoscopy remains the worldwide gold standard for CRC screening [7, 8]. Figure 1 illustrates a colonoscopy image captured under white light, highlighting abnormal cell growth located in the lower central region of the image.

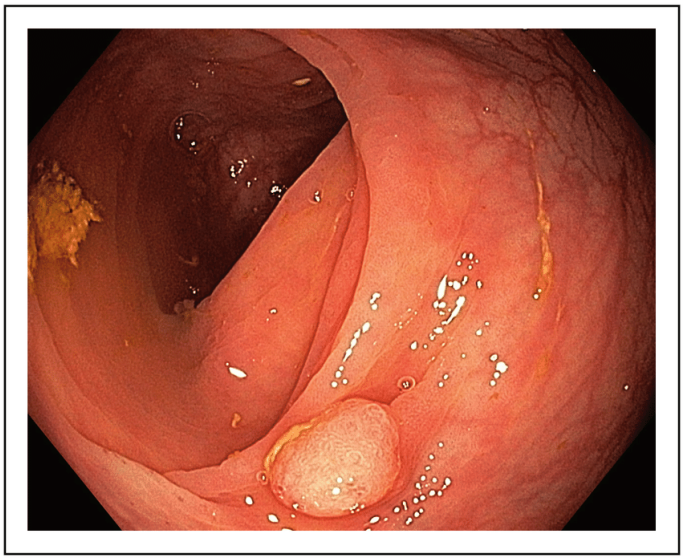


Figure 1: Colonoscopy image in white light. Taken from [9].

However, colonoscopy has some limitations, such as being an invasive procedure, the risk of missed lesions, high costs, and it causes discomfort and sometimes pain for the patient [8]. Moreover, biopsy and histopathology analysis is a time-consuming process performed by biologists, which in ambiguous cases is subject to significant inter-rater variability and, therefore, can lead to misdiagnosis if not performed by adequately trained medical personnel [7]. Furthermore, the interpretation of the data in relation to the speed of the disease progression is time-consuming and difficult [10].

Recent advancements in medical imaging and Artificial Intelligence (AI) have opened up a new era of possibilities in healthcare, significantly enhancing diagnostic capabilities and treatment outcomes [11]. The quality of the colonoscopy varies according to the endoscopist's expertise, and standardizing it by re-

ducing the number of missed polyps is a desirable goal of AI [12]. Through this process, AI can help healthcare professionals detect abnormalities, identify specific structures, and predict disease outcomes [11]. In this context, AI seems to be the ideal tool to improve quality in virtually every subdomain of endoscopy, including the classification of polyps, which is critical in CRC screening [13]. There are a lot of developments surrounding the detection and characterization of polyps during the colonoscopy procedure using AI.

While promising, many AI applications are based on black-box models that make complex decisions without easily understandable reasoning [7]. One of the main concerns about their implementation is the potential negative impact on human–AI interaction [14]. Trust is strongly influenced by how much humans understand the behavior of a model, especially when the model is used for decision prediction [15]. There is an emerging need for understanding how such decisions are furnished by AI methods when decisions derived from such systems ultimately affect humans' lives (in medicine, law, or defense) [16, 17]. This has driven the emergence of the field of XAI, which aims to provide understandable and transparent explanations for AI decisions and actions [7].

1.3 Motivation and Aims

While several proposals in the literature exploit Deep Learning (DL) for detecting cancer from bioimages, these methods often remain underutilized in the real-world clinical practice due to the lack of explainability [7]. Incorporating explainability techniques into AI models can significantly enhance decision-making processes by providing users with valuable insights into the data, allowing them to understand where and why a model made a decision [18]. This transparency is essential to build trust, especially among healthcare professionals and to improve the overall efficacy of diagnostic systems [19]. It is not uncommon for the model to focus on inappropriate information, such as artifacts in the background of the image, even when producing accurate predictions [18].

An approach to address this issue is to compare the regions highlighted by the model's explainability map with the ground truth segmentation masks of polyps (manually labelled). To quantify this alignment, overlap-based metrics such as the Dice coefficient and Intersection over Union (IoU) can be used.

Beyond explainability, challenges such as limited data availability and heterogeneous image quality also hinder robust classification. Motion blur, occlusions - partial visibility of polyps -, poor illumination, and suboptimal camera positioning can significantly degrade image quality, thereby reducing model performance. Importantly, the notion of image quality often differs between human perception and computer

vision systems [20]. Hence, developing robust metrics to automatically evaluate image quality is essential to determine which frames are suitable for reliable analysis.

Given these challenges, there is a strong need for research that integrates XAI methodologies into classification models. This dissertation addresses this gap by combining XAI techniques with polyp classification models and designing new quality metrics for colonoscopy videos. Specifically, it investigates the relationship between explanation maps and ground-truth segmentation masks (each pixel labeled polyp or background), and derives a score that quantifies how well the model's attention aligns with clinically relevant regions. Frames with low alignment can then be flagged as poor quality, allowing their careful treatment and analysis, thereby enhancing model performance and ultimately contributing to more effective CRC screening and prevention.

The specific steps to reach the goal of this dissertation are as follows:

1. Acquire and synthesize all necessary and relevant knowledge of the topic;
2. Explore and compare different explainable AI techniques and models for polyp classification;
3. Calculate metrics that evaluate the spatial overlap between explainability maps and segmentation masks;
4. Develop a new metric and conclude the correlation between the new metric and the model's classification performance;
5. Investigate the potential use of the new metrics to train a model for estimating the adequacy of images for further analysis.

Ultimately, the results of this research aim to enhance the reliability and transparency of AI models in CRC detection among healthcare professionals and improve clinical outcomes.

1.4 Dissertation Overview

This dissertation is structured into seven chapters: Introduction, Clinical Background, State of the Art, Methodology, Experiments and Results, Discussion, and Conclusions and Future Work.

Chapter 1 encompasses the context of this dissertation, states the research problem and objectives, and summarizes the study's main contributions. Chapter 2 provides the necessary medical background, including the anatomy of the colon, the colonoscopy procedure, the development of colorectal polyps and tumors, and the existing classification models for polyp detection.

The subsequent chapter reviews the technical background in AI, covering the classification architectures considered and XAI methods pertinent to this study, together with recent XAI applications in medical imaging. Chapter 4 details the methodology, detailing the approach pipeline, dataset description, preprocessing and augmentation strategies, model implementation, and evaluation metrics. Chapter 5 reports the experiments and results, comprising preliminary and main studies. Chapter 6 provides a comprehensive discussion of the findings, addressing the challenges encountered during the study and the implications of the results.

Finally, Chapter 7 concludes the dissertation by summarizing the key outcomes, reflecting on the limitations of the study, and proposing directions for future research.

Chapter 2

Clinical Background

This chapter provides the fundamental clinical background to support the technical developments described in the following chapters.

2.1 Colon Anatomy

Understanding the normal structure and function of the colon and rectum is essential to contextualize the development of CRC. Together, the colon and rectum form the large intestine, the final part of the gastrointestinal tract. The right colon absorbs water and nutrients, while the left colon mainly stores and expels feces. The colon also secretes hormones and mucus to protect the mucosa and aid transit. The rectum, connected to the sigmoid colon and ending at the anal canal, is innervated by the autonomic nervous system and controls defecation, with limited absorption of water, electrolytes, glucose, and some drugs [2].

The colonic wall consists of four layers: mucosa, submucosa, muscularis propria, and serosa. Figure 2 illustrates the histological structure of these layers. Most polyps and colorectal cancers begin in the mucosa, where abnormal cell growth arises. Malignant cells can then invade deeper layers and, once reaching lymphatic or blood vessels, spread to lymph nodes and distant organs [21].

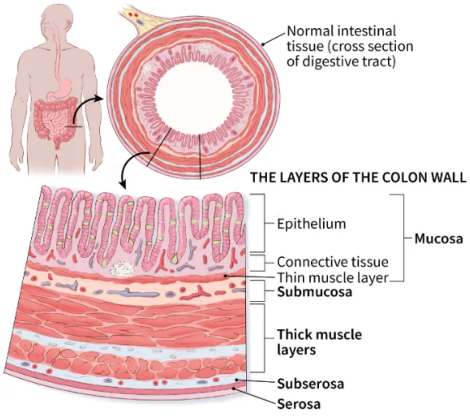


Figure 2: Structure of the human colon wall showing the main histological layers. Taken from [21].

2.2 Colonoscopy

During this visual examination, a long and flexible tube equipped with a high definition camera and light at the end (the colonoscope) is inserted through the rectum to examine the entire colon [7, 22]. Figure 3 illustrates the different stages of this procedure. At the tip of this tube, there is also a channel for the water and air supply and an instrument channel that allows instant therapeutic actions [8]. The appropriate course of action depends on their size and type: small and benign polyps may be left in situ, while larger or more concerning lesions may be removed endoscopically or referred for surgical evaluation for further examination under a microscope [7, 8, 22]. Thus, if abnormal tissue is found during colonoscopy or other imaging tests, a biopsy may be performed to analyse the cells under a microscope and confirm the presence of cancer [7].

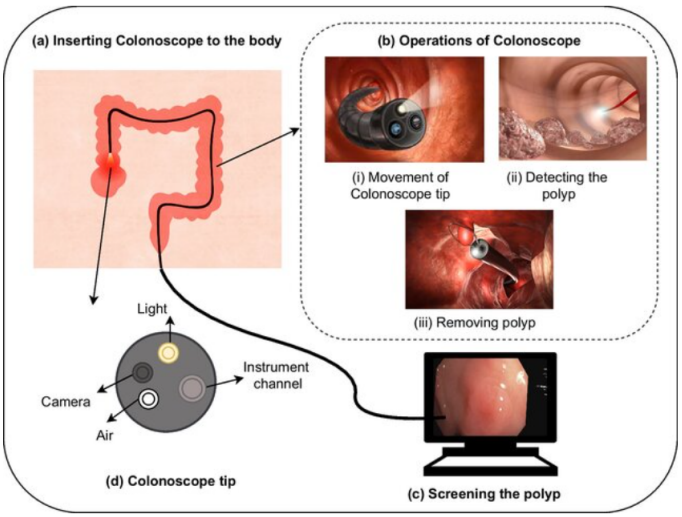


Figure 3: Different stages of colonoscopy examination. Taken from [23].

2.3 Colorectal Polyps and Tumor Development

CRC typically begins with the formation of polyps, which are abnormal growths of the inner lining of the colon or rectum that can turn into cancerous tumours over time if left untreated [7, 10, 24]. These polyps are protrusions or bumps occurring in the colon lumen, most commonly sporadic or as part of other syndromes and are mostly due to the excessive number of unwanted cells produced by the body in the lining of the bowel [25]. The risk of polyp formation is influenced by genetic predisposition as well as lifestyle factors such as age, diet, obesity, and smoking [26].

Not all colorectal polyps are alike, as they come in various shapes and sizes. Some of them are dangerous, while others are harmless bumps and can thus be respectively classified as malignant or benign [1, 27]. Polyps can be distinguished into three primary histological types: hyperplastic, adenomatous or adenomas (tubular adenomas and villous adenomas), and Sessile Serrated Polyps (SSPs) [24, 25]. Figure 4 provides a visual summary of these main categories, while Table 1 details additional subtypes and their main characteristics. Of these, hyperplastic polyps are the most common, typically small and benign, with a low risk of malignancy [25].

Table 1: Types of Colorectal Polyps and their Characteristics

Type	**Average Size**	**Cancerous Potential**	**Frequency in Patients**
Tubular adenomas	1–12 mm	Low	50%
Villous adenomas	$\geq$ 20 mm	High	$<$5%
Tubulovillous adenomas	$<$ 1 cm	High	8–16%
Sessile Serrated Lesions (SSLs)	Variable	Premalignant	15–20%
Traditional Serrated Adenomas (TSAs)	Variable	Premalignant	$<$1%
Hyperplastic polyps	$\leq$ 5 mm	Low (context-dependent)	20–40%
Inflammatory polyps	$\leq$ 2 cm	Low	10–20%

Source: Adapted from [12, 28].

Macroscopically, polyps may appear pedunculated (stalk-like) or sessile (flat). Sessile lesions are now recognized to be more common than previously thought and are particularly challenging to detect during screening endoscopy.



Figure 4: Visual summary of the three main colorectal types of polyps. Taken from [29].

2.4 Polyp Classification Models

Early detection of potentially malignant polyps is crucial. However, recognition accuracy can be challenged by factors like endoscopist fatigue, distraction, and subtle visual features of some lesions [14]. This aspect also highlights the importance of Computer Aided Detection (CAD) systems for improving detection accuracy [14]. To address polyp classification, most workflows involve inputting endoscopic images into a classification model. The model extracts features from the images and then uses a softmax function to assign probability scores to each class, with the highest score indicating the predicted class [12].

Typically, Convolution Neural Networks (CNNs) are used to distinguish between hyperplastic polyps, which are usually benign, and adenomatous polyps (tubulovillous adenomas, villous adenomas, and flat adenomas), which carry a higher risk of developing into cancer, allowing for a bimodal classification approach that aids in timely intervention.

Chapter 3

State of the Art

The state of the art contributes to the understanding of the theoretical and scientific concepts addressed throughout this dissertation. This chapter begins with an exploration of various XAI methods, including their mathematical foundations and applications in the medical domain. Furthermore, it discusses metrics for evaluating the alignment between explainability maps and ground truth maps.

3.1 Fundamentals of Artificial Intelligence

AI provides machines with the capability to perform tasks by simulating human behavior and learning, using algorithms governed by patterns recognition and self-correction on large amounts of data to narrow options to minimize errors [30].

A significant subset of AI, Machine Learning (ML), differs from traditional programming (with specific rules) by enabling systems to identify patterns and make decisions based on the data they receive [31]. Within ML, DL stands out, characterized by its ability to allow the model to automatically learn complex features from the data without requiring manual feature engineering, which is powerful in the detection and diagnosis of cancers [10]. The hierarchical relationship between AI, ML and DL is illustrated in Figure 5, highlighting how DL methods build upon ML principles to learn increasingly abstract features from data. In DL, a large number of neuron layers (neural network) automatically learn hierarchical features characterizing the input data through the application of non-linear operations together with backpropagation for training [32].

DL object recognition methods such as CNNs detect and classify polyps automatically during examinations to assist endoscopists [13]. CNNs are widely used DL architectures for solving image processing problems, offering state-of-the-art performance in computer vision tasks due to their adaptability to visual data [10, 34, 35]. Unlike traditional methods, there is no need to manually select the parameters of the dataset, CNNs automatically learn patterns and extract features [36]. Input images are represented as 3D

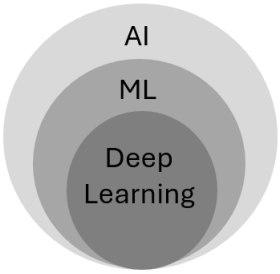


Figure 5: Hierarchical relationship between AI, ML and DL. Adapted from [33]

tensors (height × width × channels). In CNNs architectures, layers are connected in various blocks instead of being directly connected [35]. The information transfer between these blocks is similar to the visual cortex and eliminates the problems of classical techniques. A CNN architecture generally consists of convolution, nonlinear, pooling, dropout, and fully connected layers, and an example of its typical architecture is shown in Figure 6 [10].

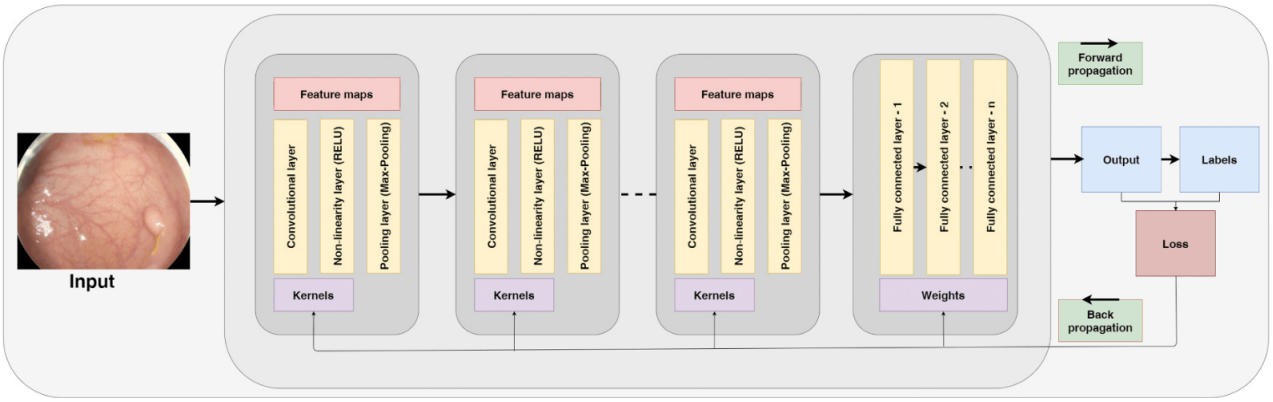


Figure 6: Typical architecture of a CNN. Taken from [10].

The convolutional layers of the CNN are responsible for the feature extraction phase by detecting local patterns using their inherent objectivity to uncover patterns that could be missed by human-designed features [36].

3.2 Classification Models

DL methods, particularly CNNs, assist endoscopists by identifying and categorizing polyps based on features such as shape, structure, and texture during colonoscopy examinations [13]. However, it still remains challenging due to various factors such as subtle appearance, differences in viewing angles, or lighting conditions [8, 13].

An algorithm that addresses polyp detection in colonoscopy videos was developed by Eelbode et al. [8].

It detects the polyps and generates a mask containing the outline of the polyp at the precise location [8]. However, their ability to automatically classify polyps into histological categories remains a more complex challenge and is still an active area of research.

3.2.1 Resnet50

One of the network architectures tried in this project is based on ResNet50 (Residual Network with 50 layers), a deep CNN, implemented using the Keras framework and initialized with weights pre-trained on the ImageNet dataset [37]. ResNet50 was introduced in 2015 by He et al., researchers at Microsoft Research [37]. It is a widely adopted DL network in computer vision tasks that can be trained on large datasets and achieve state-of-the-art results [37, 38].

The key innovation line of ResNet is the use of residual blocks/connections between layers that address performance degradation, such as accuracy saturation and vanishing gradients seen in very deep networks [37, 39, 40]. This allows layers to learn residual mappings instead of forcing them to directly learn the desired mapping [37].

Specifically, rather than learning the features directly, ResNet learns the residuals — the difference between the desired mapping and the input [37]. To achieve this, ResNet employs shortcut connections which skip one or more intermediate layers, directly connecting the input of one layer to the output of the subsequent layer [37, 40]. This mechanism simplifies optimization and accelerates convergence during training [40].

A block diagram of the ResNet model's architecture is shown in Figure 7. It connects activations of a layer via skip connections to further layers, bypassing some intermediate layers [39].

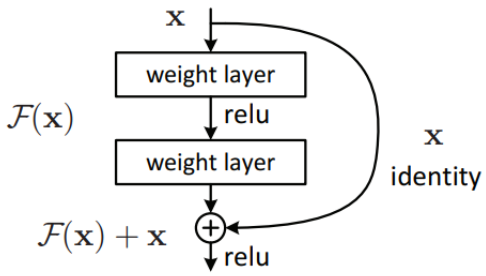


Figure 7: Residual block with a skip (shortcut) connection. Taken from [39].

Mathematically, the model learns a residual function $F(x)$, where the desired mapping $H(x)$ is expressed as:

$$F(x) := H(x) - x \quad \text{which gives} \quad H(x) := F(x) + x.$$

The overall ResNet50 architecture consists of four main parts: the convolutional layers, the identity block, the convolutional block, and the fully connected layers [38]. In Figure 8, there is a diagram illustrating the overall structure of ResNet50.

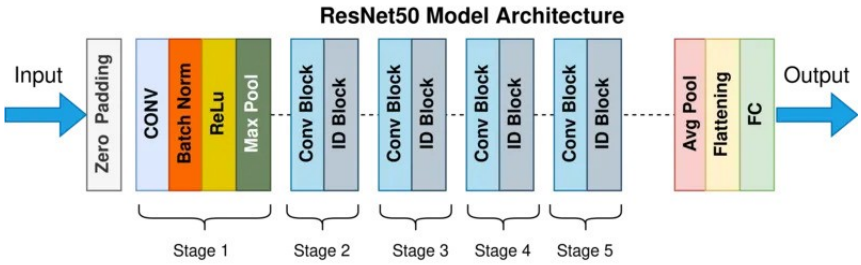


Figure 8: Overview of the ResNet50 model architecture, showing the division into five blocks. Taken from [38].

The convolutional layers in ResNet50 are several convolutional operations followed by batch normalization and ReLU activation, which help to stabilize training and introduce non-linear transformations [38]. They are responsible for extracting features such as edges, textures, and shapes, and consist of several convolution operations exploiting the local connections and shared weights [32, 38].

The identity and convolutional block are the key building blocks that process and transform these features [38]. Convolutional blocks include 1x1 convolutions to adjust feature dimensions when necessary for compatibility between input and output feature maps [38]. Lastly, the fully connected layers are connected to this vector for the final classification [32, 38]. The output of the final fully connected layer has a softmax activation function to produce the final class probabilities [38].

3.2.2 EfficientNet

EfficientNet is one of the network architectures used in this project. First proposed in 2019 by Mingxing Tan and Quoc V. Le, EfficientNet is a family of Convolution Neural Network (CNN) architectures described in the paper *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* [41]. Empirically, EfficientNet provides a superior accuracy–to–parameter trade-off compared to many earlier CNN designs, delivering higher ImageNet Top-1 accuracy for a given model size (Figure 9).

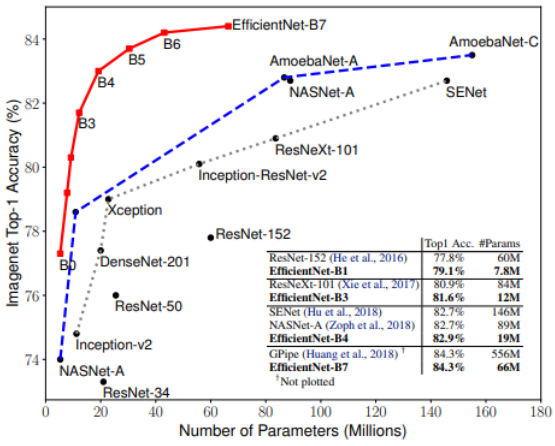


Figure 9: Model size versus ImageNet accuracy. Taken from [41].

The main idea behind EfficientNet is that increasing a network's depth, width, or input resolution on its own gives only small improvements in accuracy but greatly increases computational cost. This trade-off is shown in Figure 9, which compares model size and ImageNet accuracy. To address this inefficiency, the authors propose compound scaling, a method that scales depth (d), width (w), and input resolution (r) together. The compound scaling rules are defined as:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi},$$

where ϕ is a single scalar that controls the model size: increasing ϕ makes the network deeper, wider, and higher-resolution at the same time. The constants (α, β, γ) set how strongly each dimension grows and were chosen by the authors to keep scaling balanced.

From an architectural point of view, EfficientNet uses a small set of efficient building blocks: MobileNetV2-style inverted residual bottlenecks (MBConv), which typically include Squeeze-and-Excitation modules, together with standard normalization and activation layers that stabilize training. Figure 10 shows the baseline block layout. These components enable effective feature learning while keeping parameter counts and per-pixel computation relatively low.

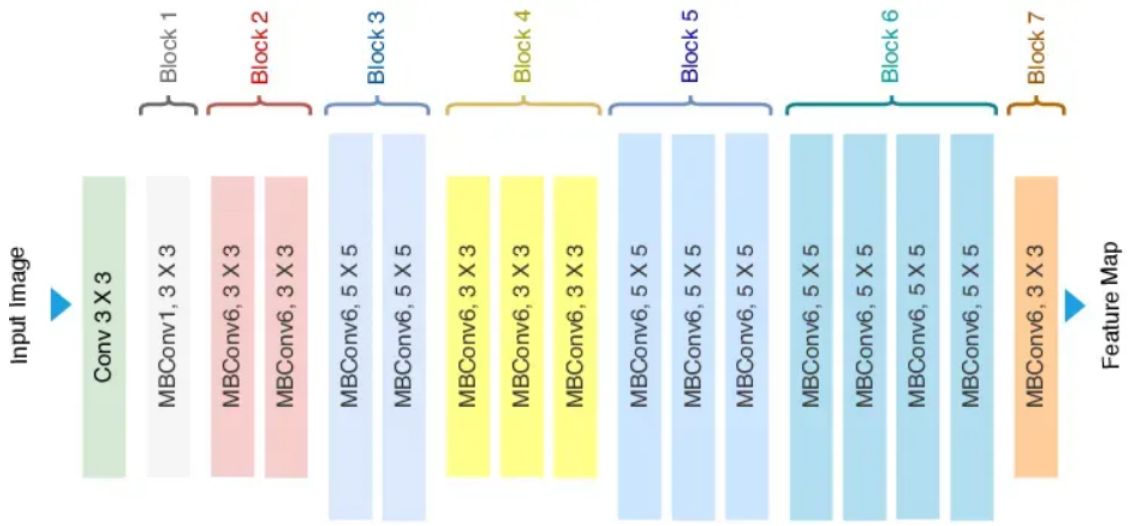


Figure 10: EfficientNet architecture using MBConv blocks. Taken from [42].

Formally, a convolutional neural network can be expressed as a composition of layer-wise operators:

$$\mathcal{N} = \bigodot_{i=1}^{s} \mathcal{F}_i^{L_i}(X_{(H_i, W_i, C_i)}),$$

where $X_{(H_i,W_i,C_i)}$ denotes the input tensor to layer i with spatial dimensions (H_i, W_i) and C_i channels, $\mathcal{F}_i$ is the operation repeated L_i times, and $\bigodot$ denotes ordered composition. Using this notation, the search for an architecture that maximises accuracy subject to resource constraints can be formulated as

$$\begin{aligned} \max_{d,w,r} \quad & \text{Accuracy}(\mathcal{N}(d, w, r)) \\ \text{s.t.} \quad & \mathcal{N}(d, w, r) = \bigodot_{i=1}^{s} \hat{\mathcal{F}}_i^{d \cdot \hat{L}_i}(X_{(r \cdot \hat{H}_i, r \cdot \hat{W}_i, w \cdot \hat{C}_i)}), \\ & \text{Memory}(\mathcal{N}) \leq \text{target_memory}, \\ & \text{FLOPS}(\mathcal{N}) \leq \text{target_flops}. \end{aligned}$$

In this formulation, the scalar factors d, w, and r scale the baseline architecture parameters (denoted with hats). The compound-scaling heuristic delivers a practical, empirically validated approach to choosing these factors so as to maximise accuracy under a specified computational budget.

The compound-scaling principle introduced by EfficientNet has been widely adopted in subsequent work on detection, segmentation, and medical-imaging applications where computational efficiency is critical. Nonetheless, model selection should always be guided by dataset characteristics, interpretability needs, and deployment constraints (latency, memory and available hardware) [41].

3.3 Explainable AI

XAI is an emerging field that has become a popular research subject within the AI field in recent years [34, 43, 44]. This reflects the strong need to have an interface between humans and a decision predictor - the model [45]. The central aim of XAI research is to make the "black-box" nature of CNNs more comprehensible and transparent to humans while maintaining a high level of performance. It seeks to minimize the trade-off between the performance of a model and its transparency [34, 43], especially for applications that take actions based on a prediction in high-risk sectors such as military, banking, and healthcare applications [15, 34].

In fact, understanding the reasons behind predictions is as crucial as the prediction's accuracy [46]. This demand for understanding the models extends beyond technical audiences to broader institutions [44]: the European Union approved the General Data Protection Regulation (GDPR) [47] and it is stated in paragraph 1 of Article 15 that the data subject has the right to be informed of different aspects related to its data and to receive "meaningful information about the logic involved" in case of automated decision-making [44]. This emphasizes the idea that a model that lacks trustworthiness is unlikely to find practical use, especially in health. Therefore, understanding the decisions made by ML models is crucial for building trust among healthcare professionals and patients [36]. Without such understanding, the model predictions cannot be legitimized and justified [44].

Besides fostering user trust, explainability offers insights for refining model performance and supports understanding of the process being modelled [46]. Only by inspecting the insights behind model predictions, domain experts can detect if the model has learned wrong associations or potential bias [44].

Figure 11 conceptually summarizes the purposes of XAI and its associated audience profiles [34]. It is important to emphasize that two key goals emerge across these groups: the need for regulatory compliance and model understanding [34].

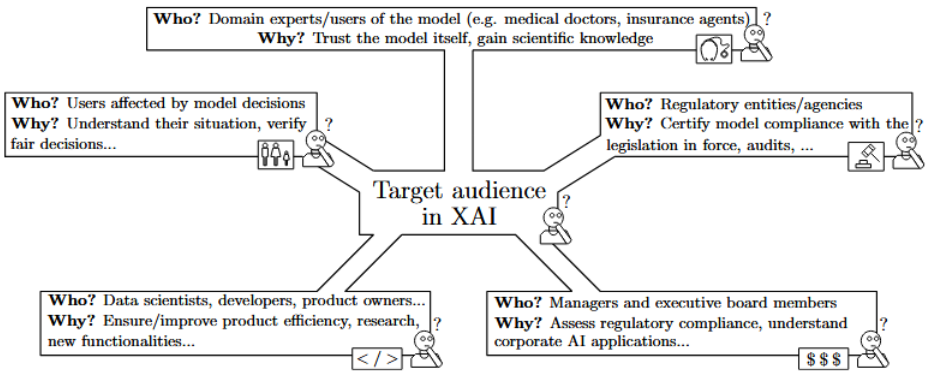


Figure 11: Diagram illustrating the different purposes of XAI sought by different audience profiles. Taken from [34]

Post-hoc explainability provides insight into the decisions of CNNs, helping end users trust that the AI makes correct, unbiased decisions grounded in evidence [43]. This category of XAI is applied after model training to identify the most relevant attributes or an explanatory set of observations, providing an informative and understandable account of the relationships between the input and the predicted outcomes [44, 48].

There are many criteria used to classify XAI methods: the degree of generalization in model explainability can be categorized into two main types: model-agnostic and model-specific methods. Furthermore, explainability can be divided based on its scope into global and local methods [44]. Global methods aim to interpret the model by providing insights into its overall decision-making structure, whereas local methods focus on understanding the rationale - the characteristic variables associated - behind a model's prediction for a single input dataset [49]. Within local methods, we distinguish sample-based explanations that provide different samples as explanations and feature-based explanations that indicate the features impacting the decision the most [36].

Figure 12 demonstrates how XAI applies to different data types: heatmaps highlight key regions in images, attention mechanisms emphasize influential words in text, and rule-based explanations identify critical features in tabular data classifications.

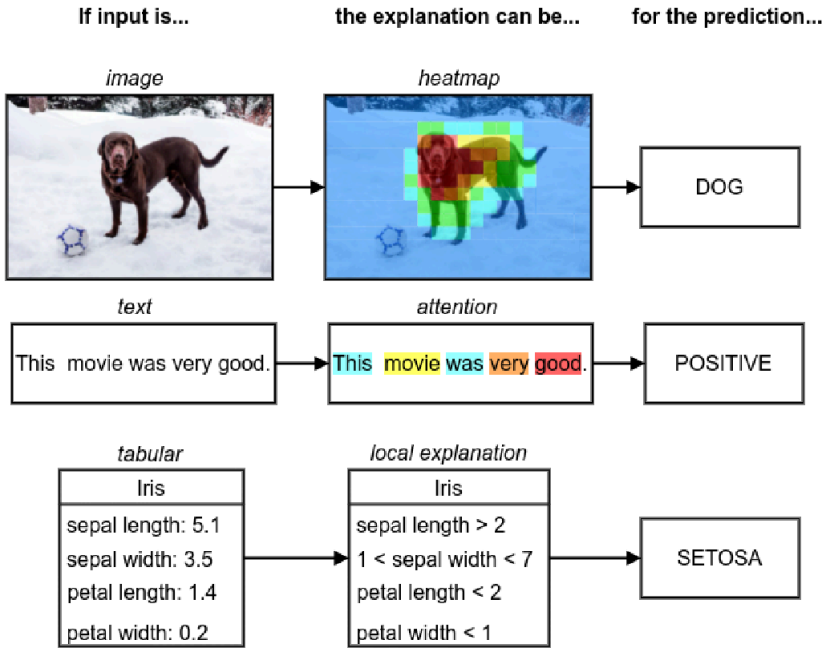


Figure 12: Examples of XAI methods applied to different types of data. Reproduced from [50].

Due to the inherent challenge of explaining how and why DL models make a particular decision, there is a recent development of solutions for visual explanations of CNNs [46, 51] that provide a qualitative understanding of the relationship between the instance's components and the model's prediction [15].

Among these, feature attribution explanations gained popularity in explaining image classification models [19] by ranking or measuring the influence and relevance of each feature on model outputs [52]. To achieve this goal, diverse algorithm approaches have been employed[52].

One of the approaches is feature attribution maps that highlight the most important pixels in an image for a given prediction of a model [19]. For color images, the attribution values for individual pixels can be computed by the average or maximum absolute values of their three color components [19]. They are represented by overlaying the original image with the saliency map, which allows us to understand which parts of an image a model is focusing on to make a decision [19, 49].

An overview of the explainability techniques is illustrated in the taxonomy in Figure 13 that was adapted from [34, 50].

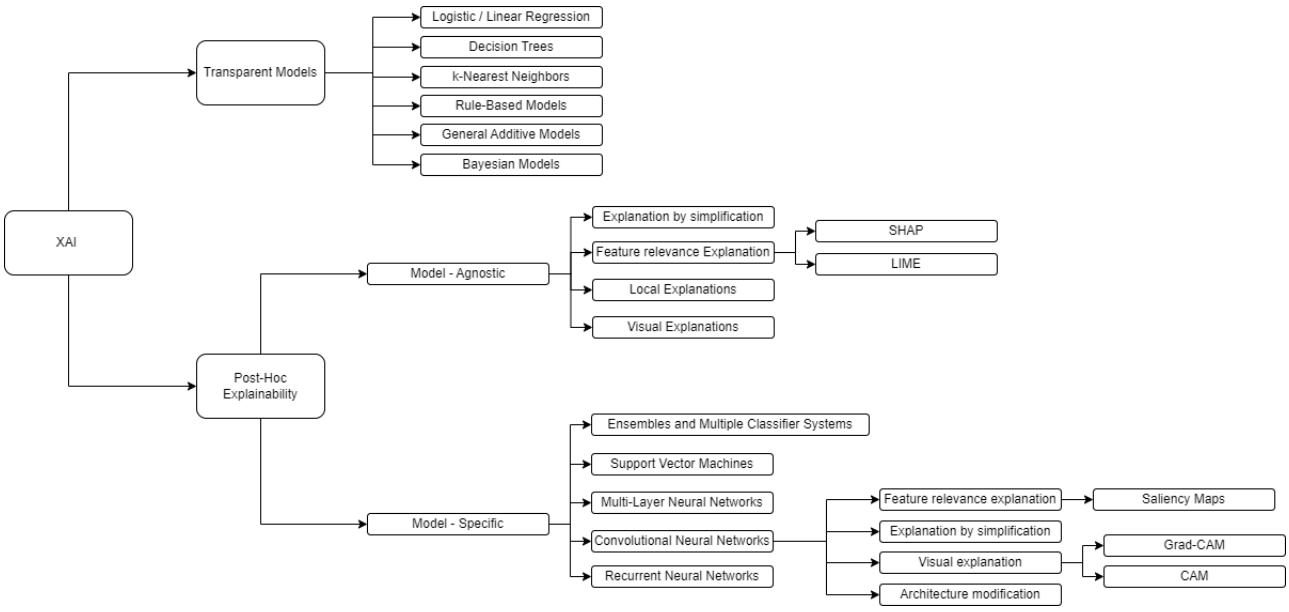


Figure 13: Taxonomy of the reviewed literature for explainability techniques. Adapted from [34, 50].

Although terminology differs between approaches (deleting information, ignoring features or using neutral values), all techniques measure a feature's influence by visualizing the impact of removing it from the model [53].

Finally, it is essential to distinguish XAI from Interpretable Artificial Intelligence (IAI). Although interpretability and explainability are often used interchangeably and lack universally agreed-upon definitions, some works emphasize key distinctions between them [54]. Rudin, for example, argues that interpretability refers to models that are inherently understandable, whereas explainability denotes post-hoc analyses that provide insight into black-box models that are not easily interpretable by humans [55]. Interpretability emphasizes the clarity of the model and its data structure (how a decision is derived), whereas explainability focuses on generating external justifications for model decisions (why a model makes that decision) [56].

3.3.1 Model-specific

Model-specific methods, as the name implies, are specific to particular kinds of models [46] [43]. By restricting our attention to specific model types, we can develop faster model-specific approximation methods [46].

3.3.1.1 Saliency Maps

Saliency maps are a class of pixel attribution methods used to identify which pixels (regions) of a certain image influence the model's classification the most [57, 58]. They are often referred to as heatmaps, where "hot" areas refer to regions that have a high impact on the model's prediction and "cold" areas highlight regions of lower relevance [59]. The brightness or intensity of a pixel in the heatmap correlates with its importance in the network's prediction [57, 58]. The resolution of the heatmap is determined by the spatial dimensions of the activation maps from the last convolutional layer of the model [58].

Saliency maps are one of the earliest explainability techniques and are part of the family of gradient-based methods [57]. To generate a saliency map, there are 5 main approaches developed:

1. Gradient Based Backpropagation, Symonian et al. 2013: computes the derivative of a specific class score with respect to the input image [60]

2. Deconvolutional Networks, Zeiler and Fergus 2013: introduces a modified back-propagation through ReLU layers [61]

3. Guided Back-propagation Algorithm, Springenberg et al. 2014: combines gradients with selective back-propagation [62]

4. Class Activation Mapping (CAM), Zhou et al. 2016 [63]

5. Grad-CAM and Guided Grad-CAM, Selvaraju et al. 2016 [64]

The first method is often referred to as "saliency mapping", although the other techniques are also ways to visualize saliency mapping. Gradient-based methods typically derive a gradient saliency map by back-propagating the gradient from a target function (such as a loss, layer or class activation) through the network and project it onto an image plane [19, 57, 58]. These methods mostly differ in how the gradient is computed and, to achieve fine-grained details, require storing intermediate data for a full backward pass on the input image, which can be computationally intensive [19, 57, 58].

To facilitate interpretation and comparison, saliency maps are typically normalized using Cumulative Distribution Function (CDF) to a scale between 0 and 1 [58]. Additionally, multiple maps can be combined by taking the weighted average: since they might differ in scale, they are upsampled using bilinear interpolation to match the dimensions of the input image [58]. Given r normalized saliency maps upsampled to the original input image size p, q, they are combined map is calculated as [58]:

$$c_{i,j} = \frac{\sum_{k=1}^{r} \gamma(s_{i,j,k}; \mu_k, \sigma_k) \cdot w_k}{\sum_{k=1}^{r} w_k} \tag{3.1}$$

Figure 14 gives a clearer visualization of some approaches to generating saliency maps.



Figure 14: Saliency maps generated using Grad-CAM, Grad-CAM++ and Smooth-Grad-CAM++, respectively. Taken from [59].

3.3.1.2 CAM

CAM is a visualization technique introduced by Zhou et al. to determine the discriminative image regions that a CNN model uses to classify images [19, 40, 63]. It involves visualizing the activation maps of the intermediate convolution layers to highlight the regions most relevant for classification [40]. The fully connected layers of the model (dense layer) are replaced by a layer named Global Average Pooling (GAP) to average the activations of each feature map [63].

To compute CAM, the selected layer is the final convolutional layer before the GAP layer or the fully connected layer [7]. This layer captures and retains spatial information about the input image and encodes

abstract high-level features learned by the network [7]. By visualizing the activation maps of this layer, CAM localizes the image regions that contribute the most to the network's prediction [7].

The process of computing CAMs is illustrated in Figure 15. The GAP layer outputs the spatial average of each feature map in the last convolutional layer, which is then combined through a weighted sum with the class-specific weights from the final fully connected layer to produce the network's prediction [63]. Similarly, a weighted sum of the feature maps in the last convolutional layer produces CAM, as formally defined for the case of a softmax classifier [63]. The same approach can extend to regression or other loss functions [63].

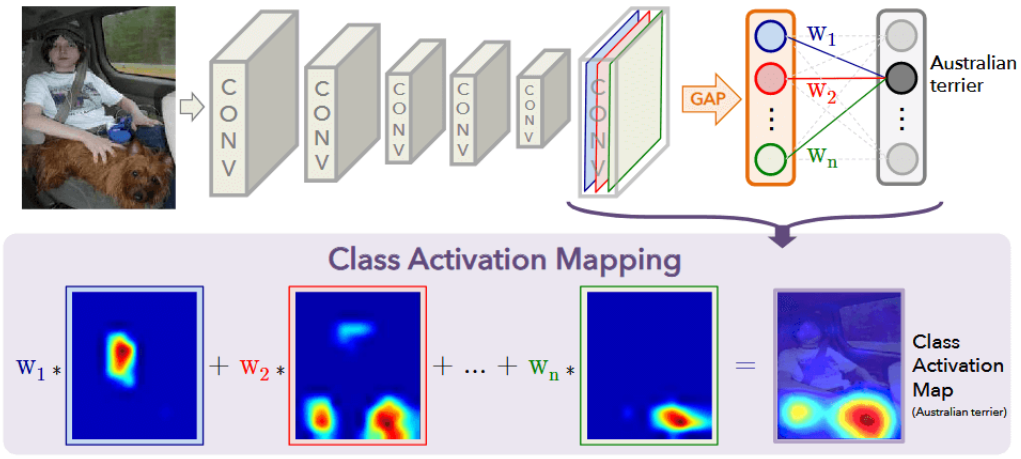


Figure 15: Class Activation Mapping. Taken from [63].

CAM algorithms depend on the learned feature representations of the neural network model, so if the relevant features are not captured correctly, the CAM-generated heatmaps will not accurately highlight the areas of interest [7].

A drawback of CAM is its requirement that the final convolutional layer directly precede a GAP layer or softmax layer, restricting its applicability to specific CNNs architectures (conv feature maps → global average pooling → softmax layer) [64].

CAM and its derivative methods, such as Grad-CAM, are based on the assumption that for each specific class c, the final score f_c of the network can be expressed as:

$$Y^c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k, \tag{3.2}$$

where Y^c is the final score, w_k^c is the weight corresponding to the k-th feature map A^k, and A_{ij}^k is the value at position (i, j) in A^k. Given this equation, the saliency map value at each location (i, j) for class c is computed as:

$$L_{ij}^{c} = \sum_{k} w_{k}^{c} A_{ij}^{k}, \tag{3.3}$$

where L_{ij}^{c} reflects the importance of location (i, j) for class c.

It is important to note that CAM assumes the penultimate layer in the CNN is global average pooling, which may not always be the case.

3.3.1.3 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) was introduced by Selvaraju et al. to overcome some limitations of CAM by producing visual explanations for any arbitrary CNN without requiring architectural changes or re-training [64, 65]. Both Grad-CAM and CAM are local backpropagation-based interpretation methods that exploit the spatial information that is preserved through convolutional layers [7, 64]. Grad-CAM is applied to trained networks with fixed weights and uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron, highlighting the regions of an input image that most influenced the model's decision [64].

Let the Grad-CAM localization map for class C be denoted by $\mathcal{M}_{\text{Grad-CAM}}^{C} \in \mathbb{R}^{b \times t}$, where b and t represent the width and height of the class C. The class score is defined by S_C before applying the softmax function. First, compute the gradient of S_C with respect to the k-th feature map m_k [43]:

$$\gamma_k = \frac{\partial S_C}{\partial m_k} \tag{3.4}$$

Next, the computed gradient γ_k is passed to the GAP layer to obtain the significance weights for the neurons of the last convolutional layer [43]:

$$w_k^C = \frac{1}{y} \sum_{p} \sum_{q} \gamma_k(p, q) = \frac{1}{y} \sum_{p} \sum_{q} \frac{\partial S_C}{\partial m_k(p, q)} \tag{3.5}$$

The weight matrix w_k^C represents the significance of the k-th feature map for class C. This weight matrix is then multiplied by the feature map m_k and passed through a ReLU activation function to obtain the Grad-CAM map [43]:

$$\mathcal{M}_{\text{Grad-CAM}}^{C} = \text{ReLU}\left(\sum_{k} w_k^C m_k \right) \tag{3.6}$$

Figure 16 represents the process of Grad-CAM and Guided Grad-CAM: given an input image and a class of interest, the image is forward-propagated through the CNN, followed by task-specific computations to obtain a raw score for the target category. The gradients are set to zero for all classes except the class of

interest, which is set to 1. This signal is then backpropagated to the rectified convolutional feature maps of interest, which are combined to compute the coarse Grad-CAM localization (blue heatmap), which highlights where the model has to look to make the particular decision. Finally, we pointwise multiply the heatmap with guided backpropagation to get Guided Grad-CAM visualizations, which are both high-resolution and concept-specific [66].

In the generated heatmap, blue colors represent minimal values, indicating that no significant features were extracted for the given class. Yellow and green colours signify intermediate values corresponding to smaller features, while red and dark red indicate higher values, representing the most relevant features for the model's decision [66].

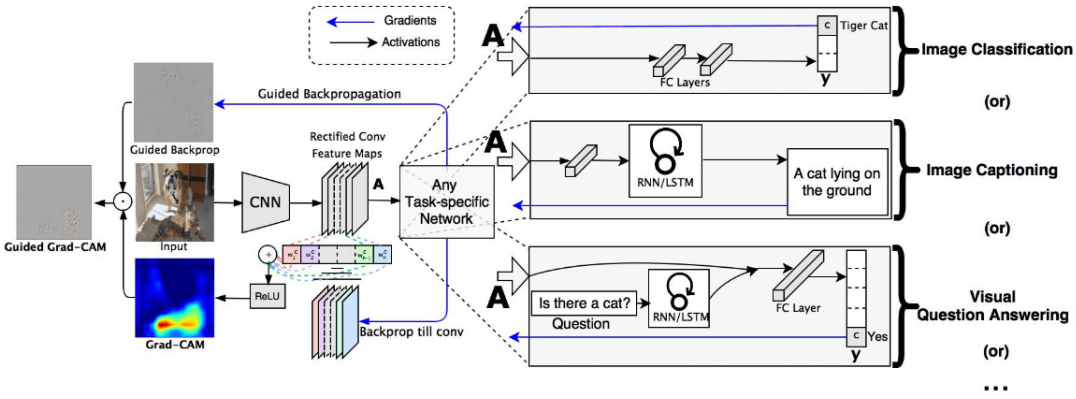


Figure 16: Overview of Grad-CAM and Guided Grad-CAM. Taken from [64].

Although Grad-CAM addresses some of the limitations of the CAM method, it still faces challenges, such as its difficulty in accurately localizing multiple instances of the same object within a single image [48, 67].

3.3.2 Model-agnostic

Model-agnostic or model-independent methods have no special requirement for the model [49]. They provide post-hoc explanations of the model behavior by treating the original machine model as a black box and can be applied to any ML model, regardless of its type [44, 46].

3.3.2.1 LIME

LIME is a technique introduced by Ribeiro et al. that provides local explanations for individual predictions of complex models, often referred to as "black-box" models, and thus helps the decision-makers in justifying the model's behaviour [15]. LIME explains the predictions of any classifier with two or more

classes by learning an interpretable model in the locality of the prediction to be explained and since it is designed to produce linear explanations, it follows a regularized additive feature attribution method [15, 46]. The coefficients of the linear model act as attribution scores for each feature [68].

Despite the overall complexity of a model, LIME simplifies the model behaviour by focusing on a small and specific area around the instance we want to explain [15, 69]. This is done by slightly perturbing the selected instance and learning a sparse linear model around it [15]. This method involves the following key components [15]:

- Interpretable Model: LIME uses a simple interpretable model $g \in \mathcal{G}$ (such as a linear model or decision tree) to approximate the predictions of the complex model f. The interpretability of g is quantified using a complexity measure $\Omega(g)$, which can include:
  - The number of non-zero weights for linear models.
  - The depth of the tree for decision trees.
- Locality Definition: To explain the prediction $f(x)$ of an instance x, LIME defines a locality around x using a proximity function $\pi_x(z)$, which measures the similarity between x and other instances z.
- Fidelity: The quality of g in approximating f locally is measured by a fidelity function $\mathcal{L}(f, g, \pi_x)$, ensuring that g represents f's behavior within the defined locality.
- Optimization Objective: The explanation $\xi(x)$ is obtained by solving the optimization problem:

$$\xi(x) = \arg\min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g). \tag{3.7}$$

  This ensures the balance of the trade-off between the fidelity of the surrogate model and its simplicity.
- Sparse Linear Models: LIME typically uses sparse linear models as g, fitting these models to perturbed versions of x. By perturbing x and observing the behavior of f, LIME builds a local approximation that highlights the most significant features contributing to the prediction.

The LIME algorithm is implemented in five main steps: first, an instance x is selected from the dataset and then a synthetic dataset is generated by sampling the original dataset and artificially producing instances, both close and far to the desired data point x. Next, the model f is applied to the new perturbed

dataset to make predictions and the weights are assigned to each feature based on proximity to data point x. Finally, a surrogate model g is used to approximate the model in the region of the selected point x [48].

For image classification tasks, LIME divides the image into contiguous superpixels that usually represent regions that share similar visual characteristics. A dataset of perturbed instances is generated by selectively turning superpixels on or off and each perturbed image is weighted based on its similarity to the original instance. Using these perturbed instances, LIME approximates the complex model with a locally weighted linear model. Finally, the superpixels with the highest positive weights are presented as an explanation. This process not only provides insights into the model's predictions but also helps optimize its performance by reducing generalization errors. During perturbation, superpixels are either retained with their original values (mapped to 1) or replaced with an average of neighbouring pixels (mapped to 0) to simulate their absence [34, 46, 67].

Figure 17 illustrates the main steps of the LIME method, including input segmentation, generation of perturbed instances, and the creation of a locally interpretable explanation using weighted regression [67].

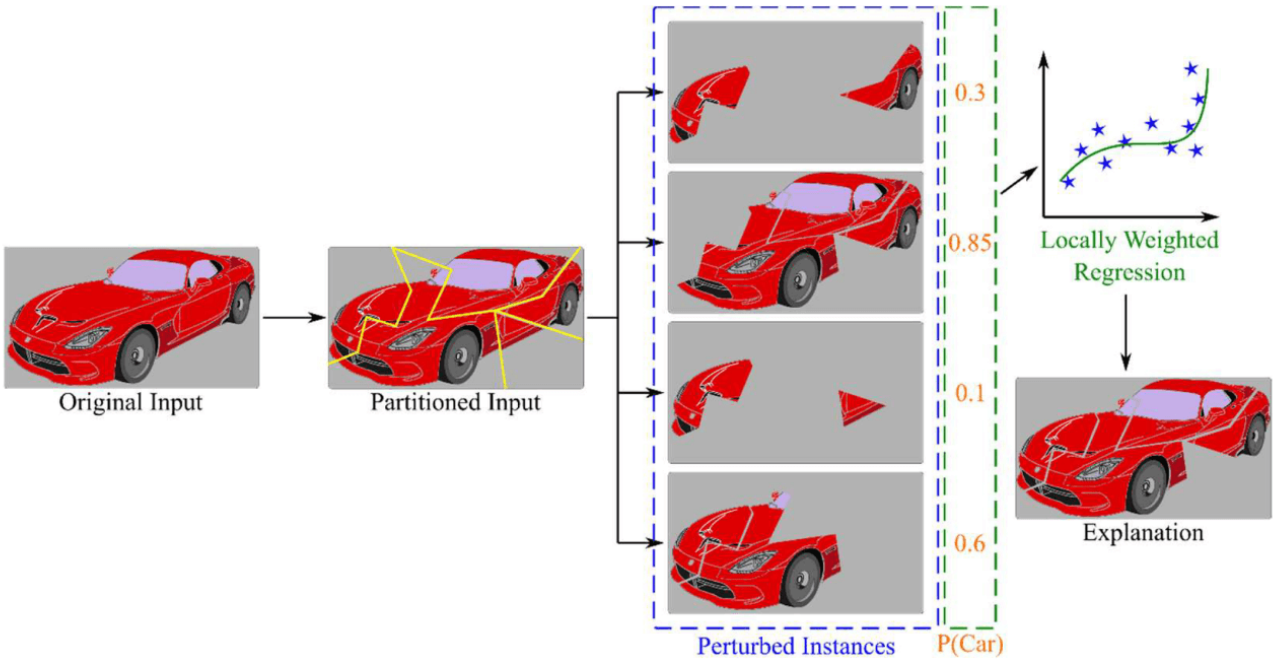


Figure 17: Steps of the LIME method. Reproduced from [67].

LIME offers several advantages and disadvantages that influence its effectiveness as an explanation method. On one hand, it is versatile and suitable for a wide range of explanatory variables, providing sparse and interpretable explanations. As mentioned before, it is model-agnostic, meaning it makes no assumptions about the underlying model, and its explanations are local, offering fidelity in the region surrounding the instance of interest. Additionally, it can generate selective and even contrastive explanations,

enhancing its utility in specific scenarios [43, 44].

On the other hand, LIME also has notable limitations. The algorithm assumes feature independence, which may not hold in many datasets, leading to potential inaccuracies. Its reliance on random perturbations introduces variability, resulting in inconsistent and unstable explanations. Furthermore, while it fits linear models to local data, many real-world relationships may remain non-linear even in localized regions, compromising the quality of explanations. LIME struggles with models that have non-linear decision boundaries. Finally, the choice of locality definition and neighbourhood generation significantly impacts the quality of the explanation, and multiple unstable explanations can arise for the same instance due to the stochastic nature of its process [43, 44, 66].

3.3.2.2 SHAP

SHAP is a game-theoretic approach introduced by Lundberg and Lee to measure feature importance and explain the predictions of any ML model [46, 70]. Rooted in the concept of SHAP values, it assigns importance scores to each feature-instance combination, providing an explanation for individual model predictions [46, 71].

Shapley values, introduced by Lloyd Shapley, are derived from cooperative game theory and are used to fairly distribute the total payoff among players in a game $v : 2^N \rightarrow \mathbb{R}$ [46, 71]. In the context of ML, features are equivalent to "players", the model is analogous to the "cost-function" and the attributions to the "cost-shares" in a coalition game [71]. The Shapley value $\phi_i(v)$ for a feature i is mathematically defined as follows [68, 72]:

$$\phi_i(v) = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} \left(v(S \cup \{i\}) - v(S)\right) \tag{3.8}$$

where $\phi_i(v)$ is the shapley value for element i, S is the subset of D excluding i, $\frac{|S|!(|D|-|S|-1)!}{|D|!}$ is the weight of subset S, $v(S)$ is the value of coalition S, and $v(S \cup \{i\}) - v(S)$ is the marginal contribution of i to subset S.

Shapley values uniquely provide both local and global explainability while satisfying key mathematical axioms that ensure fairness and consistency: efficiency, symmetry, dummy and additivity [72–74]. Shapley demonstrated that this is a unique function ϕ that satisfies these axioms across all games, thus providing a robust framework for cooperative game theory [72].

- Symmetry Axiom: Players i and j who contribute equally to all coalition subsets S receive the same

value.

$$\forall S \subseteq N \setminus \{i,j\} : v(S \cup \{i\}) = v(S \cup \{j\}) \implies \phi_i(N,v) = \phi_j(N,v) \tag{3.9}$$

- Dummy Player Axiom: A player who does not contribute to the value of any coalition receives a Shapley value of 0.

$$\forall S \subseteq N \setminus \{i\} : v(S \cup \{i\}) = v(S) \implies \phi_i(N,v) = 0 \tag{3.10}$$

- Efficiency Axiom: The feature contributions must add up to the total payoff in the game. In SHAP, this is adapted as:

$$\sum_{j=1}^{M} \phi_j = f(x) - \phi_0, \tag{3.11}$$

  where ϕ_0 is the baseline model prediction.

- Additivity Axiom: For any two games v and w, the Shapley value of their combined game satisfies:

$$\phi(v+w) = \phi(v) + \phi(w), \tag{3.12}$$

  where the combined game is defined as $[v+w](S) = v(S) + w(S)$.

There is an alternative Shapley value interpretation that relies on a permutation-based description that highlights the fairness and symmetry of the Shapley value [72]. Players are ordered randomly, and the value assigned to player i corresponds to their expected marginal contribution $v(S \cup \{i\}) - v(S)$ given this order [72].

As well as the technique LIME, SHAP is part of the family of perturbation-based methods that work by altering or masking parts of the input data and observing how the model's prediction changes [19]. They also belong to the class of additive feature attribution methods, which are those whose attributions sum to a specific value, such as the model's prediction [46].

Given the input features $x_1, \ldots, x_d$, it assigns an effect ϕ_i to each feature, and summing the effects of all feature attributions approximates the output $f(x)$ of the original model [46, 68]:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.13}$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

To unify the class of additive feature attribution methods, Lundberg and Lee introduced SHAP as a unique method that satisfies certain desirable properties (axioms) [46, 68]. However, directly calculating

Shapley values is computationally expensive and time-consuming due to the factorial complexity of evaluating all subsets. To address this, SHAP employs sampling-based methods to approximate these values, such as Monte Carlo methods, to estimate values efficiently [44, 46, 68]. This has led to distinct Shapley value approaches that differ in how they remove features [68].

Using Shapley values $\{\phi_j(v)\}_{j=1,\ldots,n}$, the SHAP explanation can be computed as shown in Eq. 3.14 and is further demonstrated in Figure 18:

$$g(z') = \phi_0 + \sum_{j=1}^{n} \phi_j z'_j, \tag{3.14}$$

where $z' = [z'_1, \ldots, z'_n] \in \{0,1\}^n$ is a vector of zeros and ones. For $z'_j = 0$, the j-th feature is not part of the coalition, whereas $z'_j = 1$ means j-th feature is present in the coalition. Moreover, $\phi_0 = \mathbb{E}[f(x)]$ is the average predicted value computed over all features [67].

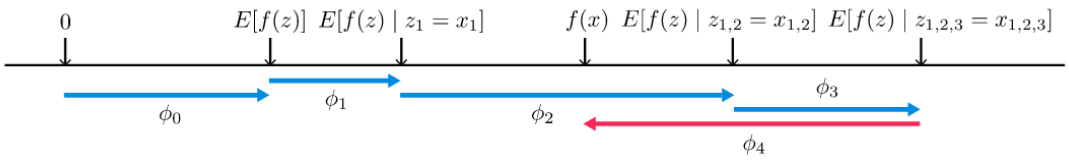


Figure 18: Shapley value computation process: decomposition of a model's prediction $f(x)$ into contributions from baseline ϕ_0 and individual feature contributions ϕ_1, ϕ_2, ϕ_3. Reproduced from [46].

The application of the SHAP technique gives a color plot visualization to explain the output of a model by assigning importance to each feature [66]. Red color regions represent the positive SHAP values, indicating the pixels that contributed to classifying that image as a specific class and negative SHAP values can be denoted by blue color regions and represent the pixels that contributed to not classifying the image as a particular class [46]. The magnitude of the values measures how strong the effect is: a "low" Shapley value suggests that the corresponding feature has a relatively smaller impact on the prediction, while a "high" Shapley value indicates a stronger influence [66]. It's important to note that the exact interpretation of "low" and "high" may depend on the specific context, dataset, and problem being analysed [66].

We can visualize in Figure 19 an example of SHAP visualizations. In the first example, the input bird image is classified as an American Egret with the next probable classes being a Crane, Heron and Flamingo [70]. The classification as an American Egret is primarily influenced by the "bump" on the bird's neck, as evidenced by the red-highlighted regions in the SHAP map [70].

In the second example, the input image of a boat is classified as a speedboat and the defining factor for its classification is the boat's shape, which is similarly highlighted in red in the SHAP map [70].

By increasing the max_evals parameter, we let SHAP execute the original model more times, resulting in a more detailed explanation. In this case, a blurring kernel was applied because its runtime is much more computationally efficient than inpainting [70].

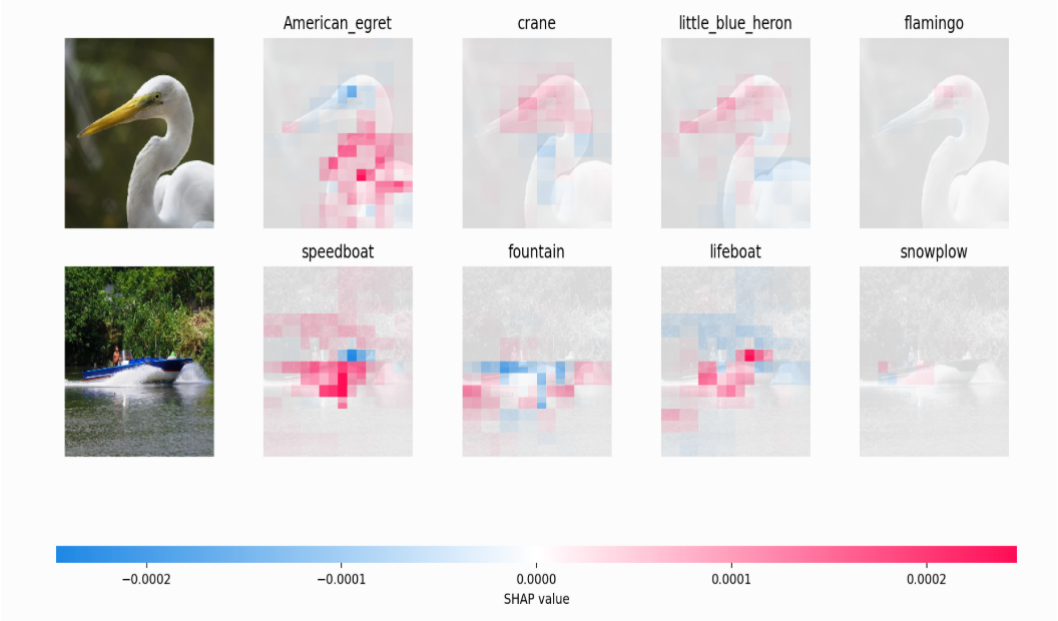


Figure 19: SHAP visualization of image classification, highlighting regions that contribute positively (red) or negatively (blue) to the predicted class. Reproduced from [70].

This model requires significant time to calculate SHAP values, so it can be combined with other techniques to accelerate the computation [49]. For example, Deep SHAP (a fast explainability technique) is often used for models with neural network-based architectures. In general, SHAP is a powerful tool for providing explanations and can be applied to a wide range of clinical topics [49].

3.4 Related Work on XAI in Medical Imaging

A significant body of research has been compiled and analyzed for this state-of-the-art review. Despite the considerable advances in DL methods and the significant progress in developing highly accurate models for various medical tasks, the challenge of explaining these models—particularly in complex applications like colorectal polyp classification—remains an underexplored area. While deep learning models have demonstrated impressive performance in medical imaging, their "black-box" nature continues to hinder their adoption in clinical practice, where transparency and interpretability are essential.

In the broader context of medical imaging, several studies have focused on enhancing the explainability of AI models to make them more suitable for clinical environments. For instance, Knapic et al. (2021) compared various post-hoc interpretable machine learning methods - LIME, SHAP, and Contextual Value

and Utility (CIU) - using human evaluations [69]. Their findings suggest that these explainable methods could be generalized to different medical datasets, highlighting their potential to provide strong decision-support in clinical settings.

Expanding on the application of XAI in medical tasks, Guah et al. (2024) focused on classifying ovarian tumors as benign or malignant using a custom ResNet60 architecture, a modified version of ResNet50, which achieved 97.5% accuracy on the test dataset [75]. Their study incorporated XAI techniques to identify the regions and features influencing the model's classification decisions, which were validated to ensure that the model's predictions were based on relevant and interpretable features.

Similarly, Aldughayfiq et al. (2024) employed LIME and SHAP to explain the predictions of a pre-trained InceptionV3 model on a dataset of retinoblastoma and non-retinoblastoma images with the high accuracy of 97% [76]. Their work provided important insights into the regions and features that influence the model's predictions, offering a more interpretable approach to the classification of medical images.

In the context of liver tumor detection, Aswin et al. (2024) proposed a novel method using ResNet architecture and XAI techniques like LIME, achieving an impressive 99.8% accuracy [77]. By providing detailed explanations for each prediction, their approach enhanced trust in the model's decision-making process, making it more applicable for use in clinical settings.

Focusing on the detection of colorectal adenocarcinoma, Di Giammarco et al. (2024) developed an automated approach for detecting cancer in colon histological images [7]. Their work demonstrated the effectiveness of CNN architectures such as MobileNet, Inception-V3, and EfficientNet, achieving 99% accuracy, precision, and recall. In addition to high performance, they employed CAM algorithms, including Grad-CAM, Score-CAM, and FastScore-CAM, to visually localize patterns in the images associated with cancerous cells in the colon. These visual explanations help clinicians interpret the model's decision, offering valuable insights into areas of the tissue that are critical for diagnosis. Similarly, Sweidan et al. (2024) utilized methods such as DeepSHAP, CoMTE, and 2-Step TSR to provide diverse perspectives on model predictions in medical diagnoses, highlighting the potential of XAI to improve model interpretability in healthcare [36].

However, while these methods offer valuable insights into model decisions and visual explanations, they do not establish a quantitative framework to correlate the model's confidence levels with the reliability of the generated visual explanations. This lack of a systematic approach to assess the consistency and trustworthiness of the explanations presents a significant gap in the literature.

3.5 Summary

Among the explainability techniques reviewed, SHAP and Grad-CAM emerge as two particularly relevant and complementary approaches for medical image analysis, especially in deep learning-based classification tasks. This selection is supported by their distinct but synergistic properties: SHAP provides a theoretically grounded, model-agnostic framework for feature attribution based on cooperative game theory, while Grad-CAM generates spatially coherent, gradient-based visualizations that preserve semantic and anatomical context. Table 2 summarizes the principal characteristics, costs, strengths, and limitations of the selected techniques.

Grad-CAM produces class-specific saliency maps that indicate the image regions most influential for a CNN's prediction, answering the "what" and "where" of model attention. When applied to convolutional architectures, Grad-CAM yields robust, spatially localized heatmaps that are useful for tasks requiring precise anatomical localization and visual validation against expert annotations. Its reliance on convolutional gradients ensures alignment with the model's internal decision signals and enables efficient per-image visualization.

By contrast, SHAP provides localized, quantitative attributions by assigning contribution values to input regions. These numeric attributions support statistical comparison across samples and enable integration into quantitative evaluation metrics. SHAP's model-agnostic formulation makes it appropriate for benchmarking and for comparing different architectures on a common interpretability scale.

Despite their advantages, each method has limitations that must be acknowledged: SHAP can be computationally demanding for high-dimensional inputs, and Grad-CAM may lack fine spatial resolution or struggle to localize multiple object instances. These trade-offs are reflected in Table 2, which aided the selection process by providing a concise comparison of interpretability and computational cost. Given these complementary properties, the two approaches are adopted in this thesis.

Table 2: Comparison of selected XAI techniques.

Attribute	**SHAP**	**Grad-CAM**
Type	Post-hoc attribution (model-agnostic; explains feature contributions)	Post-hoc gradient-based (visual explanations for convolutional networks)
Interpretability	High (local, quantitative feature attributions that can be aggregated)	Medium (visual heatmaps that are intuitive but less quantitative)
Cost	High — computationally expensive (especially exact SHAP or when many features)	Medium — requires gradients and forward/backward passes but typically cheaper than SHAP for large feature sets
Strengths	Model-agnostic; clear attribution per feature; supports global aggregation of local explanations; well-grounded theory (Shapley values)	Provides spatial/visual explanations; very useful for CNN architecture diagnostics; fast to compute for single images
Limitations	Expensive to compute at scale; can be noisy without sufficient sampling; interpretation may be nontrivial for highly correlated features	Limited to architectures with spatial structure (e.g., CNNs); heatmaps can be coarse and may not localize precisely; less directly quantitative

Chapter 4

Methodology

This chapter describes the experimental framework implemented in this project. It details the dataset and its preparation, the computational environment and software tools, the classification models and training configuration, the explainable AI techniques applied, and the evaluation protocol and metrics used to assess classification performance and explanation quality.

4.1 Approach Pipeline

Figure 20 presents the overall workflow implemented in this project.

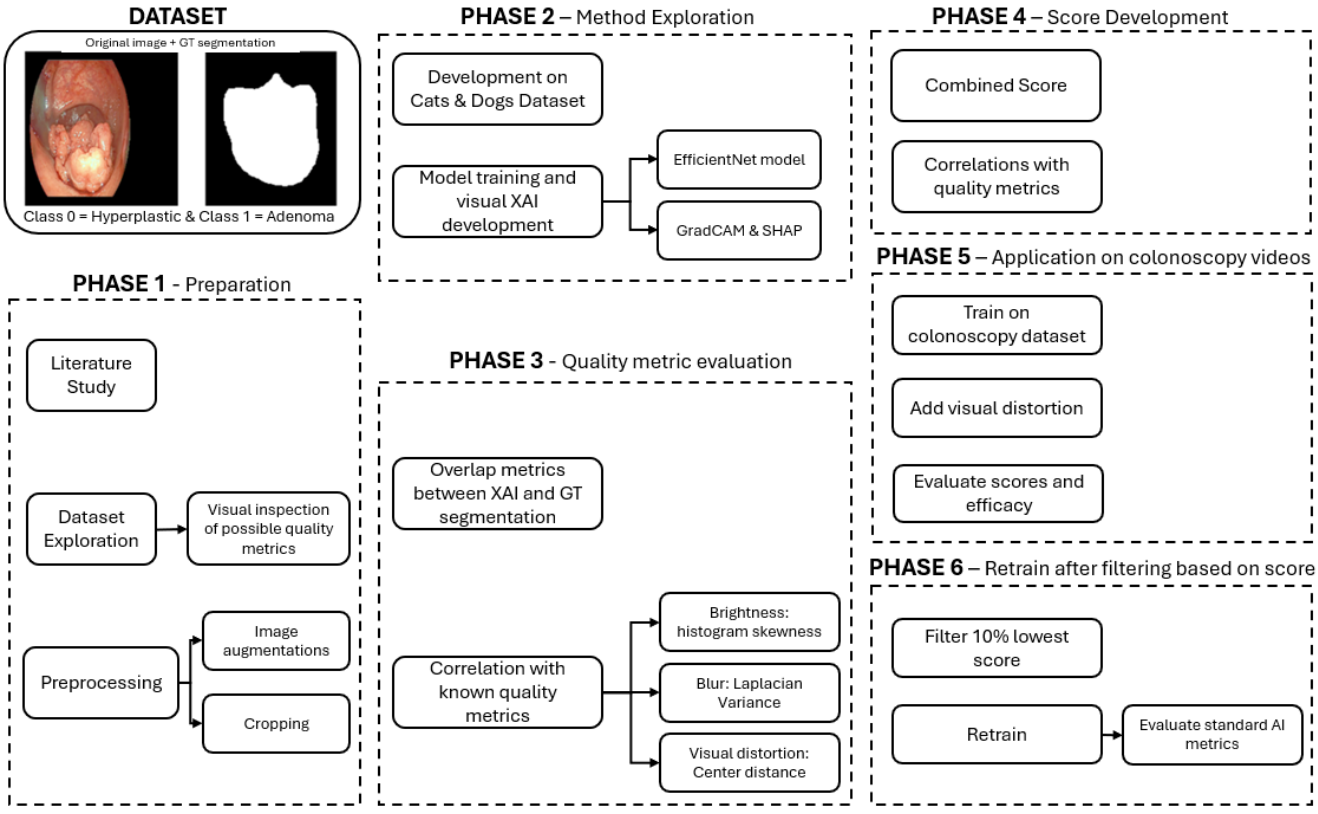


Figure 20: Overall workflow for the project.

The dataset panel (left of Figure 20) shows representative raw frames and corresponding ground-truth segmentations. This defines the principal classes and forms the basis of the preprocessing criteria employed in Phase 1.

Phase 1 includes a review of the relevant clinical and technical literature on XAI for medical imaging, followed by a visual and statistical exploration of the colonoscopy dataset to identify valid sequences and frame-level issues. This phase also involves preprocessing, which includes cropping frames with bounding boxes and padding, as well as removing invalid frames, such as those without reliable annotations or incorrect format. Image augmentations are also applied, including random rotations, flips, translations, and color distortion. Additionally, an experimental degradation branch is created to introduce controlled corruptions such as Gaussian blur, additive noise, partial occlusions, and extreme brightness shifts. These distortions are designed for later robustness testing.

In Phase 2, we validate the methodology using a simpler control dataset, cats and dogs. This step serves as an initial validation to ensure that both the model architectures and XAI techniques are functioning correctly before applying them to the complexities of the clinical data. The choice of the cats and dogs dataset is strategic as it presents a straightforward binary classification problem with highly varied and colorful images, which helps test the model's ability to handle diverse visual features. In contrast, colonoscopy images, especially those of polyps, are often limited in terms of color variation, typically showing only shades of pink, and may suffer from artifacts such as blurring, noise, and low resolution, which makes them more challenging for the model to process effectively. In clinical experiments, we apply XAI methods (Grad-CAM and SHAP) to generate per-frame saliency maps.

In Phase 3, to quantify the spatial agreement between XAI maps and ground truth masks, we calculate the percentage of the heatmap covering the object. This is done by computing the ratio between the overlap of the most important pixels in the heatmap and the object, relative to the total area of the heatmap, and multiplying by 100, as illustrated with a polyp example in Figure 21. The first panel shows the GT mask, highlighting the annotated polyp region (in red). The second panel presents the saliency map, where the p brightest pixels represent the most salient regions identified by the XAI method. The third panel combines both the GT mask and the saliency map, where the green region denotes the spatial overlap between the two.

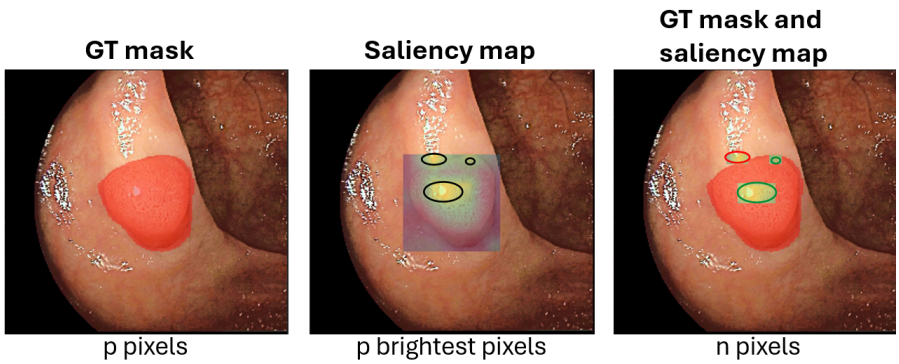


Figure 21: Thresholding and overlap strategy used to compare saliency maps with Ground-Truth (GT) segmentation.

Using the results from Phase 3, we explore and quantify some image quality features per-frames doing statistical correlations.

The Phase 4 is dedicated to the combined metric development based on the explorations of the previous Phase.

In Phases 5 and 6, we apply the full pipeline to the colonoscopy dataset. Controlled distortions from the degradation branch are injected into the sequences to explore the correlations and changes.

Lastly in Phase 6, a filtering scheme is proposed that removes the lowest-quality frames based on the composed metric. Although this filtering and retraining procedure is outlined here, its execution is not done in this project and is reserved for future work.

4.2 Preliminary Experiment

A preliminary experiment was conducted on the Oxford–IIIT Pet Dataset to validate the training pipeline, data handling and model implementation before applying the methodology to the target medical dataset. The Pets dataset provides a challenging but well-documented binary classification proxy (cat vs. dog) with large intra-class variability in scale, pose and lighting, and with pixel-level annotations useful for downstream explainability analyses.

The Oxford–IIIT Pet Dataset contains images of 37 pet categories (breeds), with roughly 200 images per class. Each image is accompanied by metadata and pixel-level annotations (trimap segmentation and head ROI), which facilitate controlled cropping and mask-based experiments. For this preliminary experiment the problem was cast as a binary classification (cat vs. dog), collapsing breed labels into two classes. Example breeds include Persian, Siamese, Bengal, Birman, Maine Coon, Ragdoll, Abyssinian, and several dog breeds such as Beagle, Boxer, German Shepherd, Pomeranian and Bulldog.

Image lists and binary labels were loaded from CSV files following a 5-fold partitioning scheme.

A pretrained EfficientNet-B0 backbone (Keras Applications) was used as the feature extractor. The backbone weights were initialised from ImageNet pretraining, and all layers were frozen except for the final blocks and the custom classification head, which was appended to produce a binary output (dog vs. cat). The classification head consisted of a global pooling layer followed by one or more dense layers and a two-unit softmax output.

During training, standard data augmentations were applied (rotations, flips, translations, zoom, brightness jitter and occasional blur) to increase robustness to common image variations. Model performance was monitored on the validation set and final metrics were computed on the held-out test set.

The dataset was split into training, validation and test partitions while maintaining class balance. The number of samples per class in each partition is summarised in Table 3.

Table 3: Distribution of samples in the *Cats vs Dogs* preliminary experiment.

	Cat (0)	**Dog (1)**
Train	1663	3481
Validation	232	503
Test	476	994

4.3 Main Experiment

The principal experiment applies the previously validated pipeline to an endoscopic polyp dataset with the dual aims of evaluating the models' ability to predict clinically relevant labels (including histological categories), and quantifying the spatial and statistical agreement between model explanations (Grad-CAM and SHAP) and ground-truth segmentations.

The clinical dataset comprises metadata such as size, location, morphology, histology, alongside with spatial annotations (presence labels, point annotations, bounding boxes and dense segmentation masks). These metadata provide essential contextual and clinical information that support the interpretation and validation of AI-based models. Data collection for this study was initiated under the supervision of Prof. Dr. Raf Bisschops at UZ Leuven in collaboration with his large network of 11 international centers and described in Eelbode's doctoral thesis where all the procedures and methodologies of data collection are detailed [8].

4.3.1 Dataset Description

The dataset was conducted in three consecutive phases between 2018 and 2022, involving multiple international endoscopy centers, namely University Hospitals Leuven (Belgium), Nuovo Regina Margherita Hospital (Italy), Solent Centre for Digestive Diseases (United Kingdom), Centre Hospitalier Universitaire Hôtel Dieu (France), University Medical Center Mainz (Germany), Krankenhaus Barmherzige Brüder Regensburg (Germany), Gastrozentrum Lippe (Germany), Centrum Onkologii–Instytut im. Marii (Poland), UZ Gent (Belgium), ULB Erasme Hospital (Belgium), AZ Delta (Belgium), and AZ Maria Middelares (Belgium) [8].

Data collection was part of a clinical trial, with explicit patient consent for the use of clinical metadata, video recordings and histological results for research purposes [8]. To ensure privacy, all personally identifiable information was removed, and ethical approval was obtained from all participating centers [8]. A semi-automatic annotation pipeline was used, where two to three manual delineations per clip were propagated to generate full-sequence segmentation masks.

The dataset was analyzed over a cumulative 9431 seconds of video recorded at 25 frames per second (FPS), resulting in 235 775 individual frames. The data acquisition process involved extracting 5-second video clips for each identified polyp where it was clearly visible. From each video clip, three representative frames, where the polyp was most prominent, were selected for detailed annotation. These annotations were further augmented by overfitting the selected frames to ensure generalization across the remaining frames. After polyp removal, histopathological evaluation was performed to confirm classification accuracy, establishing a reliable ground truth for the dataset, so the numbers of polyps reduced .

A structured summary of the dataset's content and characteristics is presented in Table 4.

Table 4: Distribution of the main dataset.

Category	**Details**
Types of Polyps	• Hyperplastic: 287 • Adenoma: 841 • Other categories: Sessile Serrated, Advanced Adenoma, Adenocarcinoma
Modality of Imaging	• White Light Imaging (WLI): 822 • Blue Light Imaging (BLI): 325 • Video Capsule Endoscopy (VCE): 489 • Other modalities: LCI, NBI, iSCAN, FICE
Video subset	• Total recording duration: 9431 s (≈ 2:37:11) • Frame rate: 25 fps • Total frames: 235 775

To create more manageable input data for classification, each 1920x1080 frame was cropped to a 256x256 region of interest around the visible polyp. This cropping allows the classification models to efficiently focus on and learn the features of the localized polyp. An example of a cropped region from one of the dataset frames is shown in Figure 22.

The quality of images in the dataset is a critical factor, as it can significantly impact the performance of the DL models. In particular, variability in the quality of the polyp images arises due to several factors such as motion blur, partial occlusion, or poor positioning during capture. These inconsistencies introduce noise, which we hypothesize may impact model effectiveness during both training and testing. To address this issue, we propose a new image quality metric based on XAI. This metric quantifies image quality and

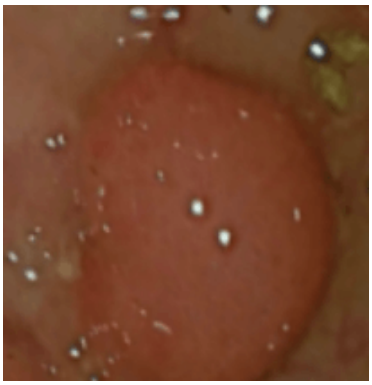


Figure 22: Example of a cropped 256x256 region of interest from the dataset.

ensures that only high-quality frames are used during the model training process.

4.3.2 Preprocessing, Augmentation and Crops

Data augmentation is an essential component in medical image analysis, particularly when training models on limited datasets. Augmentation methods increase the effective size and variability of the training data by simulating realistic acquisition variations, such as changes in orientation, illumination, and partial occlusions, thereby reducing overfitting and improving generalization on unseen data [10]. By exposing the model to diverse transformations, these techniques encourage the network to learn features that are robust to irrelevant image variations while preserving the semantic content relevant to polyp detection [8, 10].

During training, images were preprocessed using Keras/TensorFlow preprocessing utilities. The primary augmentation techniques, including rotations, flips, zooms, brightness adjustments, shear, blur, translations, and cropping, are described in Table 5. These augmentations were applied following standard practices to increase robustness and prevent overfitting.

Table 5: Summary of preprocessing steps and data augmentation techniques.

Stage	**Parameters**
Intensities Rescaling	rescale = 1./255
Geometric	Rotation $\pm 90°$; random horizontal/vertical flips; shear ≤ 0.2; translation $\leq$ 20 px
Photometric	Brightness $\pm 20\%$; color jitter factor ≤ 0.2.
Blur	Gaussian blur, $\sigma = 1.3$
Crop & resize	Crop to annotation bounding box with configurable padding; resize to $(256, 256)$.

4.3.3 Training Setup and Hyperparameters

This subsection describes the training setup, including loss function, optimizer, dataset partitioning, and the principal hyperparameters used during optimization.

Each input image was partitioned into five spatial regions (top-left, top-right, centre, bottom-left, bottom-right) to capture local contextual information across the frame. Feature maps extracted from these regions by the backbone were pooled and concatenated into a single feature vector. This concatenated vector was passed through fully connected layers with ReLU activations. Batch normalization and dropout were applied to the dense layers for regularization, and the final classification layer used a softmax activation to output class probabilities. The backbone network employed ImageNet-pretrained weights. Its output was reduced to a fixed-size representation via global average pooling (GAP), which decreases feature dimensionality, mitigates overfitting, and emphasises global feature presence rather than precise spatial location [78].

The training process follows a supervised learning approach, using labeled data to optimize the model predictions by minimizing cross-entropy loss.

Cross-entropy is a common loss function for multi-class classification tasks [79]. It is derived from information theory and it quantifies the difference between two probability distributions: the predicted probabilities and true labels [79] [80]. During training, cross-entropy loss is used when adjusting model weights to minimize the loss with smaller loss values indicating better model predictions [79]. It is mathematically defined as follows:

$$H(P, Q) = -\sum_{x \in X} P(x) \log(Q(x)) \tag{4.1}$$

where $P(x)$ is the probability of the event x in P and $Q(x)$ is the probability of event x in Q, and log is the base-2 logarithm, meaning that the results are in bits or if the base-e (natural logarithm) is used, the result will have the units called nats [79] [80].

For balanced datasets, standard loss functions like cross-entropy are sufficient, but for more complex problems such as segmentation and unbalanced tasks the selection of the loss function is more relevant [32].

Furthermore, model parameters, such as layer weights, were updated using the Adaptive Moment Estimation (Adam) optimizer to minimize the loss [81]. It combines momentum-based stochastic gradient descent and is based on the adaptive estimation of first-order and second-order moments. This is achieved by dynamically scaling the learning rates for each parameter using squared gradients, while leveraging

momentum through the moving average of the gradients [82, 83]. This can be written as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{4.2}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{4.3}$$

where m_t and v_t are the first and second moment estimates, g_t represents the gradient, and β_1 and β_2 are hyperparameters typically set to 0.9 and 0.999, respectively.

Using these moment estimates, Adam updates weights according to:

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{4.4}$$

Here, w is the weight, η is the learning rate, and ϵ is a small constant (e.g., 10^{-8}) added to prevent division by zero.

To sum up, according to Kingma et al., 2014, Adam is "computationally efficient, has little memory requirement, invariant to diagonal rescaling of gradients, and is well suited for problems that are large in terms of data/parameters" [81].

The dataset was partitioned into train/validation/test splits with proportions $[0.7,\ 0.1,\ 0.2]$. Splits were performed on a patient-wise basis to prevent information leakage: all files belonging to the same patient were assigned to the same split. Class stratification was maintained where possible to preserve class balance. In addition to the fixed split, k-fold cross-validation was used to report averaged performance and uncertainty estimates; folds were also constructed patient-wise.

To improve generalization and reduce overfitting, several strategies were employed: dropout in the dense layers, batch normalization, data augmentation (Table 6), early stopping based on validation loss, and adaptive learning-rate scheduling (ReduceLROnPlateau). Model selection was based on validation performance, while the test set remained strictly reserved for final reporting.

4.4 Grad-CAM

Grad-CAM is a widely used method for producing visual explanations of convolutional neural network (CNN) predictions. It highlights image regions that most strongly influence the model's score for a specified class by combining convolutional feature maps with the gradients of the target-class score.

The Grad-CAM pipeline proceeds as follows. First, forward activations and class scores are computed for a selected convolutional layer. Gradients of the target-class score are then backpropagated to that layer

Table 6: Summary of the planned training setup and optimization parameters.

Hyperparameter	**Value / Note**
Optimizer	Adam ($\beta_1 = 0.9,\ \beta_2 = 0.999,\ \epsilon = 10^{-8}$)
Initial learning rate	1×10^{-3} (with scheduler)
Batch size	64 (adjusted according to GPU memory)
Epochs	30
Dropout (dense layers)	0.5
Weight decay (L2)	1×10^{-5} (optional)
Loss	Categorical cross-entropy
Data split	Train / Val / Test = 0.7 / 0.1 / 0.2 (patient-wise)
Cross-validation	k-fold (patient-wise folds)

and pooled to obtain per-channel importance weights α_k^c. A weighted linear combination of the layer's feature maps is formed and passed through a ReLU to produce the class activation map (CAM).

By default the target convolutional layer is chosen automatically as the last layer that yields a four-dimensional output tensor (H, W, C). Alternative choices are supported when a different spatial resolution or localization behaviour is required. The model is configured to return both the selected convolutional feature maps and the final predictions, enabling the computation of gradients with respect to feature activations.

To sharpen saliency maps, guided masking is applied: only positive activations and positive gradients are retained prior to computing α_k^c, producing maps that are qualitatively analogous to those obtained with guided backpropagation. To enhance visualization stability, we also consider SmoothGrad-CAM: Gaussian noise $\mathcal{N}(0, \sigma^2)$ is added to the input, CAMs are computed for N noisy samples, and the maps are averaged. Typical defaults used here are $N = 10$ and σ scaled relative to the model's input preprocessing; these parameters are subject to sensitivity analysis.

Input frames follow the same preprocessing pipeline used for classification: extraction of a crop around the polyp bounding box, resizing and normalization to the model input size. Crop geometry $(x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})$ together with any resizing and padding parameters are recorded to enable exact inverse mapping. CAMs computed at model resolution are inverse-mapped to the original crop coordinates using bicubic interpolation and are masked with the ground-truth annotation prior to evaluation and visualization.

For visualization, normalized CAMs are converted to 8-bit heatmaps, colourised (for example, using

COLORMAP_VIRIDIS), resized to the original crop dimensions, and blended with the original image via alpha compositing (typical blend: 60% image + 40% heatmap). Ground-truth masks are overlaid as a semi-transparent layer (for example, red at 30% opacity) to assist qualitative assessment. For quantitative analyses, CAMs are binarized using a threshold thr (default thr $= 0.5$ on the normalized CAM); sensitivity to thr is reported.

At the frame level, we quantify the correspondence between saliency maps and annotations using the following metrics: intersection-over-union (IoU); centroid (centre-of-mass) distance normalized by the crop diagonal; sharpness measured via the variance of the Laplacian; and intensity skewness as an illumination proxy. Visual overlays, numerical summaries, and all preprocessing and inverse-mapping parameters are archived to guarantee reproducibility and to permit sensitivity analyses. Outputs are stored as paired images — (1) the heatmap overlay on the original image and (2) the original image with the ground-truth mask — and grouped by sequence to enable systematic review of frame-level explanations and their correspondence to clinical annotations.

4.5 SHAP

SHAP is employed to produce model-agnostic, spatially resolved attribution maps that indicate the contribution of input pixels to the network's output. Theoretical foundations are presented in the State of the Art. This section describes the concrete implementation used.

The SHAP pipeline implementation comprises five stages:

1. Preprocessing: load each frame, extract the annotated crop and binary mask, and record crop geometry.

2. Background construction: assemble an explainer background from annotated crops balanced by class.

3. Explainer initialization: instantiate the gradient-based explainer via `shap.GradientExplainer(model, background)`.

4. Attribution computation: compute SHAP values at the model input resolution, post-process the heatmap, inverse-map to original crop coordinates, and apply the polyp mask.

5. Metric evaluation: compute spatial and image-quality metrics and persist visual overlays and tabulated results.

Practical design choices were driven by two constraints: (1) attribution maps must be reported in original image coordinates for direct comparison with pixel-accurate annotations; and (2) computational cost must remain tractable for large-scale per-frame evaluation.

Alternative explainers are considered, including sampling-based methods (e.g., `KernelExplainer`) and tree-partition approaches. Sampling-based methods incur prohibitive runtime for dense convolutional inputs at the required scale and are therefore not adopted. Tree-partition methods are incompatible with convolutional feature hierarchies and do not produce spatially continuous attributions suited for overlap-based evaluation. Consequently, a gradient-based SHAP explainer is adopted:

`shap.GradientExplainer(model, background)`.

Gradient-based SHAP leverages model outputs and input gradients to produce dense, continuous attribution maps while remaining compatible with batched Graphics Processing Unit (GPU) inference, providing a pragmatic compromise between spatial fidelity and computational cost.

For each full-resolution frame, the annotated crop and binary mask are extracted and the crop geometry $(x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})$, together with resizing and padding parameters, is recorded. Crops are resized and normalized according to the network input specification. SHAP is computed at this resolution using the precomputed background. The resulting attribution maps are inverse-mapped to original crop coordinates using bicubic interpolation to preserve spatial coherence, and then masked with the original annotation to remove extraneous attributions.

Raw SHAP maps undergo spatial smoothing and thresholding prior to quantitative comparison with segmentation masks. A Gaussian filter with standard deviation $\sigma \in [0.8, 1.5]$ (default $\sigma = 1.0$) is applied in crop coordinates to obtain coherent attribution regions. Binarization strategies tested include percentile thresholding (retain the top p% of attribution mass, with $p \in \{10, 20, 50, 70\}$ and default $p = 10$%), absolute thresholding on normalized maps (for example 0.5), and adaptive/statistical methods such as z-score cutoffs and Otsu's method. Sensitivity to threshold choice is reported to demonstrate robustness.

For each binarized attribution map H_T the per-frame evaluation computes intersection-over-union (IoU) with the ground-truth mask, coverage percentage (fraction of H_T contained within the annotation), the normalized Euclidean distance between centroids of H_T and the ground truth, and image-quality proxies (Laplacian variance as a focus measure and intensity skewness as an illumination proxy). All visual overlays and numerical summaries are stored in standardized formats to support qualitative inspection and aggregated analysis.

To bound computational cost while providing representative coverage, a fixed sampling policy is adopted. From the test set, $N_{\text{seq}} = 150$ sequences are randomly sampled and, for each sequence,

$n_f = 100$ frames are processed, yielding approximately $N_{\text{SHAP}} = 150 \times 100 = 15{,}000$ computations. Frame selection within each sequence is either uniform in time or chosen by highest model probability; when probability-based selection is used the model is calibrated (e.g., via temperature scaling) to avoid systematic bias from overconfident predictions. The chosen configuration proves tractable with the gradient-based explainer and background size $B \approx 100$. Runtime grows approximately linearly with both the number of processed frames and the background size and therefore these parameters are balanced against the expected reduction in variance.

All implementation artifacts (explainer backgrounds, preprocessing and inverse-mapping parameters, binarization settings, visual overlays and metric tables) are retained to ensure reproducibility and to permit sensitivity analyses.

4.6 Artificial Degradations

Artificial degradations simulate common image-quality problems to evaluate model robustness and to generate corrupted variants for qualitative inspection and quantitative analysis.

Table 7 and Table 8 summarises the implemented degradations. All operators expose configurable parameters and three discrete severity levels (low, medium, high) to enable a dose–response analysis of performance versus corruption strength.

Table 7: Summary of the implemented artificial degradations and their simulation goals — Part A.

Degradation Type	**Purpose / Simulation Goal**	**Example parameters (low / medium / high)**
Global Brightening	Simulates global overexposure or uniform illumination changes affecting the entire image.	Multiplicative factor on intensity (or V channel): {1.10, 1.25, 1.50}.
Gaussian Spatial Brightening	Introduces spatially varying illumination patterns to emulate local lighting artefacts or spotlight effects.	Gaussian multiplicative mask; max gain $\in$ {1.10, 1.30, 1.50}; sigma relative to image size.
Gaussian Blur	Simulates optical defocus or mild motion blur to test resilience to spatial frequency loss.	Gaussian kernel sizes: $3 \times 3,\ 5 \times 5,\ 7 \times 7$.
Central Obstruction	Draws a black circle in the center to mimic occlusions caused by instruments, specular highlights or partial camera blockage.	Mask radius as percentage of image width: {10%, 20%, 30%}.
Contrast Adjustment (ctr)	Modifies global contrast to simulate under- or over-enhanced image appearances.	Linear contrast scaling α: {0.7 (reduced), 1.0 (neutral), 1.3 (increased)}.
Gamma Correction (gamma)	Simulates non-linear intensity mappings caused by illumination or display variations.	Gamma darken: {0.9, 0.8, 0.7}; gamma brighten: {1.1, 1.2, 1.3}.

4.7 Evaluation Metrics

The evaluation of the proposed models employs standard classification metrics together with overlap- and image-quality-based measures designed to quantify the spatial agreement between explanation maps and ground-truth polyp annotations. Classification performance is reported at the frame level using accuracy, precision, recall and F1-score. Note that confusion-matrix-based metrics can be computed at different granularities (frame-level, pixel-level). In this work, classification metrics refer to frame-level predictions; segmentation-overlap metrics (below) operate at pixel-level [32].

Spatial agreement between saliency maps (Grad-CAM, SHAP) and segmentation masks is assessed

Table 8: Summary of the implemented artificial degradations and their simulation goals — Part B.

Degradation Type	**Purpose / Simulation Goal**	**Example parameters (low / medium / high)**
Directional Shift (dir)	Applies small horizontal or vertical translations to simulate camera or scene misalignment.	Pixel shifts (px): {2, 5, 10} (or $\approx$1%, 3%, 5% of the image width).
Random Rotation	Small rotations to emulate orientation variation or probe rotation-robustness.	Rotation angles: $\{\pm 5^\circ, \pm 10^\circ, \pm 15^\circ\}$ (applied randomly).
Hue Shift (hue)	Perturbs the hue component to model color distortions or white-balance errors.	Hue shift (degrees): $\{\pm 10^\circ, \pm 20^\circ, \pm 30^\circ\}$.
Saturation Shift (sat)	Modifies color saturation to emulate sensor artefacts or compression effects.	S scale (desaturate / saturate): desaturate {0.8, 0.5, 0.3}; increase {1.2, 1.5, 1.8}.
Additive Noise (noise)	Adds Gaussian noise to replicate sensor or transmission noise, especially in low-light settings.	Gaussian noise σ (pixel intensity units): {5, 10, 30}.
Value Shift (val)	Adjusts global brightness to model underexposed or overexposed conditions.	V-channel scaling: darken {0.9, 0.8, 0.7}; brighten {1.1, 1.2, 1.3}. Histogram equalization of V applied for selected severities (e.g., medium/high).

using overlap-based metrics (Intersection over Union and Dice coefficient) as well as heatmap-specific measures (coverage and centroid distance). Finally, image-quality proxies such as intensity skewness and Laplacian variance are computed per frame and combined with the overlap metrics into a composite XAI-quality score.

4.7.1 Accuracy

The definition of accuracy refers to a model's ability to correctly distinguish between patient and healthy cases [52]. Mathematically, this can be stated as the proportion of true positive and true negative in all evaluated cases [52]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.5}$$

4.7.2 Precision

Precision is a performance metric that quantifies the proportion of correctly predicted positive instances among all instances that were predicted as positive by the model [84]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.6}$$

4.7.3 Recall

Recall can be described as the capability to determine the patient cases correctly [52]. To calculate it, we should do the proportion of true positive in patient cases [52]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.7}$$

4.7.4 F1 Score

F1 Score is a performance metric that represents the harmonic mean of precision and recall. It provides a balanced measure of a model's ability to correctly identify positive instances while accounting for both false positives and false negatives [85]:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8}$$

4.8 Quality Metrics

4.8.1 Intersection over Union

Intersection over Union (IoU), also known as the Jaccard Index, is a widely used metric for quantifying the similarity between two regions. In segmentation tasks, it measures the overlap between the predicted segmentation mask and the GT annotation [86]. Formally, it is defined as:

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{4.9}$$

The IoU takes values in the range $[0, 1]$, where 1 indicates perfect alignment between prediction and ground truth, while 0 indicates no overlap. It is strongly correlated with the Dice Coefficient, and both metrics typically lead to consistent conclusions when comparing models [87].

For binary or multi-class segmentation tasks, the mean IoU (mIoU) is commonly reported by averaging the IoU across all classes [87].

In the context of this work, IoU is adapted to evaluate the agreement between the model's saliency maps and the annotated polyp regions. Specifically, it is computed as the ratio between the area of overlap (saliency map $\cap$ polyp) and the area of their union (saliency map $\cup$ polyp), multiplied by 100 to express the result as a percentage.

4.8.2 Score (Effective Probability)

To evaluate model predictions at the frame level, we define an effective score that integrates both the network's confidence and its correctness, penalising overly confident errors. Let the network output be represented by the softmax probability vector $p = (p_1, \ldots, p_C)$ for C classes. The predicted class is given by

$$\hat{y} = \arg\max_c p_c,$$

with associated confidence

$$p_{\hat{y}} = \max_c p_c.$$

Denoting the ground-truth class by y, we introduce the indicator variable

$$\mathbb{I}_{\text{corr}} = \begin{cases} 1 & \text{if } \hat{y} = y, \\ 0 & \text{otherwise,} \end{cases}$$

which encodes whether the prediction is correct.

The effective score S is then defined as

$$S = \begin{cases} p_{\hat{y}}, & \text{if } \hat{y} = y, \\ 1 - p_{\hat{y}}, & \text{if } \hat{y} \neq y, \end{cases} \tag{4.10}$$

with a compact equivalent formulation

$$S = 1 - \left| \mathbb{I}_{\text{corr}} - p_{\hat{y}} \right|. \tag{4.11}$$

By construction, $S \in [0, 1]$. High values correspond to desirable situations, namely a correct prediction made with high confidence or an incorrect prediction made with low confidence (i.e., model uncertainty about an error). Conversely, low values occur when the model is confidently wrong, which is the

least favourable outcome. Intermediate values naturally reflect uncertain decisions, regardless of correctness. In this sense, S serves as an "effective probability" that more accurately reflects the quality of a prediction than raw confidence values alone.

4.8.3 Skewness of the intensity distribution

To quantify exposure bias within a crop (or within the annotated polyp region) we compute the skewness of the grayscale intensity distribution. Let $I = \{I_i\}_{i=1}^{N}$ denote the intensity values of the N pixels considered. Denote by $\bar{I}$ the sample mean and by σ the sample standard deviation:

$$\bar{I} = \frac{1}{N}\sum_{i=1}^{N} I_i, \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_i - \bar{I})^2}.$$

The (standardised) sample skewness is defined as the third standardized moment

$$\text{Skewness} \;=\; \frac{\frac{1}{N}\sum_{i=1}^{N}(I_i - \bar{I})^3}{\sigma^3}. \tag{4.12}$$

By construction, skewness is a dimensionless quantity: values close to zero indicate an approximately symmetric intensity distribution, positive values indicate a right (bright) tail (relative overexposure), and negative values indicate a left (dark) tail (relative underexposure). As a practical rule of thumb, $|\text{Skewness}| \lesssim 0.5$ denotes mild asymmetry, values around 1 indicate substantial skew and values $\gtrsim 2$ indicate a strongly skewed distribution; such strong skewness typically signals frames of poor visual quality that may impair model performance.

4.8.4 Laplacian variance (image sharpness)

Image sharpness is commonly estimated using the variance of the Laplacian operator applied to a grayscale image crop. Let $I = \{I_i\}_{i=1}^{N}$ denote the grayscale intensities of the N pixels considered (we recommend normalising intensities to $[0, 1]$ prior to computation). The discrete Laplacian ΔI can be obtained by convolving I with a standard 3×3 Laplacian kernel, for example

$$K_\Delta = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \Delta I = I * K_\Delta.$$

The Laplacian variance is then defined as the sample variance of the filtered image:

$$\text{LaplVar} \;=\; \frac{1}{N}\sum_{i=1}^{N}\left(\Delta I_i - \overline{\Delta I}\right)^2, \tag{4.13}$$

where $\overline{\Delta I}$ denotes the mean of ΔI over the N pixels.

Intuitively, LaplVar measures the presence of high-frequency content (edges and fine detail): sharp images exhibit stronger Laplacian responses and thus larger variance, whereas blurred images produce attenuated Laplacian responses and smaller variance. Consequently, lower values of LaplVar indicate reduced sharpness (blurrier frames).

4.8.5 Centroid distance (spatial alignment)

The centroid distance measures how closely the model's saliency is located to the annotated polyp. Compute the saliency centroid (the intensity-weighted centre of mass of the heatmap) and the mask centroid (the average position of the pixels in the GT annotation). These two points summarise where the explanation and the annotation are concentrated.

We express their separation as a scale-invariant quantity:

$$d_{\text{norm}} = \frac{\|c_S - c_M\|_2}{\sqrt{w^2 + h^2}},$$

where c_S is the saliency centroid, c_M is the mask centroid, and w, h are the crop width and height. Normalizing by the image diagonal makes the measure comparable across different image sizes.

4.8.6 Combined Metrics

The aim of the combined metric is to integrate multiple per-frame indicators—IoU, Laplacian variance, skewness, and distance-to-center—into a single, interpretable score that reflects overall image or prediction quality. This composite index facilitates filtering, quality control, and real-time monitoring by condensing multidimensional signals into a single continuous value that correlates with a reference measure.

Several methods were explored to construct the composite score, varying in complexity, interpretability, and ability to capture nonlinear relationships.

- Linear model: A linear combination of standardized metrics was fitted to predict the reference quality score:

  $$S_{\text{linear}} = \beta_0 + \sum_i \beta_i X_i$$

  where X_i denotes each individual metric. The coefficients β_i provide direct interpretability, indicating the relative contribution of each metric. But restricted to linear relationships, which may limit predictive accuracy.

- Non-linear supervised pipeline: A polynomial feature expansion (degree 2) was combined with standard scaling and a Random Forest Regressor:

$$X \xrightarrow{\text{PolynomialFeatures(2)}} X_{\text{poly}} \xrightarrow{\text{StandardScaler}} X_{\text{scaled}} \xrightarrow{\text{RandomForestRegressor}} S_{\text{RF}}$$

  Hyperparameters included $n_{\text{estimators}} = 100$ and candidate max_depth values tuned via cross-validation. This model captures interactions and simple nonlinear effects, improving predictive performance at the cost of reduced interpretability.

- Blended score: A weighted average of the linear and Random Forest outputs was computed:

$$S_{\text{blended}} = w \cdot S_{\text{linear}} + (1 - w) \cdot S_{\text{RF}}$$

  where the weight w was determined based on Leave-One-Subject-Out (LOSO) correlations with the reference quality. This approach balances interpretability and predictive accuracy.

4.9 Software and Hardware Setup

All experiments and implementations were developed in Python and executed on a dedicated compute server. The interactive development environment was Jupyter Notebook (accessed remotely). The computing environment and principal software components used for the project are summarized below to provide context for the computational resources required and to support reproducibility in Table 9. Figure 23 illustrates the development environment and remote-access workflow.

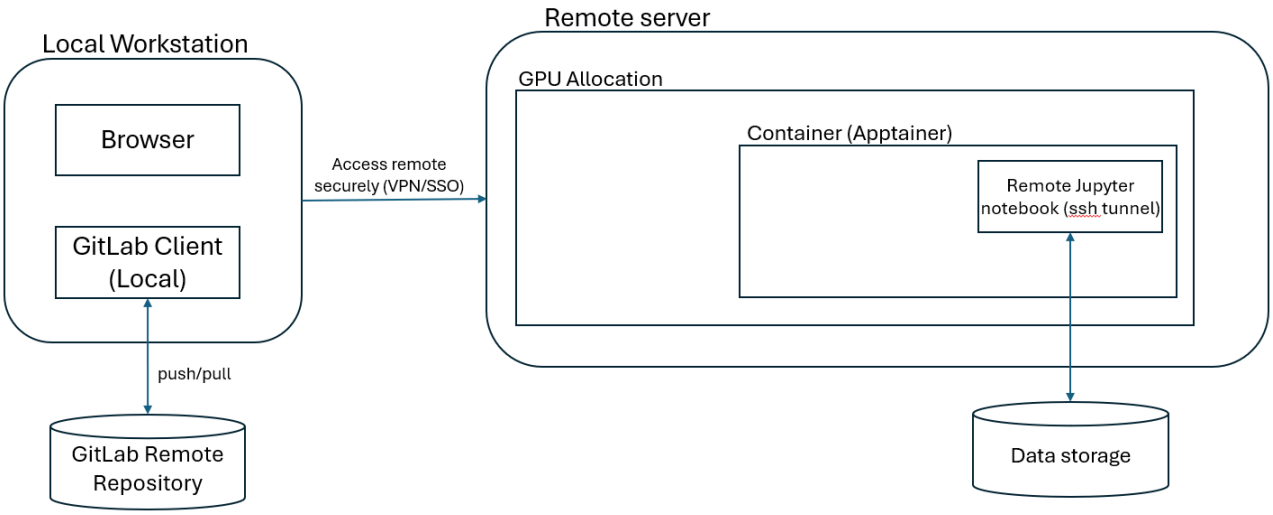


Figure 23: Development environment setup.

This scheme illustrates the remote development and execution workflow used throughout this project. Apptainer (a maintained fork of Singularity) was used to provide an isolated, reproducible runtime on

Table 9: Hardware and software summary for the experimental environment.

Operating system	Ubuntu 22.04 LTS
CPU	2 × Intel Xeon Processor E5-2690 v4 (14 cores each, 28 cores total; 2.60 GHz, 35 MB cache)
RAM	640 GB DDR4 Registered ECC (configured as 8 × 16 GB + 16 × 32 GB modules)
GPU	8 × NVIDIA GeForce GTX 1080 Ti (11 GB GDDR5X per card) — available on the server; experiments used a single GPU for training
Storage	4 × Samsung SSD PM863a 480 GB (SATA)
Programming language	Python
Interactive environment	Jupyter Notebook
Deep learning framework	TensorFlow with Keras API
Version control	Git (repository hosted on GitLab)
Containerization	Apptainer (Singularity-compatible)

the HPC server. Containers allow packaging the exact software stack (Python, libraries, drivers) without requiring a full virtual machine image, and are therefore well suited for cluster environments. The development workflow used a remote Jupyter server (accessed via secure SSH tunnelling or a VNC session when required). Container images were built from Docker images where necessary to simplify dependency management.

Chapter 5

Experiments and Results

This chapter reports the outcomes of the experimental work carried out in this dissertation.

This work aims to enhance the explainability of that classification model by incorporating XAI techniques, making it more transparent and understandable for end-users, especially in clinical settings.

5.1 Architecture Choice and Training Configuration

All models were implemented in TensorFlow using the Keras API. Some experiments used cropped polyp regions resized to 256×256 pixels. The classification task is binary (labels: 0, 1). The primary training configuration used throughout the experiments is summarised in Table 10.

The default loss was categorical cross-entropy with label smoothing (smoothing factor $= 0.4$). We also evaluated alternative losses — focal loss and a combined Dice + cross-entropy loss — to assess robustness to class imbalance and to encourage localisation-aware training. Label smoothing was applied as a regulariser to reduce overconfidence and to stabilise training.

To mitigate class imbalance and reduce sampling bias, minibatches were formed using balanced sampling at the patient level. Data augmentation (see Table 5) included geometric and photometric transformations (rotation, flipping, translation, zoom, brightness variation, shear and blur) to emulate acquisition variability and improve generalisation. Models were trained for up to 30 epochs with early stopping based on validation loss. The Adam optimizer was used with an initial learning rate of 1×10^{-3} and a ReduceLROnPlateau scheduler to decrease the learning rate upon plateaus in validation performance. Training artifacts and the best checkpoints were stored under `data/snapshots/all/` for reproducibility. TensorBoard was used for monitoring and debugging.

Experiments evaluated several convolutional backbones pre-trained on ImageNet, including ResNet variants, VGG16 and the EfficientNet family. These architectures were selected due to their strong feature-extraction capabilities and their widespread use in medical imaging tasks. During training, the pre-trained

Table 10: Summary of training configuration and hyperparameter choices.

Hyperparameter	**Value**
Batch Size	64
Number of classes	2 (labels: 0, 1)
Number of Epochs	30 (early stopping applied)
Initial learning rate	1×10^{-3}
Optimizer	Adam (with learning-rate scheduler)
Default loss	Categorical cross-entropy (label smoothing = 0.4)
Alternative losses	Focal Loss, Dice + Cross-Entropy
Regularisation	Label smoothing (0.4); no dropout applied to conv/fc layers
Data balancing	Balanced sampling during training (patient-wise)
Augmentations	Rotations, flips, zoom, brightness, shear, translation, blur (see Table 5)
Callbacks	ModelCheckpoint, ReduceLROnPlateau, EarlyStopping, TensorBoard
Evaluation metrics	Categorical accuracy (frame-level); precision, recall, F1 from confusion matrix

weights were used as initialization, followed by fine-tuning on the cropped polyp regions.

The first model evaluated was a modified ResNetV2 backbone. Instead of processing full images, each input image was cropped into five 224×224 patches: four corners (top-left, top-right, bottom-left, bottom-right) and the center of the image. These crops were preprocessed with the ResNetV2-specific preprocessing function and passed independently through the ImageNet-pretrained ResNet. Each crop was processed individually to produce a fixed-length feature vector (approximately 2048 elements for ResNet variants). The resulting five feature vectors were aggregated by calculating their elementwise average and elementwise sum, and the two vectors were concatenated to form a combined vector of size approximately 4096. This combined feature vector was then passed through a compact fully-connected head consisting of three dense layers with batch normalization and optional dropout. The model outputted a softmax distribution over the two classes. During training, the ResNet backbone was frozen, and only the top classifier layers were fine-tuned, which reduced training time but limited the adaptation of feature representations.

Following the evaluation of the modified ResNetV2, the same model was tested using cropped images instead of full-size inputs. This setup involved training the ResNet backbone using the same five 224×224 patches extracted from the polyp regions. After that, further experiments involved training on cropped polyp regions with several other backbones, including VGG16 and EfficientNet. While these models showed some improvements in performance, their results were still not satisfactory.

Table 11 summarises the comparative performance of several tested models.

Table 11: Performance comparison of the tested models in terms of accuracy, precision, recall, and F1-score.

Metric	ResNet (1080p)	ResNet (crops)	VGG16 (crops)	EfficientNet (crops)
Accuracy (%)	69.61	71.16	61.45	68.44
Precision (%)	80.57	80.95	83.76	94.94
Recall (%)	80.50	82.47	62.81	65.55
F1 Score (%)	80.54	81.17	71.79	77.59

EfficientNet (crops) was chosen for the remainder of the project due to its high precision (reducing false positives), compact model size and inference efficiency, which together make it well-suited for generating and inspecting per-frame explanations in a clinical context.

5.2 Grad-Cam Implementation

The Grad-Cam pipeline used in the experiments follows the procedure described in the Methodology sec- tion. The concrete implementation details that affect reproducibility are summarized below.

The final convolutional layer is identified automatically as the default target for Grad-CAM computation. A modified inference model is constructed that returns both the feature maps of the target convolutional layer and the final class logits in a single forward pass to minimise redundant computation.

For each evaluated sequence a representative subset of frames is sampled. Each selected frame undergoes preprocessing steps such as extracting the polyp crop according to the annotation bounding box, resize the crop to the model input size preserving aspect ratio (padding where necessary), and record the crop geometry and padding parameters to enable exact inverse mapping back to original image coordinates and normalise pixel intensities to the same scale used for training. These geometry records are saved and later used to resample and recompose the Grad-CAM heatmap precisely into original-frame coordinates.

For every preprocessed frame a Grad-CAM heatmap is computed and implemented, optionally, SmoothGrad-CAM, average Grad-CAM maps computed over N noisy perturbations of the input, and Guided Masking, element-wise multiplication of the Grad-CAM heatmap with guided-backpropagation saliency maps, to increase stability.

Computed CAMs are clipped and min–max normalised to $[0, 1]$, then converted to 8-bit $[0, 255]$, colourised with a perceptually uniform colormap (e.g. `VIRIDIS`), resized to the original crop dimensions

using bicubic interpolation, and blended with the original frame.

To enable pixel-level comparison with annotations, binary masks are derived from the normalised CAMs. The same set of thresholding strategies used for SHAP were evaluated (percentile thresholds, absolute thresholding, z-score cutoff, and Otsu).

For each evaluated frame the Grad-CAM pipeline writes a CSV row with the following header (values recorded per evaluated threshold when applicable): sequence, image_name, predicted_class, prediction_prob, ground_truth, laplacian_var, hist_skewness, com_z_x, com_z_y, com_gt_x, com_gt_y, iou.

To support qualitative inspection, the visual overlays are also saved as side-by-side images grouped by sequence (original frame with Grad-CAM overlay and original frame with GT segmentation). A few examples of the overlays are shown in Figure 24.

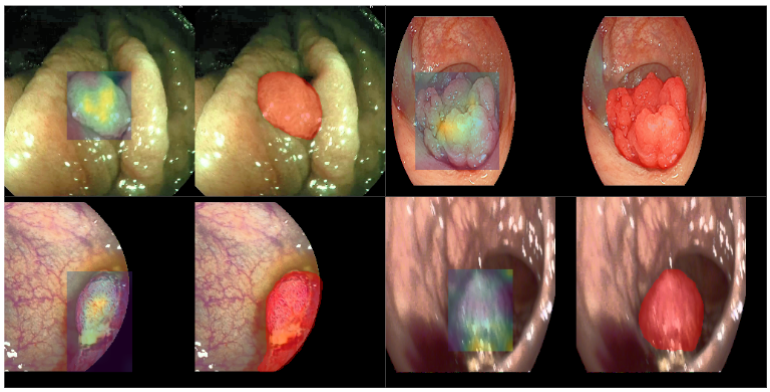


Figure 24: Example of the Grad-CAM overlay for a polyp-containing frame. The left panel of each pair of images shows the heatmap blended with the original image; the right panel shows the ground-truth segmentation mask.

5.3 SHAP implementation

This section documents the implementation details of the SHAP pipeline.

The SHAP explainer was constructed once per experimental run and immediately serialized to disk (stored as a `.pkl` file) to avoid repeated initialization and to ensure reproducible behavior across reruns. The explainer background set (used to compute SHAP baselines) consists of approximately $B \approx 100$ annotated crops, balanced across classes. To bound computational cost while maintaining representative coverage of the dataset, the attribution computations were performed on $N_{\text{seq}} = 150$ sequences with $n_f = 100$ frames per sequence, yielding approximately $150 \times 100 \approx 15{,}000$ individual attributions maps. For each preprocessed input frame the pipeline computes a SHAP attribution map for the predicted class. To reduce repeated overhead, the explainer is instantiated once and reused to compute attributions for all frames in a batch; results are flushed to disk incrementally to limit in-memory footprint.

To improve spatial coherence and reduce high-frequency attribution noise, each raw SHAP map was convolved with a Gaussian kernel with standard deviation $\sigma = 1.0$. This value was selected following a brief pilot sensitivity analysis that balanced noise reduction against preservation of spatial detail: it produced visibly smoother attributions without excessive boundary blurring. After smoothing, attribution maps were min–max normalised to make thresholding comparable across images. We evaluated several binarization strategies: percentile thresholds (e.g. $p = 50\%$ and $p = 70\%$), an absolute threshold on normalized maps ($t = 0.5$), a z-score cutoff ($z = 1.96$), and an adaptive thresholding method (Otsu's algorithm).

For each attribution (frame) the pipeline records a structured row in a CSV file to enable aggregated quantitative analysis. The CSV schema written for SHAP experiments is: sequence, image_name, predicted_class, prediction_prob, ground_truth, iou_positive, iou_negative, laplacian_var, skewness, com_pos_x, com_pos_y, com_neg_x, com_neg_y.

In addition to the quantitatively analysis, the pipeline computes all the visual SHAP maps. An example of the SHAP overlays produced by this pipeline is shown in Figure 25.

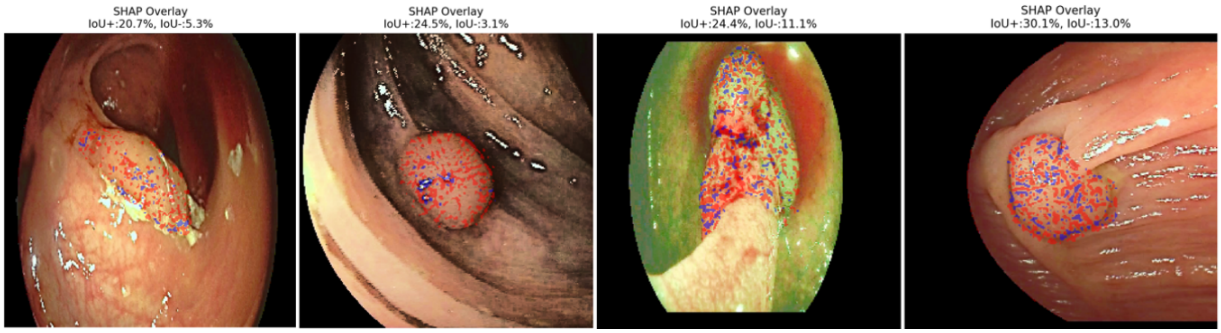


Figure 25: Example SHAP attribution overlays on polyp images.

5.4 Preliminary Experiment Implementation

The preliminary experiment uses a pretrained EfficientNet-B0 backbone from Keras Applications as the feature extractor. The backbone weights were initialised from ImageNet pretraining and all layers were frozen except for the final convolutional blocks and a task-specific classification head. The classification head was appended to the model architecture to produce a binary output, distinguishing between two classes: dog and cat. This head consisted of a global pooling layer followed by a dense block and a two-unit softmax output.

The dataset comprises labelled images of cats and dogs and Figure 26 presents examples from the

dataset for each class, illustrating the types of images fed into the model for training and testing.



Figure 26: Dataset examples of cats and dogs.

Model performance was evaluated on the hold-out test set using standard metrics and the results are summarized in Table 12. The overall accuracy for the test set was 98%.

Table 12: Classification Report.

Class	**Precision**	**Recall**	**F1-score**
Cat	0.94	0.97	0.95
Dog	0.99	0.98	0.99
Accuracy			0.98

Grad-CAM saliency maps were computed for the test-set predictions to visualize the regions of the input images that influenced the classifier predictions. Example overlays are presented in Figure 27, where the heatmaps highlight the areas of the images that were most relevant to the model's decision-making process.



Figure 27: Example of Grad-CAM for cats and dogs.

In addition, a comprehensive analysis of a set of quantitative metrics, as previously described, was conducted across all sequences and frames within the dataset. The Pearson correlation matrix for these metrics is reported in Table 13.

Table 13: Correlation Matrix

	IoU	Laplacian Variance	Skewness	Distance Center	Score
IoU	1.000	-0.047	0.202	-0.363	-0.514
Laplacian Variance	-0.047	1.000	0.042	-0.080	0.020
Skewness	0.202	0.042	1.000	-0.041	-0.005
Distance Center	-0.363	-0.080	-0.041	1.000	0.139
Score	-0.514	0.020	-0.004	0.139	1.000

The correlation analysis uncovers several relationships, particularly:

- IoU and model score ($r \approx -0.51$): A moderate negative correlation is observed between IoU and the model's confidence score, indicating an inverse association and relationship. In other words, higher model confidence tended to coincide with lower spatial overlap between the Grad-CAM activation region and the manual segmentation. This pattern may suggest that the classifier is, in some cases, relying on focused areas and resulting in lower IoU.

- IoU and distance-to-center ($r \approx -0.36$): A moderate negative correlation indicates reduced segmentation performance for objects located further from the image centre.

- IoU and skewness ($r \approx 0.20$): A weak positive correlation suggests a slight tendency for frames with greater pixel-distribution asymmetry to exhibit higher IoU.

Per-category Pearson correlation analyses were conducted to examine the relationships between IoU and the score metrics, as well as between IoU and the distance to the center metrics. The most relevant results are presented in Table 14.

To further interpret these results, the visual outputs of two representative cases with contrasting IoU–Score relationships are examined.

In the Beagle examples (Figure 28), Grad-CAM activations are narrowly concentrated on a small set of facial features (e.g., muzzle and eyes), producing relatively low spatial overlap with the manual segmentation while the classifier remains confident (IoU vs Score = -0.492). By contrast, in the Russian Blue examples (Figure 29) the saliency map covers the full extent of the animal, yielding higher overlap with the annotation and a positive IoU–Score association (IoU vs Score = 0.452).



Figure 28: Example images of the Beagle breed: IoU vs Score = -0.492 and IoU vs distance to center = -0.151.



Figure 29: Example images of the Russian Blue breed: IoU vs Score = 0.452 and IoU vs distance to center = -0.677.

Table 14: Per-breed Pearson correlations: IoU vs Score and IoU vs Distance to Center

Breed	**IoU vs Score**	**IoU vs Distance Center**
Beagle	-0.492	-0.151
Bengal	0.120	-0.637
Bombay	-0.313	-0.558
Chihuahua	-0.387	-0.463
Great	0.320	-0.670
Japanese	-0.492	-0.145
Keeshond	0.618	-0.295
Leonberger	0.410	-0.615
Persian	-0.351	-0.305
Pomeranian	0.633	-0.605
Russian	0.452	-0.677

To explore robustness, a set of artificial degradations, described before, were applied. The original image and the set of artificially degraded variants are shown in Figure 30. Table 15 summarises the predicted classification score and IoU for the *Abyssinian_101* example under varying degradation types and intensities.

Table 15: Abyssinian_101 — Predicted score and IoU across artificial degradation variants

Degradation	**Variant**	**Score**	**IoU**
Original		0.48	0.180
Blur	Kernel Size = 5	0.520	0.200
	Kernel Size = 10	0.503	0.100
	Kernel Size = 15	0.485	0.250
Directional Shift	Left = 1.5 pixels	0.585	0.250
	Left = 2 pixels	0.540	0.270
	Left = 3 pixels	0.520	0.400
	Right = 1.5 pixels	0.565	0.175
	Right = 2 pixels	0.544	0.210
	Right = 3 pixels	0.543	0.300
Saturation Shift	Factor = 0.2	0.555	0.300
	Factor = 2.0	0.540	0.600
	Factor = 3.0	0.520	0.900
Central Obstruction	Radius = 0.3	0.430	0.180
	Radius = 0.7	0.450	0.150
	Radius = 1.3	0.480	0.120
	Radius = 2.0	0.500	0.220
Hue Shift	Factor = -40	0.460	0.350
	Factor = -80	0.350	0.360
	Factor = 40	0.578	0.210
	Factor = 80	0.510	0.240
Value Shift	Factor = 0.5	0.520	0.180
	Factor = 2.0	0.530	0.220
	Factor = 3.0	0.510	0.170

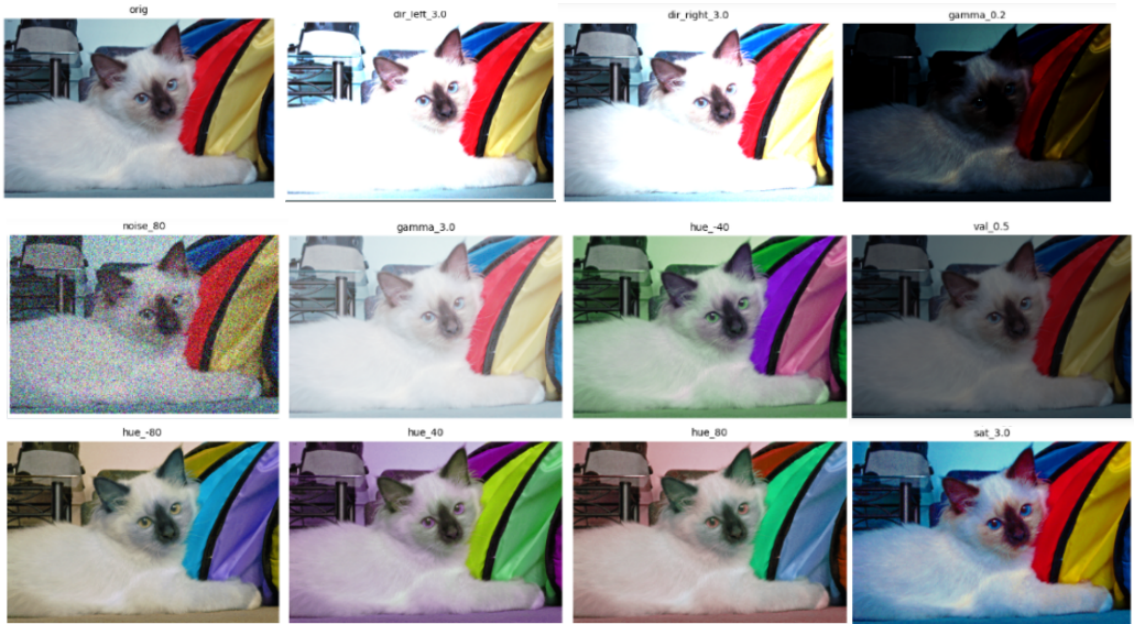


Figure 30: Original image (top-left) and 11 artificially degraded variants, including directional shifts, gamma corrections, noise, hue shifts, value shift and saturation shift applied at different intensities.

The targeted degradation study for Abyssinian_101 indicates that artificial perturbations have qualitatively distinct effects on classification confidence (score) and localization quality (IoU). As a starting point, the original (undistorted) image yields the highest IoU and thus defines the reference performance level. Broadly, the degradations split into two behavioural families. First, perturbations that remove high-frequency information—most notably Gaussian blur and extreme desaturation—produce severe reductions in IoU while causing only moderate declines in model score; for blur, IoU falls monotonically with blur intensity, indicating a loss of spatial accuracy as edge and texture cues are degraded. Second, color and geometric perturbations (hue shifts, translations/centering) produce mixed outcomes: in some cases they reduce score while leaving IoU largely unchanged, and in others they reduce score but actually increase IoU (for example, larger translations that improve object framing can raise overlap despite lowering confidence). Taken together, these patterns suggest that the model's confidence is not a uniformly reliable proxy for localization under perturbation: classification and localization appear to respond to different image cues and can become decoupled under certain degradations.

5.5 Main Experiment Implementation

This section presents the results of the main experiment using the cropped polyp dataset generated by the preprocessing pipeline outlined in the previous section.

The cropped images, the corresponding segmentation annotations, and per-frame metadata were

organized in a structured directory hierarchy. An example of a cropped image and its corresponding segmentation mask is shown in Figure 31.

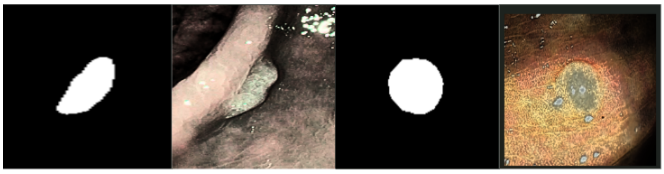


Figure 31: Example of a polyp annotation and its corresponding cropped image.

SHAP values were computed for the test dataset using the explainer configuration and parameters described in Section 5.3.

Qualitatively, negative SHAP contributions (indicated by blue regions) often coincide with image defects such as specular highlights, peripheral regions, and areas of uneven illumination, as well as regions exhibiting significant degradation or occlusion. Positive SHAP contributions (indicated by red regions) show more variability across frames: in some cases, they concentrate along lesion borders, in others, they are dispersed across raised central areas, and in a subset of frames, they co-localize with specular highlights or prominent mucosal folds. These observations align with the hypothesis that the network relies on a combination of anatomical features and image-level artifacts (lighting, texture) when making high-confidence predictions.

Table 16 reports Pearson correlations between SHAP-derived summary metrics and standard image/segmentation metrics.

Table 16: Correlation matrix of SHAP implementation

	IoU Positive	**IoU Negative**	**Laplacian Variance**	**Skewness**	**Score**
IoU Positive	1.000	0.241	−0.141	−0.182	−0.159
IoU Negative	0.241	1.000	0.152	−0.232	−0.190
Laplacian Variance	−0.141	0.152	1.000	−0.043	−0.083
Skewness	−0.182	−0.232	−0.043	1.000	0.497
Score	−0.159	−0.190	−0.083	0.497	1.000

Some correlations from Table 16 are particularly notable and inform interpretation:

- Score vs. Skewness (r = 0.497). A moderate positive correlation suggests that frames with more positively skewed intensity distributions tend to yield higher model confidence.

- Score vs. IoU Positive (r = −0.159) and Score vs. IoU Negative (r = −0.190). These weak negative correlations indicate that higher model confidence is slightly associated with a lower overlap between SHAP and ground truth regions.

SHAP, while providing rich and detailed pixel-level explanations, proved to be computationally more expensive than gradient-based saliency methods such as Grad-CAM. SHAP explanations require multiple evaluations of the model per image, which scales with the number of background samples and sampling rounds. In contrast, Grad-CAM only requires a single forward and backward pass, making it more efficient for practical purposes. Furthermore, SHAP generates separate attribution masks for positive and negative contributions. While this decomposition is useful for detailed pixel-level analysis, it complicates the creation and interpretation of an interpretable localization method. In some instances, positive and negative attributions can overlap with the ground truth region, which leads to ambiguities in the aggregate statistics.

Given these challenges, subsequent experiments prioritized Grad-CAM as the preferred method for coarse localization and per-frame quality metrics on our polyp dataset.

Consequently, Grad-CAM maps were generated from the final convolutional block of the network, and the resulting heatmaps were visually inspected. In most frames, the highest-importance Grad-CAM pixels were concentrated near the center of the polyp.

Table 17 presents Pearson correlations computed from Grad-CAM summaries.

Table 17: Correlation matrix of Grad-Cam implementation

	IoU	**Laplacian Variance**	**Skewness**	**Distance Center**	**Score**
IoU	1.00	0.12	0.12	−0.31	0.16
Laplacian Variance	0.12	1.00	0.04	−0.04	0.13
Skewness	0.12	0.04	1.00	0.09	−0.02
Distance Center	−0.31	−0.04	0.09	1.00	−0.34
Score	0.16	0.13	−0.02	−0.34	1.00

The IoU shows a moderate negative association with the activation centroid distance to the crop centre (Distance Center; $r = -0.31$), indicating that localisations are, on average, better aligned with the segmentation when activations are central within the crop. This is expected, as images experience less distortion, particularly from 'fish-eye' effects from the camera. The association between Grad-CAM IoU and the model confidence score is weak ($r = 0.16$), suggesting that high model confidence does not

consistently correlate with accurate localization as measured by Grad-CAM.

A set of controlled synthetic degradations was implemented as previously described. Representative degraded frames are shown in Figure 32 and Grad-CAM overlays in Figure 33.

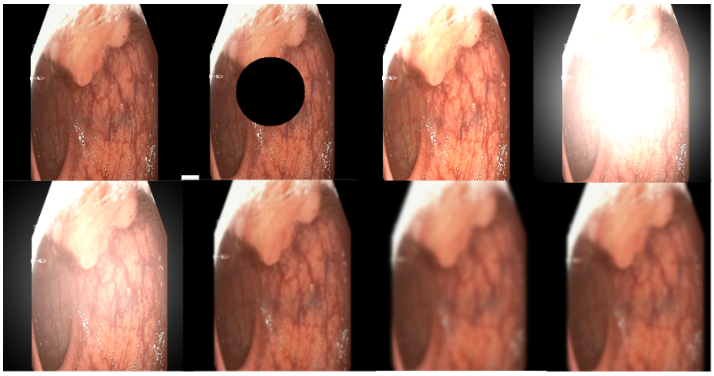


Figure 32: Original image (top-left) and seven artificially degraded variants, which include obstruction, two levels of Bright Gaussian noise, Bright Global noise, and three degrees of blur.

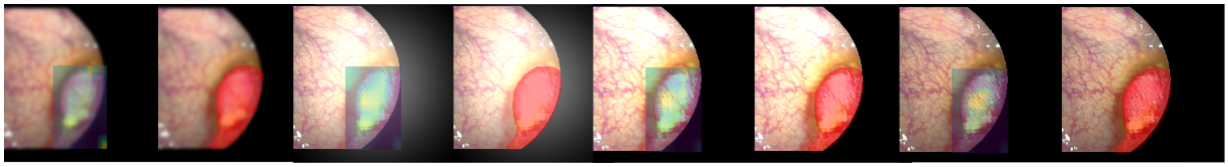


Figure 33: Application of Grad-CAM to some of the degradations and to the original image (on the right).

Table 18 reports per-sequence correlations between IoU and the model's confidence score as blur is progressively increased. The results indicate substantial heterogeneity in the correlation patterns across sequences. Some sequences maintain a positive IoU–score relationship despite the perturbation, while others exhibit sign reversals or significant fluctuations as the blur intensity increases. These findings suggest that the relationship between model confidence and IoU is highly sensitive to sequence-level visual conditions, such as polyp size, image contrast, illumination, and the presence of imaging artifacts. Consequently, identical blur perturbations do not affect all images uniformly. Specifically, certain cases, such as small polyps, noisy backgrounds, or irregular shapes, are considerably more sensitive to degradation.

Notably, a few sequences show early degradation of this relationship even at lower blur levels. This highlights the non-linear and context-dependent nature of the impact of image distortions on model performance.

Table 18: Correlation matrix of IoU–score correlation for each sequence and degradation level

Sequence	**Original**	**Blur 5**	**Blur 10**	**Blur 15**
p028_08_R_LCI	0.38	0.20	0.35	0.22
p046_19_R_WLI	-0.30	-0.40	-0.40	-0.48
p228_01_R_LCI	-0.45	-0.18	0.50	0.00
p295_02_R_BLI	0.48	0.22	0.62	0.10
p364_04	-0.28	-0.15	0.06	-0.45
p415_01	0.25	0.60	0.65	0.08
p451_05	-0.20	-0.22	0.05	-0.55
p525_08	0.15	0.58	0.78	0.72
p525_13	0.20	0.25	0.22	0.18
p526_03	-0.35	0.45	0.28	-0.10
p584_07	-0.42	-0.15	0.12	0.05
p730_02_01	0.48	0.28	-0.03	-0.10
p871_02_01	-0.20	-0.45	-0.62	-0.42

To illustrate these behaviors, we examine three representative sequences with contrasting correlation trajectories. The sequence p046_19_R_WLI consistently shows negative correlations across all blur levels (original: −0.30; blur 5: −0.40; blur 10: −0.40; blur 15: −0.48). Conversely, p295_02_R_BLI shows predominantly positive correlations (original: 0.48; blur 5: 0.22; blur 10: 0.62; blur 15: 0.10) and, finally, the sequence p526_03 exhibits mixed behavior (original: −0.35; blur 5: 0.45; blur 10: 0.28; blur 15: −0.10), with the sign of the IoU–score correlation reversing as the image is progressively blurred. The corresponding Grad-CAM heatmaps for these sequences at various blur levels are shown in Figures 34, 35, and 36.

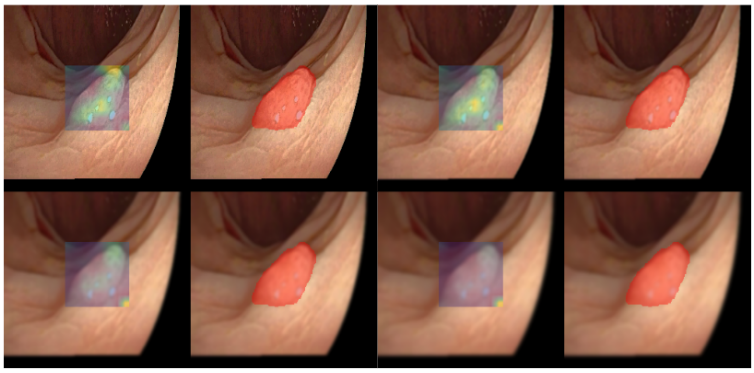


Figure 34: Application of Grad-CAM to a frame from the sequence p046_19_R_WLI. The top-left image shows the original, with blurred variants on the top-right (blur factor 5), bottom-left (blur factor 10), and bottom-right (blur factor 15).

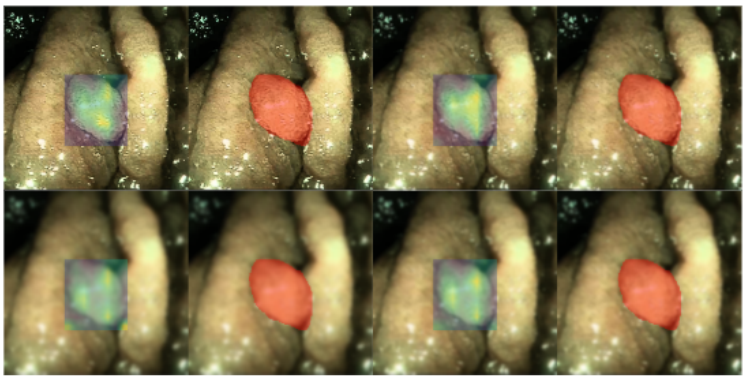


Figure 35: Application of Grad-CAM to a frame from the sequence p295_02_R_BLI. The top-left image shows the original, with blurred variants on the top-right (blur factor 5), bottom-left (blur factor 10), and bottom-right (blur factor 15).

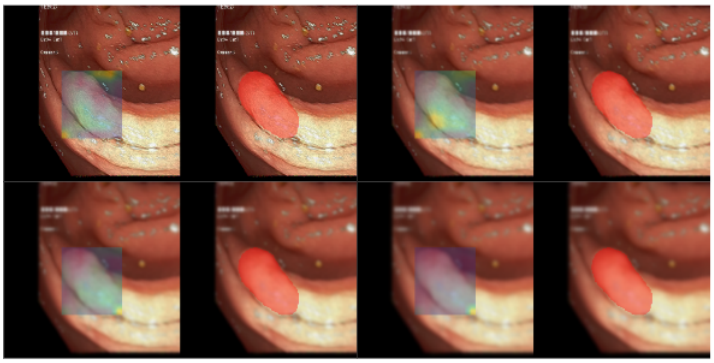


Figure 36: Application of Grad-CAM to a frame from the sequence p526_03. The top-left image shows the original, with blurred variants on the top-right (blur factor 5), bottom-left (blur factor 10), and bottom-right (blur factor 15).

For sequence p046_19_R_WLI the relationship between localisation quality and model confidence is consistently negative across blur levels and in the frame represented, all four examined cases the model produced an incorrect class label. Visual inspection of Grad-CAM maps shows that the original image's high-activation regions overlap substantially with the manual polyp mask, whereas increasing blur produces dispersion of attribution mass and emerging peaks outside the polyp, coinciding with falling IoU.

For the second sequence, the four selected frames are correctly classified (so score = probability), and the correlation is positive overall but non-linear across blur levels. Visual inspection of Grad-CAM maps shows that the dominant attribution mass remains centred on the polyp at all blur levels and as blur increases only minor, sparse peaks appear outside the annotated mask, and even at Blur 15 the principal activations still focus on the polyp.

For the third sequence, only the original frame was classified correctly. Visual inspection clarifies that the original image shows strong activations inside the polyp but with additional peaks outside it, at Blur 5 the internal activation peak strengthens and becomes more centred on the polyp, whereas at Blur 10 and Blur 15 activations are sparse and mainly off-object, matching the degradation in IoU and unstable confidence

In summary, the per-sequence analysis reveals substantial heterogeneity in how confidence relates to IoU when images are degraded.

The final step of the experiment and in an attempt to obtain a single, interpretable indicator of per-frame explanation/quality, we explored several strategies to combine multiple image- and attribution-derived descriptors into a scalar composed metric. The candidate features were

$$\mathbf{f} = \{\text{IoU},\ \text{LaplacianVar},\ \text{Skewness},\ \text{Dist}_{\text{center}}\},$$

computed on each frame. All features were normalised to ensures coefficients are comparable and to

stabilise non-linear learners.

Firstly, we fitted a linear model to the normalised features. Let $\hat{\mathbf{f}}$ denote the vector of normalised features, the linear predictor is

$$S_{\text{lin}} = \beta_0 + \sum_i \beta_i \hat{f}_i. \tag{5.1}$$

Fitting Equation (5.1) to the available reference data produced the following coefficients:

- Intercept: $\beta_0 = 0.6435$,
- $\beta_{\text{IoU}} = 0.0351$,
- $\beta_{\text{LaplacianVar}} = 0.0082$,
- $\beta_{\text{Skewness}} = 0.0329$,
- $\beta_{\text{Dist_center}} = 0.0043$.

The Pearson correlation between S_{lin} and the reference score on the validation set was $r \approx 0.3439$. This modest value points to limited predictive power from these four features under a purely linear model.

As a simple unsupervised alternative we applied principal component analysis (PCA) to the normalised feature matrix and used the first principal component (PC1) as a composite score:

$$S_{\text{PCA}} = \mathbf{u}_1^\top \hat{\mathbf{f}},$$

where $\mathbf{u}_1$ is the first principal eigenvector. Since PCA maximises variance explained in feature space rather than correlation with an external target, a lower supervised correlation is expected, $r \approx 0.2433$, lower than the supervised linear model.

To capture non-linear effects and pairwise interactions, we implemented a regression pipeline composed of polynomial feature expansion, standardisation and a tree-based regressor:

$$\text{Pipeline} = \Big(\text{PolynomialFeatures}(\text{deg} = 2) \rightarrow \text{StandardScaler} \rightarrow \text{RandomForestRegressor}\Big).$$

The key implementation details are as follows: a degree-two polynomial expansion was applied to the four base features to capture simple nonlinear effects (pairwise interactions and squared terms); the expanded feature set was standardised prior to regression to stabilise fitting and to enable comparable feature-importance measures; a `RandomForestRegressor` with $n_{\text{estimators}} = 100$ (scikit-learn default) was used, with other hyperparameters left at their default values—these can be optimised using nested cross-validation if required; and model generalisation was assessed using leave-one-sequence-out (LOSO) cross-validation, in which each sequence is held out in turn to preserve sequence-level independence.

When applied to clean - original - images, the composed metric (from the trained pipeline) correlates strongly with the model's score on a per-sequence basis: Table 19 reports the per-sequence Pearson correlation between the composed metric and the model score. Correlations are uniformly high (typically $r > 0.90$), indicating that the composed metric is an excellent proxy for the model score under clean acquisition conditions.

Table 19: Pearson Correlation by Sequence of Composed Metric vs Score

Sequence	**Original**
p028_08_R_LCI	0.949
p046_19_R_WLI	0.942
p228_01_R_LCI	0.942
p295_02_R_BLI	0.956
p364_04	0.949
p415_01	0.905
p451_05	0.971
p525_08	0.956
p525_13	0.927
p526_03	0.964
p584_07	0.949
p730_01_01	0.947
p871_02_01	0.964

Tables 20 and 21 present contingency matrices comparing the binarised composed metric (High/Low) with the model score (High/Low) for distorted and original frames, respectively. For the full set of distorted images (Table 20, $N = 9{,}499$) the two measures agree in $2{,}800 + 2{,}799 = 5{,}599$ cases, yielding an overall accuracy of approximately $5599/9499 \approx 58.9\%$. The off-diagonal entries are nearly equal (1950 vs. 1950), indicating a loss of discriminative power under perturbation rather than a systematic bias toward one class. By contrast, for original, unmodified frames (Table 21, $N = 1{,}357$) agreement occurs in $650 + 649 = 1{,}299$ cases (accuracy $\approx 95.7\%$), with only 29 false positives and 29 false negatives, which demonstrates that the composed metric is an excellent proxy for the model score under clean acquisition conditions.

Table 20: Score vs. Composed Metric — distortion images

	Metric High	**Metric Low**
Score High	2800	1950
Score Low	1950	2799

Table 21: Score vs. Composed Metric — only for originals

	Metric High	**Metric Low**
Score High	650	29
Score Low	29	649

To further evaluate the robustness of the composed metric, we computed its correlation with the true score across different sequences and image modifications. Table 22 reports the Pearson correlation coefficients obtained per sequence and modification type. The results show that, for the *original* images, the correlation between the composed metric and the true score is consistently very high (ranging from 0.85 to 0.99), indicating that the composed metric is a reliable proxy under clean, unaltered conditions. This confirms that, in the absence of distortions, the proposed metric captures the underlying factors that drive the model's effective confidence. In contrast, the correlation values become highly variable under image modifications such as blur, brightness perturbations, and artificial obstructions. For example, while some sequences maintain moderate to strong positive correlations (e.g., blur5 in sequence `p451_05` with $r = 0.816$), others display strong negative correlations (e.g., blur10 in sequence `p028_08_R_LCI` with $r = -0.597$).

Table 22: Correlation of the composed metric versus the true score, by sequence and modification.

Sequence	Blur5	Blur 10	Blur 15	Bright Gaussian = 0.5	Bright Global	obstructed	original
p028_08_R_LCI	−0.123	−0.597	−0.598	−0.033	−0.190	0.516	0.953
p046_19_R_WLI	0.223	0.669	0.643	−0.167	0.027	−0.287	0.953
p228_01_R_LCI	0.173	0.263	0.100	0.077	−0.033	−0.023	0.955
p295_02_R_BLI	0.181	−0.273	−0.393	0.566	0.256	0.591	0.968
p364_04	0.417	0.567	0.422	0.183	0.370	−0.008	0.968
p415_01	−0.147	0.374	−0.082	0.280	0.035	0.111	0.908
p451_05	0.816	0.557	0.621	−0.451	−0.313	−0.419	0.986
p525_08	0.452	0.619	0.728	−0.298	−0.311	0.292	0.974
p525_13	0.335	0.096	0.241	0.111	−0.036	−0.205	0.942
p526_03	−0.289	−0.212	0.188	0.008	0.112	0.172	0.982
p584_07	- 0.053	0.291	0.250	−0.746	0.124	0.324	0.974
p730_01_01	0.208	−0.372	−0.509	0.334	−0.034	0.127	0.958
p871_02_01	0.508	0.551	0.555	0.106	0.334	0.387	0.983

Figure 37 illustrates the application of Grad-CAM to a frame from sequence p295_02_R_BLI: the left panel shows the original image with the polyp visible, and the right panel shows the same frame with the polyp completely artificially obscured. Despite the occlusion, the frame was correctly classified and the Grad-CAM heatmap is concentrated on the obstructing region.

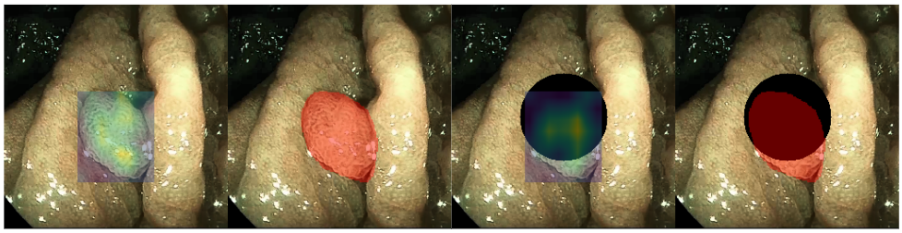


Figure 37: Application of Grad-CAM to a frame from the sequence p295_02_R_BLI. Left: original image; right: obstructed version.

5.6 Summary

This section provides a concise synthesis of the experimental work.

5.6.1 Classification model

The selected classifier achieved approximately an accuracy of 68%, a precision of 95%, a recall of 66% and F_1 of 78%. It was perceived that region-focused inputs (cropped frames) consistently improved predictive performance relative to full-frame inputs, ResNet full-frame 69.61% versus cropped 71.16%, indicating that explicit localisation of the polyp region reduces distracting background information and facilitates feature learning. It should be noted that the proposed methodology is inherently dependent on the underlying model performance, which directly influences the obtained results.

5.6.2 Preliminary experiment (pipeline validation)

A controlled experiment on the Oxford–IIIT Pet dataset was used as a fast methodology pipeline-validation and to validate the integration of explainability modules. The saliency maps revealed two distinct patterns. In some cases, the model focused on the entire animal or in big regions, resulting in higher IoU values. In other instances, the model became hyper-focused on discriminative features, such as the eyes, ears, or nose, leading to lower IoU. Despite these variations, the model's confidence remained high in both cases.

5.6.3 Polyp Experiment

Both SHAP and Grad-CAM provide useful visual explanations for cropped polyp frames, though each reveals different strengths and weaknesses. SHAP generates high-resolution, signed attribution maps. Negative attributions often coincide with specular highlights and peripheral artifacts, while positive attributions tend to align with lesion borders or raised central regions, though occasionally they also align with local artifacts. In the case of Grad-CAM, the most important activation pixels were typically concentrated within the polyp body and along the lesion borders; however, in a subset of frames the most salient pixels corresponded to imaging artefacts (e.g., specular reflections from illumination).

Quantitatively in SHAP there is a moderate positive correlation between model confidence and attribution intensity skewness (Score vs. Skewness $r \approx 0.497$) and weak negative correlations between confidence and SHAP–GT overlap (Score vs. Positive IoU $r \approx -0.159$; Score vs. Negative IoU $r \approx -0.190$).

In practice, higher confidence is associated with more concentrated, point-like SHAP attributions; however, increased confidence does not imply improved spatial alignment with the clinical mask. In Grad-CAM, IoU correlates negatively with the activation centroid distance from the crop centre (IoU vs. Distance Center $r \approx -0.31$) and only weakly positively with model confidence (IoU vs. Score $r \approx 0.16$). Activations nearer the crop centre tend to overlap more with the expert mask.

5.6.4 XAI comparison

The two methods are complementary but inherently different: SHAP yields fine-grained, signed pixel attributions at high computational cost and with overlapping positive/negative maps that complicate localisation metrics, while Grad-CAM produces coarse, regional heatmaps from a single forward/backward pass, which are computationally efficient and easier to threshold and compare with segmentation masks. For these reasons we selected Grad-CAM as the primary XAI method for the quantitative experiments, while retaining SHAP for targeted, fine-grained inspections where pixel-level signed attributions were specifically required.

5.6.5 Robustness to image modifications and sequence effects

Frequency-destroying perturbations (Gaussian blur, extreme desaturation) produced substantial reductions in IoU between Grad-CAM maps and segmentation masks while often causing only moderate decreases in score (the model confidence for the grount-truth class). This dissociation suggests that peak Grad-CAM activations do not necessarily capture all discriminative features exploited by the classifier, which may rely on alternative, non-localized cues. Colour and geometric perturbations yielded heterogeneous effects across sequences. The IoU–confidence relationship varied markedly between sequences, indicating that robustness is conditioned on sequence-specific visual factors (polyp size, lesion–background contrast, illumination and specular artefacts).

5.6.6 Composed Metric

An exploratory composed metric, defined from a small set of image- and attribution-based features $\text{IoU},\ \text{LaplacianVar},\ \text{Skewness},\ \text{Dist}_{\text{center}}$ was evaluated for its correlation with the model's performance on original frames. This composed metric exhibited a strong correlation with the model's score, as indicated by Pearson correlation coefficients generally exceeding $r > 0.90$ on a per-sequence basis. When applied to artificially degraded sequences, the metric's correlation with model score became unstable, ranging

from positive to negative values depending on degradation type and severity. These results suggest the metric is informative for high-quality images and may be useful as a lightweight filter to flag frames with severe deformations, but it requires further formal validation.

Chapter 6

Discussion

This chapter provides a critical interpretation of the experimental results, practical implications and directions for future work.

6.1 Experiments

Based on the literature, previous research indicates that image quality can significantly impact the performance of deep learning classifiers, with a fundamental difference in how humans and AI systems perceive it.

This study tested the hypothesis that explanation maps correlate with classification results, specifically focusing on whether Grad-CAM activations overlap substantially with the ground truth annotation of the objects being classified.

We first implemented a binary classification model to differentiate hyperplastic polyps from adenomatous polyps. Various architectures were tested, and the final model achieved a precision of 95%, which reflects its strong ability to minimize false positives. However, the accuracy (68%) and recall (66%) values indicate that there is still considerable room for improvement, particularly in the model's ability to identify all positive instances. This could be attributed to the dataset's heterogeneous image conditions, including motion blur, instrument occlusions, and poorly illuminated regions, which likely hinder the model's ability to consistently learn discriminative polyp features. These challenges are common in medical imaging and present practical obstacles that can significantly affect classification performance.

Performance improved when using cropped frames (71.16%) compared to full-frame inputs (69.61%). This improvement underscores the importance of focusing on the region of interest, as it reduces the impact of background noise and enhances the model's ability to learn discriminative features. These results highlight the critical role of preprocessing steps, such as cropping, in medical image classification, which can significantly enhance model performance, especially in clinical applications.

To validate our approach, we tested the same methodology on a simpler classification problem: the preliminary experiment, classifying cats and dogs in the Oxford–IIIT Pet dataset. Using a similar training approach, we concluded that the classifier itself was unlikely to be the primary source of the problems observed. However, this experiment revealed differences in the model's feature localization. In some sequences, the model focused on large regions, such as the entire body of the animal, which resulted in higher IoU values. In other cases, the model hyper-focused on smaller discriminative features, such as eyes, ears, or noses, leading to lower IoU values.

With these results in mind, we applied this reasoning to our main focus: classification in colonoscopy. Similar factors influencing the correctness of the model's predictions were observed. Notably, the relationship between IoU and confidence varied across different sequences of polyp images. This inconsistency suggests that model robustness is not uniform and may depend on sequence-specific factors, such as polyp size, lesion-background contrast, and illumination conditions. These findings highlight the need for further improvements in model robustness to better handle real-world image variations.

Our qualitative evaluation of Grad-CAM results confirmed that the model attended to relevant features in the image, which is crucial for ensuring that the model's decisions can be interpreted and trusted by clinicians. Experiments using SHAP and Grad-CAM provided complementary insights into the model's decision-making process. Grad-CAM offered a more intuitive understanding of the model's focus, particularly in terms of spatial alignment with the lesion. This makes it a useful tool for clinicians, as it produces heatmaps that are easier to interpret and directly correlate with regions of interest in the image, aiding them in confirming the model's predictions. In contrast, SHAP attributions revealed a strong association between classifier confidence and global image photometric asymmetry, such as skewness, and it is more sensitive to input-level perturbations. While this makes SHAP highly valuable for AI experts or researchers, who can leverage the fine-grained, pixel-level attributions to understand underlying model behavior, its relevance may be less direct for clinicians.

The analysis of the model's robustness to image modifications revealed a heterogeneous relationship between IoU and score. When images underwent perturbations like Gaussian blur, the IoU values decreased significantly, while model confidence often remained relatively high. This dissociation suggests that the Grad-CAM activations do not fully capture all discriminative features used by the classifier, and the model might rely on non-localized cues (such as global image patterns) to make its decisions.

In some cases, we observed that augmentations similar to those used during training did not yield the expected improvements or degradations. This outcome may be due to the model already being trained on similar types of distortions.

We also explored the use of a composed metric, designed to assess the alignment between the quality of images (with the features) and the system (the correctness of the classification prediction). This was suggested because no single quality metric could fully define the model's score. While neither linear approximation nor Principal Component Analysis (PCA) yielded reliable results, a Random Forest model demonstrated promise, with Pearson correlation coefficients consistently exceeding 0.90 on a per-sequence basis. However, when applied to sequences with artificial degradation, the correlation between the metric and the model's score became unstable, varying from positive to negative values depending on the type and severity of the degradation. These results suggest that the metric is particularly useful for identifying high-quality images and could serve as a lightweight tool for flagging frames with severe distortions or deformations. Nonetheless, to ensure its broader applicability, further formal validation is required, particularly in more diverse and realistic image conditions.

In general, we can conclude, based on observations, that some of the selected quality factors are an indicator of the final correctness of the classification, but we infer that IoU is informative but insufficient by itself to judge explanation quality or to compare methods with fundamentally different attribution characteristics.

This thesis started out with the goal to use explainable AI techniques to identify the quality of the image and the classification. During this project, we noticed that this approach was overly simplistic, which made the thesis evolve towards an overall analysis of quality, explainable AI techniques and classification results. Both the model and the dataset have inherent limitations. AI models may not be sufficiently robust to handle all the variations in the data, and datasets themselves often contain imperfections, such as noisy annotations, variations in image quality, and environmental factors like lighting or patient motion. In this sense, it is essential to recognize that both components are not ideal and align the model's capabilities with the full variability of the dataset.

6.2 Challenges

Throughout this study, several challenges emerged that influenced the results and the methodology applied.

One of the primary challenges in evaluating model explanations was the dependence of IoU on thresholding and heatmap normalization. The threshold selection process is crucial for a fair and consistent evaluation of explanation methods, as it significantly impacts the IoU score.

Although our study contributes valuable insights by evaluating commonly used model explanation

methods, it is important to note some limitations. Specifically, only two XAI methods—Grad-CAM and SHAP—were explored. While numerous other XAI methods have been proposed in the literature, it was infeasible to investigate all of them within the scope of this study. However, the choice of methods was not arbitrary. Grad-CAM and SHAP were selected because they are among the most popular techniques used for explaining deep neural networks in image analysis. Additionally, we aimed to compare two distinct families of post-hoc explanation methods: Grad-CAM and SHAP, making them natural choices for inclusion. Future work could expand this investigation by exploring other explanation methods, such as LIME, Integrated Gradients, or counterfactual and concept-based explanations, to offer a more comprehensive understanding of the model's behavior. Notably, the metric IoU tends to favor Grad-CAM, as it produces larger, smoother activation regions (blobs), while penalizing SHAP, which generates sparse, point-wise attributions. This difference in the nature of attributions—Grad-CAM's larger regions versus SHAP's precise points—means that IoU alone is insufficient for a fair comparison between methods.

Another challenge is the low availability of high-quality, annotated datasets. In frames from colonoscopy videos, the images have poor contrast and the colors are not uniformly distributed which increases the error rate. A further limitation of this study is the reliance on manual segmentation for GT annotations. It introduces subjectivity and variability into the methodology. Finally, the fact that the analysis is sequence-dependent presents another challenge. The results are only directly applicable in relative cases, meaning that the findings are specific to the sequences examined in this study. This limits the generalization of the distortion analysis, highlighting the need for further exploration.

To conclude, since this study was focused on gastrointestinal diseases, particularly polyp classification in colonoscopy images, one important challenge is the generalization of the findings to other areas of medicine. Differences in medical specialties—such as variations in image modalities, diagnostic techniques, and the specific characteristics of diseases—can introduce challenges in applying the results of this study to other domains.

6.3 Prospect for future work

While this study provided useful insights into deep learning for polyp classification, several lines of future work would strengthen robustness and clinical readiness.

First, a key challenge in medical image analysis is ensuring that models generalize well across a wide range of real-world variations in image quality. To address this, advanced augmentation strategies should be explored to improve generalization. Techniques such as Mixup, CutMix, GAN-based synthesis, and

adversarial augmentation can increase dataset diversity and help models cope with realistic degradations common in endoscopy, such as motion blur, occlusion, and uneven illumination.

Second, the current study focused on two popular explainability methods, Grad-CAM and SHAP, but there is potential to expand the set of explainability methods. Counterfactual explanations can reveal decision boundaries and model sensitivity to targeted changes, while concept-based approaches (e.g., Concept Activation Vectors) can identify higher-level visual constructs relevant to clinicians. Comparing methods with metrics suited to both dense and sparse attributions will yield a fairer evaluation than IoU alone.

Third, strengthen statistical validation of the proposed composite quality metric. Use frame-level bootstrap confidence intervals (CIs), cross-sequence validation, and hypothesis tests to quantify stability and generalizability of correlations between the metric and classifier correctness. Investigate how the composite score behaves under different realistic degradations.

Further exploration is required to refine metrics that offer a more comprehensive assessment of both image quality and classification performance. Such investigations could include the development of a weighting system that differentiates the significance of pixels within the polyp region from those outside the region of interest.

Fourth, a promising direction is to further investigate the composite quality metric as a decision tool for handling low-quality frames. Low-scoring frames should be automatically flagged for clinician review and categorized (for example: corrupted/unusable, mislabeled, or representative of realistic clinical variation). Depending on that categorization, one potential approach is to investigate re-training the model using sequences that exhibit the lowest values on the composed metric, for example, the bottom 10% of frames. By focusing on the worst-performing frames, the model can be forced to learn how to handle lower-quality images more effectively. This could significantly improve the model's robustness in real-world clinical scenarios where image quality varies.

Finally, for clinical translation, validate models prospectively with clinician feedback and diverse, multi-center datasets.

Chapter 7

Conclusions and future work

7.1 Conclusions

This dissertation explores the integration of XAI techniques into polyp classification systems and examines how image-quality metrics relate to model-based explanation metrics, with the ultimate aim of improving model robustness and clinical trust in colorectal cancer screening.

For this achievement, it was developed a deep learning model for colorectal polyp classification that achieved strong performance in terms of accuracy, precision, and recall, reaffirming the feasibility of using deep learning for CRC detection in colonoscopy images.

The application of XAI techniques, specifically Grad-CAM and SHAP, offered complementary insights into the model's decision-making. Grad-CAM produced coarse, clinically intuitive heatmaps that often aligned with relevant polyp regions, making it particularly useful for clinician-facing explanations. SHAP produced fine-grained, pixel-level attributions that supported expert-level analysis of model behavior. Together, these methods increased the interpretability and transparency of model predictions. The study also demonstrated that image quality strongly affects classification performance. Artifacts such as motion blur, occlusions, and uneven illumination were associated with reduced accuracy.

Another contribution of this work is the evaluation of the alignment between model explanation maps and GT segmentation masks, and the derivation of a composite quality metric that combines these alignment measures with image-quality features. This metric showed promise for identifying high-quality frames and flagging low-quality frames likely to hinder model performance. Although further validation on larger and more diverse datasets is required, the metric may be valuable for pre-screening colonoscopy frames and focusing clinician attention on the most diagnostically relevant data.

In summary, this study demonstrates that integrating explainability into classification models is both feasible and clinically valuable. It underscores the importance of transparency in medical AI and highlights the need to ensure that such systems are closely aligned with clinical practice.

7.2 Future Work

This dissertation provides a foundation for several promising future research avenues, as summarized in Section 6.3.

Future work should focus on enhancing the robustness of AI models to real-world image variations, expanding the exploration of additional XAI methods, and broadening the validation of the composite metric.

Furthermore, validating the proposed methods in real-world clinical settings, with input from healthcare professionals, will be crucial for ensuring the successful translation of these tools into practice.

In conclusion, integrating explainable AI techniques into CRC detection has significant potential to enhance the transparency, reliability, and accuracy of automated diagnostic systems. By addressing the current challenges and refining the methods presented in this thesis, we can contribute to ongoing efforts to improve early CRC detection and, ultimately, patient outcomes.

Bibliography

[1] World Health Organization. Colorectal cancer, 2024. Available at: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer. [Accessed: 25 October 2024].

[2] B. Duan, Y. Zhao, J. Bai, J. Wang, X. Duan, X. Luo, R. Zhang, Y. Pu, M. Kou, J. Lei, and S. Yang. Colorectal cancer: An overview. In J.A. Morgado-Diaz, editor, *Gastrointestinal Cancers*. Exon Publications, Brisbane (AU), 2022. Online first 09 Jul 2022.

[3] Rebecca L. Siegel, Angela N. Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1):12–49, January 2024.

[4] Milcamps A, Alexandru A, Baldacchino S, Stănică L, O'grady T, Stepien M, Caldeira S, and Rodriguez Rasero F. Uncovering inequalities: Colorectal cancer screening in europe. Scientific analysis or review, European Commission, Ispra (Italy), 2024.

[5] World Health Organization International Agency for Research on Cancer. Cancer today: Portugal fact sheet. https://gco.iarc.who.int/media/globocan/factsheets/populations/620-portugal-fact-sheet.pdf, 2022. Accessed: 2025-01-04.

[6] M. Abd El-Salam, S. Reda, S. Lotfi, T. Refaat, and E. El-Abd. Imaging techniques in cancer diagnosis. In *Cancer Biomarkers: Minimal Noninvasive Early Diagnosis and Prognosis*, pages 19–38. CRC Press, 2014.

[7] Marcello Di Giammarco, Fabio Martinelli, Antonella Santone, Mario Cesarelli, and Francesco Mercaldo. Colon cancer diagnosis by means of explainable deep learning. *Scientific Reports*, 14(1):15334, July 2024.

[8] Tom Eelbode. *Automated Data-driven Online Detection and Characterization of Endoscopic Lesions Using Deep Learning*. Phd thesis, KU Leuven, Leuven, March 2023.

[9] Ilja Tacheci and Marcela Kopacova. Peutz-jeghers syndrome. *Current opinion in gastroenterology*, Publish Ahead of Print, 02 2021.

[10] Ishak Pacal, Dervis Karaboga, Alper Basturk, Bahriye Akay, and Ufuk Nalbantoglu. A comprehensive review of deep learning in colon cancer. *Computers in Biology and Medicine*, 126:104003, November 2020.

[11] L. Pinto-Coelho. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering (Basel)*, 10(12):1435, December 2023.

[12] Khaled ELKarazle, Valliappan Raman, Patrick Then, and Caslon Chua. Detection of Colorectal Polyps from Colonoscopy Using Machine Learning: A Survey on Modern Techniques. *Sensors*, 23(3):1225, January 2023.

[13] Adrian Krenzer, Stefan Heil, Daniel Fitting, Safa Matti, Wolfram G. Zoller, Alexander Hann, and Frank Puppe. Automated classification of polyps using deep learning architectures and few-shot learning. *BMC Medical Imaging*, 23(1):59, April 2023.

[14] Giulio Antonelli. Current and future implications of artificial intelligence in colonoscopy. *Annals of Gastroenterology*, 2023.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA, August 2016. ACM.

[16] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[17] Alexander Stevens, Peter Deruyck, Ziboud Van Veldhoven, and Jan Vanthienen. Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1241–1248, Canberra, ACT, Australia, December 2020. IEEE.

[18] Karolina Szczepankiewicz, A. Popowicz, Kamil Charkiewicz, Katarzyna Nałęcz-Charkiewicz, Michal Szczepankiewicz, Lasota Slawomir, Pawel Zawistowski, and Krystian Radlak. Ground truth based comparison of saliency maps algorithms. *Scientific Reports*, 13, 10 2023.

[19] Arne Gevaert, Axel-Jan Rousseau, Thijs Becker, Dirk Valkenborg, Tijl De Bie, and Yvan Saeys. Evaluating Feature Attribution Methods in the Image Domain, August 2024. arXiv:2202.12270 [cs].

[20] Samuel Dodge and Lina Karam. Understanding How Image Quality Affects Deep Neural Networks, April 2016. arXiv:1604.04004 [cs].

[21] American Cancer Society. What is colorectal cancer?, 2024. Available at: https://www.cancer.org/cancer/types/colon-rectal-cancer/about/what-is-colorectal-cancer.html. [Accessed: 2024-09-08].

[22] C. M. Stauffer and C. Pfeifer. Colonoscopy. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island (FL), January 2024. [Updated 2023 Jul 24].

[23] Raseena T.P, Jitendra Kumar, and S. Balasundaram. Deepcpd: deep learning with vision transformer for colorectal polyp detection. *Multimedia Tools and Applications*, 83:1–24, 02 2024.

[24] Fiona Foss, Kenneth P. West, and Alastair H. McGregor. Pathology of polyps detected in bowel cancer screening programme. *Diagnostic Histopathology*, 17, 2011.

[25] Mira Meseeha and Mona Attia. *Colon Polyps*. StatPearls [Internet]. StatPearls Publishing, Treasure Island (FL), 2024. [Updated 2023 Aug 8].

[26] American Cancer Society. Colonoscopy, n.d. Available at: https://www.cancer.org/cancer/diagnosis-staging/tests/endoscopy/colonoscopy.html. [Accessed: 2024-09-08].

[27] N. Shussman and S.D. Wexner. Colorectal polyps and polyposis syndromes. *Gastroenterology Report*, 2:1–15, 2014.

[28] Iris D Nagtegaal, Robert D Odze, David Klimstra, Valerie Paradis, Massimo Rugge, Peter Schirmacher, Kay M Washington, Fatima Carneiro, and Ian A Cree. The 2019 who classification of tumours of the digestive system. *Histopathology*, 76(2):182–188, 2020.

[29] F. Sierra-Jerez and F. Martínez. A deep representation to fully characterize hyperplastic, adenoma, and serrated polyps on narrow band imaging sequences. *Health Technology*, 12:401–413, 2022.

[30] Thomas G. Myers, Prem N. Ramkumar, Benjamin F. Ricciardi, Kenneth L. Urish, Jens Kipper, and Constantinos Ketonis. Artificial intelligence and orthopaedics. *Journal of Bone and Joint Surgery*, 102(9):830–840, February 2020.

[31] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers*, 12(5), 2023.

[32] Luisa F. Sánchez-Peralta, Luis Bote-Curiel, Artzai Picón, Francisco M. Sánchez-Margallo, and J. Blas Pagador. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artificial Intelligence in Medicine*, 108:101923, August 2020.

[33] Ksenia Yadav, Serge Bidnyk, and Ashok Balakrishnan. Artificial intelligence and machine learning in optics: tutorial. *Journal of the Optical Society of America B*, 41:1739–1753, 07 2024.

[34] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

[35] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[36] Jana Sweidan, Mounim A. El-Yacoubi, and Anne-Sophie Rigaud. Explainability of CNN-based Alzheimer's disease detection from online handwriting. *Scientific Reports*, 14(1):22108, September 2024.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[38] Nitish Kundu. Exploring resnet50: An in-depth look at the model architecture and code implementation, n.d. Retrieved from Medium.

[39] GeeksforGeeks. Residual networks (resnet) – deep learning, 2024. Available at: https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/. [Accessed: 2024-11-06].

[40] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in Biology and Medicine*, 128:104115, 2021.

[41] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[42] Petru P. What is efficientnet? the ultimate guide. Roboflow Blog, August 2023. Available at: https://blog.roboflow.com/what-is-efficientnet/. [Accessed: 2024-09-24].

[43] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

[44] Eliana Pastor. Pattern-based algorithms for Explainable AI.

[45] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2018.

[46] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, November 2017. arXiv:1705.07874 [cs, stat].

[47] Council of the European Union and European Parliament. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, 2016.

[48] Dara Varam, Rohan Mitra, Meriam Mkadmi, Radi Aman Riyas, Diaa Addeen Abuhani, Salam Dhou, and Ayman Alzaatreh. Wireless capsule endoscopy image classification: An explainable ai approach. *IEEE Access*, 11:105262–105280, 2023.

[49] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2), 2023.

[50] Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated, 2021.

[51] Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods, June 2022. arXiv:2201.13291 [cs].

[52] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emerg (Tehran)*, 3(2):48–49, 2015. Spring 2015.

[53] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation, 2022.

[54] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021.

[55] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[56] Subrato Bharati, M. Rubaiyat Hossain Mondal, and Prajoy Podder. A review on explainable artificial intelligence for healthcare: Why, how, and when? *IEEE Transactions on Artificial Intelligence*, 5(4):1429–1442, April 2024.

[57] Christoph Molnar. Pixel attribution (saliency maps), 2021.

[58] T. Nathan Mundhenk, Barry Y. Chen, and Gerald Friedland. Efficient Saliency Maps for Explainable AI, March 2020. arXiv:1911.11293 [cs].

[59] GeeksforGeeks. What is saliency map?, 2024. Available at: https://www.geeksforgeeks.org/what-is-saliency-map/?ref=ml_lbp. [Accessed: 2024-12-02].

[60] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. arXiv:1312.6034 [cs].

[61] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.

[62] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[63] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.

[64] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[65] Harvard IACS. Lecture 11: Saliency maps and explainable ai, 2020. Available at: https://harvard-iacs.github.io/2020F-AC295/lectures/lecture11/presentation/lecture11.pdf. [Accessed: 2024-12-02].

[66] Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, and Huey-Wen Liang. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.

[67] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif CIFCI, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladík, Saeid Nahavandi, and Panos M. Pardalos. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, 2024.

[68] Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, June 2023. Feature attributions based on the Shapley value are popular for explaining machine learning models, but their estimation is complex from both theoretical and computational standpoints.

[69] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. Explainable artificial intelligence for human decision-support system in medical domain, 2021.

[70] Scott Lundberg and Su-In Lee. Shap (shapley additive explanations) documentation, 2024. Available at:https://shap.readthedocs.io/en/latest/. [Accessed: 2024-11-06].

[71] Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation, February 2020. arXiv:1908.08474 [cs].

[72] A.E. Roth, editor. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[73] Christoph Molnar. Interpretable machine learning - a guide for making black box models explainable, 2024. Available at: https://christophm.github.io/interpretable-ml-book/shapley.html. [Accessed: 2024-11-16].

[74] The Modern Scientist. What is the shapley value?, 2024. Available at: https://medium.com/the-modern-scientist/what-is-the-shapley-value-8ca624274d5a. [Accessed: 2024-11-16].

[75] Srirupa Guha, Ashwini Kodipalli, Steven L. Fernandes, and Santosh Dasar. Explainable ai for interpretation of ovarian tumor classification using enhanced resnet50. *Diagnostics*, 14(14), 2024.

[76] Bader Aldughayfiq, Farzeen Ashfaq, N. Z. Jhanjhi, and Mamoona Humayun. Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap. *Diagnostics*, 13(11), 2023.

[77] R Aswin, H Kumar, and U Varma. Resnet-based deep learning framework for liver cancer detection with explainable ai (xai) technique. In *Proceedings of the 2024 International Conference on Computing, Communication, and Networking Technologies (ICCCNT)*, pages 1–6, 06 2024.

[78] The Deep Learning Intern. Diving deep into pooling: The backbone of cnn efficiency, 2023. Available at: https://thedeeplearning-intern322.medium.com/diving-deep-into-pooling-the-backbone-of-cnn-efficiency-75d6b300025a. [Accessed: 2024-11-16].

[79] unpackAI. Cross-entropy loss in ml, 2023. Available at: https://medium.com/unpackai/cross-entropy-loss-in-ml-d9f22fc11fe0. [Accessed: 2024-11-06].

[80] Jason Brownlee. Cross-entropy for machine learning, 2023. Available at: https:https://machinelearningmastery.com/cross-entropy-for-machine-learning/. [Accessed: 2024-11-05].

[81] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[82] GeeksforGeeks. Adam optimizer in tensorflow, 2024. Available at: https://www.geeksforgeeks.org/adam-optimizer-in-tensorflow/. [Accessed: 2024-11-26].

[83] TensorFlow Authors. Adam optimizer - tensorflow core v2.13.0, 2024. Available at: https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam. [Accessed: 2024-11-27].

[84] Amber Roberts. Precision: Understanding this foundational performance metric. https://arize.com/blog-course/precision-ml/, November 2022. Published 28 November 2022. Accessed 22 October 2025.

[85] Amber Roberts. F1 score: How it combines precision and recall. https://arize.com/blog-course/f1-score/. Accessed 22 October 2025.

[86] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Submitted on 25 Feb 2019 (v1), last revised 15 Apr 2019 (this version, v2).

[87] Towards Data Science. Metrics to evaluate your semantic segmentation model. *Towards Data Science*, 2021. Available at: https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2. [Accessed: 2024-11-05].