

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Explainable Automatic Chest Radiography Pathology Detection and Reporting

Pedro Daniel Couto Pereira

DISSERTATION



Mestrado em Engenharia Biomédica

Supervisor: João Manuel Patrício Pedrosa

Second Supervisor: Ana Maria Rodrigues de Sousa Faria de Mendonça

October 9, 2023

Explainable Automatic Chest Radiography Pathology Detection and Reporting

Pedro Daniel Couto Pereira

Mestrado em Engenharia Biomédica

October 9, 2023

Abstract

Chest radiography, also known as Chest X-Ray (CXR), is a widely used medical procedure for studying the lungs and heart, areas that are constantly the target of life-threatening diseases and trauma. Despite CXR's cost-effectiveness and accessibility, the high demand for CXR examinations places a significant burden on radiologists. Automatic support systems can help alleviate the workload by providing aid to radiologists in analyzing CXR images.

The digitalization of clinical practices has led to a large number of CXR being stored, while advancements in computational power and algorithms lead to improvements in deep learning (DL) approaches in computer vision, to the point where models can identify and diagnose pathologies in CXR with performances comparable to expert radiologists. However, the models' increasing complexity and opacity in decision-making raise concerns among medical professionals, hindering their integration into real-world scenarios. This demand for understanding DL model decision-making has driven the development of various explanation techniques that help elucidate model reasoning. Yet, a gap remains in assessing the practical applicability and precision of these methods due to the subjective nature of explainability and challenges in quantitative comparisons.

In this work, different visual explanations methods were implemented and evaluated. In the first experiment, the VinDr-CXR dataset was trained to classify different pathological classes and multiple visual explanations were extracted. The explanations achieved comparable results to the state-of-the-art, with the Grad-CAM++ and Saliency standing out. A new method, that mimics written reports, was created to convey visual explanations, where the locations of the findings are given with respect to the lung's position. The method standardized visual explanations allowing for comparison with ground truth annotations, giving rise to the Zones Score evaluation metric.

The second experiment compared deep learning architectures and image sizes, revealing that DenseNet121 outperformed ResNet50 in both classification and explanation performance and that reducing image size from 512×512 to 256×256 minimally affected classification performance but had a notable impact on explanations. A statistical analysis was conducted in the same chapter to understand which variables influence the explanations, revealing that the y-position, area, and frequency of the annotation as well as the confidence of the model and the contrast and energy of the pathology had a direct relationship with the explanation score. Finally, the third experiment examined explanations across the VinDr-CXR and ChestX-ray14 datasets and the results showed that the explanation behavior is dependent on the dataset properties and annotations.

This work shows that Grad-CAM++ and Saliency offer more accurate visual explanations than other state-of-the-art methods, highlighting that the effectiveness of visual explanations is contingent on the model, image size, and is subject to dataset bias. Furthermore, a novel explanation method is introduced, mimicking a written report, which facilitates standardization and comparison of visual explanations.

Keywords: Automatic Medical Reports, Chest X-Ray, Deep Learning, Explainability

Resumo

A radiografia do tórax, também conhecida como Raio-X do Tórax (CXR), é uma técnica de imagem médica amplamente utilizada para estudar os pulmões e o coração, órgãos que são constantemente alvo de doenças e acidentes. Apesar da acessibilidade e custo-benefício do CXR, a elevada procura por estes exames coloca os radiologistas sob bastante pressão. Sistemas de suporte automático podem ajudar a aliviar a carga de trabalho fornecendo ajuda aos radiologistas na análise de imagens CXR.

A digitalização das práticas clínicas levou ao armazenamento de um grande número de radiografias de tórax, enquanto os avanços no poder computacional e nos algoritmos conduziram a melhores abordagens de aprendizagem profunda (DL) na visão computacional, ao ponto de os modelos conseguirem identificar e diagnosticar patologias em CXR com desempenhos comparáveis aos de radiologistas. No entanto, a crescente complexidade dos modelos e a falta de transparência na tomada de decisões suscitam preocupações entre os especialistas, dificultando a sua integração no mundo real. Esta necessidade de melhor explicar o processo de tomada de decisão em modelos DL levou ao desenvolvimento de várias técnicas de explicação que ajudam a elucidar o raciocínio dos modelos. No entanto, continuam a existir problemas na avaliação da aplicabilidade prática e precisão destes métodos devido em parte, à natureza subjetiva da explicabilidade, mas também à dificuldade em efetuar comparações quantitativas entre os resultados de diferentes métodos.

Neste trabalho, foram implementados e avaliados diferentes métodos de explicação visual. Na primeira experiência, o *dataset* VinDr-CXR foi treinado para classificar diferentes classes patológicas e foram extraídas diferentes explicações visuais. As explicações obtiveram resultados comparáveis ao estado da arte, destacando-se a *Grad-CAM++* e a *Saliency*. Além disso, foi criado um novo método para converter explicações visuais num relatório médico em que as localizações dos achados são indicadas relativamente à posição do pulmão. O mesmo foi então utilizado para padronizar as explicações visuais num formato que poderia ser comparado com anotações radiológicas convertidas para o mesmo formato, levando a uma nova métrica de avaliação denominada *Zones Score*.

A segunda experiência comparou arquiteturas de aprendizagem profunda e tamanhos de imagem, revelando que a *DenseNet121* superou a *ResNet50* tanto na tarefa de classificação como na explicação e que a redução do tamanho da imagem de 512×512 para 256×256 afetou minimamente o desempenho na classificação, mas teve um impacto notável nas explicações. Foi realizada uma análise estatística para compreender quais as variáveis que influenciam as explicações, revelando que a posição no eixo dos *yy*, a área e a frequência das anotações, bem como a confiança do modelo e o contraste e a energia da patologia tinham uma relação direta com a pontuação da explicação. Por fim, a terceira experiência examinou as explicações nos *datasets* VinDr-CXR e ChestX-ray14 e os resultados mostraram que o comportamento da explicação depende das propriedades e anotações do *dataset*.

Este trabalho mostra que a *Grad-CAM++* e a *Saliency* oferecem explicações visuais mais

exactas do que outros métodos do estado da arte, evidenciando ainda que a eficácia das explicações depende do modelo, do tamanho da imagem e está sujeita ao enviesamento do *dataset*. Além disso, é introduzido um novo método de explicação, imitando um relatório escrito, que facilita a normalização e a comparação de explicações visuais.

Palavras-chave: Explicabilidade, Radiografia do Tórax, Aprendizagem Profunda, Relatórios Médicos Automáticos.

Acknowledgements

Esta dissertação é o culminar de cinco anos académicos recheados de aventuras, aprendizagens e pessoas incríveis, que foram essenciais no desenvolvimento da minha pessoa.

Gostaria de começar por expressar a minha gratidão para com os meus pais e irmã, cujo amor e confiança nunca me faltou. São um reflexo dos valores que prezo e uma fonte de inspiração constante.

Felizmente, sempre tive a sorte de estar acompanhado de pessoas incríveis fora do meu agregado familiar. Aos meus amigos antes da faculdade e que permaneceram ao meu lado após a mesma, que continuemos a criar momentos como aqueles até hoje. À malta de Aveiro, Sara, Vera, Raúl e Rúben, um muito obrigado por fazerem de Aveiro uma memória que aquece. À Vera, por se rir das minhas piadas e ser a melhor roommate, ao Raúl, por fazer de mim uma pessoa mais liberal e ao Rubén, por me ajudar a manter a alma nortenha acesa. Um especial obrigado à Sara pelos momentos vividos, pelo carinho e por ser fonte de motivação.

Aos meus compinchas da ANEEB, com quem partilhei dois anos de muita aprendizagem e crescimento, o meu sincero obrigado. À Rebecca e Mathias, que fizeram do meu Erasmus uma das experiências mais incríveis, thank you!

Não posso não deixar palavras de apreço ao meu orientador e supervisor da dissertação, João Pedrosa, e à minha coorientadora, Ana Maria Mendonça, do INESC TEC. A vossa disponibilidade e apoio incessante durante toda a dissertação foram inestimáveis, estou convicto de que fui sortudo por ter tido oportunidade de trabalhar e acima de tudo aprender com vocês.

Pedro Daniel Couto Pereira

"He who has a why to live for can bear almost any how."

Friedrich Nietzsche

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	1
1.3	Goals	2
2	Fundamental Concepts	3
2.1	Radiography	3
2.1.1	Fundamental Physical Concepts	3
2.1.2	Acquisition System and Image Formation	4
2.2	Chest Imaging	6
2.2.1	Chest Radiography	6
2.2.2	Chest Radiography in Clinical Practice	6
3	Explainability in Automatic CXR Analysis	11
3.1	Towards Automatic Chest Radiography Analysis	11
3.2	Abnormality Detection and Classification	13
3.3	Explainable AI	14
3.3.1	Model-transparent Methods	15
3.3.2	Model Agnostic Methods	20
3.3.3	Case-based Explainability	23
3.3.4	Evaluation of Explainability Methods	25
4	Materials and methods	29
4.1	Datasets	29
4.1.1	VinDr-CXR	29
4.1.2	ChestX-ray14	30
4.1.3	VinDr-CXR and CXR14 Label Match	31
4.2	Models and Training Details	32
4.2.1	ResNet50	33
4.2.2	DenseNet121	33
4.3	Explainable AI	34
4.3.1	Visual based Explainability	34
4.3.2	Structured Report based Explanation	35
4.4	Evaluation Metrics	39
4.4.1	Model Performance	39
4.4.2	Explainability Evaluation Metrics	40
4.5	Statistical Analysis - Variable Selection	43

5	Evaluating XAI in Chest Radiography Pathology Detection	45
5.1	Experiment Setup	45
5.1.1	Training Details	45
5.1.2	Data Pre-processing	47
5.2	Results and Discussion	49
5.2.1	Model Performance	49
5.2.2	Selection of Visual Explanations	49
5.2.3	Visual Explanations Performance across Models	52
5.2.4	Visual Explanation Performance across Classes	54
5.2.5	Statistical Analysis Result	58
6	Evaluation of XAI in Chest Radiography Pathology Detection across datasets	61
6.1	Experiment Setup	61
6.1.1	Training Details	61
6.1.2	Data Pre-processing	62
6.2	Results and Discussion	63
6.2.1	Model Performance	63
6.2.2	Visual Explanations Performance across Datasets	64
7	Conclusions and Future Work	69
	References	71
A	Supplementary Figures	81

List of Figures

2.1	X-ray tube schema.	4
2.2	Different X-ray exams.	5
2.3	Overall total frequencies per 1000 of the population for different countries. The relative contributions of the four main groups (plain radiography without dental, fluoroscopy, computed tomography and interventional radiology) are also shown.	7
2.4	Chest X-ray main organs and structures.	7
2.5	Examples of CXR pathologies.	8
2.6	Radiologist written report examples from different sources.	10
3.1	CNN main architecture overview.	13
3.2	XAI methods overview.	15
3.3	Application of DeepLiFT on deep model regularization using CXR reports. Left, original image. Right, original image with overlaid DeepLiFT.	17
3.4	Grad-CAM pipeline.	18
3.5	Grad-CAM used in EIPH detection. Left, original CXR image. Right, Grad-CAM heatmap overlaid on top of the CXR.	19
3.6	Example of LRP used to explain a correct classification of pleural effusion. Left, is the original CXR image. Right, LRP heatmap for the pleural effusion class.	20
3.7	Example of LIME explanation of COVID-19 classification. Left, is the original CXR image. Right, LIME super-pixels overlaid on top of the original image.	22
3.8	SHAP explanation on multiple pathology detection. The first image is the original CXR image, followed by CXR images for each labeled class with the SHAP heatmap overlaid on top of the original image.	22
3.9	Example-based-explanations on ship image.	23
3.10	An overview on Image-based retrieval systems on CXR images.	24
3.11	Example-based explanations on CXR images, comparing different retrieval methods. LXIR stands for lesion-based CXR image retrieval.	25
3.12	Example of Saporta et al. benchmark study and comparison metrics. Left, CXR with ground-truth and saliency method annotations for pleural effusion. The segmentations have a low IoU of 0.078, but it was a "hit" because the most representative point identified by the saliency method is within the ground-truth segmentation. In contrast, in the case of an enlarged cardiomeastinum, the segmentations have a high IoU of 0.682, but it was not a successful "hit" because the most representative point identified by the saliency method is outside of the ground-truth segmentation.	26
4.1	Examples of VinDR-CXR dataset images with radiologist's annotations.	30
4.2	Class frequencies in the VinDR-CXR dataset and number of labels per image.	30

4.3	Class frequencies in the ChestX-ray14 dataset and number of labels per image. . .	31
4.4	Overview of the ResNet50 residual block.	33
4.5	Overview of the DenseNet architecture and the dense block.	34
4.6	Difference between saliency maps before and after applying the Gaussian filter. .	35
4.7	CXR segmentation process using U-Net.	36
4.8	Description of the lung zones and possible combinations.	37
4.9	Example of a structured report using radiologist bounding box annotations. The left portrays the original CXR with the bounding box plotted on top while the right shows the report extracted.	38
4.10	Bounding box extracted from a heatmap.	38
4.11	Example of a Precision-Recall curve. The point 2 and 4 represent the threshold for Perfect Model and Good Model examples, respectively.	39
4.12	Visual examples of the IoU (0.6) and SA-IoU metrics. A random CXR image and heatmap were selected, as well as a sample ground truth bounding box.	41
4.13	Visual examples of the hit rate. A random CXR image and two different heatmaps. The black "x" points to the highest intensity pixel of the heatmap.	41
4.14	Bounding box converted into a one-hot encoded array.	42
5.1	Bar plot of each fold class division.	46
5.2	Fold division.	46
5.3	Examples of augmentation techniques applications.	47
5.4	Bounding box weighted match and fusion example.	48
5.5	Example of CXR with the PuF class and corresponding heatmaps after applying the size adaptive threshold for different explanations, along with the ground truth bounding box. The first row displays all the CAM-based explanations, while the second row contains gradient-based explanations.	51
5.6	Example of CXR with the AoE class and corresponding Grad-CAM++ and Saliency heatmaps after applying the size adaptive threshold for different models and image sizes, along with the ground truth bounding box.	54
5.7	Metrics results for Grad-CAM++ and Saliency on the DenseNet121-512 model across all classes. The dotted lines represent the average values for all classes. . .	55
5.8	Average explanations heatmaps of the test set and bounding box annotations, per class. The first column represents Grad-CAM++, followed by Saliency and average bounding box annotations in the last column.	57
5.9	Correlation matrix results.	58
6.1	Frequency of each class of the ChestX-ray14 dataset with bounding box annotations.	62
6.2	Grad-CAM++ and Saliency SA-IoU results for all classes and dataset combinations. The bars that have no pattern were trained on the ChestX-ray14 and those that have a pattern were trained on VinDr-CXR. Orange bars were evaluated on VinDr-CXR, and green were evaluated on the ChestX-ray14 dataset.	65
6.3	Example of CXR with the Inf class and corresponding heatmaps after applying the size adaptive threshold for different models and datasets combinations. The first row displays all the Grad-CAM++ explanations while the second row contains the Saliency heatmaps.	66

- 6.4 Average Grad-CAM++ heatmaps for all configurations. The first two columns represent configurations evaluated on the VinDr-CXR dataset and the third, GT-V, represents the VinDr-CXR average bounding box annotations. The fourth and fifth column represents configurations evaluated on the ChestX-ray14, while the last column, GT-C, represents the ChestX-ray14 average bounding box annotations. 67
- A.1 Example of CXR with the PIE class and corresponding heatmaps after applying the size adaptive threshold for different explanations, along with the ground truth bounding box. The first row displays all the CAM-based explanations, while the second row contains gradient-based explanations. 81
- A.2 Average Saliency heatmaps for all configurations. The first two columns represent configurations evaluated on the VinDr-CXR dataset and the third, GT-V, represents the VinDr-CXR average bounding box annotations. The fourth and fifth column represents configurations evaluated on the Chest X-ray 14, while the last column, GT-C, represents the ChestX-ray14 average bounding box annotations 82

List of Tables

3.1	Sporta et al. results for the IoU metric.	26
3.2	Sporta et al. results for the Hit rate metric.	27
4.1	Datasets information.	29
4.2	Label matching between datasets.	32
4.3	Structured report entries.	36
5.1	Mean AUC values for the VinDr-CXR test set across different models.	50
5.2	Mean XAI performance metrics for the VinDr-CXR test set.	50
5.3	RLM results.	59
6.1	Mean AUC values for the ChestX-ray14 and VinDr-CXR datasets.	63
6.2	Mean SA-IoU, Hit Rate and Zones Score values for Grad-CAM++ and Saliency explanations across the VinDr-CXR and ChestX-ray14 datasets.	64

Acronyms and Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
CAD	Computer Aided Diagnostic
CAM	Class Activation Map
CNN	Convolutional Neural Network
CR	Computed Radiography
CT	Computed Tomography
CTR	Cardio-Thoracic Ratio
CXR	Chest X-Ray
DL	Deep Learning
DR	Digital Radiography
DDR	Digital Direct Radiography
DeepLiFT	Deep Learning Important FeaTures
GT	Ground Truth
IoU	Intersection over Union
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer-wise Relevance Propagation
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
PET	Positron Emission Tomography
ReLU	Rectified Linear Unit
SFR	Screen Film Radiography
SSIM	Structural Similarity Index
SHAP	Shapley Additive Explanations
XAI	Explainable Artificial Intelligence
YOLO	You Only Look Once

Chapter 1

Introduction

1.1 Context

According to the World Health Organization [1], 4 out of the top 5 causes of death are cardiac or respiratory-related diseases. Given that the thorax cavity is a key location for trauma and health complications, there is a crucial need to examine this area of the body thoroughly.

Of the different methods used to assess the interior of the thorax cavity, Chest radiography (CXR) is the most common, accounting for almost 70% of all the imaging exams, whereas in Portugal that number is around 84% [2]. The popularity of CXR is associated with its cost-effectiveness and accessibility. Moreover, it is very easy to use, fast at providing results, non-invasive, and requires low levels of radiation when compared with other imaging techniques. In addition, radiologists can analyze the differences in brightness, shape, and format present in the CXR acquiring crucial insights to detect different heart and lungs related pathologies. Therefore, CXR plays an essential role in screening, diagnosis, and disease management.

Furthermore, digital radiography allows for a computer representation of the exam that can be post-processed and analyzed which leads to the creation of large X-ray datasets with high-quality images.

1.2 Motivation

Despite CXR's many advantages, interpreting the exam can be a laborious and time-consuming task, representing a significant workload for radiologists. There are dozens of parameters to be analyzed and a high variability of diseases which leads to uncertainty and possible errors in the medical diagnosis. This opens up the possibility of using deep learning (DL) models to assist healthcare professionals, shortening the diagnosis time and helping them prioritize urgent cases. Despite the challenges, automatic systems have shown promising results in detecting and correctly classifying most of the diseases present in the CXR. These automatic systems can provide the same information as radiologists' imaging reports, which include the main findings, their location, nature, probability, and relevance [3].

Even though these systems can achieve high accuracy levels, medical applications that assist radiologists are not commonly used. This is primarily due to radiologists' lack of trust in the predictions made by the model, as they are unable to comprehend the reasoning behind them. To address this issue, several explainability methods have been proposed, providing insight into the functioning of DL systems. However, despite the number of available methods, little attention has been devoted to comparing different explainability methods in real cases and validating which method(s) can provide more precise and complete explanations. This is partly due to the subjective nature of explainability but also because different methods provide different outputs making quantitative comparisons challenging.

This work aims to implement and validate state-of-the-art artificial intelligence explainable methods that would unveil the reasoning behind DL models trained on CXR abnormality detection. Different visual explainability methods will thus be compared quantitatively, based on radiologists' annotations, on multiple architectures, and across datasets. In addition, a novel explainable method will be developed that is based on medical reports structure, thus ensuring it "speaks the same language" as radiologists.

1.3 Goals

The project has the following goals:

- Development of automatic structured reports of CXR images based on deep learning automatic systems.
- Implementation of CXR visual explainability methods.
- Quantitative comparison of visual explainability methods applied to different deep learning architectures and evaluated on multiple datasets.

Chapter 2

Fundamental Concepts

In this chapter key concepts of X-ray imaging modality are introduced, which is central to the thesis. First, a brief history of the X-ray discovery, then the fundamental physical concepts that are behind the X-rays as well as the different types of X-ray imaging applications that are used by the medical community. Following this, a particular type of radiography is covered, Chest Radiography, and the common radiological findings are described in detail.

2.1 Radiography

In 1895, while testing whether cathode rays - radiations produced by a cathode tube - could pass through a volume of glass, Wilhelm Röntgen, a professor of physics in Wurzburg, Germany, discovered the X-rays. He called the rays "X" because this is the letter commonly used to represent "unknown" [4]. Röntgen realized that even when the cathode tube was covered, to prevent visible or ultraviolet light from escaping, an incandescent green light still leaked out and projected onto a nearby fluorescent screen. After conducting additional tests, he discovered that the enigmatic light would penetrate most materials but cast shadows on solid objects. The same effect was observed in the human body by Röntgen when he projected X-rays into his wife's hand creating what is considered to be the first "medical" X-ray. The absorption of X-rays by tissues and organs varies due to their atomic composition, which is the foundation of the different imaging techniques utilized in radiography.

2.1.1 Fundamental Physical Concepts

X-rays are produced using an X-ray tube that evolved from the cathode tube. The tube is placed in vacuum and has a cathode filament that emits electrons into the vacuum as well as an anode plate, normally made of tungsten, that collects these electrons, creating an electrical current known as the beam in the tube, as represented in Figure 2.1. The electrons are emitted from the cathode filament due to the thermionic effect, which happens when a metal is heated by flowing electric

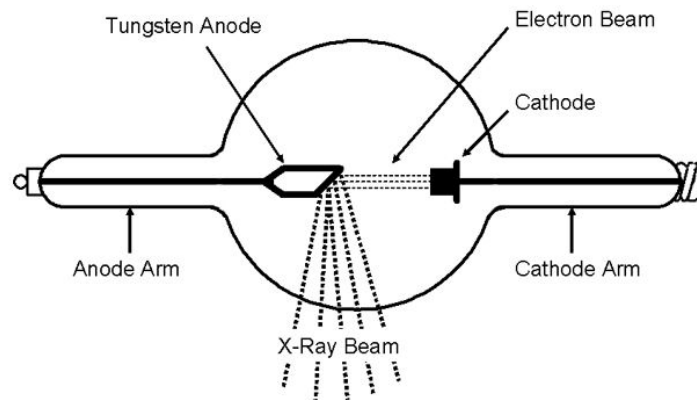


Figure 2.1: X-ray tube schema [5].

current through it. The high-temperature filament produces high-speed electrons, which are attracted to the positive anode by the acceleration voltage between the negative cathode and the positive anode-tube voltage. When a high energy electron hits the target plate it is decelerated and interacts with the atoms of the anode, which results in the emission of an X-ray photon. The energy of emissions can be adjusted by altering the applied voltage and installing aluminum filters of different thicknesses. These filters are placed in the X-ray beam's path to remove low-energy non-penetrating radiation. The dose or amount of emitted X-ray photons can be regulated by adjusting the current intensity as well as the exposure time. Therefore, X-rays have a broad spectrum of wavelengths, going between 0.01 and 10 nm and energies between 10 and 100 keV, which makes it possible for the human body to sustain the radiation exposure. In comparison with other ionizing radiation, X-rays energy is higher than UV rays and lower than gamma [6].

Characteristic X-rays, Bremsstrahlung X-rays, and heat production are the main outcomes of electron acceleration from the cathode to the anode. During the ionization process, there is an inner electron-shell emission and subsequent orbital void filling which results in the production of characteristic X-rays, while when an electron passes near the nucleus of the target, it decelerates and changes paths with the loss of energy throughout the form of Bremsstrahlung X-rays [7]. The human body receives this radiation and the tissues, having different properties, will attenuate more, or less, the radiation that crosses the body. This is the key property that then allows the image reconstruction by counting the photons that can reach the other side. As an example, soft tissues are more transparent to X-rays, which leads to a higher number of photons to pass, on the other hand, bones are more opaque, which reduces the number of X-rays that are able to cross these structures, as exemplified in Figure 2.2.

2.1.2 Acquisition System and Image Formation

X-ray images are created through a process similar to photography, where X-rays serve as the illuminating source instead of light. In an X-ray image, areas of the film that are exposed to the

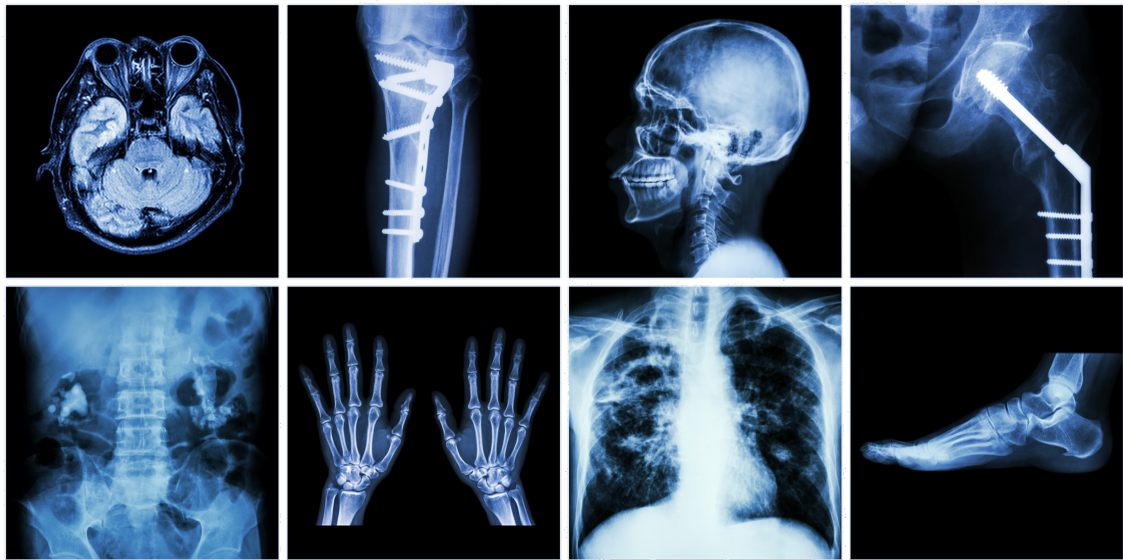


Figure 2.2: Different X-ray exams, adapted from [8].

X-ray beam will appear dark, while areas that are protected from the beam due to the absorption of the X-rays by soft tissues will appear light.

The traditional X-ray acquisition method, Screen Film Radiography (SFR), produces images using cassettes that are exposed to X-rays. These image receptor cassette contains a radiography film, made up of a base covered by film emulsion, placed between two phosphor screens. Silver bromide crystals, a light-sensitive material, are present in the film emulsion component [7]. A radiographic image is created when light photons contact with silver bromide crystals and transfer their information into a latent image in the film. The film is then taken out of the cassette and developed, turning the latent image into a permanent image that is deposited as silver in the film's emulsion layer [9]. However, SFR presents some drawbacks, such as taking a long time to process, using expensive films made of dangerous materials alongside the problem of not having a digital version of the image making it impossible to do any kind of image processing or analysis. Due to these issues, this approach is considered to be inconvenient in nowadays practice.

The development of digital receptor systems and technological advancements in hardware led to the emergence of Digital Radiography (DR). Within DR, there are two main methods: Computed Radiography (CR) and Direct Digital Radiography (DDR). Although both methods allow for the acquisition of digital images, they do so using different techniques.

CR makes use of a computed radiographic image receptor that, in the same way as the SFR, is an X-ray-sensitive plate, while in this case using storage phosphors and trapped electrons in metastable states to produce an image instead of a scintillator used in SFR. A stimulating laser beam is used to release the metastable electrons, and light is emitted when the electrons return to their ground state. This light emission is used to form an image, using an analog electrical system that is later digitized. This type of radiography is the only one that uses reusable plates [9]. However, the image quality of CR is poorer than DDR and requires a longer exposure time.

DDR is a more modern approach to Digital Radiography that uses no intermediate processing step, allowing for faster acquisition and higher spatial resolution. It uses a flat panel and is divided into two conversion methods: direct and indirect. The first method uses a semiconductor material to directly convert absorbed X-rays into electrical charge without an intermediate step, while the second method employs a scintillator to first convert X-rays to light before the conversion to electrical charge for readout. Specifically, the common scintillator used is cesium iodide which has a needle-like structure to minimize scatter during the light conversion, and the light is then converted to electrical charge by a low-noise photodiode array [9]. Each photodiode corresponds to a single pixel and the electrical charge is digitally read out and sent to the image processor.

2.2 Chest Imaging

In spite of plain radiography's enormous success in thorax analysis, other imaging modalities that do not use X-rays, have also become quite popular in the last decades namely Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET). This breakthrough in imaging techniques was only possible due to the advances in both hardware and software. However, one can say that plain X-ray is still a great asset taking into account that it is less expensive when compared with the other exams, it is the one with the most availability (X-ray machines are widely available in hospitals, clinics, and other medical facilities), plain X-ray exams are quick and can be performed in just a few minutes and even though there is radiation exposure, X-ray exams involve low levels of radiation and are generally considered safe for most people. In addition, it is a versatile technique, being used to image a wide range of body parts and help identify problems such as fractures, infections, and tumors.

2.2.1 Chest Radiography

In Europe, plain radiography is the leading imaging technique in every country, even when dental X-ray procedures are excluded, as Figure 2.3 illustrates, accounting for almost 70% of all the imaging exams, whereas in Portugal the number is around 84% [2]. Radiography is a widely utilized and well-known diagnostic imaging technique often serving as the initial step in the diagnostic process, even when more advanced exams are necessary. From the different areas to which it may be applied, on average 48% of the requested X-rays are Chest/Thorax X-rays, followed by Mammography with an average of 12.6%. These numbers are explained by the fact that the thorax cavity is the place where important structures are located, from bones to organs such as the heart and the lungs, which are crucial to our well-being and are constantly the target of diseases and physical accidents.

2.2.2 Chest Radiography in Clinical Practice

The ability to discern normal appearances and structures on CXR is essential for accurately interpreting any abnormal findings. To accomplish this, it is necessary to understand what are the main

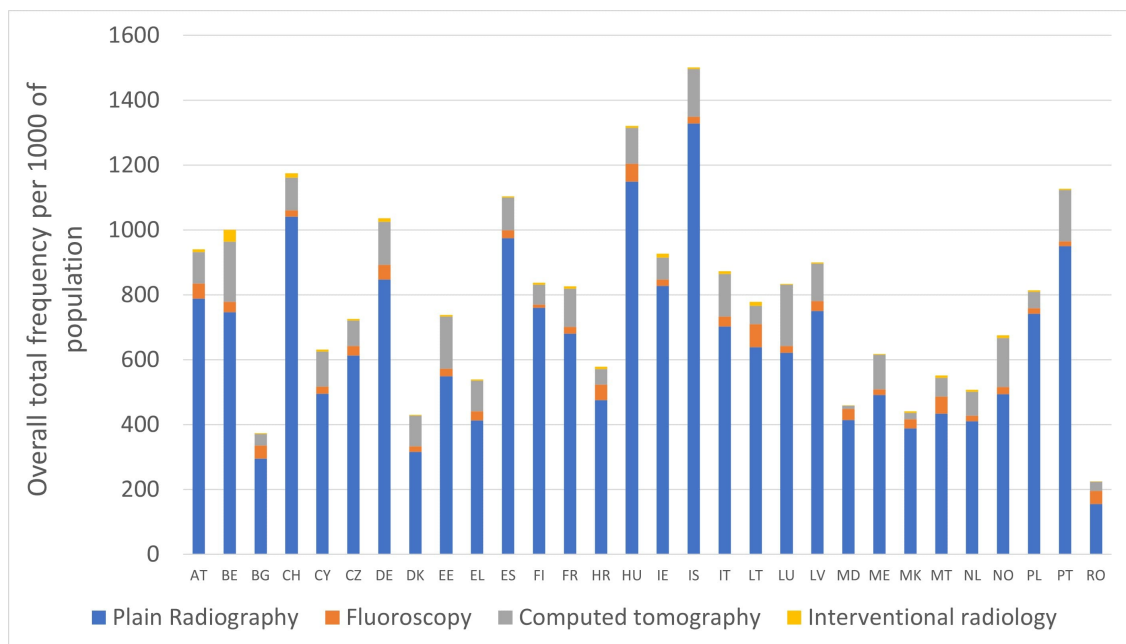


Figure 2.3: Overall total frequencies per 1000 of the population for different countries. The relative contributions of the four main groups (plain radiography without dental, fluoroscopy, computed tomography and interventional radiology) are also shown, based on [2].

structures present in the CXR and how the radiologist analyzes them daily. CXR can be taken from different chest projections, such as Postero-anterior (PA) - the ray enters the posterior aspect of the patient and exits the anterior, Antero-posterior (AP) – the ray enters the anterior aspect and exits the posterior - or Lateral [10].

Normally, in a CXR it is possible to observe different structures and organs, such as the lungs, the ribs, the sternum bone, the clavicles, the heart, the diaphragm, and the lung pleura, as shown in Figure 2.4. The thoracic box holds the lungs and heart being fully comprised by the curvature of the ribs, between the external bone and the vertebral column.

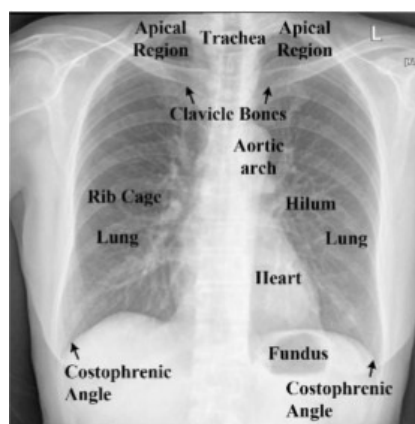


Figure 2.4: Chest X-ray main organs and structures, from [11].

The adoption of a systematic methodology for the interpretation of all films is of paramount importance to ensure that no abnormality is neglected. It is advisable to cultivate the habit of scrutinizing the review or "check" sections each time a physician interprets a chest X-ray. These sections are, according to Joarder et al. [10], the following: behind the heart, cardiophrenic angles, costophrenic angles, apices, the periphery of pleura, bones, and below diaphragms. For each section, the radiologist will ask different questions and compare the exam with what is known to be normal. From such examination, several discoveries can be made and utilized by the physician to arrive at a diagnosis. In this section, some of these findings will be mentioned, including but not limited to lung opacities, cardiomegaly, pleural effusion, and pneumothorax, as shown in Figure 2.5.

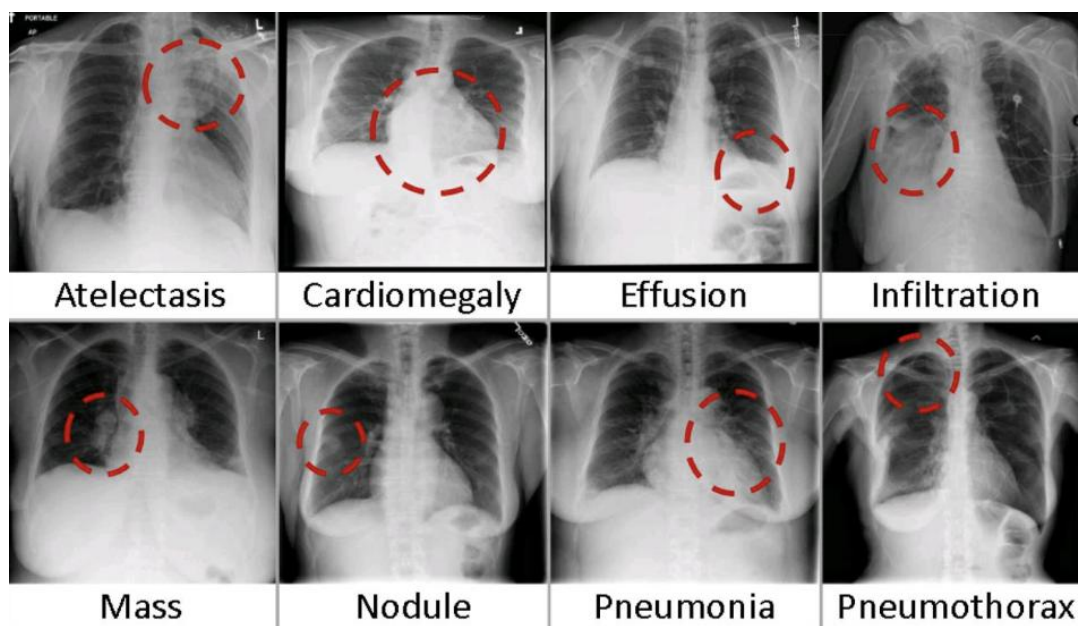


Figure 2.5: Examples of CXR pathologies, adapted from [12].

Lung opacity or pulmonary opacification occurs when there is a reduction in the gas-to-soft tissue (such as blood, lung tissue, and connective tissue) ratio inside the lungs. When looking at the CXR, as these tissues are denser than the air, there is a loss of the vascular pattern common in the lung and the image will appear brighter in those areas. The opacity may originate from a different condition such as infections, pulmonary edema, alveolar hemorrhage, pulmonary embolism pneumonia, and pulmonary consolidation [13] and is characterized by taking into consideration factors such as size, shape, and brightness.

Enlarged heart, or cardiomegaly, is a condition where the heart reaches abnormal dimensions and can be a consequence of a variety of factors, including stress, high blood pressure, coronary artery disease, and certain types of heart muscle disorders. The normal size of the heart can vary based on a person's age, sex, and body size, so the definition of cardiomegaly can vary somewhat, however, it is common to use the cardio-thoracic ratio (CTR) as a parameter to diagnose this condition, where the normal measurement is 0.42-0.50 [14]. CTR is the ratio of maximal

horizontal cardiac diameter to maximal horizontal thoracic diameter (inner edge of ribs/edge of pleura). The numerator is usually calculated by adding the greatest perpendicular diameter from the midline to the right heart with the greatest perpendicular diameter from the midline to the left heart border.

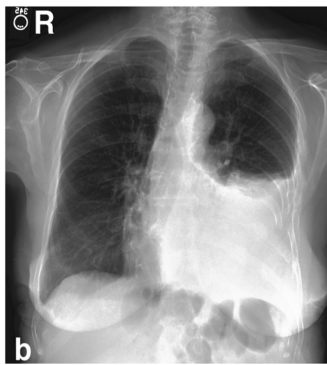
Other common findings are pleural abnormalities, such as pleural effusions, pneumothorax, and pleural thickening [10]. Pleural effusions occur when excess fluid accumulates in the pleural space. It is normally caused by congestive cardiac failure and cancer [15]. On the other hand, pneumothorax occurs when there is gas (air, normally) in the pleural space. This condition is caused by a penetrating chest injury or after a medical procedure and may put in danger the patient's life, as in some cases the presence of air leads to the lung collapsing. When comes to pleural thickening, it refers to the thickening involving either the parietal or visceral pleura, due to the presence of scar tissue, which may be caused by an infection, cancer, or recurrent pneumothoraces [16].

Another important topic of attention is the radiological reports. In addition to understanding the common pathologies, it is crucial to get an idea of how the physician performs the diagnosis and how he annotates his findings.

According to [17, 18], radiological reports vary from hospital to hospital, but some steps are shared by the majority of them. The report normally starts with the patient personal information, including the patient's name, date of birth, and other identifying information, followed by a summary of the patient's symptoms and medical history relevant to the imaging study. Then, the imaging findings, a free-text description of the results of the imaging study, including any abnormal findings or observations.

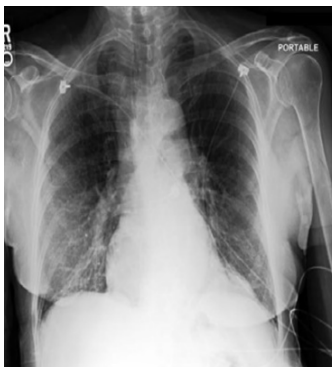
When preparing a CXR report, radiologists may have a preliminary idea of the potential diagnosis based on the patient's medical history or previous pathologies. In such cases, it is common for them to begin the report by addressing the specific case or relevant findings related to the previous conditions. Alternatively, if there is no prior indication, radiologists typically start by evaluating the heart, followed by the aortic and pulmonary arteries. They then proceed to assess the presence of nodules, masses, pneumothorax, and other common pathologies, as shown in Figure 2.6 based on the MIMIC-CXR [19] dataset, a database of CXR written reports used to train NLP models.

After this, the professional impression or interpretation should be given, containing a summary of the radiologist's conclusions and recommendations for further testing or treatment. From the above-mentioned steps, it is particularly useful for a radiologist to get assistance in the imaging findings step. Normally, this section of the radiological should be focused on providing clear, concise, and factual observations. Interpretations that go beyond the scope of the imaging study should be avoided, as well as using overly subjective language and repeating information unnecessarily [20]. Such properties are capable of being automated by a computer and, therefore, the development of a structured report that can display the information in a close way to what an expert would do can be considered a solid explanation method.



Findings: Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT.

(a) Example of a written report with references to past exams, adapted from [19].



Findings: The heart size appears to be borderline and similar to normal. No evidence of pneumothorax is observed. Segmental elevation noted in the left hemidiaphragm. Minimal basilar atelectasis. Probable scarring in the right costophrenic angle.

(b) Example of a written report adapted from [21].

Figure 2.6: Radiologist written report examples from different sources.

Chapter 3

Explainability in Automatic CXR Analysis

The next section covers the main advances in chest X-Ray abnormality detection, the common technologies used, as well as introducing the explainability concept in the Artificial Intelligence (AI) world. Furthermore, this chapter delves into the latest and trending explanation methods and their applications in the medical field, with a particular emphasis on CXR.

3.1 Towards Automatic Chest Radiography Analysis

Chest-X-Ray is one of the most common medical imaging techniques due to its affordability and non-invasive nature, playing a key role in screening, diagnosing, and treatment planning of a disease, as explained in 2.2. However, the radiologist's task is not straightforward as the CXR is an arduous image to analyze, which contributes to elevated variability in the reports. Not only that but when we take into account the fact that an expert radiologist takes years to train with the high number of exams to report, it becomes clear that CXR analysis is an extremely time and resource-consuming task. Therefore, there is a need to develop tools that can aid the physician in this first stage, mitigating the complexity of the task, reducing variability, and saving both time and resources. Hence, multiple studies to develop computer-aided diagnosis (CAD) systems for CXR assessment have been done, trying to attenuate the time and complexity of this task [3].

The latest CAD systems are mainly based on AI. AI is a field of computer science that aims to understand and replicate the functions of intelligent entities in order to create software programs that mimic their operation [22]. In the beginning, AI models were focused on specific tasks, such as interpreting electrocardiogram signals or helping in the selection of the appropriate treatment taking into consideration patient characteristics. Today, AI has reached a point where it is also used to help manage patient information, health services, and aid the diagnostic stage, adding a lot of value in the predictive medicine field [23].

In the realm of healthcare applications, two primary AI technologies are utilized: machine learning (ML), a subset of AI, and deep learning (DL), a subfield of ML. ML is seen as the traditional approach, it uses algorithms to analyze large amounts of data and identify patterns that can be used to make predictions. ML is particularly effective at handling large numbers of predictors, which are normally pre-extracted features, and combining them in complex, nonlinear ways. This approach is similar to the way a doctor might learn and grow through exposure to various cases and data [24]. Due to the "curse of dimensionality," which hinders the ability of ML algorithms to learn from high-dimensional data such as images, feature engineering, and selection are often necessary prerequisites [25]. DL solves most of the problems of ML, using multiple layers, therefore the "deep", to learn the features directly from the data and achieve overall better generalization and performance. Unfortunately, large datasets are required as well as time and high-performance workstations. With the newest hardware advances, providing the necessary computing power, DL has observed enormous growth which enabled unprecedented breakthroughs in a variety of computer vision tasks, making these networks an excellent choice for CXR CAD systems.

A CXR CAD system incorporates several steps to reach a diagnosis, one of the most important being abnormality detection, which focuses on prioritizing the more urgent abnormal cases. State-of-the-art research mentions that having beforehand knowledge of which cases are more critical would allow the clinicians to better manage their time and resources, which ultimately helps reduce the damages, considering that cardiothoracic and pulmonary abnormalities are one of the leading causes of morbidity and mortality [26]. Some authors add that this step could be further complemented with another crucial task, which is the identification of the pathologies present in the exam at hand. Here, one must consider a multi-label thoracic pathology classification approach, given that it is possible to have more than one in the same image/patient [26].

The detection and classification of cardiothoracic and pulmonary abnormalities often resort to Convolutional Neural Networks (CNN), due to their great ability to handle data with a strong spatial relationship. These particular architectures are made of three layers, that are repeated throughout the network in different arrangements, a convolutional layer, a pooling layer, and a fully connected layer, as represented in Figure 3.1. The convolutional layer, which names this particular class of networks, is comprised of a dot product between two matrices. One matrix, known as the kernel, comprises the learnable parameters, while the other matrix is a circumscribed section of the input features, such as part of the pixels of an image. These layers are responsible for extracting features from the input data, being highly effective in learning patterns that are normally disregarded by the human eye. In addition, to help with non-linear features, it is also common to find non-linearity layers, such as Sigmoid or rectified linear unit (ReLU). Non-linear functions allow the model to learn complex and more accurate representations of the data, which can improve its performance. After such processes, it is common to find a fully connected layer that is accountable for the final prediction. The pooling layers help reduce the spatial size of its input, decreasing the amount of computation power needed.

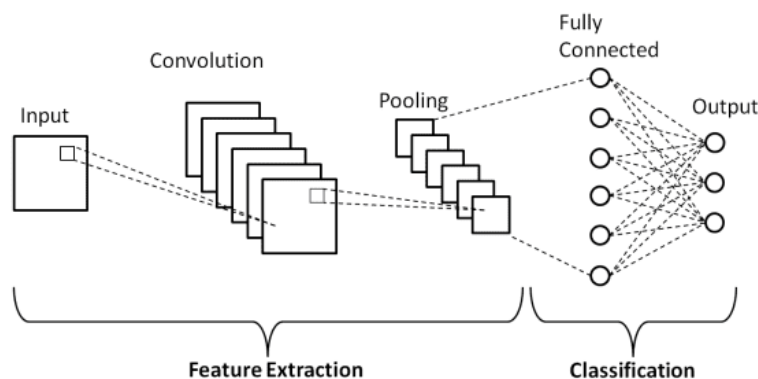


Figure 3.1: CNN main architecture overview [27].

3.2 Abnormality Detection and Classification

Recently, various studies have employed DL models to identify and categorize abnormalities in CXR. One of two lines of study is normally followed, either considering a binary classification scenario, where the images are either normal or abnormal, or multi-label pathology classification, in which the authors develop systems to automatically identify multiple pathologies.

Attending the binary classification studies, both Tang et al. [28] and Dunnmon et al. [29] compared different well-known CNN architectures and concluded that the DenseNet architecture stands out, even though all CNNs performed well, achieving an area under the curve (AUC) score of 0.881 on the public ChestX-ray14 [30] dataset.

When it comes to the automatic identification of multiple pathologies in CXR, Rajpurkar et al. [31] proposed the CheXNet, considered to be the classical example of a DenseNet implementation for multi-label detection, achieving an AUC of 0.837 on the ChestX-ray14. In the next years, both Kumar et al. [32] and Uribanbayev et al. [33] tried to improve the work presented in CheXNet by using a boosted cascaded context and an improved optimizer, respectively, with AUC values of 0.775 and 0.830. Gündel et al. [34] proposed the DNetLoc which is a variant of DenseNet that addresses class imbalance by incorporating additional weights within the loss function, achieving an AUC score of 0.840. Chen et al. [35] proposed a dual asymmetric architecture, that combines the DenseNet and ResNet to build the DualCheXNet. It adaptively captures more discriminative features of several pathologies since the DenseNet and ResNet capture different and unique features, leading to mean AUC values of 0.863.

Some authors noticed that the networks that were performing better were capturing more discriminative characteristics of each pathology present in the images, therefore Guan et al. [36] implemented an attention-guided CNN to combine the features extracted from both the global and local image for thoracic disease classification, which leads to a mean AUC score of 0.883. Kabir et al. [37] used a weakly supervised approach that combines DenseNet layers with a saliency sampling approach to extract the most informative parts or pixels of the input images, with a mean AUC score of 0.890 in ChestX-ray14.

Another trend is using object detection networks on CXR datasets with bounding box annotations. Under this line of study, Luo et al. [38] combined the YOLOv5 [39] network, state of the art when it comes to object detection with a ResNet on the VinDr-CXR [40] dataset, managing to improve the precision, but no score of AUC was provided. Pedrosa et al. [41] performed a more intensive study comparing the YOLOv5 with DenseNet on different class groups and on multiple datasets, including VinDr-CXR and the ChestX-ray14, achieving AUC values of 0.993 and 0.742, respectively, on the YOLOv5 model.

3.3 Explainable AI

A DL model can estimate complex and arbitrary functions and classify with great accuracy cardiothoracic and pulmonary abnormalities on CXR images. However, it does not come with forthright explanations. Given that neural networks rely on very low-level and unconventional structures – a large number of different filters and operations distributed across a stack of layers – it is very challenging for medical professionals to understand how the model arrives at certain conclusions – the black box problem. It's hard to look "into" the network and ascertain which features produced a given output. Although the impressive state-of-the-art performance cannot be disputed, it is increasingly insufficient to place confidence in inferences without justification. In particular, when considering deployments where models generate high-stakes predictions, sensitive decisions often mandate a sufficient accompanying explanation [42]. Such lack of transparency creates a barrier of understanding and distrust, preventing AI from being widely adopted in the clinic, despite its promising performance [43]. Moreover, explanation methods are essential to validate the models. A certain model may be achieving the right conclusion with astonishing accuracies, but based on the wrong information which has no clinical value. This proves that interpretability matters, and in order to build trust in intelligent systems and move towards their meaningful integration into our everyday lives, it is clear that we must create "transparent" models that are capable of supporting their decision with an explanation [44].

To overcome the need described, a new AI subfield has emerged – eXplainable Artificial Intelligence (XAI). XAI is defined as the set of features that explain how an AI model arrived at its prediction [45]. These features can be seen as explanations, such as images, text, or other visual aid that accompanies a prediction to offer intuition into the underlying reasons for the model output [42], aiming to make its behavior more understandable to humans.

The literature presents different ways of grouping the approaches when trying to explain DL models, from the type of input to the scope, stage, and nature of the explanation [46]. Jeyakumar et al. [42] divides them into three groups according to the nature of the explanation: model-transparent methods, model-agnostic methods, and case-based methods. Model-transparent approaches enhance specific locations that triggered activations within the model's hidden layers, therefore being specific for each model. Model-agnostic techniques do not rely on any specific part of the model architecture and instead aim to understand the connection between the input data

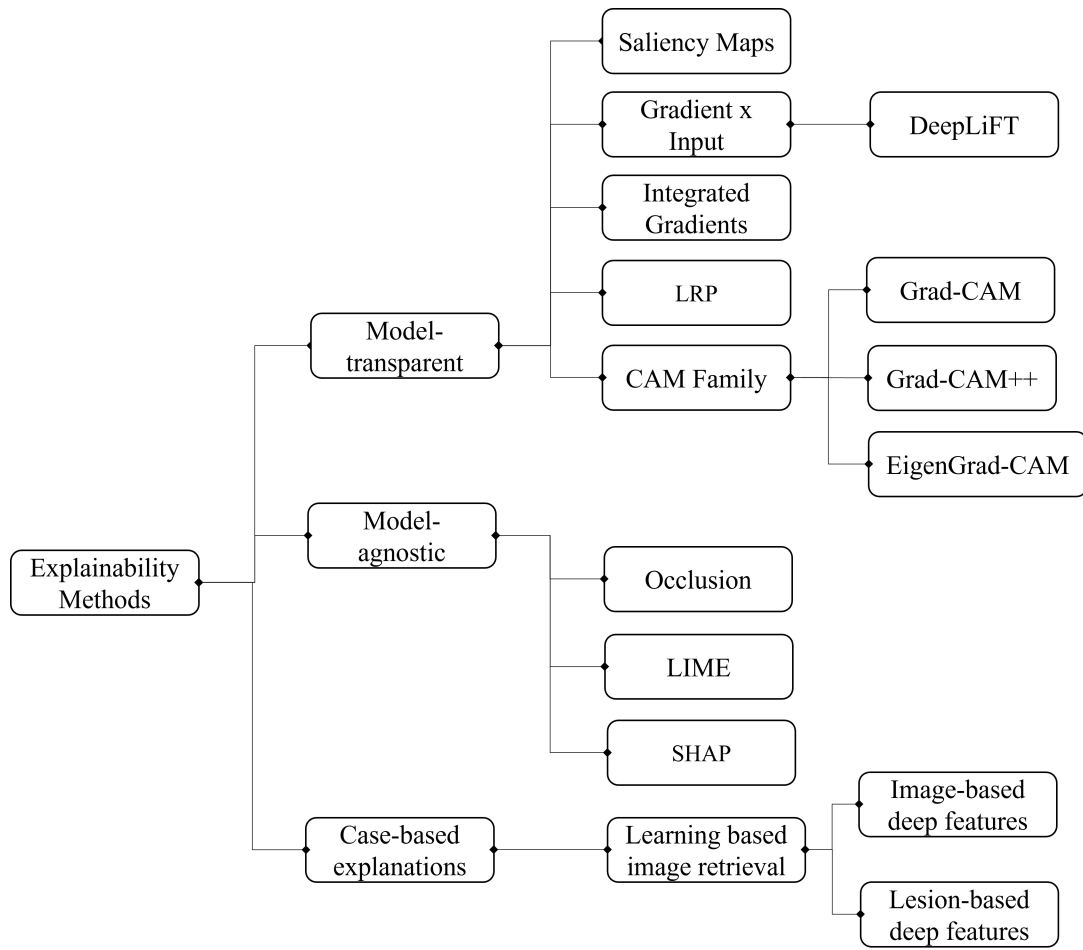


Figure 3.2: XAI methods overview.

and the output prediction by perturbing the input and studying the changes in the output. Case-based methods use examples from the training dataset that appear to portray the same features as the input sample and use them as explanations. There are other types of divisions such as the one proposed by Ras et al. [47], where the methods are divided into visualization, distillation, and intrinsic methods. In addition, the methods can also be divided into *Ante-hoc* approaches, which address the issue of explainability during both the implementation and training stages of the model, while *Post-hoc* methods are applied to an already trained model. However, in this section, the methods will be explained according to the division proposed in [42], and Figure 3.2 provides an overview of the main methods that will be covered.

3.3.1 Model-transparent Methods

Model-transparent methods aim to emphasize the critical regions of the input by creating heatmaps that highlight those regions. Normally, the methods use the information present in the last layer of the CNN architecture, therefore, being specific for each model and dependent on the network

architecture. The first method to be considered is the **Saliency Map** [48] where saliency maps are generated by computing the gradient of the output prediction with respect to the input image and then using these gradients to highlight the pixels that have the greatest impact, therefore quantifying the sensitivity of the network's output with respect to input features.

Gradient x Input

Shrikumar et al. [49] proposed a small variation to improve the sharpness of the attribution maps, by multiplying the gradient of the output with the input and named this variation **Gradient x Input**. Then Sundararajan et al. [50] introduced **Integrated Gradients**, similar to Gradient x Input, which calculates the effect of each input feature on the output by determining the partial derivatives. However, while the Gradient x Input calculates the derivative at a specific input, Integrated Gradients calculates the average gradient over a linear path between a baseline input and the provided input [47].

Deep Learning Important FeaTures (DeepLiFT) [51] is an improved version of the Gradient x Input technique, from the same author, that uses a reference input along with the normal input to explain the model predictions. It assigns a relevance score to each input feature of a neural network, based on the difference in output between an input and a reference input. The difference in output between a certain input and the reference input can be represented by $\Delta t = f(x) - f(x')$, and the difference between the input and reference input, represented by $\Delta x = x - x'$. Then, for each input feature, a relevance score $R_{\Delta x_i \Delta t}$ is assigned, which represents the influence of that feature on the difference in output. The overall difference in output is then computed as the sum of the relevance scores for all the input features that are necessary to compute the output:

$$\Delta t = \sum_i^N R_{\Delta x_i \Delta t} \quad (3.1)$$

The relevance score can be thought of as a weight that represents how much influence each input feature has on the overall difference in output. It has been shown to be a good approximation to Integrated Gradients [52].

When it comes to DeepLiFT applications on CXR, Johnson et al. [53] used it to learn domain knowledge from deep network saliency attributions on cardiomegaly detection, while Mata et al. [54] used it to assess the difference in deep models saliency maps that were regularized with CXR clinical report information, as shown in Figure 3.3.

Class Activation Maps Family

Another typical family of saliency methods is the **Class Activation Maps (CAM)**, first introduced by Zhou et al. [44] after understanding the potential of Global Average Pooling layers for object localization in CNN architectures. To generate the CAM, the feature maps of the convolutional layers are first multiplied by the weights of the last fully connected layer. Then, a global average pooling operation is applied to the resulting weighted feature maps, which generates a single

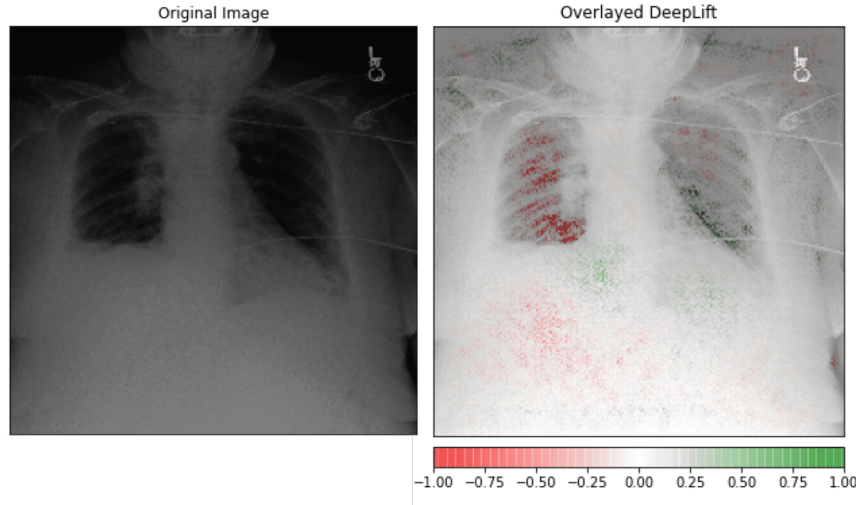


Figure 3.3: Application of DeepLiFT on deep model regularization using CXR reports. Left, original image. Right, original image with overlaid DeepLiFT, from [54].

output map for each class. Finally, the output map that corresponds to the predicted class is thresholded to generate the final CAM.

Grad-CAM [55] is a generalization of CAM, as it uses the gradients of the output of the last convolutional layer with respect to the weights of the last fully connected layer, Figure 3.4. In contrast to CAM, Grad-CAM has wider applicability as it can be used with a broader range of CNNs, as long as the final activation function used for network prediction is differentiable [47, 55].

Grad-CAM authors based their method on a fundamental assumption that the final score Y^c for a particular class c can be written as a linear combination of its global average pooled last convolutional layer feature maps A^k . The weights w_k^c for a particular feature map A^k and class c are defined as:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{i,j}^k} \quad (3.2)$$

where Z is a constant (number of pixels in the activation map).

This approach, however, may fail to properly localize objects in an image if the image contains multiple occurrences of the same class. This is a serious issue as multiple occurrences of the same object in an image are very common in the real world. Another consequence of an unweighted average of partial derivatives is that often, the localization does not correspond to the entire object, but only parts of it. This can hamper the user's trust in the model, and prevent Grad-CAM's premise of making a deep CNN more transparent.

To overcome this problem, Grad-CAM++ [56] applies a more sophisticated normalization, by explicitly coding the structure of the weights w_k^c as:

$$w_k^c = \frac{1}{Z} \sum_i \sum_j \alpha_{ij}^{kc} \cdot \text{ReLU}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right), \quad (3.3)$$

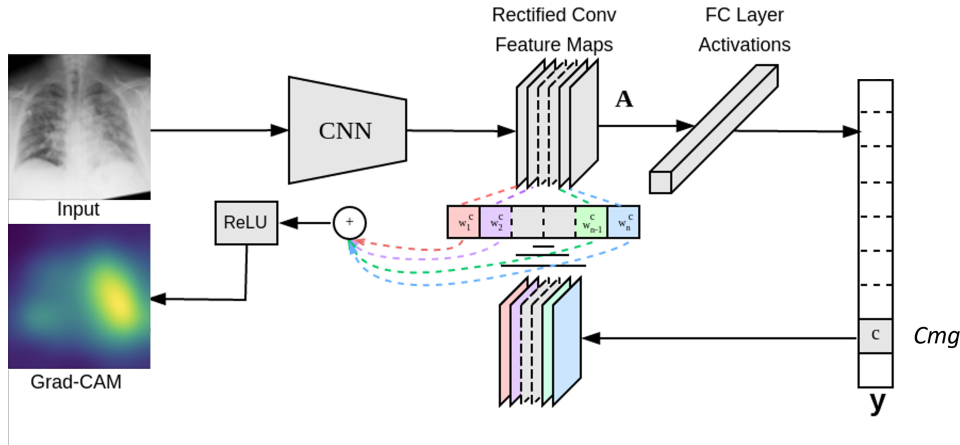


Figure 3.4: Grad-CAM pipeline, adapted from [55].

where the $ReLU$ is the Rectified Linear Unit activation function. Here the α_{ij}^{kc} are weighting coefficients for the pixel-wise gradients for class c and convolutional feature map A^k , defined as:

$$\alpha_{ij}^{kc} = \frac{1}{\sum_{lm} \frac{\partial Y^c}{\partial A_{lm}^k}} \quad \text{if} \quad \frac{\partial Y^c}{\partial A_{ij}^k} = 1, \quad 0 \quad \text{otherwise} \quad (3.4)$$

This way the presence of objects in all feature maps is highlighted with equal importance. w_k^c captures the importance of a particular activation map A^k , and there is a preference for positive gradients to indicate visual features that increase the output neuron's activation, rather than suppress the output neuron's activation.

To conclude the CAM family, it is also worth mentioning the method proposed by Muhammad et al. [57], the **EigenGrad-CAM**. One of the main limitations of Grad-CAM++ is that it uses only the gradient information of the final convolutional layer, while EigenGrad-CAM also uses gradient information from multiple layers before the last one and applies principal component analysis to obtain a more accurate and interpretable heatmap [57].

At the time of writing, Grad-CAM is the most popular model-transparent method used in the medical image analysis field, with application in a variety of imaging modalities. CXR is naturally no exception with multiple studies using Grad-CAM or small variations of it to explain model decisions on COVID-19 screening or classification [58, 59, 60, 61]. Devnath et al. [62] used multiple variations of Grad-CAM to explain a deep model trained to detect Pneumoconiosis, while Suryani et al. [63] used it as a tool to visualize and localize lung tumor on a deep fusion model trained on a dataset including normal CXR images and images with pulmonary tumors. Kusunose et al. [64] applied it to the detection of Exercise-Induced Pulmonary Hypertension (EIPH), as shown in Figure 3.5. On the other hand, Chen et al. [65] proposed a lesion location attention-guided network to focus on the discriminative features from lesion locations for multi-label thoracic disease classification on CXRs and used the Grad-CAM to enhance the results.

Beyond their use as explainability modules in second opinion systems, the use of Grad-CAM to enhance the performance of radiologists has been proposed by Dissez et al. [67], where a full

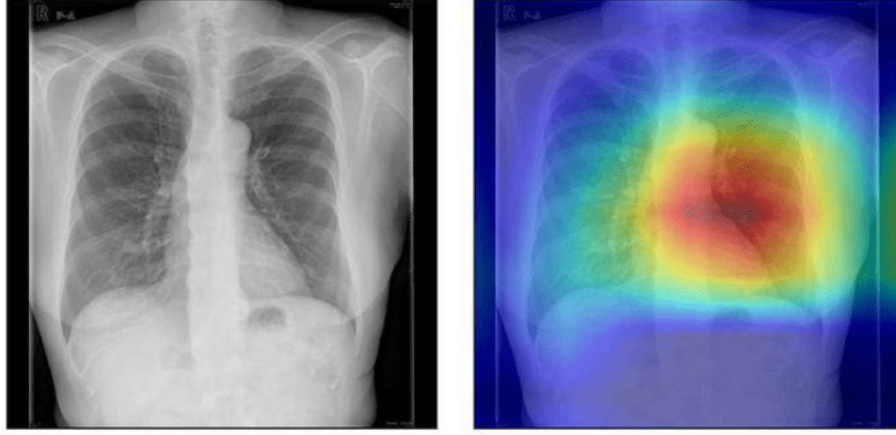


Figure 3.5: Grad-CAM used in EIPH detection, from [66]. Left, original CXR image. Right, Grad-CAM heatmap overlaid on top of the CXR.

end-to-end AI system for lung cancer detection with a built-in Grad-CAM explanation is used by radiologists to improve their performance, achieving an overall increase of 17.4% lung cancers being identified on CXRs.

Layer-wise Relevance Propagation

Some argue that methods that are gradient-based only reflect the output's sensitivity to changes in the input and may not necessarily correspond to the regions from which the network derives its decisions [68]. To overcome this problem, Bach et al. [69] developed the **Layer-wise relevance propagation (LRP)**. LRP methods decompose the output value $f(x)$ of a deep network f across input features $x = (x_1, x_2, \dots, x_N)$, such that $f(x) = \sum_i r_i$ where r_i is the relevance score of feature x_i . The method is based on the fact that f is differentiable and hence can be approximated by a Taylor expansion of f at some root $\hat{x}$ for which $f(\hat{x}) = 0$. Therefore, the relevance score can be written as:

$$r_i = \frac{\partial f}{\partial x_i}(\hat{x}_i)(x_i - \hat{x}_i) \quad (3.5)$$

To extend this idea to a CNN the method considers a conservative decomposition of relevance scores across layers of the network requiring that the relevance score of a node i at layer l , denoted r_i^l , be decomposable into:

$$r_i^l = \sum_j^M r_{i,j}^l \quad (3.6)$$

where M represents all the nodes and, therefore, the summation is done over M in the layer $l + 1$ that node i in layer l is connected to or contributes to. The relevance score of the later layers can be used to calculate the relevance score of the earlier layers by backpropagation.

In comparison to CAM methods, LRP heatmaps tend to be more accurate and detailed because LRP assigns a relevance score to each individual pixel, whereas CAM only examines activation maps in the final layer [47, 45]. It has been used for explaining Alzheimer’s disease classification [68] on MRI images and also for morphological and molecular breast cancer profiling [70] using histological images. Particularly in CXR, Alam et al. [71] investigated the applicability of LRP and Grad-CAM on the CXR multi-label dataset CheXpert [72], shown in Figure 3.6. Karim et al. [73] proposed an automatic detection system trained on CXR images with covid-19, pneumonia, and normal cases, and on top of the abnormality detection, LRP is used to explain the model decision.

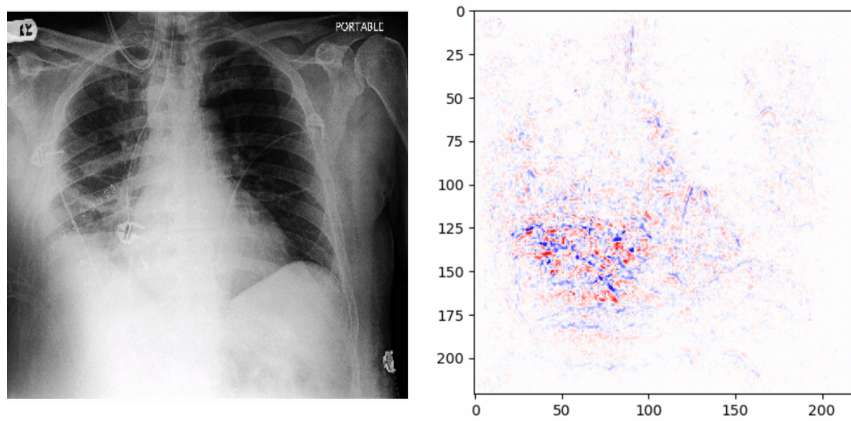


Figure 3.6: Example of LRP used to explain a correct classification of pleural effusion, from [71]. Left, is the original CXR image. Right, LRP heatmap for the pleural effusion class.

3.3.2 Model Agnostic Methods

Model-agnostic techniques do not rely on any specific part of the model architecture, which means they do not consider weights or structural information [46] and instead aim to understand the connection between the input data and the output prediction, and are therefore able to be applied to any black-box model.

Occlusion

One of the earliest and simplest interpretability techniques for image classifications, proposed by Zeiler et al. [74], is the **Occlusion** method. The key idea involves occluding parts of the input with a grey square to observe how the model’s prediction changes as the patch are placed over different regions [47]. If the model’s performance decreases when certain areas are covered, those areas are likely to contain important information. The differences in performance can be associated with areas (or pixels) and a sensitivity heatmap can be created using that information. Islam et al. [75] used Occlusion to interpret the results obtained after training DL models on abnormality detection tasks, using the location capabilities of the method to explain the performance on classes such as

cardiomegaly and pulmonary edema. Nigri et al. [76] proposed an explanation method that is based on Occlusion to interpret the results of models trained to diagnose Alzheimer's diseases.

Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME) is a popular perturbation-based method created by Ribeiro et al. [77]. To generate explanations, LIME alters the input by introducing noise to continuous features, removing words (in natural language processing (NLP) datasets) or parts of images [78], and then trains an interpretable model, such as a linear regression model, on the modified input data. It then explains the predictions of the original classifier by showing the contribution of each perturbed feature to the interpretable model's predictions.

LIME defines an interpretable model g from a class of inherently interpretable models with a different domain that approximates f well in the local area around x . The model g operates over an interpretable representation x_0 of the input data presented to the unexplainable model f , which, in image examples, is normally a binary vector denoting if a certain pixel or color pattern exists in an image input. It is common to write the linear model g as $g(z') = w_g z'$, where z' can be given by a perturbed sample. To fit the model one needs first to get the predictions for all the perturbations and besides that, as different perturbations will have different weights according to their similarity with the original image, one also needs to calculate the model starting weights. These weights are acquired after calculating the cosine distance between each prediction vector and a non-perturbed vector, and using the cosine distance-vector output and passing it to the kernel function $\pi_x(z)$:

$$\pi_x(z) = \exp(-D(x, z)^2 / \sigma^2) \quad (3.7)$$

where the exponential kernel is defined on some distance function D - cosine in this case - with width σ , representing a measure of proximity between the perturbation z and the original perturbation vector, x . The weights of the fitted model translate the contribution of each element in the perturbed array to the prediction. A downside, however, is the time needed. As the model creates perturbed samples and classifies them, it is low for interactive use with complex models.

LIME has been used to help explain the results of a DL model trained to detect early Parkinson's Disease [78] as well as explain the prediction of depressive symptoms in textual datasets [79]. Regarding CXR applications, different studies have applied LIME on models trained on abnormality detection, namely pneumonia by Schneider et al. [80] and Rajaraman et al. [81], while [82, 83, 84] focus their LIME explanations only in COVID-19, Figure 3.7.

Shapley Additive Explanations

Another state-of-the-art agnostic method is **SHapley Additive exPlanations (SHAP)** [85] which is based on the concept of Shapley values from cooperative game theory. Shapley values provide a way to fairly distribute a value among a group of players based on their contributions to the outcome of the game [86]. In the context of XAI, Shapley values are used to attribute feature importance in a prediction by a model and are calculated similarly. SHAP assigns a value to

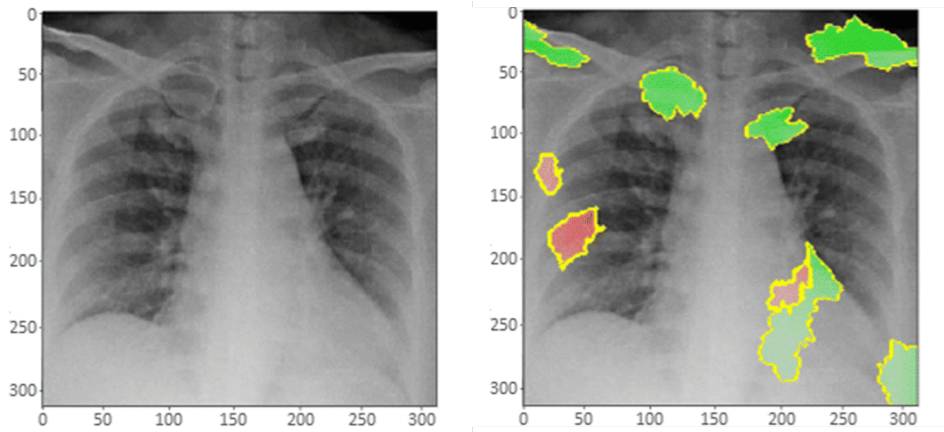


Figure 3.7: Example of LIME explanation of COVID-19 classification, adapted [84]. Left, is the original CXR image. Right, LIME super-pixels overlaid on top of the original image.

each feature, which represents the average contribution of that feature to all possible coalitions of features. This way, it takes into account the interactions between features and assigns a relevance value to each one based on its contribution to the final prediction. In the healthcare environment it is used, for example, to analyze the contribution of clinical features to assess the risk of in-hospital mortality [87], dementia [88], hepatitis [89] and stroke [90]. Bhandari et al. [91] used it to explain results obtained by a model trained to detect multiple pathologies, as shown in Figure 3.8, while Ong et al. [92] applied only on COVID-19 screening in a study comparing it to LIME.

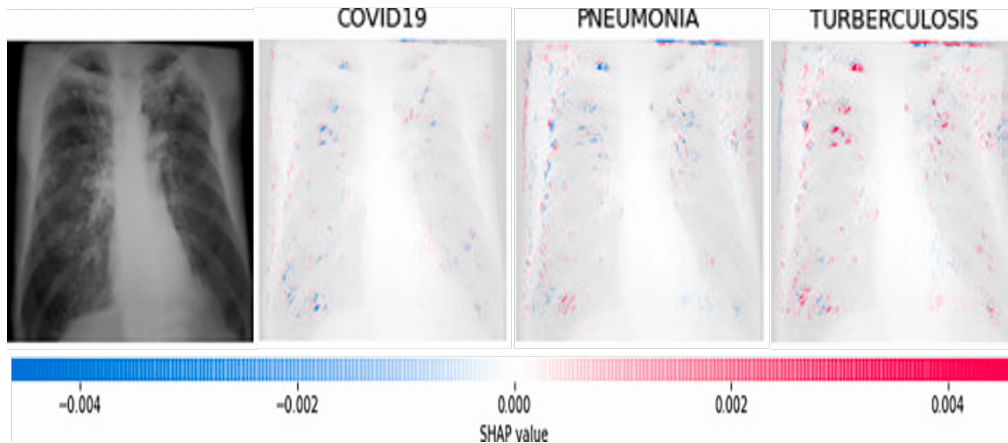


Figure 3.8: SHAP explanation on multiple pathology detection, adapted from [91]. The first image is the original CXR image, followed by CXR images for each labeled class with the SHAP heatmap overlaid on top of the original image.

3.3.3 Case-based Explainability

The third group is inspired by cognitive science and the way humans learn based on past experiences and examples. A Radiologist, for example, goes through samples of the database looking for cases that are similar to a new one in analysis because they know that visiting a solved example with similar characteristics normally helps solve new cases that share, at least, part of those characteristics. In the same way, Case-based methods utilize examples from the training dataset to try to understand the relationship between the input and the training data that influenced the model's decision.

For a given problem, such as image classification, the most similar previous examples are retrieved before being used to make a prediction [93], as shown in Figure 3.9. The retrieval process typically involves matching the features of the current problem with the stored cases using a predetermined distance function to find the nearest neighbors. The success of retrieval is highly influenced by the assigned weights of the features, which indicate their significance in the specific domain. These features can be clinical data measures, such as hypertension or age, or patterns in images as well as specific word arrangements in NLP cases. Therefore, the set of features extracted for each data sample must be an element of discrimination of its particular class in the best possible way as they are what determines the order of the cases retrieved [93].



Figure 3.9: Example-based-explanations on ship image, from [42].

In cases where the dataset consists of clinical data and sequential data, the features are straightforward. However, when dealing with imaging, selecting these representative features is not trivial. Early methods heavily relied on features that were manually crafted, a task that was both tedious and required specialized knowledge. Subsequently, metrics that attempt to align the similarity of two images with human perception have gained considerable attention, with the most prevalent being the Structural Similarity Index (SSIM). In recent years, after the proven success of CNN architectures in detecting patterns in images, the community started to look at image features as a non-linear representation that can be fetched from the last convolutional layer [42, 94]. Thus, in the presence of annotated datasets with enough representation of the different classes, the feature-extraction phase can be automated, working as the foundation of Content-Based Image Retrieval (CBIR) systems. This type of CBIR system creates a library of features that were extracted from

the penultimate layer, before classification, of a pre-trained model, and uses them to search for and retrieve similar images, as shown in Figure 3.10.

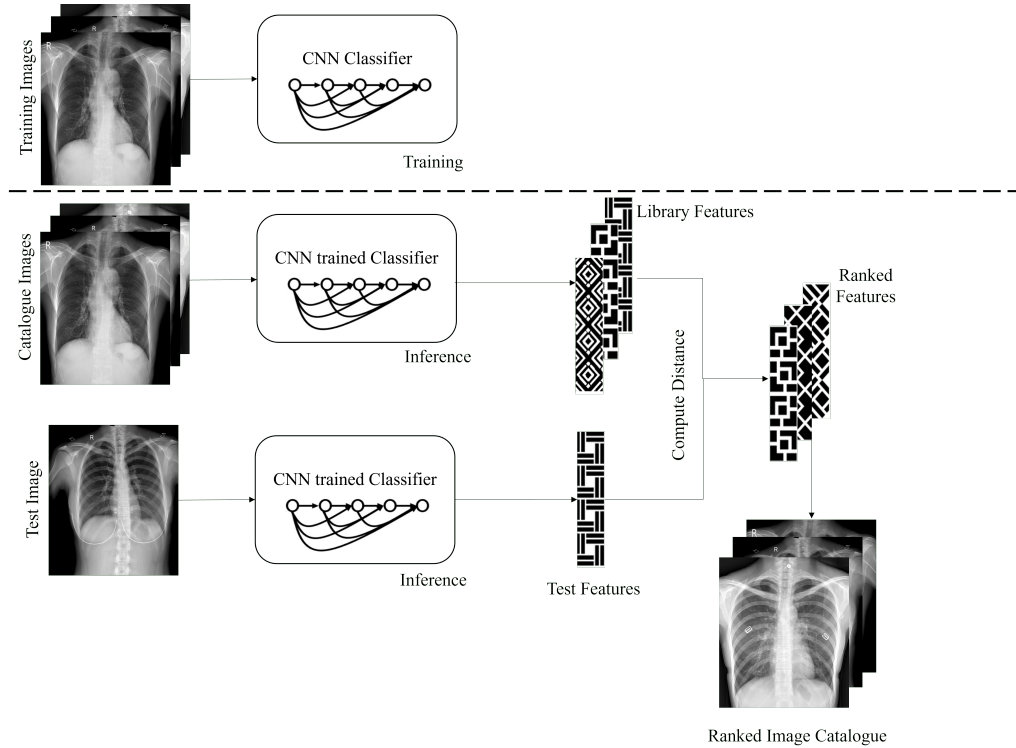


Figure 3.10: An overview on Image-based retrieval systems on CXR images.

Jeyakumar et al. [42] developed an open-source implementation of an explanation-by-example system, *ExMatchina* and performed an experiment with 455 individuals where they applied various XAI techniques, such as Grad-CAM++, LIME, SHAP and *ExMatchina* on datasets containing different types of data (text, images, audio and sensor data) and concluded that when available, the explanation based on examples methods was largely preferred, with the participants selecting it more than 70% of the times.

Regarding CXR applications, Silva et al. [94] assessed the quality of the retrieval by comparing the results obtained when using the SSIM and when using deep-learned features extracted using saliency maps as an attention mechanism to focus the feature representations to image regions that characterize the class to which they belong - **Image-based deep features**. To do so, the authors first trained a CNN to detect Pleural Effusion and then fine-tuned it with the saliency maps. The results presented showed that this approach leads to an improvement in medical image retrieval, with the most significant improvement being related to the ranking quality of the retrieval. Zhong et al. [95] also created a system for retrieving CXR images that utilize a spatial attention component integrated alongside the feature extraction component, resulting in disease-specific image embeddings for patients that proved highly efficient in the CXR retrieval.

Pedrosa et al. [3] proposed an automatic image retrieval method based on the YOLOv5 for the

detection of multiple lesions in CXRs images - **Lesion-based deep features**. The paper demonstrates that the method proposed can effectively identify and extract features from lesions, accurately describing their high-order characteristics, and enabling the retrieval of lesions of the same pathological class. Additionally, the methodology was found to be nearly twice as effective as SSIM, according to the evaluations conducted by radiologists. Figure 3.11 compares the lesion-based retrieval method with SSIM and a random extraction.

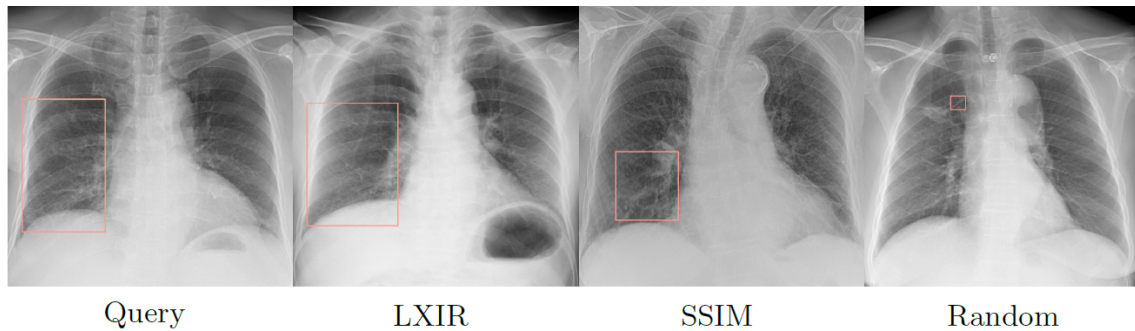


Figure 3.11: Example-based explanations on CXR images, comparing different retrieval methods, from [3]. LXIR stands for lesion-based CXR image retrieval.

However, explanations based on examples can also be used the other way around, by presenting counter-examples to a query test, instead of the samples that are the most similar. In both cases, one may argue that these methods make use of data (mainly images) that belong to some patients and, therefore, fall under data protection policies. The topic of data protection in the medical field is not a novel one, however, with the expansion of digital healthcare systems, it has become a particularly pressing concern [96]. Therefore, one possible solution is the development of artificially created images that are capable of reliably representing the different pathologies. This way, the generated images would have no "owner" associated with them but would still carry the feature that represents the different pathologies.

3.3.4 Evaluation of Explainability Methods

The previous section covered a variety of XAI techniques that provide insights into the model behavior or a justification for its output. However, in spite of the number of available methods, little attention has been devoted to comparing different explainability methods in real cases and validating which method(s) provide more precise and complete explanations. This is partly due to the subjective nature of explainability but also due to the fact that different methods provide different outputs making quantitative comparisons challenging.

Sporta et al. [97], on "Benchmarking saliency methods for chest X-ray interpretation", developed a new standard for comparison of XAI methods applied to CXR multilabel classification. Using the CheXpert [72] dataset that contains manual segmentations of different pathology made by expert radiologists, they tested different DL architectures and a variety of XAI methods based on saliency maps such as Grad-CAM, DeepLiFT, and LRP, as shown in Figure 3.12. In their

work, two important metrics were used to quantitatively capture the localization accuracy of the explainable methods: mean Intersection over union (mIoU), which measures the overlap of the method segmentation, and the expert segmentation and the Hit rate, which assesses if the point in the saliency map with the highest probability falls within the boundary of the expert segmentation. It is a starting point for comparing the methods, but datasets with radiologist-annotated segmentations are scarce.

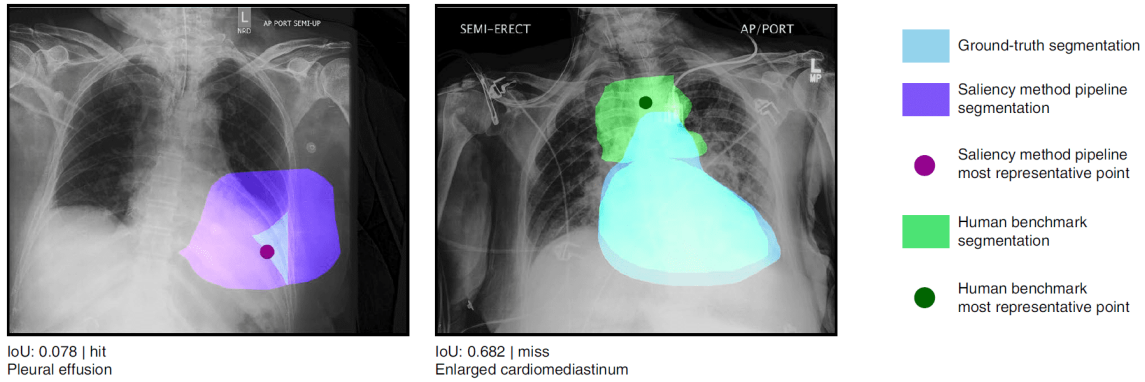


Figure 3.12: Example of Saporta et al. [97] benchmark study and comparison metrics. Left, CXR with ground-truth and saliency method annotations for pleural effusion. The segmentations have a low IoU of 0.078, but it was a "hit" because the most representative point identified by the saliency method is within the ground-truth segmentation. In contrast, in the case of an enlarged cardiomeastinum, the segmentations have a high IoU of 0.682, but it was not a successful "hit" because the most representative point identified by the saliency method is outside of the ground-truth segmentation.

Tables 3.1 and 3.2 display the results obtained from their best-performing model, DenseNet121, using various visual explanation methods and considering different classes within the CheXpert dataset, for both IoU and Hit rate metrics. While most existing literature uses visual explanations only to supplement specific findings, this study uniquely introduces a comparative analysis of their explanatory abilities.

Table 3.1: Sporta et al. [97] results for the IoU metric.

Pathology	Grad-CAM	Grad-CAM++	IG	Eigen-CAM	DeepLIFT	LRP	Occlusion
Airspace opacity	0.248	0.234	0.123	0.293	0.111	0.112	0.242
Atelectasis	0.254	0.245	0.116	0.267	0.126	0.109	0.250
Cardiomegaly	0.452	0.346	0.160	0.379	0.167	0.150	0.312
Consolidation	0.408	0.296	0.177	0.332	0.088	0.099	0.212
Edema	0.362	0.388	0.073	0.370	0.059	0.047	0.347
Enlarged cardiom.	0.379	0.400	0.154	0.372	0.109	0.117	0.363
Lung lesion	0.101	0.089	0.107	0.089	0.072	0.088	0.087
Pleural effusion	0.235	0.195	0.088	0.249	0.090	0.082	0.250
Pneumothorax	0.213	0.216	0.077	0.218	0.084	0.066	0.214
Support devices	0.163	0.133	0.099	0.116	0.086	0.052	0.126
Overall mean	0.282	0.254	0.117	0.269	0.099	0.092	0.240

Table 3.2: Sporta et al. [97] results for the Hit rate metric.

Pathology	Grad-CAM	Grad-CAM++	IG	Eigen-CAM	DeepLIFT	LRP	Occlusion
Airspace opacity	0.498	0.558	0.606	0.566	0.528	0.566	0.367
Atelectasis	0.501	0.620	0.520	0.530	0.415	0.468	0.343
Cardiomegaly	0.903	0.732	0.697	0.709	0.610	0.644	0.515
Consolidation	0.738	0.708	0.624	0.626	0.571	0.283	0.338
Edema	0.746	0.781	0.300	0.758	0.468	0.156	0.469
Enlarged cardiom.	0.818	0.630	0.704	0.612	0.469	0.594	0.767
Lung lesion	0.290	0.290	0.423	0.146	0.497	0.356	0.072
Pleural effusion	0.507	0.347	0.332	0.439	0.408	0.283	0.291
Pneumothorax	0.392	0.489	0.801	0.195	0.801	0.697	0.297
Support devices	0.355	0.364	0.491	0.216	0.598	0.264	0.189
Overall mean	0.575	0.552	0.550	0.480	0.537	0.431	0.365

Moreover, despite their value and potential, most of the current techniques for visual explanations in medical imaging still face challenges in aligning with physicians’ day-to-day diagnostic practices, which often adhere to structured layouts and specific guidelines, as previously discussed. For instance, when a radiologist needs to report the presence of a pleural effusion CXR, they typically clarify the location, specifying the lung involved, and in some cases, providing further details like the superior right lobe. In this context, receiving direct information on the precise location of the finding holds greater relevance than visual aids enhancing the area in question. Thus, ensuring that the explanations align with established medical practices becomes crucial to their successful integration into the diagnostic workflow.

Chapter 4

Materials and methods

This chapter provides an in-depth overview of the CXR datasets used in this study. It further covers the explanation of the underlying architectures used to develop models dedicated to CXR pathology classification and subsequent extraction of explanations.

4.1 Datasets

A variety of datasets for CXR classification are available online. However, only a small percentage of these datasets come with spatial annotations of the pathologies, primarily due to the highly time-consuming nature of the task. The datasets used in this study are summarized in Table 4.1 and publicly available online, which facilitates reproducibility and further comparison with other studies.

Table 4.1: Datasets information.

Dataset	Year	Images	Bounding box annotated	View	Labels	Annotations
VinDr-CXR [40]	2020	18,000	15,000	PA	14	manual
ChestX-ray14 [98]	2015	112,120	880	PA/AP	14	NLP

It is important to highlight the contrasting labeling processes employed for both datasets. The VinDr-CXR dataset was manually labeled by multiple experienced radiologists. On the other hand, the ChestX-ray14 dataset utilized NLP techniques for labeling. Furthermore, while both datasets share the same number of labels per dataset (14), it is important to note that the labels themselves are not the same, as detailed in subsequent sections.

4.1.1 VinDr-CXR

In the literature, the VinDr-CXR dataset [40] is the largest CXR dataset available with spatial annotations, created by the VinBigdata Institute. The dataset consists of 18,000 PA CXRs with 15,000 that were manually annotated by different radiologists using a web-based framework called VinDr Lab [99].

The CXR annotation includes the location of each finding through bounding boxes, where each bounding box is assigned a pathology class from 14 possible labels: Aortic Enlargement (AoE), Atelectasis (Atl), Calcification (Clc), Cardiomegaly (Cmg), Consolidation (Cns), Infiltration (Inf), Interstitial Lung Disease (ILD), Lung Opacity (LOp), Nodule/Mass (Nod), Other Lesion (OtL), Pleural Effusion (PLE), Pleural Thickening (PIT), Pneumothorax (Pnm) and Pulmonary Fibrosis (PuF). The images, originally in DICOM format, have resolutions varying from 823 to 3400 pixels. Figure 4.1 shows examples of images present in the dataset with the radiologist's annotations.

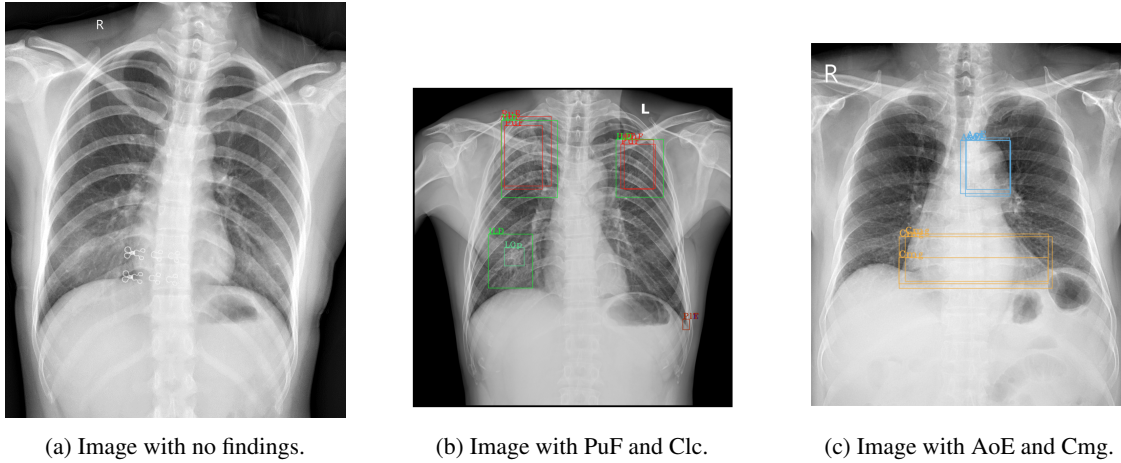


Figure 4.1: Examples of VinDR-CXR dataset images with radiologist's annotations, from [40].

In addition, the histograms present in Figure 4.2 illustrate the number of samples of each class in the dataset, including the "No findings" label that refers to the images with no pathology and the frequency of pathologies per image, respectively.

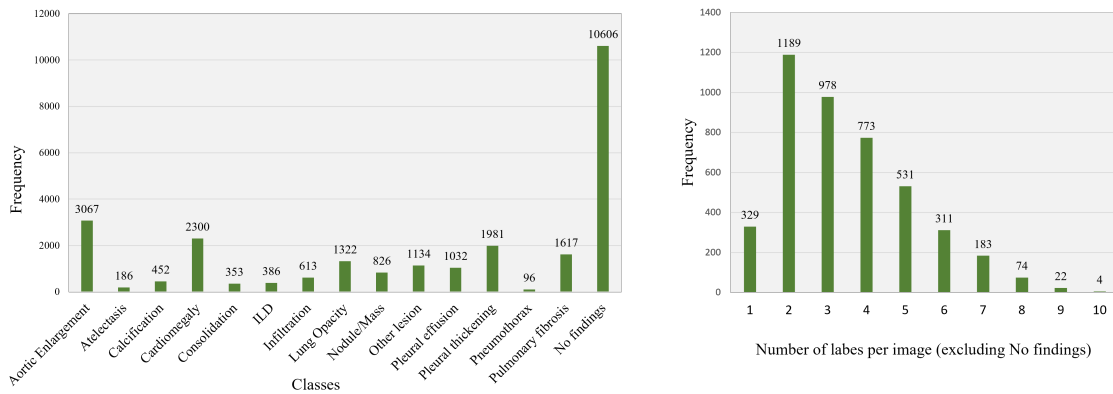


Figure 4.2: Class frequencies in the dataset and number of labels per image.

4.1.2 ChestX-ray14

The ChestX-ray14 [98] dataset, also referred to as CXR-14, is a very well known dataset in the field of medical imaging. It stands as one of the largest datasets, comprising 112,120 images ob-

tained from over 30,000 unique patients. The dataset encompasses a diverse set of 14 pathological classes and is an extension of the ChestX-ray8, which contains images collected from 1992 to 2015. These classes include Pleural Thickening (PIT), Mass (Mas), Pneumonia (Pno), Effusion (Efu), Hernia (Her), Emphysema (Emp), Cardiomegaly (Cmg), Infiltration (Inf), Nodule (Ndl), Edema (Edm), Pneumothorax (Pnm), Atelectasis (Atl), Consolidation (Cns), Fibrosis (Fib), as well as normal images exhibiting no findings. The histograms present in Figure 4.3 highlight the different label counts, as well as the number of labels per image.

Over the years, the ChestX-ray14 dataset has emerged as a valuable resource, supporting a wide range of research endeavors focused on automatic disease detection and classification from CXR images. Within this extensive collection, there are 67,310 images captured in the PA view. Only these images will be used for the upcoming experiments to allow for result comparisons with the VinDr-CXR dataset, which exclusively consists of PA images.

Most of the labeling was acquired from radiological reports via NLP techniques. For this particular dataset, only 880 images have bounding box annotations, of which 529 are PA.

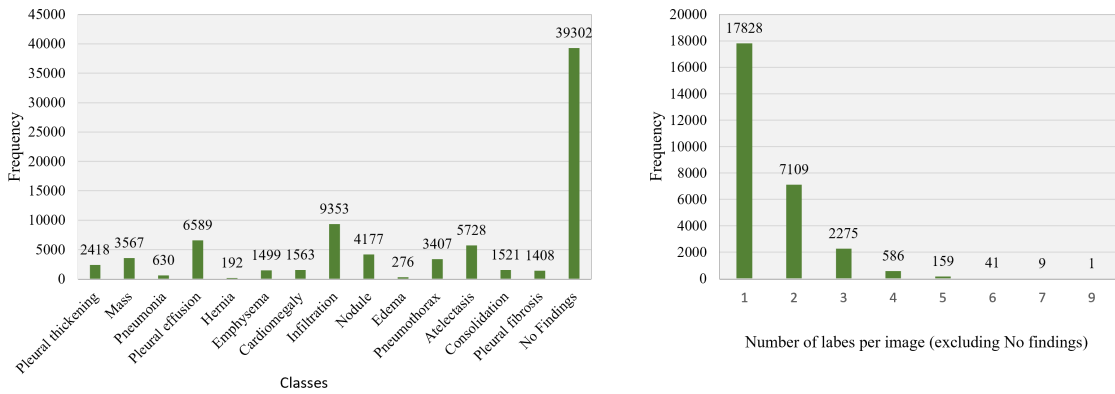


Figure 4.3: Class frequencies in the dataset and number of labels per image.

4.1.3 VinDr-CXR and CXR14 Label Match

In this study, various comparisons will be made using the two different datasets. Therefore, it is crucial to be able to map the labels of each dataset to the other whenever possible. To ensure accurate matching of the different labels, the assistance of a professional radiologist was sought, who possesses an understanding of the diverse pathological terms, their multiple denominations, and groups.

Table 4.2 shows the label conversion between the two datasets, VinDr-CXR and ChestX-ray14. Each row represents a label from the VinDr-CXR dataset, while the corresponding label from the ChestX-ray14 dataset is shown in the adjacent column. The VinDr-CXR dataset includes labels such as AoE, LOp, Clc, and ILD, which do not have direct matches in the ChestX-ray14. In addition, labels such as Cmg, Inf, PuF, Atl, PIT, PIE, and Pnm have a one-to-one correspondence between the two datasets, as the labels remain the same. The Nod label in the VinDr-CXR dataset is split into two separate labels in the ChestX-ray14 dataset, Mas and Ndl, while the Cns in the

Table 4.2: Label matching between datasets.

VinDr-CXR [40]	ChestX-ray14 [98]
AoE	-
Cmg	Cmg
LOp	-
Inf	Inf
Clc	-
PuF	Fib
ILD	-
Atl	Atl
Nod	Mas Ndl
Cns	Cns Pno
-	Edm
PIT	PIT
PIE	Efu
Pnm	Pnm
OtL	Her Emp

VinDr-CXR dataset match both Cns and Pno labels in the ChestX-ray14 dataset. The same goes for the OtL label in the VinDr-CXR dataset, being associated with both Her and Emp labels in the ChestX-ray14 dataset. The Edm label in the ChestX-ray14 dataset does not have a corresponding label in the VinDr-CXR dataset.

4.2 Models and Training Details

As this work is focused on explaining the results of deep learning models, different state-of-the-art architectures were tested. All models were trained on a computer with the following specifications: Intel Core i9-10900 @ 2.80GHz, NVIDIA TITAN RTX, and 128 GB of usable RAM.

The training hyperparameters included a batch size of 32, a maximum of 100 epochs with early stopping if there were no loss improvements after 10 epochs, and a learning rate of 0.001. These hyperparameters were chosen based on preliminary experiments, which also involved testing other options. Additionally, the optimizer used was the Adaptive Moment Estimation optimizer, commonly known as Adam, due to its efficiency in the training process. For the loss function, Binary Cross-Entropy was employed. This loss function allows the model to interpret the classification of each label individually and then concatenates the output for all classes to form a single output per image. The entire implementation was carried out using the PyTorch library [100], a powerful tool for training deep learning models.

4.2.1 ResNet50

The ResNet [101], an advanced image classification family of networks, was developed in 2015 to address the challenge of vanishing gradients. As neural networks become deeper with multiple layers, effectively propagating gradients throughout the network during training becomes increasingly difficult. This issue often leads to a performance decline as network depth increases, contradicting the expectation that deeper networks should yield better results. To overcome this obstacle, ResNet introduces a unique approach called residual learning.

Instead of directly learning the mapping functions between input and output, ResNet focuses on acquiring knowledge about residual functions. These functions capture the discrepancies between the input and output, allowing the network to learn the residual mapping more efficiently [101]. ResNet achieves this by incorporating residual connections, also known as skip connections. These connections act as shortcuts, enabling the direct flow of information from earlier layers to later layers within the network. By doing so, ResNet facilitates a better approximation of the residual mapping, as depicted in Figure 4.4.

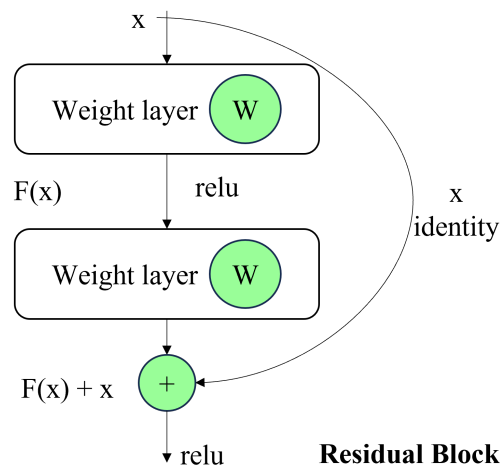


Figure 4.4: Overview of the ResNet50 residual block.

ResNet architectures have demonstrated remarkable success across various computer vision tasks, including image classification, object detection, and semantic segmentation. For this case, the most common ResNet was selected, the ResNet50, where 50 stands for the total number of layers.

4.2.2 DenseNet121

The DenseNet [102] was introduced in 2016 and has since become a widely used architecture in deep learning, particularly in computer vision tasks. This family of network architectures aims to deepen the neural networks while simultaneously enhancing their training efficiency. It achieves this by leveraging feature reuse and incorporating shorter connections between layers.

Unlike traditional CNNs, which restrict each layer's access to features from only the immediately preceding layer, DenseNet introduces a unique concept. It combines feature maps from all preceding layers within a dense block, enabling each layer to utilize the features extracted from all previous levels.

In this architecture, every layer is connected to deeper layers in the network, facilitating maximum information flow. For instance, the first layer is connected to the 2nd, 3rd, 4th, and subsequent layers, while the second layer is connected to the 3rd, 4th, 5th, and so on. This interlinkage promotes effective communication within the network, allowing the 'i' layer to receive inputs from 'i' previous layers and incorporate feature maps from all preceding convolutional blocks, as illustrated in Figure 4.5. Considering the task at hand, the DenseNet121 model with a total of 121 layers was chosen. Due to its substantial size and a large number of training images, the model's pre-trained weights from the ImageNet dataset [103] were utilized.

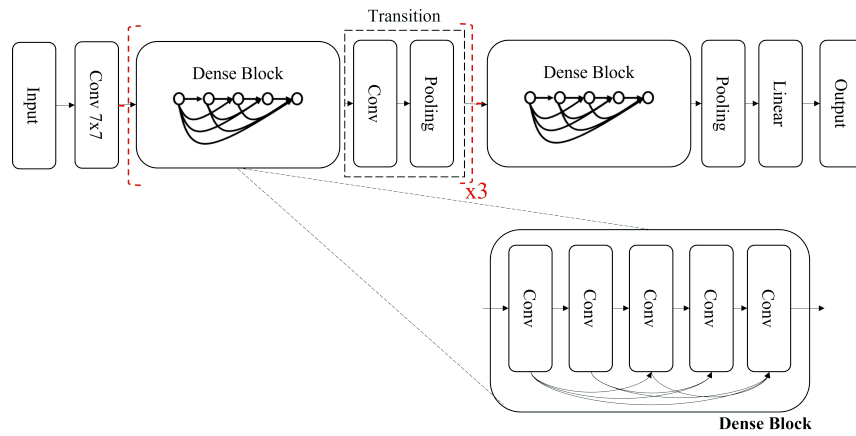


Figure 4.5: Overview of the DenseNet architecture and the dense block.

4.3 Explainable AI

4.3.1 Visual based Explainability

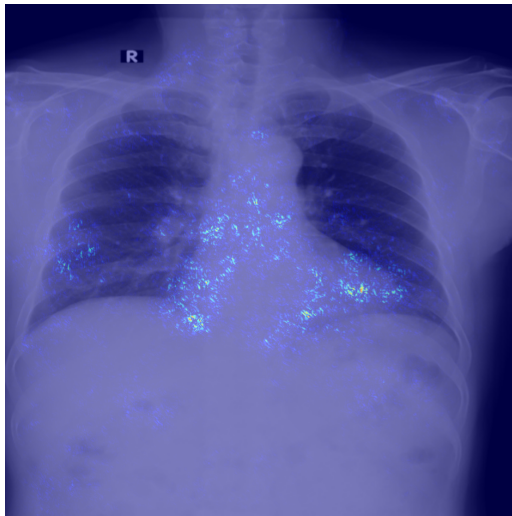
From the different visual explanations previously presented, this work is going to focus on model-transparent explanations. Model-agnostic techniques were considered, but due to the time constraint that they require, they ended up being left out.

Therefore, the GradCAM, Grad-CAM++, EigenGrad-CAM, Saliency maps, LRP, and DeepLift methodologies were selected and implemented. The CAM methods were based on PyTorchGrad python library [104] while the gradient-based explanations (Saliency, LRP, and DeepLift) were based on the Captum library [105], both famous libraries for visual explanations and model interpretability. Each visual explanation was acquired for individual pathologies, focusing exclusively on scenarios where the model made accurate classifications (true positives only).

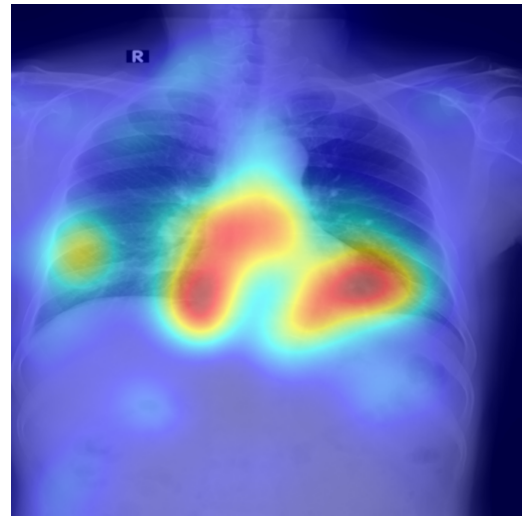
When implementing CAM, apart from specifying the model, one must also select the layer from which the maps are going to be extracted. The choice of layer depended on the specific model

architecture, typically opting for the last layer before flattening the output to obtain predictions. For example, in DenseNet121, the last layer of the fourth dense block was selected, whereas, for ResNet50, the last layer of the layer4 group was chosen.

In the case of gradient-based methods, no specific post processing needed to be met. However, the resulting heatmaps underwent a Gaussian filter operation, using a size equivalent to one-quarter of the original image dimensions. This additional step aimed to enhance the quality and smoothness of the heatmaps, as shown in Figure 4.6, which will greatly help to make more suitable comparisons with the CAM-based heatmaps.



(a) Example of Saliency map without post-processing.



(b) Example of Saliency map with Gaussian filter applied.

Figure 4.6: Difference between saliency maps before and after applying the Gaussian filter.

4.3.2 Structured Report based Explanation

Radiologists play a crucial role in analyzing medical images and documenting their observations through free-text notes during routine clinical care. However, there is no universal standard for CXR reports, and they can vary even within the same hospital, as they are subjective to each individual radiologist's expertise, as detailed in section 2.2.2.

The next section will focus on proposing a structured report that does not follow a free-text approach, but rather a fixed-entry approach which implies standardization. By looking at different written reports and using the labels present in the VinDr-CXR dataset, one can summarize the report entries in Table 4.3.

Now that a structure has been established, the main objective is to effectively convey the information about the positions of the pathologies in the report. This information is presented with reference to the lungs' position, as it is usually done in written reports. To determine the lungs' position, the subsequent step involves segmenting the CXR using a U-Net model developed by [106] to isolate the lungs. The ultimate goal is to compare the bounding box annotations' positions, which provide the location of the pathologies, with their position within the lungs.

Table 4.3: Structured report entries.

Entry	Labels
Cardiac Pathologies	AoE and Cmg
Parenchymal Lesions	LOp, Atl, Cns, Inf, ILD, Nod and PuF
Pneumothorax	Pnm
Pleural Lesions	PIE and PIT
Other Lesion	OtL

As the segmentation model expects rectangular images, the CXR images are padded instead of stretched to avoid any potential deformation of anatomical structures. After the segmentation, post-processing techniques are applied to retain only the two largest areas, representing the lungs, and remove any potential missegmentations. Figure 4.7 provides a visual representation of the segmentation process.

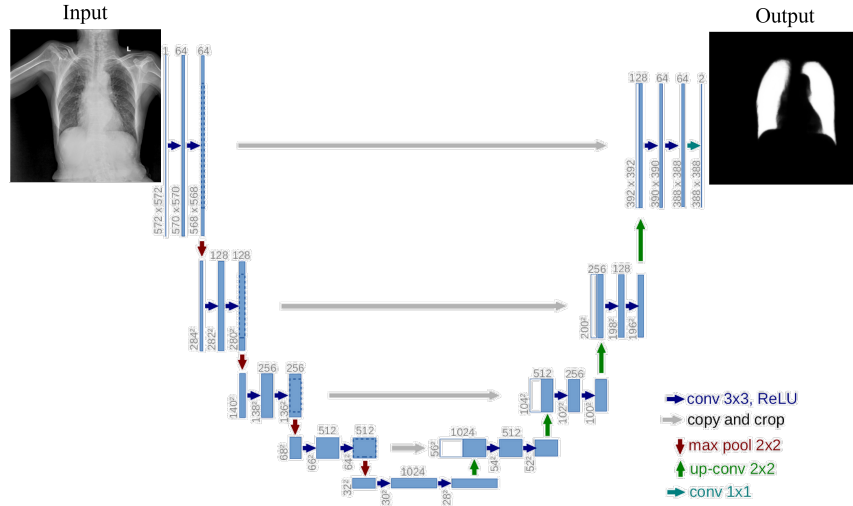


Figure 4.7: CXR segmentation process using U-Net.

After segmenting the lungs, they can be divided into six main zones. However, it is common for pathologies to occupy an extensive area of the lungs, meaning that the merging of vertical zones across each lung must also be considered. For instance, zones 1 and 3, representing the Upper Right and the Middle Right, can combine to form the Top Middle Right zone. Similarly, other possible combinations are illustrated in Figure 4.8.

Finally, in order to determine if a given bounding box of a certain pathology is considered to be inside of a given zone, the following equation is applied:

$$(h_{zone,bb} > \min(\frac{h_{zone}}{2}, \frac{h_{bbox}}{2})) \quad \wedge \quad (w_{zone,bb} > \min(\frac{w_{zone}}{2}, \frac{w_{bbox}}{2})) \quad (4.1)$$

where $h_{zone,bb}$ stands for the intersection of heights between the zone under analysis and the bounding box of the pathology, h_{zone} stands for the height of the zone and h_{bbox} corresponds to the height

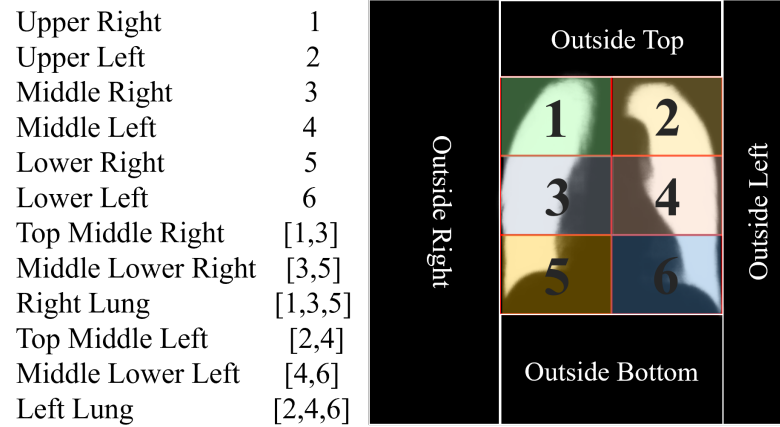


Figure 4.8: Description of the lung zones and possible combinations.

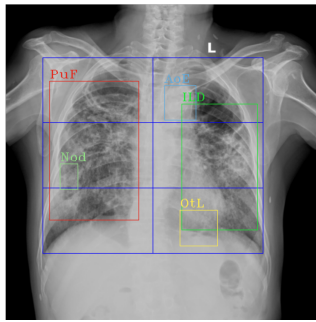
of the bounding box. Similarly, $w_{zone,bb}$, w_{zone} and w_{bbox} denote the intersection of widths, width of the zone, and width of the bounding box, respectively.

For bounding boxes that have sizes larger than the zone, determining the presence of pathology within a specific zone is based on the proportion of the intersected height compared to the height of the zone itself. Conversely, if the bounding box is too small, the comparison is made with the height of the bounding box rather than the zone. The second condition applies the same reasoning but for the width instead of the height. These two conditions allow us to fully cover the different scenarios that arise from the bounding box position and dimensions.

After verifying the presence of the pathology's bounding box in each individual zone, if two or more zones contain the pathology, they are merged. Figure 4.9 displays two examples of a structured report along with the corresponding image that serves as its origin. These images were randomly selected from the VinDr-CXR dataset, where the locations of pathologies are indicated by bounding boxes.

For entries associated with multiple pathologies, such as Parenchymal Lesions, the different pathologies are specified in the report. Otherwise, only the number of pathologies and their position are referenced. The Cardiac Pathologies entry does not require a location since these pathologies have fixed positions.

In contrast to a typical heatmap, this method of organized reporting significantly improves interpretability. By presenting findings in a structured manner that mirrors the format of traditional medical reports, it offers a level of familiarity and standardization that is often lacking in other visualization approaches. This approach has the added benefit of rendering the reports more accessible and comprehensible to a broader audience, making them not only easier for experts to grasp but also more straightforward to communicate to colleagues.



Cardiac Pathologies: AoE.

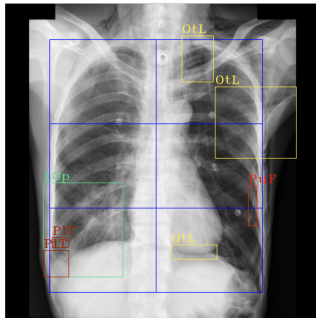
Parenchymal Lesions: 1 ILD in Middle Lower Left, 1 PuF in Right Lung and 1 Nod in Middle Right.

Pneumothorax: No findings.

Pleural Lesions: No findings.

Other Lesion: 1 in Lower Left.

(a) First example of structured report.



Cardiac Pathologies: No findings.

Parenchymal Lesions: 1 LOp in Lower Right and 1 PuF in Middle Left.

Pneumothorax: No findings.

Pleural Lesions: 2 PIT in Lower Right.

Other Lesion: 1 in Top Left, 1 in Top Middle Left and 1 in Lower Left.

(b) Second example of structured report.

Figure 4.9: Example of a structured report using radiologist bounding box annotations. The left portrays the original CXR with the bounding box plotted on top while the right shows the report extracted.

Up to this point, the report has relied on the bounding box annotations provided by the radiologists. However, it can also incorporate various types of visual explanations. These visual explanations offer insights into the regions that influence the model's predictions, enabling the identification of positions associated with the pathologies.

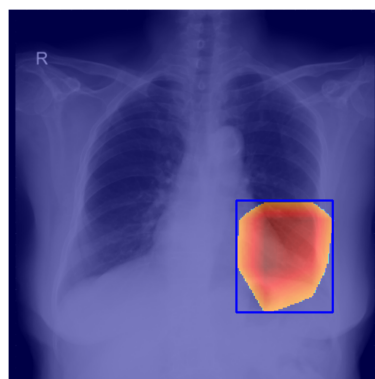


Figure 4.10: Bounding box extracted from a heatmap.

To achieve this, heatmaps are extracted from the given visual explanations, and a bounding box that encompasses the visual map is generated, as illustrated in Figure 4.10. Bounding boxes offer a more convenient format compared to heatmaps, which can exhibit diverse shapes. Extracting bounding boxes ensures a consistent and uniform representation of the pathology's location,

facilitating subsequent comparison with ground truth annotations provided by the radiologists.

4.4 Evaluation Metrics

4.4.1 Model Performance

To study and compare the classification performance of the trained models, the AUC was selected. It provides an overall measure of the model's ability to differentiate between classes. The AUC is calculated using the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds, and the AUC represents the area under this curve.

In addition, because the data is highly unbalanced, the Precision-Recall curve was utilized to determine optimal thresholds for each class. Precision-Recall curves provide insights into the trade-off between precision and recall at different classification thresholds. The Precision and Recall values are given by the equations 4.2 and 4.3.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (4.2)$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (4.3)$$

By examining these curves, thresholds that maximize both precision and recall can be identified, considering the specific characteristics of each pathology and potential variations in data distribution. The selected thresholds are crucial in achieving more accurate test predictions.

For example, Figure 4.11 illustrates a Precision-Recall curve where the best thresholds are selected by maximizing the harmonic mean of precision and recall - F1-score. By carefully choosing the optimal thresholds, the model's overall performance and accuracy in handling different pathologies can be improved.

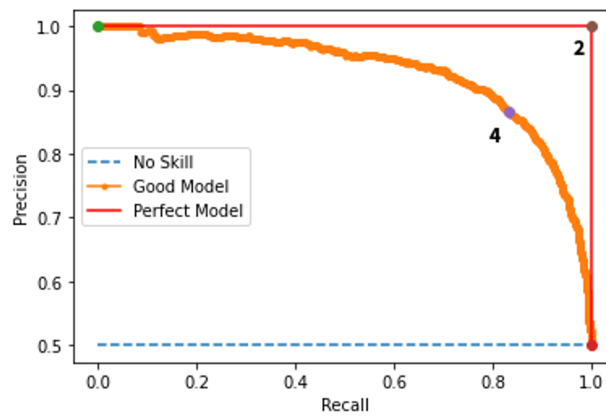


Figure 4.11: Example of a Precision-Recall curve. The point 2 and 4 represent the threshold for Perfect Model and Good Model examples, respectively.

4.4.2 Explainability Evaluation Metrics

To gain a deeper understanding of how visual explanations enhance the area that contains the pathology, it is crucial to define quantitative metrics. These metrics serve as a means to evaluate and compare the performance of different explanation methods. By defining such metrics, one can systematically analyze the results and measure the effectiveness of each explanation method.

Intersection over Union

Intersection over Union (IoU) is a fundamental evaluation metric used in object detection and image segmentation tasks. It is commonly employed to measure the degree of overlap between two objects, although it can be applied to other types of annotations as well. The IoU score quantifies the similarity between the predicted and ground truth areas by calculating the ratio of their intersection area over their union area.

In this study, the method's performance is evaluated by comparing the ground truth bounding box with the visual explanation generated by the explainability method. To facilitate this comparison, a fixed threshold of 0.6 is applied to the heatmap, converting it into a binary mask, as shown in Equation 4.4, where $P_{0.6}$ stands for the predicted explanation using the fixed threshold of 0.6 and G the ground truth. Various threshold values were tested, but 0.6 appeared to be the most suitable for all classes. This binarization process allows the identification of regions of interest within the heatmap. By computing the IoU between the binary mask and the ground truth bounding box, the accuracy of the model's visual maps can be assessed. A higher IoU score indicates a better alignment between the predicted regions of interest and the ground truth.

$$IoU = \frac{P_{0.6} \cap G}{P_{0.6} \cup G} \quad (4.4)$$

In addition to IoU, the Size Adapted Intersection over Union (SA-IoU) metric is utilized in this study. Unlike the fixed threshold approach, SA-IoU incorporates the size of the ground truth bounding box in the binarization process. By creating a binary mask that matches the dimensions of the bounding box, the SA-IoU metric ensures that the relative size and spatial extent of the object of interest are considered during evaluation. This modified approach aims to improve the alignment between the predicted regions of interest and the ground truth, leading to a more accurate assessment of the explanation performance.

$$SA-IoU = \frac{P_t \cap G}{P_t \cup G} \quad (4.5)$$

Equation 4.5 calculates the SA-IoU by utilizing the bounding box mask. Instead of using a fixed threshold as in IoU, it uses a t value that is calculated so that the number of pixels in the heatmap matches the area of the ground truth. Figure 4.12 provides a visual illustration of these two metrics using a practical example.

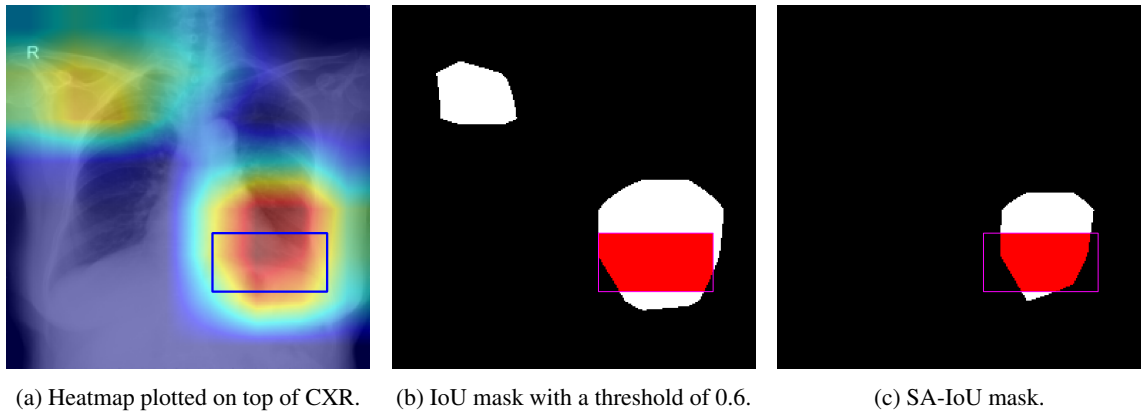


Figure 4.12: Visual examples of the IoU (0.6) and SA-IoU metrics. A random CXR image and heatmap were selected, as well as a sample ground truth bounding box.

Hit Rate

Hit Rate is another valuable metric for assessing the performance of visual explanations. The Hit Rate focuses on the intensity values of the visual explanation to determine if the point with the highest intensity is present within the ground truth covered area.

To compute the Hit Rate, the visual explanation is examined, and the point with the highest intensity value is identified. This point is then compared to the ground truth area to determine if it falls within the annotated region. If the highest intensity point aligns with the ground truth area, it is considered a "hit". At the same time, if the point falls outside the ground truth area, it is classified as a "miss", as shown in Figure 4.13.

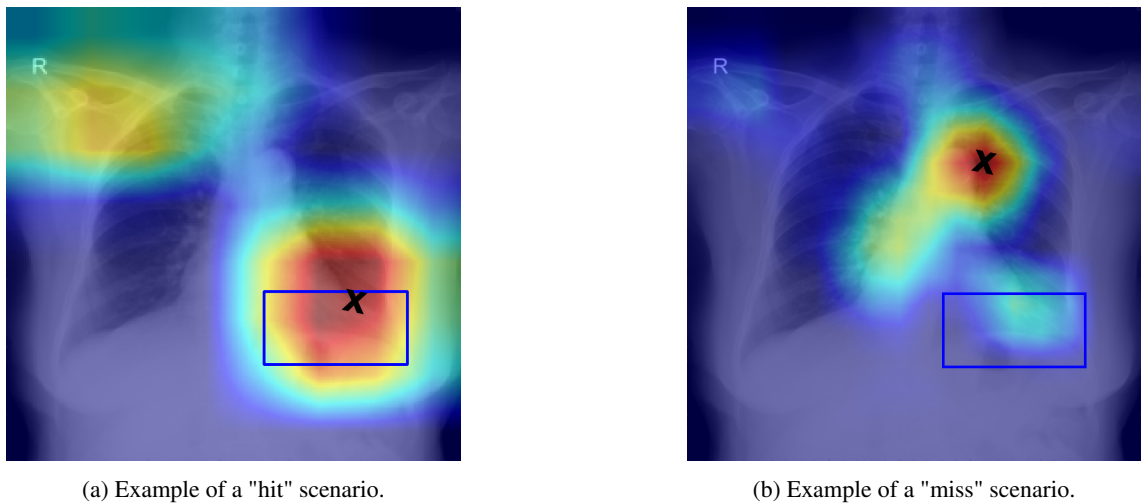


Figure 4.13: Visual examples of the hit rate. A random CXR image and two different heatmaps. The black "x" points to the highest intensity pixel of the heatmap.

A higher Hit Rate indicates that the visual explanation successfully identifies the most important area, aligning with the ground truth annotations. This metric helps evaluate the reliability and effectiveness of the explanation in highlighting the relevant regions of interest.

Zones Score

The structured reports, as first introduced in Section 4.3.2, provide a comprehensive framework for reporting the CXR pathological findings in a way that is easier for the radiologist to understand since it is based on the way radiological reports are written. In contrast to metrics like SA-IoU or Hit Rate, a metric based on the structured report would allow us to extend the analysis beyond spatial properties by generalizing the pathology position to fixed zones. This methodology offers a more holistic evaluation of the explanations, considering not only their spatial accuracy but also their alignment with a more generalized approach that is commonly used by experts.

To access the structured reports as a quantitative metric, one must compare the structured reports derived from the ground truth annotations provided by radiologists with the structured reports generated from the visual explanation maps. By mapping each pathology to its corresponding zone, the task can be treated as a multi-label classification problem, allowing for a comparison of the classification results with the ground truth annotations. To comprehensively represent the presence or absence of pathologies in different zones, one approach is to create an array of one-hot encoded values. In this encoding scheme, a value of 0 indicates the absence of pathology in that particular zone, while a value of 1 indicates its presence, as shown in Figure 4.14. This methodology can be applied to both the ground truth annotations and the visual explanation maps, resulting in two corresponding one-hot encoded arrays.

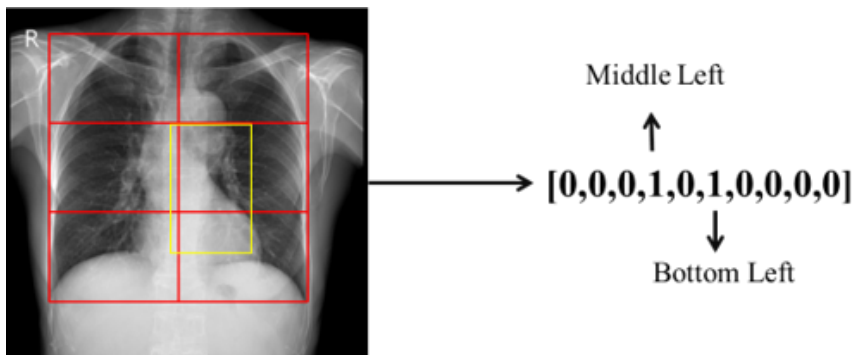


Figure 4.14: Bounding box converted into a one-hot encoded array.

Furthermore, framing the evaluation as a classification task allows us to leverage well-established metrics such as precision, recall, and F1-score to quantitatively assess the classification performance of visual explanations. Among these metrics, the F1-score particularly stands out in the context of medical reports. It provides a balanced assessment by taking into account both precision and recall, without favoring either metric over the other. Achieving a high F1-score is crucial as it indicates a good balance between correctly identifying positive instances - pathologies that were assigned to the correct zone - and minimizing false positives and false negatives. For that reason, the F1-score was used as our base to calculate the structured reports. Equations 4.2 and 4.3 are then used to calculate the Zones Score as shown in equation 4.6:

$$\text{Zones Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.6)$$

4.5 Statistical Analysis - Variable Selection

The analysis of visual explanations extracted from deep learning models could assist us in uncovering how models acquire their features and provide insights into the rationale behind their predictions. However, many variables can significantly influence the efficacy of visual explanations. Consequently, a specific pathology might seem easily explainable by the model not due to its capability to detect it, but possibly because certain colors or textures facilitate its identification, or because the annotations themselves contribute to an higher score.

The variables to be utilized in the model have mainly been derived from both the annotation properties - which aim to describe patterns in the positions and frequencies of bounding boxes - and the texture properties of the images and from the model. In addition to these, the probability variable was also included, which translates the confidence of the model for a given prediction. The annotation-based variables retrieved are the following:

- Frequency (represents the number of bounding boxes for each particular case)
- Bounding box center (x and y position)
- Bounding box area

Image-based metrics rely on the expectation that certain pathologies exhibit distinct grayscale patterns, characterized by variations in pixel intensity and texture. To analyze the texture of an image, the widely used Grey Level Cooccurrence Matrix (GLCM) method was employed on the areas of the image where the bounding box annotations were present. This technique involves constructing a matrix that captures the frequency of different combinations of pixel intensities arranged in specific patterns within the image or a region of interest. In this case, the region of interest was the bounding box, so the image was cropped, and the GLCM was generated specifically for that region. From the GLCM, four distinct texture features can be computed:

- Homogeneity
- Contrast
- Correlation
- Energy, also known as Angular Second Moment

Homogeneity, derived from the GLCM, quantifies the degree of uniformity or similarity in pixel intensity values within the image or sub-image. Contrast measures the local variations in pixel intensities, emphasizing the presence of sharp transitions between different textures. Correlation evaluates the linear relationship between the pixel intensities of neighboring pixels in the

image or sub-image. Lastly, Energy characterizes the overall intensity distribution of the image or sub-image and provides insights into the amount of energy or emphasis present in different textures. All of the above were extracted using the Python library Skimage [107].

Taking these aspects into consideration, a statistical analysis will be carried out in the search for meaningful variables. For this purpose, a Robust Linear Model (RLM) will be employed to investigate the linear relationship between variables and the scores of visual explanations. In this specific study, the SA-IoU metric has been chosen as the independent variable, and the RLM was implemented using the Python library statsmodels [108]. The decision to use a Robust Linear Model over a conventional multiple linear regression stems from the data distribution properties. RLMs are preferred due to their robust nature and ability to provide more accurate coefficients in cases where data distribution might deviate from normality.

Chapter 5

Evaluating XAI in Chest Radiography Pathology Detection

To evaluate the different XAI methods, firstly a series of experiments were conducted involving various models and image sizes. This chapter presents a description of those experiments and provides comparisons between the models and image sizes. Furthermore, the visual explanations extracted from each model will be compared using the different metrics previously introduced. Lastly, the chapter includes a statistical study to further understand what factors influence the explanation's performance.

5.1 Experiment Setup

The VinDr-CXR dataset was the selected dataset for this series of experiences based on the high number of images with bounding box annotations present. The initial experiment involves training and evaluating the classification performance of various deep learning architectures using different image sizes. Subsequently, a careful selection of the two most promising visual explanations is performed, followed by a comprehensive comparison across different architectures.

5.1.1 Training Details

For the experiments ahead, two different architectures were employed, DenseNet121 and ResNet50, and the training hyperparameters are detailed in [4.2](#), the sole modification was the input size of the samples, which was set to 512×512 and 256×256 , depending on the specific experiment. The output classes were set to 14, corresponding to the number of labels in our dataset.

To ensure a thorough evaluation of each experiment, a 5-fold cross-validation strategy was employed, resulting in the creation of five models. This approach not only addresses data imbalance and potential bias from a single division but also provides more reliable results. This strategy involved dividing the dataset into five equal folds, with each fold containing a balanced representation of all the classes, as shown in [Figure 5.1](#).



Figure 5.1: Bar plot of each fold class division.

The training of each model followed a consistent procedure, with three folds used for training, one for validation, and one for testing, as shown in Figure 5.2. After each epoch during training, the model's performance on the validation set was measured, and if no improvement was observed for ten consecutive epochs, the training process was halted. The model with the lowest validation loss was then chosen and evaluated on the test set. This process was carried out five times, using different data splits each time. The final metrics for the specific task were obtained by averaging the results from all five runs.

This training strategy offers notable benefits. Firstly, the implementation of early stopping effectively prevents overfitting, ensuring that the model's performance is not solely tailored to the training data but also capable of generalizing to new, unseen data. Secondly, by repeatedly conducting the training with diverse data subsets, the models' inherent variability is accounted for, leading to more robust and informed analyses.

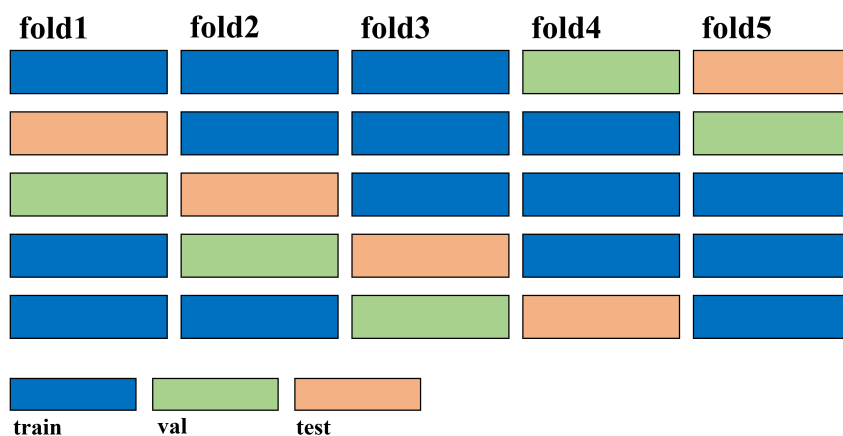


Figure 5.2: Fold division.

5.1.2 Data Pre-processing

CXR Images Pre-processing

An important step in training a deep learning model is data augmentation, which involves applying various transformations to the existing data to artificially create new images. During each iteration, the loaded image undergoes a random process where a set of transformations is applied. This ensures that the model "sees" slightly different images in each iteration, thereby increasing the complexity of the training task and aiding the model in becoming more robust.

After loading the image, the initial step is to crop any black padding that commonly occurs in these types of images and resize the image to the desired size. Subsequently, several key augmentation methods are applied. These include random rotations of up to 6 degrees, random contrast and brightness changes of up to 15%, as well as random border cropping of up to 5% on each side, as exemplified in Figure 5.3.

Following these augmentation steps, the image is normalized between 0 and 1 using the mean and standard deviation. This normalization is essential as the operations inside the model tend to perform better with normalized images. Additionally, the image is converted into an RGB format by replicating the grayscale image, as the architecture expects an RGB image as input.

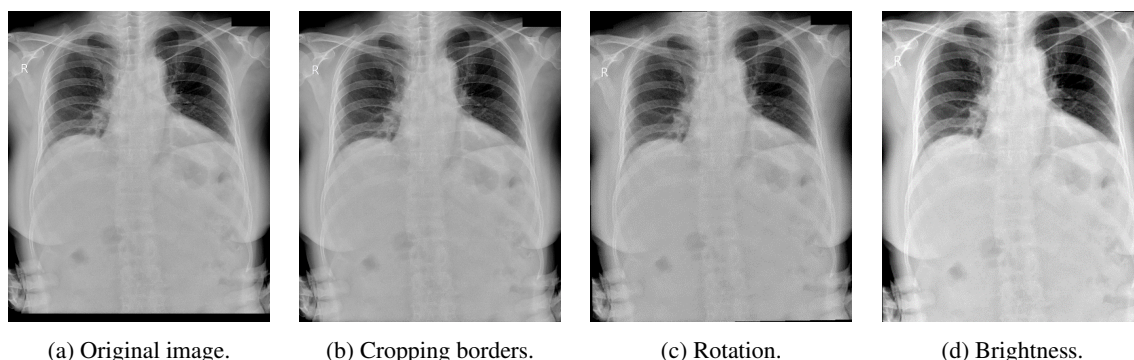


Figure 5.3: Examples of augmentation techniques applications.

Bounding Boxes Pre-processing

In addition, it is important to address the processing of annotations. Although they are not used for training, they play a crucial role in comparing explanations and calculating performance metrics. Since the pathological images were annotated by different radiologists, a majority vote was used to determine whether a given bounding box should be considered. This approach helps reconcile any discrepancies or inconsistencies in the annotations.

Furthermore, due to the variations in annotations from different radiologists, the bounding box locations may not always align perfectly, leading to potential false positive or negative cases. This is particularly evident in cases of diffuse pathologies, where abnormalities can extend across multiple areas in the lungs. One radiologist may annotate multiple small areas, while another may annotate a single, larger abnormal region. Both approaches are considered valid from an

explanatory perspective, and penalizing either one could be misleading. Therefore, there is the need to fuse bounding boxes that essentially reference the same pathology. Figure 5.4 portrays the before and after the fusing is applied, using the method described below, based on [3].

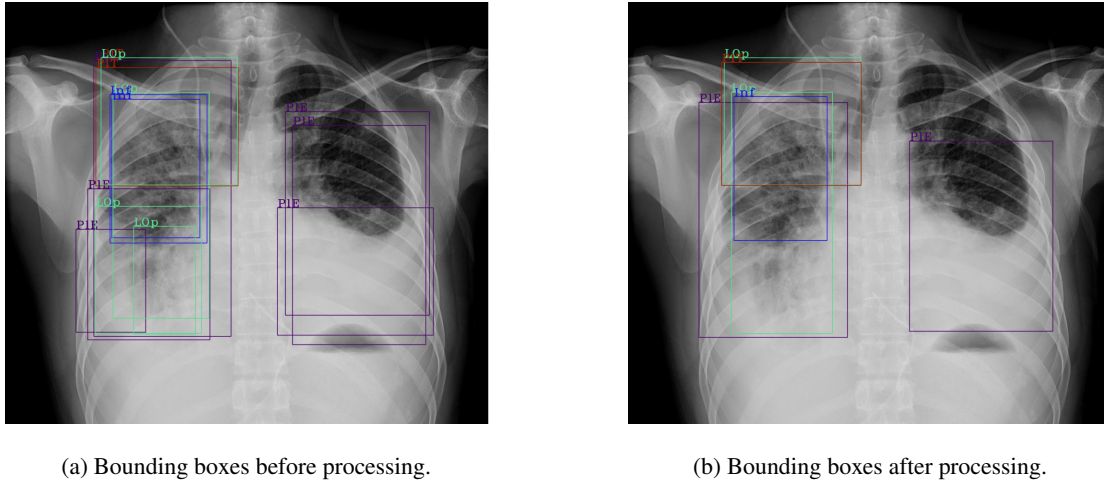


Figure 5.4: Bounding box weighted match and fusion example.

The process of fusing bounding boxes follows a two-phase approach. In the first phase, the primary objective is to combine boxes with similar sizes and positions. To achieve this, the algorithm calculates the IoU between boxes and considers those with IoU greater than or equal to 0.4 as matching candidates. The fusion of these matched boxes is performed by averaging their coordinates and obtaining a new box that represents their combined region. The matched bounding boxes are excluded from the original set of bounding boxes and the process is repeated for every bounding box in the list that was previously ordered following a decreasing area criteria.

In the second phase, the focus shifts to merging boxes that potentially represent the same object but may have different sizes. The goal is to obtain a bounding box that encompasses all abnormalities effectively. In this phase, a different fusion strategy is applied, as further detailed in [3].

It is important to note that first, phase one is applied to all the bounding boxes of the image, and only after that, with a new set of bounding boxes that are the output of phase one, the algorithm proceeds to the second phase. This approach helps reconcile the variations in annotations, facilitating a more comprehensive analysis of the image data.

5.2 Results and Discussion

5.2.1 Model Performance

Table 5.1 presents the results of the multi-label classification on the test set, using the AUC evaluation metric. The results from the table indicate that the DenseNet models, particularly DenseNet121-512, perform exceptionally well across most classes, showing the best AUC values. Likewise, the ResNet50-512 model also performs well, with mean AUC values close to or slightly lower than the DenseNet models. However, ResNet50-256 consistently exhibits slightly lower mean AUC values compared to the other models, though still within an acceptable range.

Remarkably, the image size criteria, initially thought to hold significant importance due to the level of information it provides, had a relatively minimal impact on performance, decreasing, at most, 1.3% of the mean AUC for the ResNet50. This not only suggests that these architectures are robust and can handle variations in image sizes effectively but also indicates their ability to maintain stable performance across different input dimensions.

Regarding the performance of each class individually across the different models, both the AoE and Cmg classes stand out as the top performers, consistently achieving high mean AUC values, with all models scoring above 0.950. The fact that these pathologies always appear in the same location is most likely related to these results, in addition to the fact that they are the most common pathologies in the dataset, as shown in Figure 5.1. Notably, classes such as Cns, Inf, and PIE also achieve relatively high AUC values despite being less common in the data, suggesting that their distinct pathological properties might make them easier to discriminate.

On the contrary, the Clc, Nod, and OtL classes exhibit relatively subpar performance, as evidenced by consistently low AUC values below 0.9 across all model and size combinations. A potential contributing factor to the challenges faced by these classes lies in the nature of the annotations themselves. The dataset exhibits some ambiguity concerning the distinction between the Nod and Clc classes, prompting uncertainty about the appropriate classification of calcified nodules. Moreover, the inclusion or exclusion of aortic calcifications remains unresolved, adding further ambiguity to the dataset annotations. Additionally, the non-specific nature of the OtL class may have introduced greater variability in the annotations, contributing to the observed performance discrepancies. It is also worth noticing that the least common classes fall within this group of lower AUC values.

5.2.2 Selection of Visual Explanations

In this initial experiment, six different visual explanation methods, namely Grad-CAM, Grad-CAM++, EigenGrad-CAM, Saliency, DeepLift, and LRP, were extracted from the DenseNet121-512 model, which exhibited the best performance. The objective of this experiment was to identify the most effective CAM-based and gradient-based methods for future comparisons.

Table 5.2 provides a comprehensive overview of the mean XAI performance metrics for each method on the VinDr-CXR test set. Among the six different explanations, Grad-CAM++ emerges

Table 5.1: Mean AUC values for the VinDr-CXR test set across different models.

Class	DenseNet121-512	DenseNet121-256	ResNet50-512	ResNet50-256
AoE	0.973 $\pm$ 0.003	0.969 $\pm$ 0.004	0.963 $\pm$ 0.005	0.951 $\pm$ 0.012
Atl	0.908 $\pm$ 0.039	0.892 $\pm$ 0.040	0.904 $\pm$ 0.038	0.886 $\pm$ 0.023
Clc	0.857 $\pm$ 0.030	0.845 $\pm$ 0.012	0.832 $\pm$ 0.032	0.817 $\pm$ 0.040
Cmg	0.973 $\pm$ 0.002	0.975 $\pm$ 0.003	0.971 $\pm$ 0.003	0.960 $\pm$ 0.008
Cns	0.944 $\pm$ 0.016	0.935 $\pm$ 0.025	0.917 $\pm$ 0.026	0.922 $\pm$ 0.020
ILD	0.904 $\pm$ 0.011	0.919 $\pm$ 0.024	0.872 $\pm$ 0.024	0.862 $\pm$ 0.019
Inf	0.927 $\pm$ 0.008	0.930 $\pm$ 0.007	0.905 $\pm$ 0.010	0.886 $\pm$ 0.026
LOp	0.931 $\pm$ 0.007	0.913 $\pm$ 0.014	0.907 $\pm$ 0.019	0.898 $\pm$ 0.019
Nod	0.874 $\pm$ 0.026	0.876 $\pm$ 0.014	0.868 $\pm$ 0.026	0.845 $\pm$ 0.025
OtL	0.864 $\pm$ 0.021	0.869 $\pm$ 0.015	0.866 $\pm$ 0.020	0.852 $\pm$ 0.025
PIE	0.967 $\pm$ 0.012	0.970 $\pm$ 0.003	0.959 $\pm$ 0.004	0.945 $\pm$ 0.007
PIT	0.912 $\pm$ 0.012	0.917 $\pm$ 0.009	0.907 $\pm$ 0.010	0.895 $\pm$ 0.013
Pnm	0.933 $\pm$ 0.028	0.934 $\pm$ 0.029	0.903 $\pm$ 0.023	0.892 $\pm$ 0.030
PuF	0.909 $\pm$ 0.003	0.908 $\pm$ 0.013	0.891 $\pm$ 0.010	0.878 $\pm$ 0.021
Overall Mean	0.924	0.922	0.909	0.896

as the top-performing method, achieving the highest values in every metric. This result indicates that Grad-CAM++ generates more accurate and spatially consistent heatmaps compared to other techniques. Although Grad-CAM++ excels, the other two CAM-based methods, namely Grad-CAM and EigenGrad-CAM, closely follow in performance. These results suggest that Class Activation Mapping approaches in general are effective in explaining model predictions for the given dataset.

Regarding the other three explanations, the Saliency method outperforms both the DeepLift and LRP explanations by a significant margin. LRP operates in a backward propagation manner, distributing relevance from the output to the input, while DeepLift computes relevance scores based on the difference in activations with a reference input. However, these approaches seem to render highly ineffective for CXR images, particularly when trying to improve the impact of each original pixel of the input image. As a result, the Saliency method emerges as the more reliable and effective technique for interpreting CXR images among the gradient-based approaches.

Table 5.2: Mean XAI performance metrics for the VinDr-CXR test set.

Method	IoU	SA-IoU	Hit Rate	Zones Score
Grad-CAM	0.214 $\pm$ 0.024	0.313 $\pm$ 0.030	0.410 $\pm$ 0.032	0.612 $\pm$ 0.033
Grad-CAM++	0.223 $\pm$ 0.029	0.325 $\pm$ 0.039	0.426 $\pm$ 0.073	0.626 $\pm$ 0.034
EigenGrad-CAM	0.218 $\pm$ 0.025	0.318 $\pm$ 0.031	0.405 $\pm$ 0.029	0.608 $\pm$ 0.029
Saliency	0.188 $\pm$ 0.016	0.278 $\pm$ 0.022	0.353 $\pm$ 0.038	0.601 $\pm$ 0.037
DeepLift	0.126 $\pm$ 0.013	0.196 $\pm$ 0.020	0.148 $\pm$ 0.014	0.460 $\pm$ 0.021
LRP	0.085 $\pm$ 0.014	0.136 $\pm$ 0.019	0.134 $\pm$ 0.032	0.424 $\pm$ 0.019

When compared with state-of-the-art XAI benchmark results for CXR images, the provided

explanations demonstrate competitiveness and, in certain cases, even outperform benchmark results for some classes. However, it's crucial to recognize that these comparisons encompass distinct datasets, implying varying image counts and data distributions, as well as differing annotations.

Figure 5.5 shows the heatmaps for the different methods present, in these cases for a CXR with PuF class, that support the previous results. Another example can also be found in Figure A.1 of the Appendix A for the PIE class. One can directly understand that the visual explanations acquired from CAM-based methods have a different format when compared with the gradient-based explanations, where the last is more diffuse, even after the Gaussian filter is applied. In addition, is also easy to understand that the CAM methods share some of their heatmap properties which have to do with the way they are calculated. When analyzing the second row of the figure, one can observe that both the Saliency and DeepLift methods share a common characteristic of having some dispersion in the heatmap. However, the LRP explanation stands out as it exhibits a concentrated heatmap with barely any diffusion, even though it is completely miss-aligned with the ground truth.

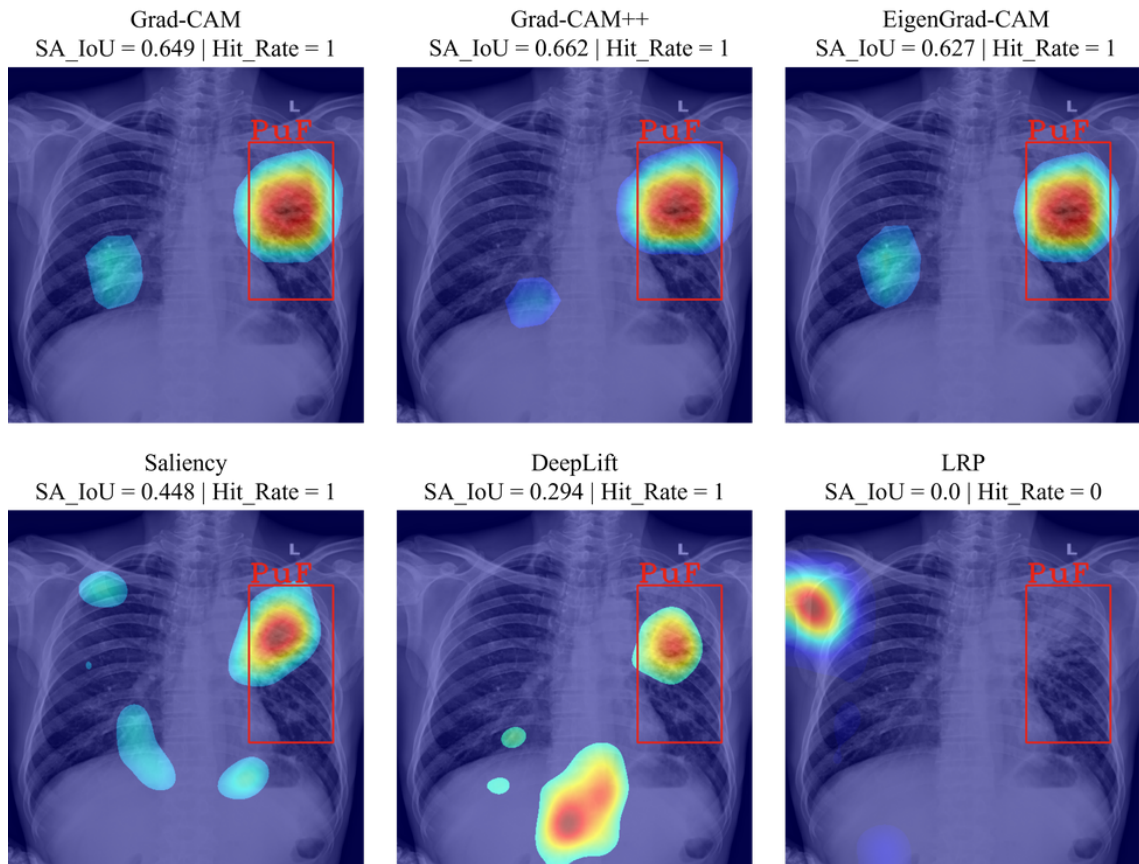


Figure 5.5: Example of CXR with the PuF class and corresponding heatmaps after applying the size adaptive threshold for different explanations, along with the ground truth bounding box. The first row displays all the CAM-based explanations, while the second row contains gradient-based explanations.

Additionally, to facilitate future comparisons, GradCAM++ was selected as a representative of all the CAM-based methods, given that they achieve close results and the heatmaps are also very similar. When it comes to the methods that are based only on the gradients, the Saliency method was chosen once it is the best-performing technique. It is also worth noticing that the IoU and SA-IoU metrics are very similar regarding what they represent, so the IoU metric will not be shown in future comparisons. Instead, the SA-IoU metric will be utilized due to its superior performance and ability to offer a more flexible threshold for enhanced interpretability and evaluation.

5.2.3 Visual Explanations Performance across Models

In the following section, the mean results across all classes of the dataset are shown for the different architectures and image sizes.

Model	SA-IoU	Hit rate	Zones Score
Grad-CAM++			
DenseNet121-512	0.325 $\pm$ 0.039	0.426 $\pm$ 0.073	0.626 $\pm$ 0.034
DenseNet121-256	0.282 $\pm$ 0.052	0.331 $\pm$ 0.063	0.503 $\pm$ 0.065
ResNet50-512	0.158 $\pm$ 0.026	0.179 $\pm$ 0.053	0.460 $\pm$ 0.041
ResNet50-256	0.156 $\pm$ 0.033	0.171 $\pm$ 0.042	0.424 $\pm$ 0.052
Saliency			
DenseNet121-512	0.278 $\pm$ 0.022	0.353 $\pm$ 0.038	0.601 $\pm$ 0.037
DenseNet121-256	0.320 $\pm$ 0.036	0.383 $\pm$ 0.057	0.607 $\pm$ 0.022
ResNet50-512	0.198 $\pm$ 0.026	0.238 $\pm$ 0.044	0.448 $\pm$ 0.019
ResNet50-256	0.231 $\pm$ 0.024	0.268 $\pm$ 0.055	0.468 $\pm$ 0.039

From the results present in Table ??, it can be observed that the performance of Grad-CAM++ varies depending on the model and image size. Generally, larger image sizes (512×512) tend to lead to higher SA-IoU, Hit rate, and Zones Score, indicating that Grad-CAM++ benefits from more detailed input images for accurate segmentation and region identification. In addition, the explanation generated from the DenseNet121 architectures behaves much more accordingly with the annotations, with both models being superior to any of the ResNet50 ones.

It is also interesting to note that for the DenseNet121-512 model, in more than 40% of the cases, the explanations' highlighted spots coincided with the bounding box, indicating that the most important area in the image was effectively highlighted. Additionally, the Zones Score consistently achieved high values, with a score of 62.6% for this particular model. The remarkable performance of the Zones Score, which measures the agreement between the structured report generated by the explanations and the structured report generated by the ground truth bounding box, suggests that the structuring of the explanation to a set of zones in the image closely aligns with the ground truth annotations represented by the bounding box. This indicates that the generalization

that occurs in the structured reports can effectively capture and match the ground truth regions, serving as compelling evidence that the adoption of a structured report can effectively bridge the gap between the visual explanations generated by the models and the ground truth annotations. Overall, it supports the idea that structured reports are valuable not only as explanations on their own but also for evaluating and validating the visual explanations produced by the models.

The analysis of the Saliency results, also present in Table ?? reveals that Saliency methods tend to perform better with smaller image sizes. Across both model architectures (DenseNet121 and ResNet50), the model utilizing the smallest image size (256×256) consistently demonstrated superior results across all evaluation metrics when compared to the models with larger image sizes (512×512). Interestingly, this observation contradicts the trend observed with the Grad-CAM++ method, where larger image sizes consistently yielded better visual explanations.

This observed trend is attributed to the inherent nature of the explanation methods rather than variations in model performance. As mentioned before upon examination of the results presented in Table 5.1, the advantage gained from using bigger image sizes was relatively modest. The Saliency methods exhibited greater sensitivity to smaller image sizes, potentially due to their emphasis on capturing individual pixel importance, resulting in a more effective representation of fine details in smaller images. Conversely, the Grad-CAM++ method, which highlights class-specific discriminative regions, exhibited improved performance with larger image sizes. This can be attributed to the richer contextual information available in larger images, enabling the method to capture more extensive regions that are relevant to the target class.

Figure 5.6 illustrates the comparison between Grad-CAM++ and Saliency heatmaps for the different models. In this example, it is evident that the ResNet50 architecture can not properly highlight the correct areas, with SA-IoU scores below 0.1 for every case and no "hits". Both heatmap's position is more scattered in this architecture when compared with the DenseNet121. This, however, does not imply that the model cannot make reasonable predictions. It signifies that the model estimates and calculates the output using different approaches to interpret the image.

Furthermore, the figure shows that Grad-CAM++ performs better with a bigger image size, achieving an SA-IoU score of 0.415 and a "hit". However, its performance decreases when transitioning to a smaller image size, reaching a score of 0.154. Conversely, for Saliency maps, the trend is the opposite, with the SA-IoU score increasing from 0 to 0.726 as the image size decreases from (512×512) to (256×256).

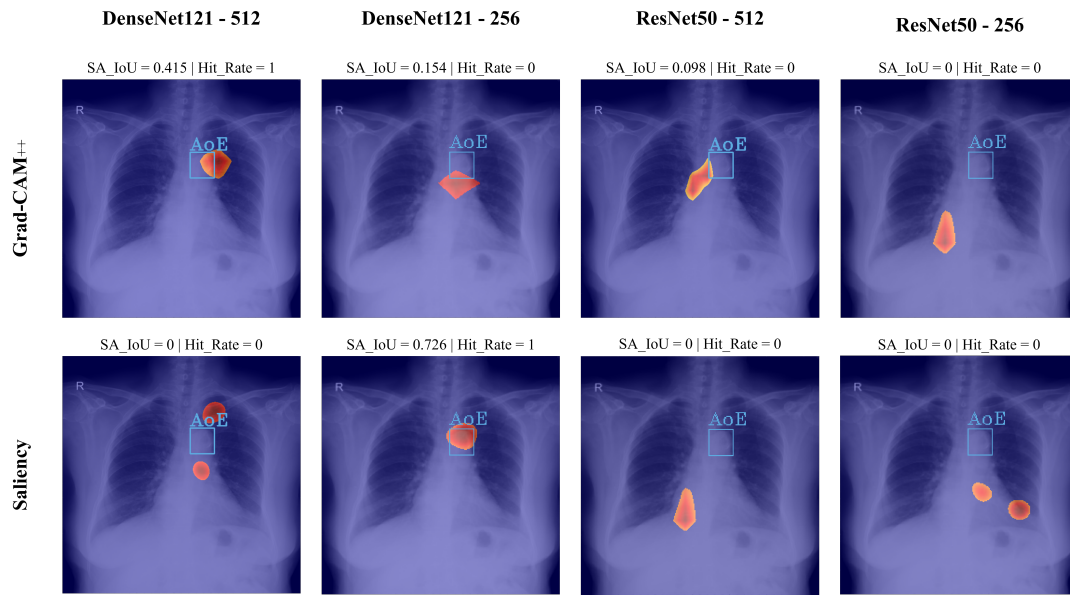


Figure 5.6: Example of CXR with the AoE class and corresponding Grad-CAM++ and Saliency heatmaps after applying the size adaptive threshold for different models and image sizes, along with the ground truth bounding box.

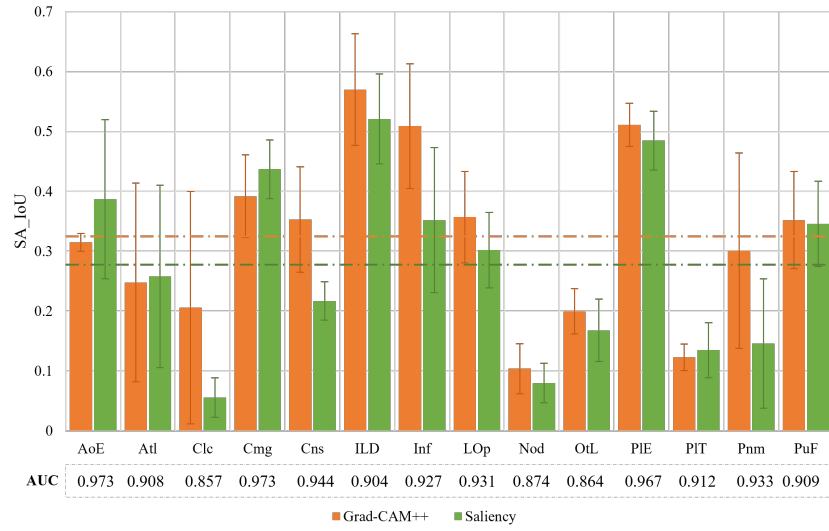
5.2.4 Visual Explanation Performance across Classes

The next section goes deeper into the analysis and looks at each class individually, trying to understand their behavior. To do so, the DenseNet121-512 results were selected for both Grad-CAM++ and Saliency explanations.

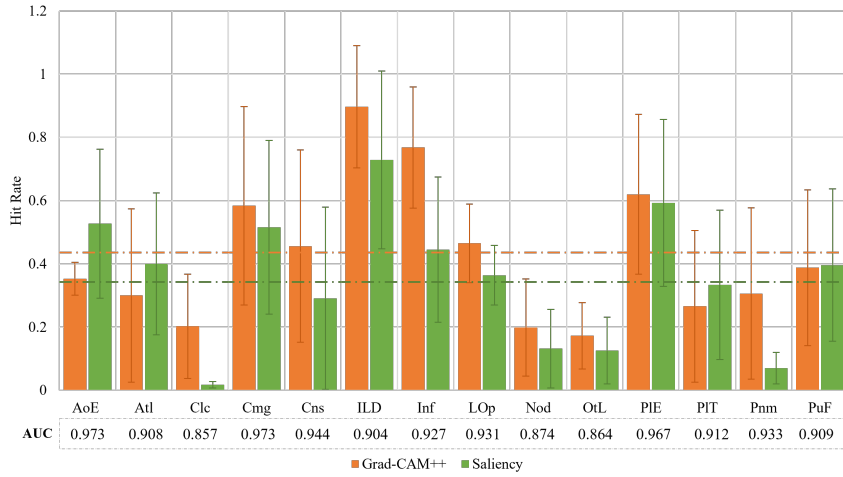
The results present in Figure 5.7a show that some classes perform exceptionally well with the Grad-CAM++ method, while others exhibit lower scores compared to the overall average. The same goes for the Hit rate and Zones Score, present in Figure 5.7b and Figure 5.7c.

Among the classes assessed, ILD, PIE, and Inf stand out as the best performers, boasting SA-IoU scores significantly higher than the overall average (0.325) with values of 0.570, 0.511, and 0.509, respectively. And, the same pattern is found when looking into the Hit rate. It is, however, interesting to notice that the AUC values for these classes vary, with only the PIE demonstrating particularly high AUC, indicating that a good explanation is not solely reliant on good AUC but also depends on the unique features and characteristics of each class.

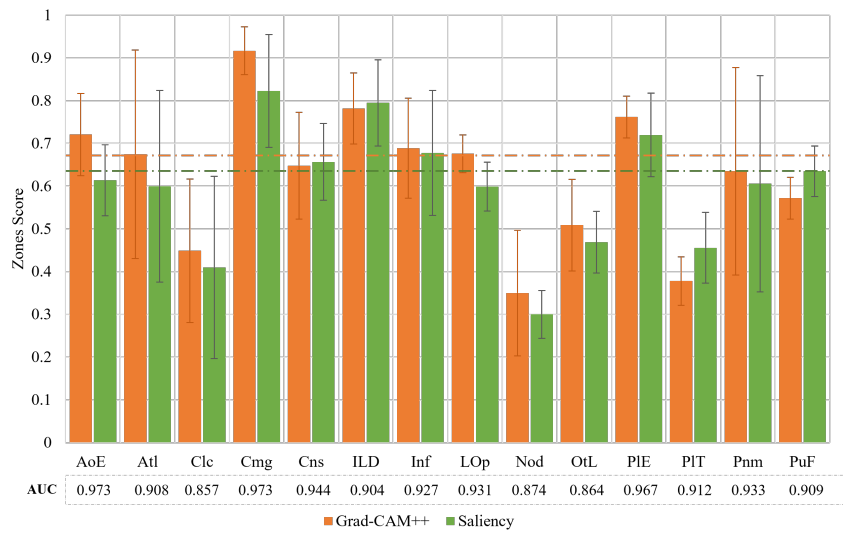
On the other hand, classes like Cmg, LOp, Cns, PuF, AoE, and Pnm exhibit SA-IoU scores very close to the average, ranging from 0.392 to 0.301. Once again, the Hit rate metric portrays the same information. However, the standard deviation appears to be higher when compared with the SA-IoU, which may be related to the fact that the Hit rate is a binary metric. In addition, the presence of the two best-performing classes, AoE and Cmg, within this group, is noteworthy, both attaining an AUC value of 0.973. Notably, these classes represent the most common pathologies in the dataset and are associated with the cardiovascular system, making them "stationary" pathologies. As a result, the model's ability to detect these pathologies may be facilitated by their higher occurrence and characteristic localization.



(a) SA-IoU results across all classes.



(b) Hit rate results across all classes.



(c) Zones Score results across all classes.

Figure 5.7: Metrics results for Grad-CAM++ and Saliency on the DenseNet121-512 model across all classes. The dotted lines represent the average values for all classes.

On the contrary, certain classes, including *Atl*, *Clc*, *OtL*, *PIT*, and *Nod*, display scores significantly below the average, ranging from 0.248 to 0.104. The Grad-CAM++ relative underperformance for these classes suggests challenges in accurately identifying and highlighting relevant features and regions for these specific pathologies. The corresponding AUC values for these classes further validate the method's limitations in providing accurate and meaningful explanations for these challenging cases. Moreover, it is worth noting that the pathologies with fewer occurrences are also present in this low-performing group, indicating that the model may have had less opportunity to adequately learn the distinctive characteristics of these rarer pathologies.

Analyzing the Saliency results, a more pronounced influence of AUC for each class can be observed, as exemplified by the varied scores of *Cns* and *Nod* classes. In comparison to Grad-CAM++, the Saliency method exhibits a wider range of differences between the best-performing class (0.521) and the worst-performing class (0.056). This suggests that the Saliency method's performance is more sensitive to the unique characteristics of each class, which may impact its ability to provide accurate explanations.

Despite these variations, there is a notable similarity in the performance patterns of both methods. The best-performing classes remain consistent across both Grad-CAM++ and Saliency, reinforcing the idea that certain pathologies possess intrinsic features that facilitate their explanation. Following the same idea, the classes with the lowest scores persist as the weakest performers in both methods, indicating shared limitations in accurately explaining these specific classes. Regarding the Hit rate results, the only class that presents a different behavior is the *Cmg*, where the Saliency method has a lower Hit rate, but slightly higher SA-IoU.

It is also safe to say that the Zones Score follows a pattern very close to both the SA-IoU and Hit rate, as expected. In summary, the analysis of Grad-CAM++ and Saliency results uncovers a heterogeneous performance pattern across different classes. While some classes benefit from the methods' ability to capture discriminative features, others encounter challenges in achieving accurate and informative explanations.

To get a better understanding of how each class heatmap compares with the other, Figure 5.8 illustrates the average heatmaps for both Grad-CAM++ and Saliency across all classes, accompanied by the average bounding box annotations for both datasets. This comparison notably accentuates the disparity between the Grad-CAM++ and Saliency heatmaps. The average Saliency maps demonstrate a considerably diffuse distribution, essentially encompassing the entire image, a sharp contrast to the Grad-CAM++ heatmaps. Furthermore, the Grad-CAM++ heatmaps exhibit a pattern that closely follows the annotations, which is logical considering it produced the most accurate explanations. However, despite these differences, the highest intensity regions remain consistent across the majority of the classes.

Focusing now on the annotations, the *AoE* and *Cmg* pathologies are present in the same location for almost all CXR, which helps the model identify them, as proved by their high AUC values. Conversely, all the other pathologies exhibit a more sporadic distribution across the CXR.

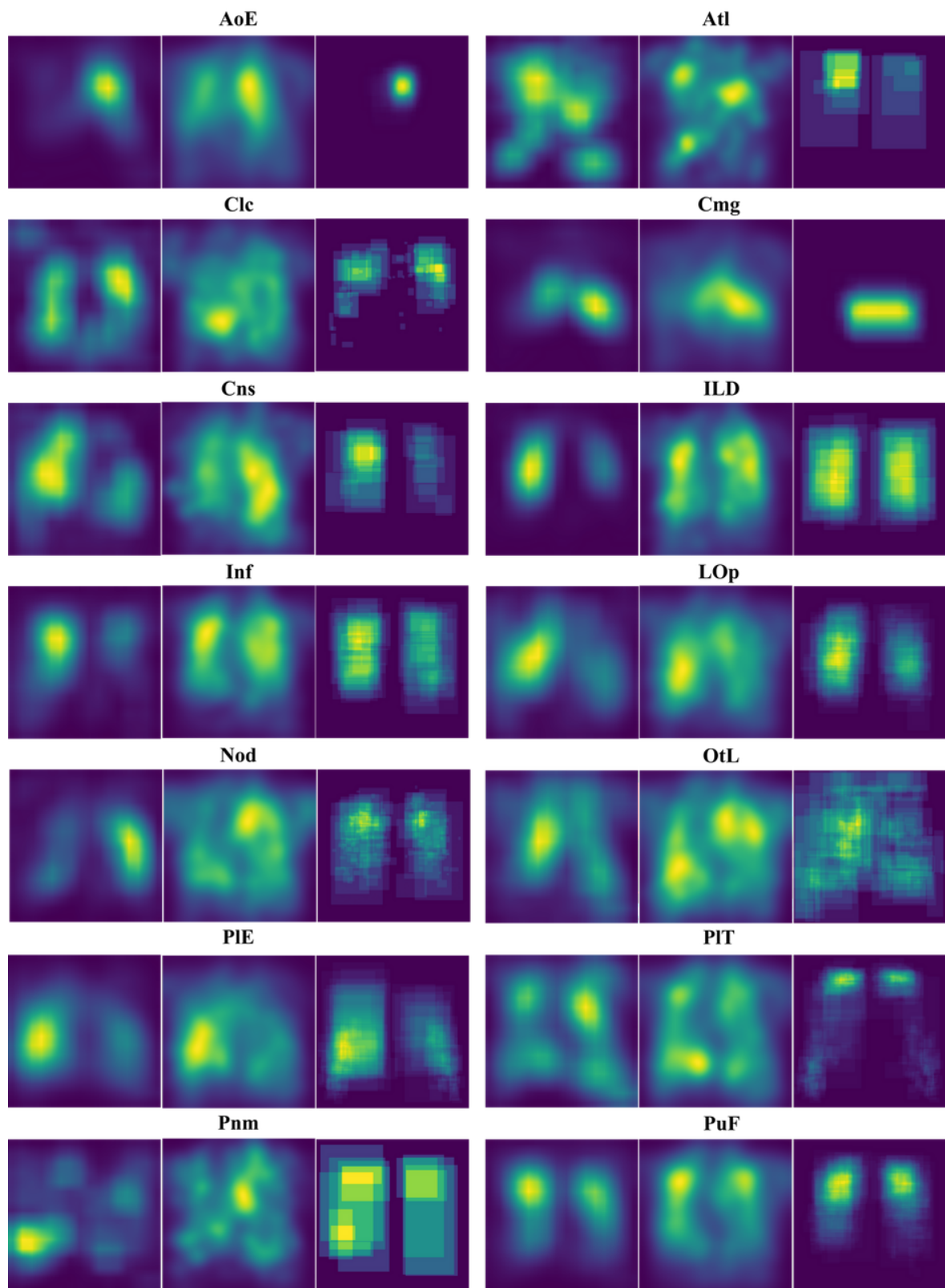


Figure 5.8: Average explanations heatmaps of the test set and bounding box annotations, per class. The first column represents Grad-CAM++, followed by Saliency and average bounding box annotations in the last column.

5.2.5 Statistical Analysis Result

In the analysis conducted in the previous section, the focus was primarily on AUC values and the frequency of each class in the dataset. However, this approach proved insufficient to fully comprehend the results. The goal of this section is to explore the relationship between the chosen variables and their potential discriminatory nature concerning each pathology's explanation. The aim is to find a linear relation between these variables and the dependent variable (SA-IoU), thereby identifying attributes playing a pivotal role in distinguishing different pathologies. Any of the previously introduced metrics (Hit rate and Zones Score) could be used as the independent variable in this analysis given that they follow the same trends.

The variables considered include annotations-based variables, such as frequency, bounding box position (x and y), area, and image-based features related to the texture of the pathology, namely homogeneity, contrast, energy, and correlation, as defined on 4.5. In addition, the model confidence in a given prediction was also included as the probability variable. To ensure a fair and meaningful comparison among these variables, normalization using the MinMax method was applied. This normalization process rescaled all metrics to a common range, enabling a direct comparison of their relative contributions to the explanations. By doing so, the significance of each variable in explaining the variations observed in the dependent variable can be better understood.

To address multicollinearity, which arises when two independent variables exhibit high correlation, signifying that they provide similar information to the model, the correlation matrix presented in Figure 5.9 was analyzed. Upon examination, a significant correlation coefficient of 0.93 was identified between the homogeneity and energy. Consequently, the decision was made to remove the homogeneity variable from the data to circumvent potential issues associated with multicollinearity and to ensure the robustness of the analysis.

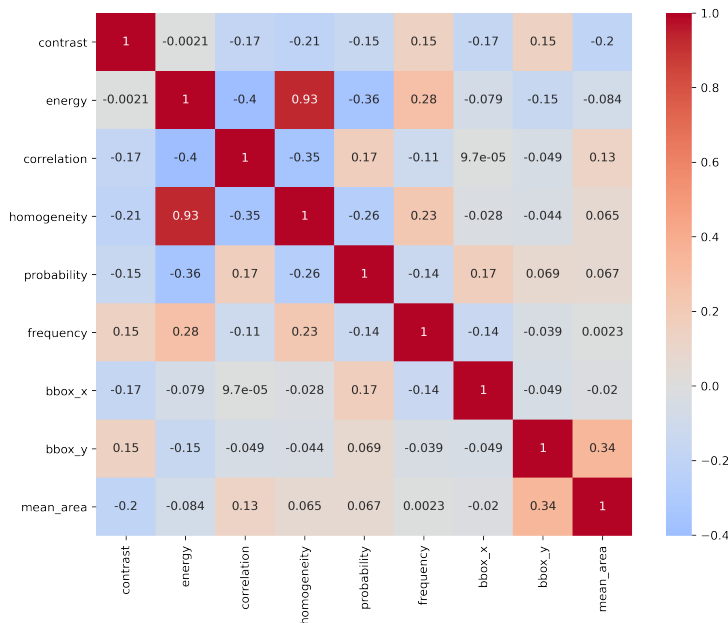


Figure 5.9: Correlation matrix results.

The RLM using the remaining variables proved capable of explaining 29% of the variability present in explanations. Table 5.3 presents the results regarding each metric and their significance. In the context of image-based metrics, the analysis revealed that only energy and contrast showed statistical significance (p-values less than 0.05). Both parameters displayed negative coefficients, with values of -0.0881 and -0.423, respectively which indicate that higher energy and contrast values correspond to lower SA-IoU values.

Table 5.3: RLM results.

Variable	Coef	Std. Err	z	P-value
energy	-0.0881	0.034	-2.629	0.009
contrast	-0.423	0.078	-5.424	<0.001
correlation	0.1428	0.107	1.34	0.180
probability	0.1239	0.022	5.665	<0.001
frequency	0.2246	0.092	2.44	0.015
bbox_X	-0.022	0.034	-0.639	0.523
bbox_Y	0.0904	0.038	2.373	0.018
area	0.8268	0.054	15.179	<0.001

Regarding the metrics related to the annotations, all but one displayed statistical relevance. Specifically, probability, frequency, bounding box y-positioning, and area exhibited positive influences on explanations, implying that an increase in any of these metrics leads to improved explanations.

Notably, probability and area seem particularly important, given their very low p-values (< 0.001), highlighting their significant impact. Regarding the probability variable, the positive coefficient of 0.1239 indicates that an increase in probability results in a small increase in the explanation scores. This suggests that classes better learned by the model correspond to, overall, improved explanations. Regarding the area, it demonstrates a substantial positive coefficient of 0.8268, suggesting a considerable impact on the results. Such results are related to the fact that if the pathology covers a bigger area, then it is easier for the explanation to be correct, given that there is a higher chance of matching the bounding box annotation.

The x-position of the bounding box showed a non-significant effect, as indicated by its p-value of 0.523. Conversely, the y-axis demonstrated statistical relevance, with a positive coefficient of 0.0904. A possible reason for such result would be that pathologies with good explanations scores tend to be posited in the middle and lower portion of the CXR, as happens with Cmg and PIE.

As hypothesized at the beginning of the section, there is a significant amount of information laying withing the image features and the properties of the pathology, such as its position and area. This analysis helped uncover some patterns and understand some of the previously presented results.

Chapter 6

Evaluation of XAI in Chest Radiography Pathology Detection across datasets

The central objective of this chapter is to assess the ability of deep learning models to provide visual explanations on domains different from their training data. Firstly, a new model will be trained on the ChestX-ray14 dataset. Subsequently, experiments will be conducted by evaluating the models trained on one dataset using data from another dataset. Through classification and visual explanation extraction across diverse datasets, this study aims to evaluate the models' generalization and adaptation capabilities to varying data distributions, while ensuring the provision of accurate explanations.

6.1 Experiment Setup

6.1.1 Training Details

In this experiment, the ChestX-ray14 [98] dataset was utilized along with the DenseNet121 architecture, which has demonstrated the best performance in prior studies. To ensure a fair comparison with the VinDr-CXR dataset, the same training conditions were adopted and are detailed in 4.2. Additionally, the image size was standardized to 512×512 for consistency throughout the study.

Given the considerable size of the dataset and the limited number of images containing bounding box annotations, implementing traditional cross-validation would have consumed a significant amount of time and computational resources. However, since the main objective of this study was to explore the performance of the explanations rather than solely creating a highly sophisticated model, a practical approach was adopted. The dataset was divided into 5 folds, as before, but cross-validation was not applied. Instead, a single iteration was conducted, utilizing 4 folds for training and 1 fold for validation. The testing phase exclusively focused on images with annotations, allowing for a focused evaluation of the explanations on the available labeled data.

6.1.2 Data Pre-processing

CXR Images Pre-processing

The pre-processing techniques remain consistent with previous experiments to maintain standardization and facilitate future comparisons. After loading the image, any black padding commonly present in these images is cropped, and the image is resized to 512×512 . Subsequently, augmentation methods are applied, including random rotations of up to 6 degrees, random contrast and brightness changes of up to 15%, and random border cropping of up to 5% on each side. Following these augmentation steps, the image is normalized between 0 and 1 using the mean and standard deviation. Additionally, to meet the model's input requirements, the image is converted into an RGB format by replicating the grayscale image, as the architecture expects RGB input.

For every possible combination of model and dataset, the final outputs are consistently converted into classes present in the VinDr-CXR dataset. If the model was trained on the VinDr-CXR dataset, the output is given in the correct classes. However, for models trained on the ChestX-ray14 dataset, the output needs to be converted to the VinDr-CXR classes. Not only that but the comparison is restricted to classes that are either present in both datasets or can be mapped using the Table 4.2.

Among the 14 different classes in the ChestX-ray14 dataset, annotations are available only for 8 classes, namely Atl, Cmg, Eff, Inf, Mas, Ndl, Pno, and Pnm, totaling 529 bounding boxes annotations for PA CXR, as shown in Figure 6.1. To ensure consistency in comparisons across datasets, Table 4.2 is used to convert these specific classes into the corresponding VinDr-CXR classes. Upon examination, it becomes evident that Mas and Ndl do not have direct matches. Instead, they are encompassed in the class Nod. Consequently, the outcomes for Mas and Ndl are merged under this unified class, and the performance evaluation, both for classification and visual explanations, is an average of both individual classes. Likewise, the class Pno lacks a direct match but is included in the VinDr-CXR dataset under Cns. For all the other classes, there is a direct match, Atl, Inf, and Pnm keep the same name, and Eff becomes PIE.

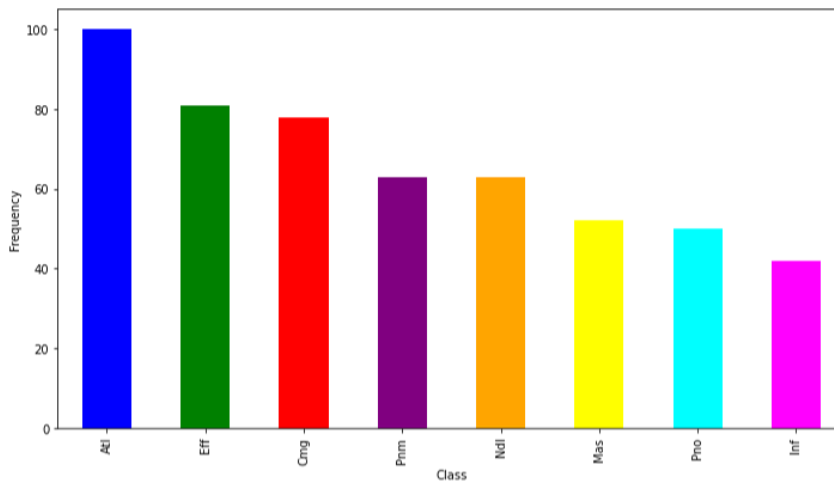


Figure 6.1: Frequency of each class of the ChestX-ray14 dataset with bounding box annotations.

Regarding bounding box processing, no additional steps were necessary since each label had only one bounding box. This approach ensures a seamless evaluation process across datasets and allows for a direct comparison of model performance using only the classes common to both datasets.

6.2 Results and Discussion

6.2.1 Model Performance

To better understand and compare the classification performance across datasets, Table 6.1 summarizes the mean AUC values for every possible combination of train and evaluation, across all common classes. From now on, VV refers to the model trained on VinDr-CXR and evaluated on VinDr-CXR, VC refers to the configuration of the model trained on VinDr-CXR and evaluated on ChestX-ray14, CV refers to the model trained on ChestX-ray14 and evaluated on VinDr-CXR and CC to the model trained and evaluated on the ChestX-ray14.

Table 6.1: Mean AUC values for the ChestX-ray14 and VinDr-CXR datasets.

Train	VinDr-CXR		ChestX-ray14	
Evaluation	VinDr-CXR	ChestX-ray14	VinDr-CXR	ChestX-ray14
Atl	0.908	0.425	0.847	0.814
Cmg	0.973	0.817	0.810	0.956
Cns	0.933	0.493	0.545	0.728
Inf	0.927	0.602	0.346	0.695
Nod	0.874	0.528	0.830	0.841
PIE	0.967	0.703	0.782	0.881
Pnm	0.933	0.721	0.703	0.913
Overall Mean	0.931	0.613	0.695	0.833

Upon analyzing the results from the table, it becomes evident that the model's performance varies significantly across different classes and datasets. For the VV configuration, the mean AUC value for the common classes is 0.931, while the CC is 0.833, representing a notable 10% difference. This difference in performance is likely attributed to the varying data availability and the dataset size. Specifically, the ChestX-ray14 dataset is 4.5 times larger than the VinDr-CXR dataset, resulting in a higher degree of variability and complexity in the data. Supporting this observation, when the model trained on ChestX-ray14 is evaluated in the VinDr-CXR dataset (CV), it manages to achieve higher AUC values compared to the opposite scenario (VC). This suggests that the model trained on the larger ChestX-ray14 dataset exhibits greater robustness and generalization capabilities.

When considering each class individually, it becomes apparent that the model's performance deteriorates for every pathology when evaluated on a different dataset than the one it was trained on. However, certain classes exhibit more pronounced differences. For instance, in the case of the model trained on the VinDr-CXR dataset, the Atl, Cns, and Nod classes are notably affected.

This is attributable to the fact that these classes are in the minority within the VinDr-CXR dataset, leading to greater difficulty in their accurate classification. On the other hand, the model trained on the ChestX-ray14 dataset encounters more challenges with the Cns and Inf classes.

In conclusion, the analysis of AUC values sheds light on the model's performance across various classes and datasets. It emphasizes the significance of dataset-specific characteristics and the need to address challenges in generalization.

6.2.2 Visual Explanations Performance across Datasets

The performance of Grad-CAM++ and Saliency explanations extracted from the model trained on the ChestX-ray14 dataset shows very similar results, as shown in Table 6.2. Even when evaluated on the VinDr-CXR dataset, the model yields very similar results.

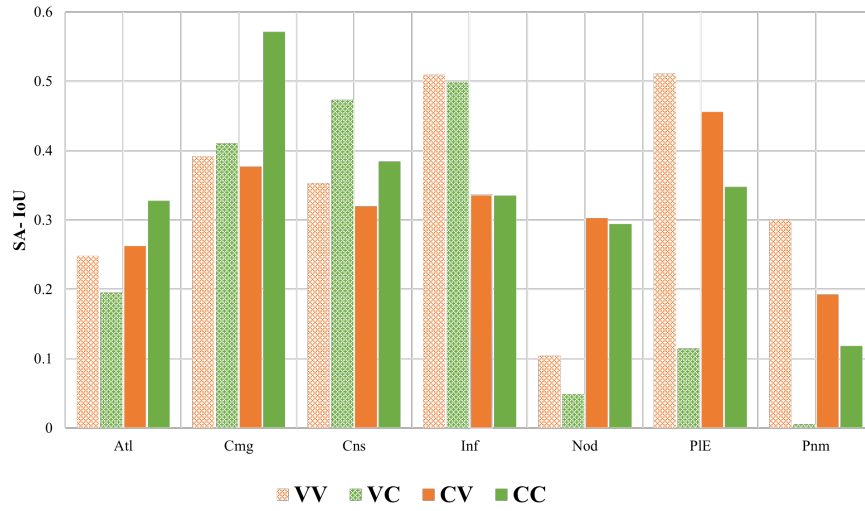
However, this tendency is not observed in the model trained on the VinDr-CXR dataset, as it displays substantial differences in the scores of the two explanations (SA-IoU goes from 0.345 to 0.282 and Hit Rate from 0.462 to 0.350). This observation suggests that visual explanations are influenced by the data properties and even datasets with images that portray the same pathological findings, as in this case, exhibit considerable dissimilarities in the way the models learn their features, resulting in significant variations in the production of visual explanations.

Table 6.2: Mean SA-IoU, Hit Rate and Zones Score values for Grad-CAM++ and Saliency explanations across the VinDr-CXR and ChestX-ray14 datasets.

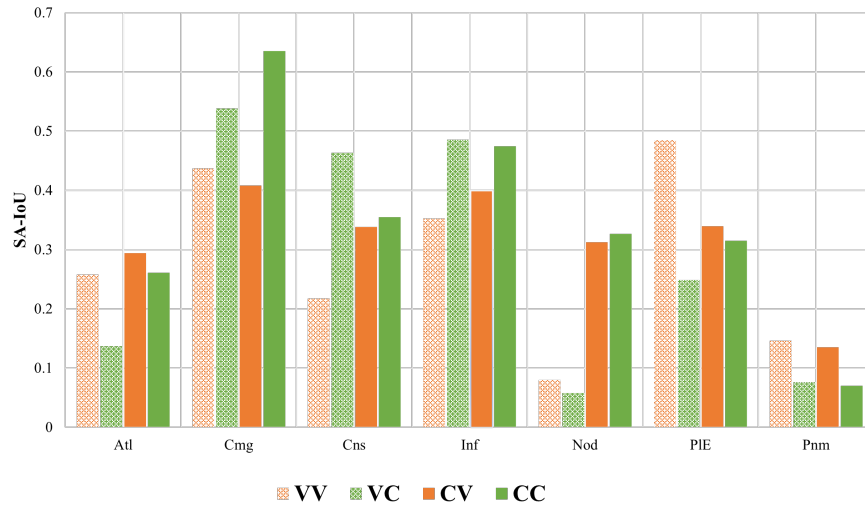
Train	VinDr-CXR		ChestX-ray14	
Evaluation	VinDr-CXR	ChestX-ray14	VinDr-CXR	ChestX-ray14
Grad-CAM++				
SA-IoU	0.345	0.249	0.321	0.340
Hit Rate	0.462	0.301	0.431	0.444
Zones Score	0.668	0.550	0.552	0.700
Saliency				
SA-IoU	0.282	0.287	0.318	0.348
Hit Rate	0.350	0.344	0.427	0.401
Zones Score	0.626	0.689	0.545	0.704

Another important point to consider is that the VinDr-CXR dataset has multiple radiologist inputs and, in most cases, there is more than one bounding box annotation per image. On the other hand, ChestX-ray14 has a single bounding box per annotation. Such differences may also explain some of the disparities in results, particularly the increase in the zones score when the models are evaluated on ChestX-ray14, as the model has fewer annotations which leads to simpler reports and, therefore higher Zones scores.

To study each class individually, Figure 6.2 showcases the SA-IoU results for Grad-CAM++ and Saliency explanations, across all classes.



(a) Grad-CAM++ SA-IoU results for all classes and dataset combinations.



(b) Saliency SA-IoU results for all classes and dataset combinations.

Figure 6.2: Grad-CAM++ and Saliency SA-IoU results for all classes and dataset combinations. The bars that have no pattern were trained on the ChestX-ray14 and those that have a pattern were trained on VinDr-CXR. Orange bars were evaluated on VinDr-CXR, and green were evaluated on the ChestX-ray14 dataset.

Upon evaluating the SA-IoU values across all configurations, it becomes evident that the Cmg, Cns, and Inf pathologies generally achieve higher values compared to other classes, both for Grad-CAM++ and Saliency. However, the quality of explanations varies from dataset to dataset. For configurations trained on VinDr-CXR - VV and VC - the class Nod proves to be the most challenging to explain. This class had one of the worst AUC values on VV and VC, and, due to the small size of typical nodules and masses, obtaining good explanations becomes more difficult.

On the other hand, for the model trained on Chestx-ray14, the worst-performing class was Pnm, followed by Atl. Surprisingly, when compared with top performers such as Cns and Inf, Pnm and Atl achieved better AUC values. This suggests that the model learns to classify these classes

differently from how humans identify them, leading to differences in heatmaps and, consequently, lower SA-IoU values.

Figure 6.3 showcases the Grad-CAM++ and Saliency explanations on a CXR with the Inf class. The first two cases use an image from the VinDr-CXR dataset, with two different bounding box annotations, while the other two cases use an image from the ChestX-ray14 dataset.

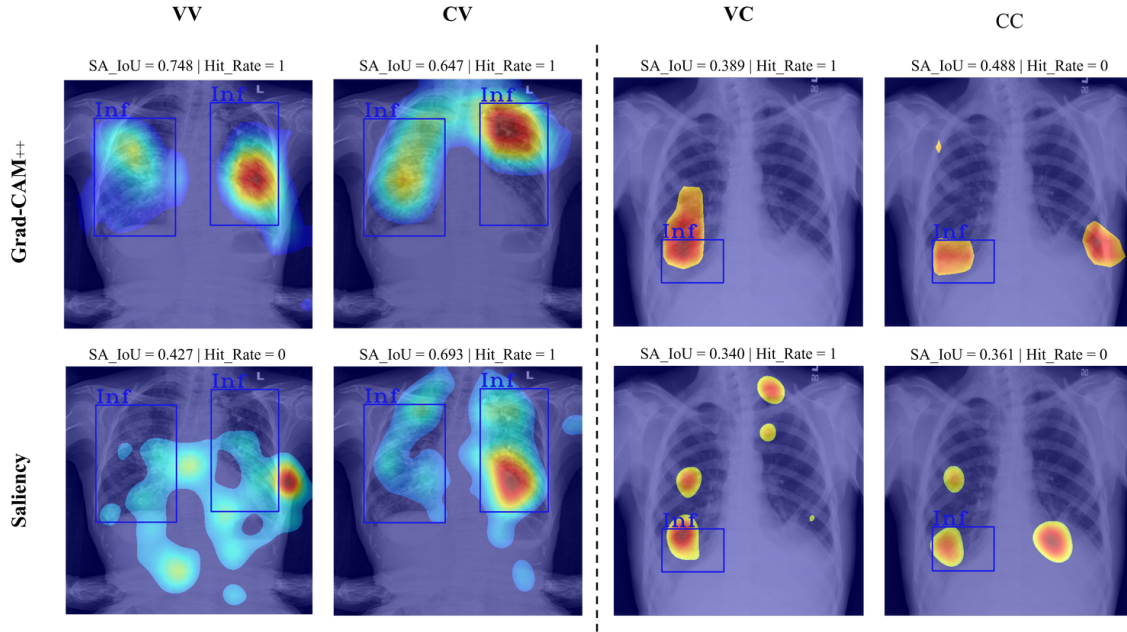


Figure 6.3: Example of CXR with the Inf class and corresponding heatmaps after applying the size adaptive threshold for different models and datasets combinations. The first row displays all the Grad-CAM++ explanations while the second row contains the Saliency heatmaps.

The differences in the Grad-CAM++ and Saliency explanations are evident in these examples, with Saliency showing a more diffuse pattern and Grad-CAM++ exhibiting a more focused representation. It is also interesting to note that the differences between the two explanation techniques are mitigated in the configurations trained on the ChestX-ray14 dataset, namely CV and CC. Specifically, when observing the CV column, both the Grad-CAM++ and Saliency explanations exhibit their strongest areas on the left lung, with a considerable component on the right lung as well. However, for the VV case, while the Grad-CAM++ is even more accurate than the CV Grad-CAM++, the Saliency explanation is significantly worse than the CV one. Thus, if for the studies using exclusively the VinDr-CXR dataset, the Grad-CAM++ explanations achieved, overall, better scores, the same can not be said for the ChestX-ray14 dataset, which leads to the conclusion that the best explanations are highly dependent on the dataset properties such as class distribution and quality of the annotations.

Figure 6.4 displays the average Grad-CAM++ heatmaps for every class and configuration, as well as the average bounding box annotation coordinates for both datasets. The Saliency results are present in the Figure A.2 of the Appendix A.

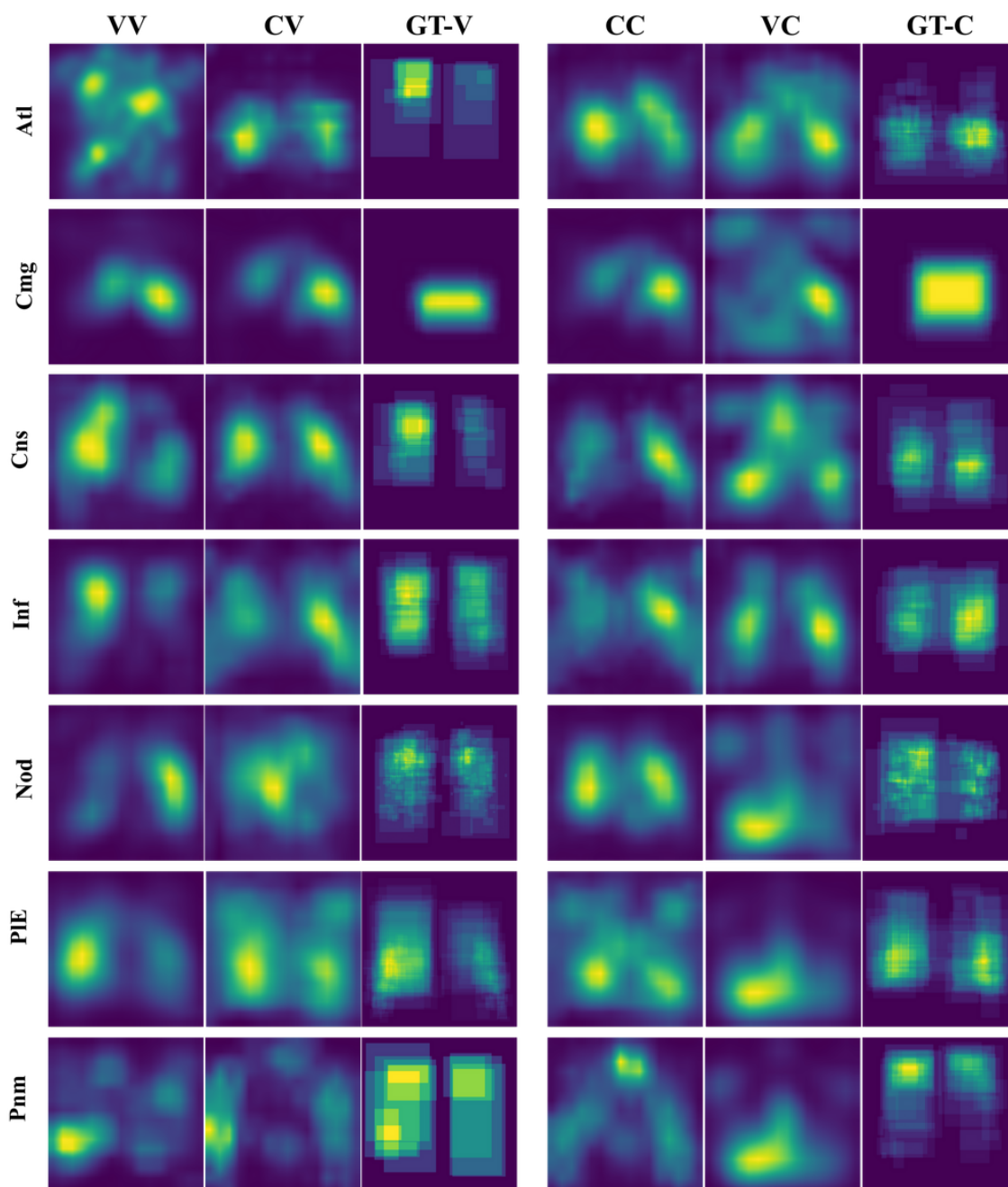


Figure 6.4: Average Grad-CAM++ heatmaps for all configurations. The first two columns represent configurations evaluated on the VinDr-CXR dataset and the third, GT-V, represents the VinDr-CXR average bounding box annotations. The fourth and fifth column represents configurations evaluated on the ChestX-ray14, while the last column, GT-C, represents the ChestX-ray14 average bounding box annotations.

Several of the previously discussed outcomes can also be inferred through an analysis of the heatmaps. For example, it becomes readily apparent that the Cmg class ranks among the most successful, whereas the explanation for the Pnm and Nod classes appears to be more challenging. Furthermore, it was previously stated that the VC models exhibit poor performance in the Nod,

PIE, and Pnm classes, consistently ranking the lowest among all configurations in these classes. These conclusions manifest in the heatmaps generated for these specific classes, consistently congregating around the lower-right corner. This trend serves to emphasize the model's limitations in accurately depicting the unique attributes of these respective pathologies, leading to poor results.

It's worth highlighting that the bounding box annotations in the ChestX-ray14 dataset tend to be larger in comparison to those in the VinDr-CXR dataset. Such tendency can be seen in the annotations for Cmg, a class that always appears in the location. In addition, there is a lot of variety regarding the annotation positions between datasets. For example, for the Alt class, the annotations are concentrated in the top left, while for the ChestX-ray14 they are more present in the bottom right.

Chapter 7

Conclusions and Future Work

This study successfully developed a novel explanation technique based on structured reports and compared various visual explanations among different models and datasets.

The proposed structured report explanation converts visual explanations into a standardized format resembling a medical report, aiding interpretability and facilitating communication with medical staff. It organizes findings in a structured way while allowing for direct comparison with ground truth annotations. As a further improvement, the reports could include information about the relative size of the pathology, thus being more specific and precise. Nonetheless, for comprehensive validation and improvement, this method should be generated at a larger scale and analyzed by radiologists to fully assess its advantages and disadvantages.

The implemented classification models showed competitive performance, and the extracted visual explanations offered valuable insights. Chapter 5 emphasized the model-specific nature of these visual explanations and the impact of reducing image size on their generation. Moreover, the outcomes across all metrics remained notably distant from the maximum attainable results, underscoring that there is still considerable ground to cover on the path toward achieving truly explainable deep learning models. Future research should involve expert radiologists' opinions and explore explanations beyond the heatmap to further enhance interpretability. In the same chapter, the statistical analysis revealed that some pathologies are easier to explain given their texture features and ground truth annotation properties.

The comparison in Chapter 6 highlighted that the explanations are significantly influenced by the dataset properties and annotations, even when dealing with the same pathology. This emphasizes one of the major challenges in visual explanations extracted from deep learning architectures: their lack of reproducibility.

Overall, this work effectively benchmarked various visual explanations and showcased their potential as valuable tools to gain insights into model reasoning. However, obtaining perfect explanations from deep learning models remains challenging due to their stochastic feature learning, leading to diverse outcomes that may or may not approximate the way humans learn and identify visual features. Hence, there is a need to delve deeper into understanding how these models learn and interpret pathologies. A first step towards that goal would involve annotating a greater number

of images and establishing standardized criteria for these annotations in order to reduce their variability and allow for more accurate conclusions. Studying these models, in turn, may then help the human expert to learn new patterns, improving detection and diagnosing.

References

- [1] World Health Organization Team, “The top 10 causes of death.” <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>, 2020. accessed on November 10, 2022.
- [2] NHS England and Performance Analysis Team. Diagnostic Imaging Dataset 2022-23 Data, “Statistics - diagnostic imaging dataset 2022-23 data.” <https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/diagnostic-imaging-dataset-2022-23-data/>, 2022. accessed on January 5, 2023.
- [3] J. Pedrosa, P. Sousa, J. Silva, A. M. Mendonça, and A. Campilho, “Lesion-based chest radiography image retrieval for explainability in pathology detection,” in *Iberian Conference on Pattern Recognition and Image Analysis*, (Aveiro, Portugal), pp. 81–94, Springer, 2022.
- [4] Columbia Surgery, “History of medicine: X-ray imaging | columbia surgery.” <https://columbiasurgery.org/news/2015/09/17/history-medicine-dr-roentgen-s-accidental-x-rays>. accessed on November 15, 2022.
- [5] D. Allard, “William D. Coolidge, Inventor of the Modern X-Ray Tube,” *Medical Physics*, vol. 43, 10 2016.
- [6] C. Landeros, “Electromagnetic spectrum 101: X-rays.” <https://gamma-sci.com/2021/10/13/electromagnetic-spectrum-101-x-rays/>, 10 2021. accessed on December 3, 2022.
- [7] A. Tafti and D. W. Byerly, “X-ray image acquisition.” <https://www.ncbi.nlm.nih.gov/books/NBK563236/>, 1 2022.
- [8] ImageCare Radiology, “Top 5 uses of x-ray | imagecare radiology.” <https://imagecarecenters.com/blog/top-5-uses-of-x-rays/>. accessed on November 22, 2022.
- [9] S. Christofides, A. Maidment, I. M. K.-H. Ng, and D. R. Dance, *Diagnostic Radiology Physics: A Handbook for Teachers and Students*. IAEA, 2014.
- [10] R. Joarde and N. Crundwell, *Chest X-Ray in Clinical Practice*. Springer London, 2009.
- [11] T. Peng, Y. Gu, Z. Ye, X. Cheng, and J. Wang, “A-lugseg: Automatic and explainability-guided multi-site lung detection in chest x-ray images,” *Expert Systems With Applications*, vol. 198, p. 116873, 2022.

- [12] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jul 2017.
- [13] M. J. R. de Matos, M. E. E. Rosa, V. M. Brito, L. T. W. Amaral, G. L. Beraldo, E. K. U. N. Fonseca, R. C. Chate, R. B. D. Passos, M. M. A. Silva, P. Yokoo, R. S. Neto, G. B. da Silva Teles, M. C. B. da Silva, and G. Szarf, “Differential diagnoses of acute ground-glass opacity in chest computed tomography: pictorial essay,” *Einstein*, vol. 19, p. eRW5772, 2021.
- [14] R. Sharma and F. Gaillard, “Cardiothoracic ratio.” <https://radiopaedia.org/articles/cardiothoracic-ratio>, 10 2011. accessed on December 10, 2022.
- [15] J. Jones, “Pleural effusion.” <https://radiopaedia.org/articles/6159>, 5 2009.
- [16] W. Brant and C. Helms, *Fundamentals of Diagnostic Radiology*, vol. 4. Wolters Kluwer Health, 2012.
- [17] D. Bell and C. Hacking, “The four ds of radiology reporting,” *Radiopaedia.org*, 4 2015.
- [18] S. Meystre, R. Gouripeddi, J. Tieder, J. Simmons, R. Srivastava, and S. Shah, “Enhancing comparative effectiveness research with automated pediatric pneumonia detection in a multi-institutional clinical repository: A PHIS pilot study,” *Journal of Medical Internet Research*, vol. 19, p. e162, 5 2017.
- [19] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng, R. G. Mark, and S. Horng, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Scientific Data*, vol. 6, 12 2019.
- [20] M. Hartung, I. Bickle, F. Gaillard, and J. Kanne, “How to create a great radiology report,” *RadioGraphics*, vol. 40, no. 6, pp. 1658–1670, 2020. PMID: 33001790.
- [21] J. H. Moon, H. Lee, W. Shin, and E. Choi, “Multi-modal understanding and generation for medical images and text via vision-language pre-training,” *CoRR*, vol. abs/2105.11333, 2021.
- [22] R. Burton, M. Albur, M. Eberl, and S. Cuff, “Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections,” *BMC Medical Informatics and Decision Making*, vol. 19, 9 2019.
- [23] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, “Concepts of artificial intelligence for computer-assisted drug discovery,” *Chemical Reviews*, vol. 119, pp. 10520–10594, 7 2019.
- [24] Z. Obermeyer and E. J. Emanuel, “Predicting the future — big data, machine learning, and clinical medicine,” *New England Journal of Medicine*, vol. 375, pp. 1216–1219, 9 2016.
- [25] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, “Machine learning and integrative analysis of biomedical big data,” *Genes*, vol. 10, p. 87, 1 2019.
- [26] J. Rocha, A. M. Mendonça, and A. Campilho, “A review on deep learning methods for chest x-ray based abnormality detection and thoracic pathology classification,” *U. Porto Journal of Engineering*, vol. 7, no. 4, pp. 16–32, 2021.

- [27] V. H. Phung and E. J. Rhee, “A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets,” *Applied Sciences (Switzerland)*, vol. 9, 11 2019.
- [28] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers, “Automated abnormality classification of chest radiographs using deep convolutional neural networks,” *npj Digital Medicine*, vol. 3, 5 2020.
- [29] J. A. Dunnmon, D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren, “Assessment of convolutional neural networks for automated classification of chest radiographs,” *Radiology*, vol. 290, pp. 537–544, 2 2019.
- [30] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, Hawaii), IEEE, 7 2017.
- [31] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” *arXiv:1711.05225 [cs]*, 2017.
- [32] P. Kumar, M. Grewal, and M. M. Srivastava, “Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs,” in *Lecture Notes in Computer Science*, pp. 546–552, Springer International Publishing, 2018.
- [33] K. Urinbayev, Y. Orazbek, Y. Nurambek, A. Mirzakhmetov, and H. A. Varol, “End-to-end deep diagnosis of x-ray images,” *arXiv:2003.08605 [eess.IV]*, 2020.
- [34] S. Gündel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, “Learning to recognize abnormalities in chest x-rays with location-aware dense networks,” in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 757–765, Springer International Publishing, 2019.
- [35] B. Chen, J. Li, X. Guo, and G. Lu, “DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest x-rays,” *Biomedical Signal Processing and Control*, vol. 53, p. 101554, 8 2019.
- [36] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, “Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification,” *arXiv:1801.09927 [cs.CV]*, 2018.
- [37] S. Kabir, L. Farrokhvar, and A. Dabouei, “A weakly supervised approach for thoracic diseases detection,” *Expert Systems with Applications*, vol. 213, p. 118942, 2023.
- [38] Y. Luo, Y. Zhang, X. Sun, H. Dai, and X. Chen, “Intelligent solutions in chest abnormality detection based on YOLOv5 and ResNet50,” *Journal of Healthcare Engineering*, vol. 2021, pp. 1–11, 10 2021.
- [39] G. Jocher, Ayush Chaurasia, A. Stoken, J. Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, Imyhxy, , Lorna, Zeng Yifu), C. Wong, Abhiram V, D. Montes, Zhiqiang Wang, C. Fati, Jebastin Nadar, Laughing, UnglvKitDe, V. Sonck, Tkianai, P. Skalski, A. Hogan, Dhruv Nair, M. Strobel, and M. Jain, “Yolov5 sota realtime instance segmentation,” *Zenodo*, 2022.

- [40] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu, “Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations,” *arXiv:2012.15029 [eess.IV]*, 2020.
- [41] P. Pedrosa, João, J. Sousa, A. M. Mendonça, and A. Campilho, “Location-aware chest radiography abnormality classification with object detection yolov5,” in *submitted*, 2022.
- [42] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, “How can i explain this to you? an empirical study of deep neural network explanation methods,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4211–4222, 2020.
- [43] Y. Xie, M. Chen, D. Kao, G. Gao, and X. Chen, “Chexplain: enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (Honolulu, Hawaii), pp. 1–13, 2020.
- [44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 2921–2929, IEEE Computer Society, 6 2016.
- [45] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, “Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022),” *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107161, 11 2022.
- [46] G. Vilone and L. Longo, “A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods,” *Frontiers in Artificial Intelligence*, vol. 4, 11 2021.
- [47] G. Ras, N. Xie, M. van Gerven, and D. Doran, “Explainable deep learning: A field guide for the uninitiated,” *arXiv:2004.14545 [cs.LG]*, 2020.
- [48] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv:1312.6034 [cs.CV]*, 2013.
- [49] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” *arXiv:1605.01713*, 2016.
- [50] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *arXiv:1703.01365 [cs.LG]*, 2017.
- [51] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, (Sydney, Australia), pp. 3145–3153, PMLR, 8 2017.
- [52] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methods for deep neural networks,” *arXiv:1711.06104*, 2017.

- [53] E. Johnson, S. Mohan, A. Gaudio, A. Smailagic, C. Faloutsos, and A. Campilho, "Heartspot: Privatized and explainable data compression for cardiomegaly detection," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, (Ioannina, Greece), pp. 01–04, 2022.
- [54] D. Mata, W. Silva, and J. S. Cardoso, "Increased robustness in chest x-ray classification through clinical report-driven regularization," in *Pattern Recognition and Image Analysis* (A. J. Pinho, P. Georgieva, L. F. Teixeira, and J. A. Sánchez, eds.), (Cham, Switzerland), pp. 119–128, Springer International Publishing, 2022.
- [55] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, (Venice, Italy), pp. 618–626, 2017.
- [56] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.
- [57] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Visual explanations for deep convolutional neural networks," *SN Computer Science*, vol. 2, 1 2021.
- [58] M. Chetoui, Moulay, Akhloufi, B. Yousefi, and E. M. Bouattane, "Explainable COVID-19 detection on chest x-rays using an end-to-end deep convolutional neural network architecture," *Big Data and Cognitive Computing*, vol. 5, p. 73, 12 2021.
- [59] J. Hou and T. Gao, "Explainable DCNN based chest x-ray image analysis and classification for COVID-19 pneumonia detection," *Scientific Reports*, vol. 11, 8 2021.
- [60] R. K. Singh, R. Pandey, and R. N. Babu, "COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest x-rays," *Neural Computing and Applications*, vol. 33, pp. 8871–8892, 1 2021.
- [61] W. Shi, L. Tong, Y. Zhu, and M. D. Wang, "COVID-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2376–2387, 7 2021.
- [62] L. Devnath, Z. Fan, S. Luo, P. Summons, and D. Wang, "Detection and visualisation of pneumoconiosis using an ensemble of multi-dimensional deep features learned from chest x-rays," *International Journal of Environmental Research and Public Health*, vol. 19, p. 11193, 9 2022.
- [63] A. I. Suryani, C.-W. Chang, Y.-F. Feng, T.-K. Lin, C.-W. Lin, J.-C. Cheng, and C.-Y. Chang, "Lung tumor localization and visualization in chest x-ray images using deep fusion network and class activation mapping," *IEEE Access*, vol. 10, pp. 124448–124463, 2022.
- [64] K. Kusunose, Y. Hirata, N. Yamaguchi, Y. Kosaka, T. Tsuji, J. Kotoku, and M. Sata, "Deep learning for detection of exercise-induced pulmonary hypertension using chest x-ray images," *Frontiers in Cardiovascular Medicine*, vol. 9, 2022.
- [65] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion location attention guided network for multi-label thoracic disease classification in chest x-rays," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2016–2027, 2020.

- [66] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. Ebert, S. Pomerantz, J. Romero, S. Kamalian, R. Gonzalez, M. Lev, and S. Do, “An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets,” *Nature Biomedical Engineering*, vol. 3, pp. 173–182, 12 2018.
- [67] G. Dissez, N. Tay, T. Dyer, M. Tam, R. Dittrich, D. Doyne, J. Hoare, J. J. Pat, S. Patterson, A. Stockham, Q. Malik, T. N. Morgan, P. Williams, L. Garcia-Mondragon, J. Smith, G. Pearse, and S. Rasalingham, “Enhancing early lung cancer detection on chest radiographs with ai-assistance: A multi-reader study,” *arXiv:2208.14742 [eess.IV]*, 2022.
- [68] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, “Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based alzheimer's disease classification,” *Frontiers in Aging Neuroscience*, vol. 11, 7 2019.
- [69] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLOS ONE*, vol. 10, p. e0130140, 7 2015.
- [70] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert, K.-R. Müller, and F. Klauschen, “Morphological and molecular breast cancer profiling through explainable machine learning,” *Nature Machine Intelligence*, vol. 3, pp. 355–366, 3 2021.
- [71] M. U. Alam, J. R. Balvinsson, and Y. Wang, “Exploring LRP and grad-CAM visualization to interpret multi-label-multi-class pathology prediction using chest radiography,” in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, address = Shenzhen, China, IEEE, 7 2022.
- [72] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” *arXiv:1901.07031 [cs.CV]*, 2019.
- [73] M. R. Karim, T. Dohmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, “DeepCOVIDExplainer: Explainable COVID-19 diagnosis from chest x-ray images,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, address = Seoul, South Korea, IEEE, 12 2020.
- [74] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision – ECCV 2014*, pp. 818–833, Springer International Publishing, 2014.
- [75] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, “Abnormality detection and localization in chest x-rays using deep convolutional neural networks,” *arXiv:1705.09850 [cs.CV]*, 2017.
- [76] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, “Explainable deep CNNs for MRI-based diagnosis of alzheimer's disease,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, (Glasgow, UK), IEEE, 7 2020.
- [77] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you? explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (California, USA), pp. 1135–1144, 2016.

- [78] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An explainable machine learning model for early detection of parkinson's disease using LIME on DaTSCAN imagery," *Computers in Biology and Medicine*, vol. 126, p. 104041, 11 2020.
- [79] Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, "Deep learning for prediction of depressive symptoms in a large textual dataset," *Neural Computing and Applications*, vol. 34, pp. 721–744, 8 2021.
- [80] P. Schneider, D. Müller, and F. Kramer, "Classification of viral pneumonia x-ray images with the aucmedi framework," *arXiv:2110.01017 [cs.CV]*, 2021.
- [81] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," *Applied Sciences*, vol. 8, p. 1715, 9 2018.
- [82] M. M. Ahsan, R. Nazim, Z. Siddique, and P. Huebner, "Detection of covid-19 patients from ct scan and chest x-ray data using modified mobilenetv2 and lime," *Healthcare*, vol. 9, no. 9, 2021.
- [83] E. G. Cervantes and W.-Y. Chan, "Lime-enabled investigation of convolutional neural network performances in covid-19 chest x-ray detection," in *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, (Canada), pp. 1–6, 2021.
- [84] N. S. Punna and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks," *Applied Intelligence*, vol. 51, pp. 2689–2702, 10 2020.
- [85] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv:1705.07874 [cs.AI]*, 2017.
- [86] L. S. Shapley, "17. a value for n-person games," in *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318, Princeton University Press, 12 1953.
- [87] Z. Jiang, L. Bo, Z. Xu, Y. Song, J. Wang, P. Wen, X. Wan, T. Yang, X. Deng, and J. Bian, "An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission," *Computer Methods and Programs in Biomedicine*, vol. 204, p. 106040, 6 2021.
- [88] N. Beebe-Wang, A. Okeson, T. Althoff, and S.-I. Lee, "Efficient and explainable risk assessments for imminent dementia in an aging cohort study," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2409–2420, 7 2021.
- [89] J. Peng, K. Zou, M. Zhou, Y. Teng, X. Zhu, F. Zhang, and J. Xu, "An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients," *Journal of Medical Systems*, vol. 45, 4 2021.
- [90] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach, and D. Frey, "Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome," *PLOS ONE*, vol. 15, p. e0231166, 4 2020.
- [91] M. Bhandari, T. B. Shahi, B. Siku, and A. Neupane, "Explanatory classification of cxr images into covid-19, pneumonia and tuberculosis using deep learning and xai," *Computers in Biology and Medicine*, vol. 150, p. 106156, 2022.

- [92] J. H. Ong, K. M. Goh, and L. L. Lim, “Comparative analysis of explainable artificial intelligence for covid-19 diagnosis on cxr image,” in *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 185–190, 2021.
- [93] E. M. Kenny and M. T. Keane, “Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ann-cbr twins for xai,” in *Twenty-Eighth International Joint Conferences on Artificial Intelligence*, (Macau, China), pp. 2708–2715, 2019.
- [94] W. Silva, A. Poellinger, J. Cardoso, and M. Reyes, “Interpretability-guided content-based medical image retrieval,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Lima, Peru), pp. 305–314, Springer, 2020.
- [95] A. Zhong, X. Li, D. Wu, H. Ren, K. Kim, Y. Kim, V. Buch, N. Neumark, B. Bizzo, W. Y. Tak, S. Y. Park, Y. R. Lee, M. K. Kang, J. G. Park, B. S. Kim, W. J. Chung, N. Guo, I. Dayan, M. K. Kalra, and Q. Li, “Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in covid-19,” *Medical Image Analysis*, vol. 70, p. 101993, 2021.
- [96] World Health Organization Office, “The protection of personal data in health information systems-principles and processes for public health.” <https://apps.who.int/iris/handle/10665/341374>, 2021. accessed on January 4, 2023.
- [97] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. H. Truong, C. D. T. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, “Benchmarking saliency methods for chest x-ray interpretation,” *Nature Machine Intelligence*, vol. 4, pp. 867–878, 10 2022.
- [98] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017.
- [99] N. Nguyen, P. Truong, V. Ho, and T. Nguyen, “VinDr Lab: A Data Platform for Medical AI,” 2021.
- [100] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.
- [101] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv:1512.03385 [cs.CV]*, 2015.
- [102] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (Hawaii, USA), pp. 4700–4708, 2017.
- [103] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

- [104] J. Gildenblat and contributors, “Pytorch library for cam methods.” <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- [105] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for pytorch,” 2020.
- [106] B. Ricardo, P. João, M. A. Maria, and C. Aurélio, “Semi-supervised multi-structure segmentation in chest x-ray imaging,” in *IEEE CBMS*, (Venice, Italy), 2023.
- [107] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.
- [108] S. Seabold and J. Perktold, “Statsmodels: Econometric and statistical modeling with python,” in *9th Python in Science Conference*, 2010.

Appendix A

Supplementary Figures

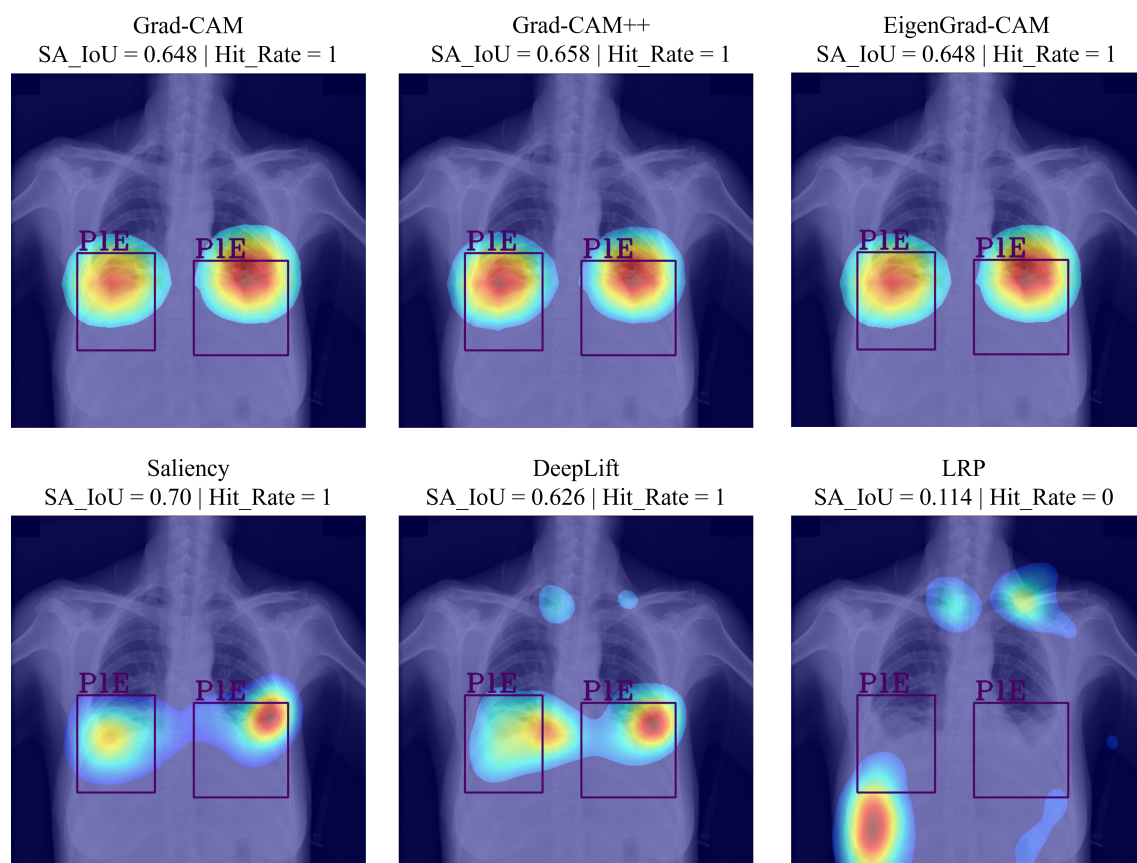


Figure A.1: Example of CXR with the PIE class and corresponding heatmaps after applying the size adaptive threshold for different explanations, along with the ground truth bounding box. The first row displays all the CAM-based explanations, while the second row contains gradient-based explanations.

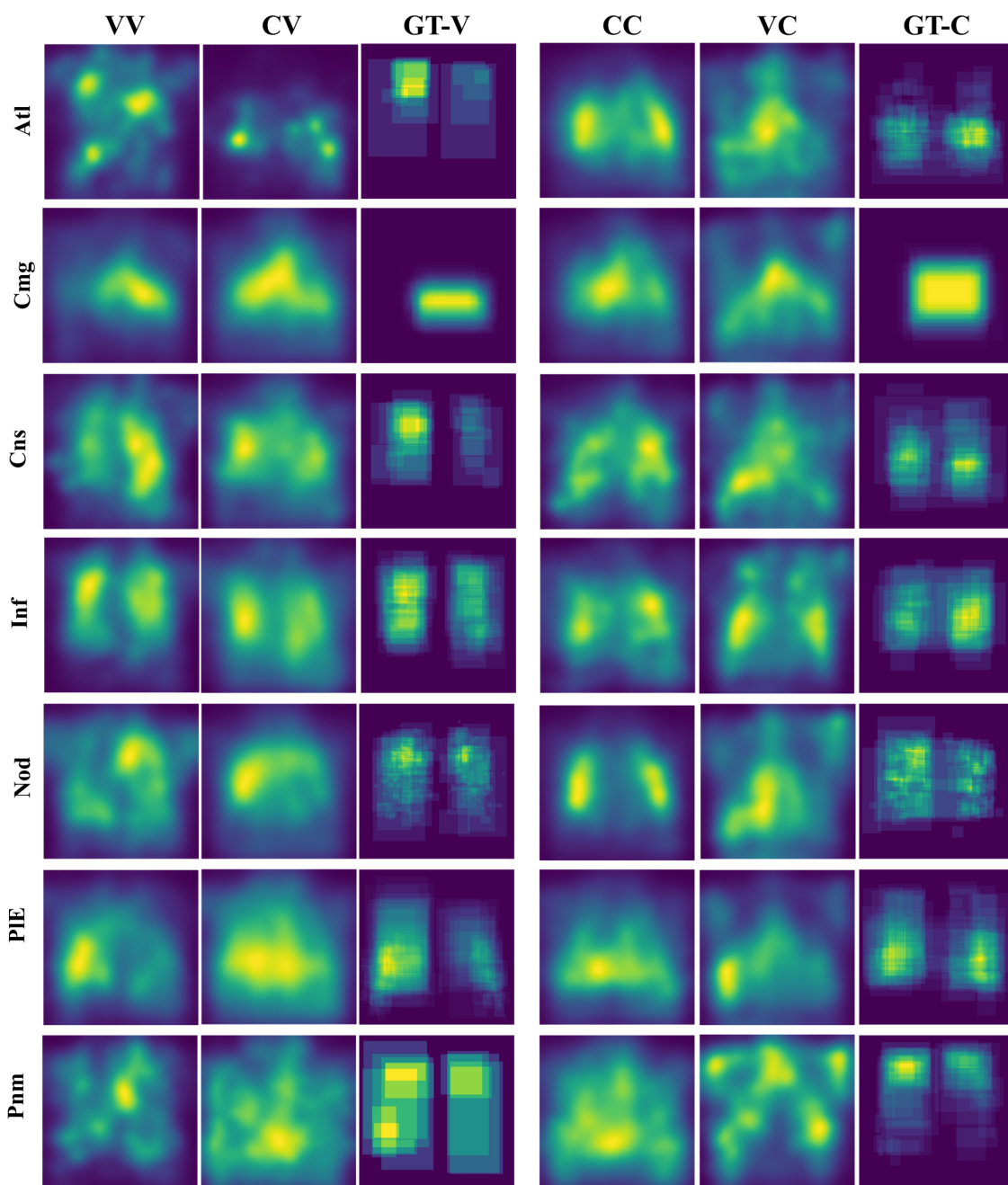


Figure A.2: Average Grad-CAM++ heatmaps for all configurations. The first two columns represent configurations evaluated on the VinDr-CXR dataset and the third, GT-V, represents the VinDr-CXR average bounding box annotations. The fourth and fifth column represents configurations evaluated on the ChestX-ray14, while the last column, GT-C, represents the Chest X-ray 14 average bounding box annotations.