Automatic Eyetracking-Assisted Chest Radiography Pathology Screening

Rui Manuel Azevedo dos Santos

Dissertation

Supervisor: Prof. Dr. João Manuel Patrício Pedrosa
Co-Supervisor: Prof. Dr. Ana Maria Rodrigues de Sousa Faria de Mendonça

Mestrado em Bioengenharia

2022-07-03

Abstract

The lungs and heart are targets of a myriad of life-threatening diseases. Chest radiography, one of the most common medical procedures worldwide, is a vital tool to study such conditions, since it is easy to perform and affordable. However, the high demand for Chest X-Ray (CXR) examinations creates a huge burden on radiologists, who could benefit from the support of new Artificial Intelligence (AI) methods.

With the ongoing digital transformation in clinical practice chest radiographs are being stored in massive amounts, which can be collected and annotated to create large datasets with the goal of developing Deep Learning (DL) models. Such models can be applied in the task of pathology classification, for instance, to provide a second opinion or for patient stratification. Nevertheless, DL models are undergoing a limited implementation in real world scenarios, due to the lack of trustworthy explanations that provide insights about their reasoning. Consequently, DL generates reluctance in the medical community.

The need for explanations generated a change of paradigm in DL, and an increasing focus in the topic of explainability. Several methods have been developed to accommodate this need, including model-agnostic methods that can be used with virtually any DL model. This also prompted the design of models capable of rendering better explanations, since providing further evidence alone is insufficient.

In this work, Eye-Tracking Data (ETD) was used to train models with the goal of performing CXR pathology classification in a more explainable way. ETD consists, essentially, in a set of data points relative to the gaze fixations of radiologists, including the respective coordinates and time duration. ETD was converted to images, namely heatmaps, in order to be used with Convolutional Neural Networks (CNN). The first stage consisted in reconstructing the heatmaps using CXR images as inputs. This has the main advantage of making the proposed framework independent of the ETD during inference. Then, two approaches were developed. The first one consists in using the reconstructed heatmaps to segment the thorax in CXR images. By providing only the thorax area to a subsequent classifier, it is guaranteed that the model does not use irrelevant information present in other parts of the images. The second approach leverages the fact that reconstructed heatmaps contain higher pixel intensities in pathological areas, and multiplies them by the corresponding CXR images. Ultimately, the objective is to guide the classifier into focusing on important features for the prediction.

Three ETD datasets available in the literature, EGD, REFLACX, and CXR-P, were used to train and/or test the models. Furthermore, several experiments were performed to select the best architectures for both the heatmap reconstruction and pathology classification tasks. A UNet with a pretrained DenseNet121 encoder was chosen for the first task, and a pretrained EfficientNet-b0 as the classifier for both approaches. The developed models were compared with the state-of-the-art in CXR pathology classification using ETD, and with a baseline classifier that received as input the original images.

The results obtained show that the developed models are comparable to the state-of-the-art

and to the baseline classifier. For the thorax segmentation approach, mean Area Under the Curve (AUC) values of 0.82 and 0.72 were achieved for the EGD and REFLACX datasets, respectively. For the second approach, the mean AUC values were 0.81 and 0.74. Nevertheless, the explainability method used proved to be unreliable, demanding further experiments to demonstrate the link between the use of heatmaps and quality explanations. A more extensive optimization is also warranted for the proposed models to surpass the current state-of-the-art.

Keywords: Chest X-Ray, Deep Learning, Explainability, Eye-Tracking Data.

Resumo

Os pulmões e o coração são alvos de uma grande variedade de doenças que podem colocar em risco vidas humanas. A radiografia do tórax é um dos exames médicos mais comuns, e constitui uma ferramenta valiosa no estudo dessas mesmas doenças, dado que é um método prático e de reduzido custo. No entanto, o grande número de radiografias efetuadas cria dificuldades aos radiologistas, que poderiam beneficiar com a ajuda da inteligência artificial.

Com a contínua transformação digital que está a ocorrer na medicina, torna-se cada vez mais fácil armazenar radiografias em grandes quantidades que, ao serem agrupadas e anotadas, podem originar novos datasets que suportam o desenvolvimento de modelos de aprendizagem profunda. Esses modelos podem ser aplicados na classificação de patologias em radiografias do tórax, quer seja para dar uma segunda opinião aos médicos, quer seja para realizar a estratificação de pacientes. No entanto, estes modelos ainda apresentam uma implementação limitada em cenários reais, principalmente devido à sua incapacidade de fornecerem explicações credíveis relacionadas com as suas previsões. Isto gera relutância nos profissionais médicos em promover a aplicação destes modelos.

A necessidade de explicações gerou uma mudança de paradigma no campo da aprendizagem profunda, que se foca cada vez mais na explicabilidade dos modelos. Vários foram os métodos desenvolvidos até hoje, incluindo alguns que são independentes do modelo em si e que por isso podem ser usados em praticamente todas as situações. A par do desenvolvimento destes métodos, os investigadores também se focam cada vez mais em modelos capazes de fornecerem melhores explicações, dado que apenas a existência de uma explicação é insuficiente para fomentar a confiança numa dada decisão.

Neste trabalho, dados de *eye-tracking* foram usados com o objetivo de treinar modelos capazes de, para além da classificação de patologias, gerarem explicações de confiança. Os dados de *eye-tracking* consistem essencialmente em coordenadas relativas às fixações do olhar de radiologistas, e a respetiva duração temporal. Estes dados foram convertidos para *heatmaps*, tendo em vista a sua aplicação em modelos de aprendizagem profunda. A primeira fase do trabalho consistiu em reconstruir estes *heatmaps* a partir das radiografias do tórax. Esta fase inicial oferece a vantagem de, durante a inferência, os modelos serem independentes dos dados de *eye-tracking*. De seguida, duas abordagens são propostas em relação à classificação de patologias. A primeira realiza a segmentação do tórax em radiografias a partir dos *heatmaps* reconstruídos. Ao fornecer apenas a área do tórax a um classificador, é dada a garantia de que o modelo não se foca em partes irrelevantes das imagens, como marcas comuns nos cantos. A segunda abordagem aproveita o facto dos *heatmaps* conterem uma maior intensidade nas áreas onde se verificam patologias, e multiplica estes *heatmaps* pelas radiografias do tórax. O intuito desta abordagem é de guiar o classificador de forma a que este preste atenção a zonas importantes das imagens para a previsão.

Três *datasets* disponíveis na literatura, EGD, REFLACX, e CXR-P, foram usados para treinar e/ou testar os modelos. Além disso, várias experiências foram realizadas para selecionar as melhores arquiteturas tanto para reconstruir os *heatmaps* como para classificar as imagens. Uma

UNet com uma DenseNet121 pré-treinada como *encoder* foi escolhida para a primeira tarefa, e uma EfficientNet-b0 pré-treinada como classificador para ambas as abordagens. Os modelos desenvolvidos foram comparados com o estado da arte na classificação de radiografias do tórax com recurso a dados de *eye-tracking*, e com uma *baseline* que recebia como *input* as imagens originais.

Os resultados obtidos foram comparáveis ao estado da arte e à *baseline* utilizada. Para a abordagem correspondente à segmentação do tórax, valores médios de *area under the curve* de 0.82 e 0.72 foram obtidos para os *datasets* EGD e REFLACX. Para a segunda abordagem, os valores foram de 0.81 e 0.74. No entanto, o método utilizado para analisar a explicabilidade dos modelos não funcionou da forma desejada, o que implica a realização de mais experiências, com métodos alternativos, para averiguar a qualidade das explicações fornecidas. Para além disso, é necessária uma otimização mais extensa dos modelos propostos, de forma a que seja possível superar o estado da arte.

Palavras-chave: Radiografia do Tórax, Aprendizagem Profunda, Explicabilidade, Dados de *Eye-Tracking*.

Agradecimentos

Começo estes agradecimentos por demonstrar a minha gratidão para com a minha mãe. Desde bem cedo que incutiu em mim a ideia de que a escola era importante, o que me tornou um melhor aluno, e me permitiu desenvolver as capacidades que possibilitaram a minha entrada na faculdade. Obrigado ao meu pai, que durante a minha vida me proporcionou experiências que moldaram o meu caráter. Obrigado à minha irmã, companheira de brincadeiras durante a minha infância, e com quem posso partilhar tudo. E obrigado ao meu irmão, que desde sempre foi um modelo a seguir em termos de rigor e dedicação.

Estes cinco anos na universidade, que passaram a seis fruto da minha impulsividade e vontade de obter conhecimento, foram inesquecíveis e marcaram sem dúvida uma nova fase na minha vida, em que me tornei melhor como pessoa. Passei por experiências incríveis (e inconcebíveis), que certamente ficarão gravadas na minha memória para todo o sempre. Agradeço a todos os meus amigos que sempre me acompanharam, e que me continuarão a acompanhar mesmo estando espalhados pelo mundo inteiro.

Foi também durante esta aventura que adquiri o gosto pela ciência, nas suas várias vertentes. Sinto-me uma pessoa mais repleta e preparada para o que aí vem, independentemente do que o futuro me reserva. Em especial, quero agradecer à professora Perpétua Pinto do Ó, que me proporcionou a minha primeira experiência científica.

Relativamente a esta tese, não posso deixar de agradecer ao meu orientador, João Pedrosa, que sempre foi impecável, acessível, e que me ajudou em tudo o que foi preciso. Agradeço também à professora Ana Maria Mendonça, pelos seus comentários sempre construtivos. E às pessoas do grupo de investigação onde me inseri, pela sua imensa simpatia.

Não me esqueci de ti, Beatriz, que me tens acompanhado mais de perto nestes últimos anos. Um especial obrigado pelo apoio e confiança que me deste durante estes meses em que realizei a dissertação, e por todas as aventuras que temos passado juntos.

Por último, não posso deixar de agradecer a mim próprio (pressupondo a existência de livre-arbítrio). Aguardo com curiosidade pelo que o próximo capítulo me reserva.

Rui Manuel Azevedo dos Santos

"Na vida, nada se descobre.
As coisa, sim, se revelam."

Mia Couto

Contents

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
BCE	Binary Cross Entropy
CHF	Congestive Heart Failure
CNN	Convolutional Neural Network
CR	Computed Radiography
CT	Computed Tomography
CXR	Chest X-Ray
DL	Deep Learning
DR	Digital Radiography
DSC	Dice Similarity Coefficient
ET	Eye-Tracking
ETD	Eye-Tracking Data
FPR	False Positive Rate
GradCAM	Gradient-weighted Class Activation Mapping
GT	Ground Truth
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
NS	Noisy-student
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SFR	Screen Film Radiography
TL	Transfer Learning
TPR	True Positive Rate
VR	Virtual Reality

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

Cardiac and respiratory diseases are amongst the main causes of severe illness and death worldwide. The heart, and the major vessels around it, are suscetible to a myriad of life-threatening conditions. The lungs, on the other hand, constitute the internal organ more sensitive to infection and injury due to the daily contact with pathogens and chemicals [18]. This may lead to several pathologies, such as respiratory infections, with an estimated death toll of almost 4 million people every year [19], or lung cancer, which is the most common and deadly neoplasm in men and the second deadliest in women [20]. Since both these organs are located in the thoracic cavity, a necessity arises to study this area of the human body in great detail.

Chest radiography is one of the most commonly prescribed medical examinations. It is relatively inexpensive, noninvasive, quick to perform and uses only a small dose of radiation. It allows physicians to look into the thoracic cavity and diagnose diseases characterized by changes in contrast or shape of the represented structures. This makes it an invaluable resource in every clinic or hospital around the world.

Furthermore, with the advent of digital radiography (DR), it is now possible to store Chest X-Ray (CXR) images in great quantity. This allowed for the creation of extensive datasets that, alongside the increasing computational power available, were a key factor in the development of Deep Learning (DL) models capable of detecting and classifying different health problems in this type of medical images.

1.2 Motivation

Given the high number of patients requiring chest radiographs and the difficulty in diagnosing certain pathologies, such a high number of examinations creates a huge burden on radiologists all over the world. This creates room for the introduction of DL models capable of performing a variety of tasks in order to assist healthcare professionals. However, despite the constant advances in DL, its actual applications in the medical imaging field are still scarce. One of the main reasons

for this is the fact that DL models constitute black-boxes, i.e, their underlying processes leading to a prediction are hard to interpret and to explain. Conversely, medical decisions have to be explainable - they have to be based on medical facts that can be explained to the patient and others. Therefore, if a DL model is to be trusted, it needs to explain its decision. In the medical imaging field, one of the most important and straightforward means of explainability consists in indicating what parts of the image were relevant for the decision, and what features were used.

This work aims to design DL models that can classify CXR pathologies in a more explainable way. In order to do so, datasets containing Eye-Tracking Data (ETD) are used. In essence, ETD contains all the locations a radiologist visited with his sight during a medical image analysis, and the time spent at each point. Since the radiologists focus on relevant parts of the image, it is possible to use this type of data to guide DL models. However, the ETD is not directly used to guide the classifiers. Instead, it is first used to train other models in the heatmap reconstruction task. Ultimately, the goal is to shift the attention of classifiers towards meaningful areas of a CXR image, using the reconstructed heatmaps, while being independent of ETD during inference.

1.3 Objectives

The objectives of this work are:

- Creation of models capable of performing ETD heatmap reconstruction using only CXR images as input;
- Creation of models that, by using the reconstructed heatmaps as a guide, are capable of outputting more reliable and explainable predictions;
- Validation of these models in different datasets.

1.4 Structure

The remainder of this work is composed by eight chapters. Chapter 2 describes the performed literature review, from the production methods of a chest radiograph and its medical relevance, to automatic methods of pathology detection and classification using DL. In Chapter 3, the Eye-Tracking (ET) datasets used, alongside the DL architectures integrated in the developed models, are presented in detail. Chapter 4 reports on the ETD analysis performed for the various datasets, and Chapter 5 describes the methods used to generate the heatmaps. Chapter 6 depicts the results for the baseline experiments, both for heatmap reconstruction and for pathology classification. In Chapter 7, the proposed approaches are described, and their corresponding results are presented. Chapter 8 contains the discussion of the results, and comments regarding future work. Finally, Chapter 9 presents key considerations regarding the outcome of this work.

Chapter 2

Chest Radiography

This chapter starts by explaining how radiography works and how an image can be obtained. Then, the clinical relevance of a chest radiograph is discussed, namely what diseases can be diagnosed from a CXR image. Afterwards, a description of the contributions of DL in chest radiography analysis is performed, and relevant models that have surfaced over the years are shown. In the final sections, examples regarding the use of ETD in the x-ray imaging field are presented, alongside the main contributions of this work.

2.1 Physics of Radiography

X-rays were discovered in 1895 by Wilhelm Conrad Röntgen while experimenting with electrical current emissions in vacuum. Since then, a lot more is known about the nature of x-rays, which allowed its incorporation as the basis for many medical imaging techniques. In radiography, x-rays are produced from vacuum tubes made of glass, containing a cathode and an anode. The cathode consists of a filament that, upon passage of an electrical current, increases in temperature. This heating process will lead to the emission of electrons from the filament. By regulating either the voltage or the intensity of the current, it is possible to control the energy contained in the electrons or the number of released electrons, respectively. The electrons are then directed through the vacuum tube into the metallic anode due to the existence of an acceleration voltage between the negative cathode and the positive anode. When the electrons hit the anode they interact with its atoms, which leads to a deceleration and concomitant loss of energy from the electrons. The energy lost is then converted to x-rays. In Figure 2.1 there is a schematic of a vacuum x-ray tube.

The generated x-rays have a broad energy spectrum. Since the very low energy x-ray photons do not contribute to the formation of an image, they are filtered using a thin metallic plate. This prevents the patient from absorbing needless radiation. On the other hand, the unfiltered part of the spectrum is preserved and the photons are emitted towards the patient. Upon contact with the human body, there is a process of attenuation, where x-rays may disappear, change direction,

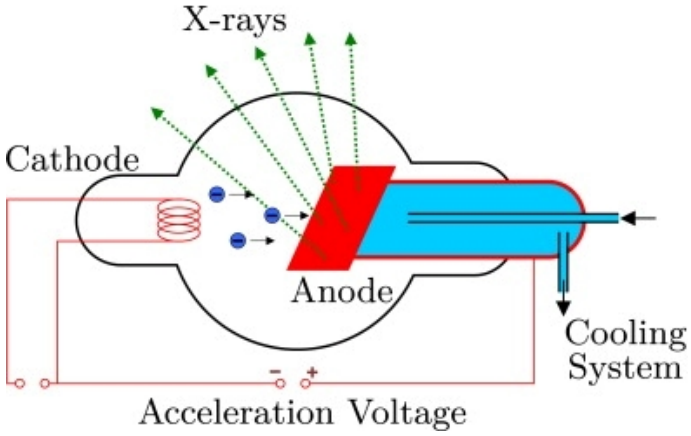


Figure 2.1: Schematic representation of a vacuum x-ray tube [1].

or lose energy. Different tissues have distinct properties, which will affect the attenuation of the photons. Bone, for instance, is more opaque to x-rays, while soft tissues are more transparent. By recording the number of photons capable of passing through the patient, it is possible to get a high resolution image regarding the existing structures.

2.1.1 Screen Film Radiography

Currently there are different ways available to acquire an x-ray image. Conventional Radiography, or Screen Film Radiography (SFR), makes use of a cassette containing a pair of screens and a film in between them. Each new radiograph requires the loading of a different screen-film combination. The screens are composed of phosphor, and the film by a photosensitive emulsion. Upon exposure of the screen to x-rays, part of the deposited energy is converted to light, which stimulates the film into forming a latent image. This latent image is then converted to a permanent image through the deposit of silver in the emulsion layer of the film. Only one screen may be used for a higher resolution (e.g. mammography), or just the film (in the case of dental radiography) [21].

SFR presents some drawbacks regarding its use. It presents a fixed dose latitude, which limits the ability to reduce the patient exposure to x-rays. The film used is expensive, it is composed of hazardous materials, it takes longer to process, and it needs to be stored physically. Furthermore, the fact that the images are not stored digitally prevents any post-processing [22].

2.1.2 Computed Radiography

In Computed Radiography (CR) the imaging plate is composed of photostimulable phosphor, similarly to SFR. However, instead of using the light produced by the phosphor in response to X-rays, this material contains traps for the excited electrons. The number of electrons trapped at each point of the plate structure is proportional to the incident radiation. A subsequent scanning with a red laser beam releases the electrons, causing the emission of light. This light is guided towards a photomultiplier tube which detects and amplifies the signal, leading to the formation of a digital

image. After this process, the plate is irradiated with light to delete any remaining information, and is available for repeated use.

CR, in comparison with SFR, has the obvious advantages of delivering digital images and employing reusable plates. On the other hand, the fact that it only uses one screen in the cassette decreases its absorption efficiency when compared to dual screen strategies [21].

2.1.3 Digital Radiography

DR is a more modern technique, allowing for direct digital readout. This speeds up the process of obtaining an image by decreasing the number of steps required, while increasing the spatial resolution. DR uses flat panel detectors, and is composed of two variants: direct and indirect conversion. In indirect conversion, a phosphor layer is still used to produce light after stimulation by x-rays. The generated light is then converted into electrical charges by a matrix of photodiodes, which are stored in capacitors. Indirect conversion flat panel detectors also contain thin-film transistors that act as switches during the readout process. In the case of direct conversion, there is no phosphor layer and thus no intermediate production of light. Instead, flat panel detectors include a layer of sensitive photoconducters on the thin-film transistor matrix, which allows for the direct generation of charged particles.

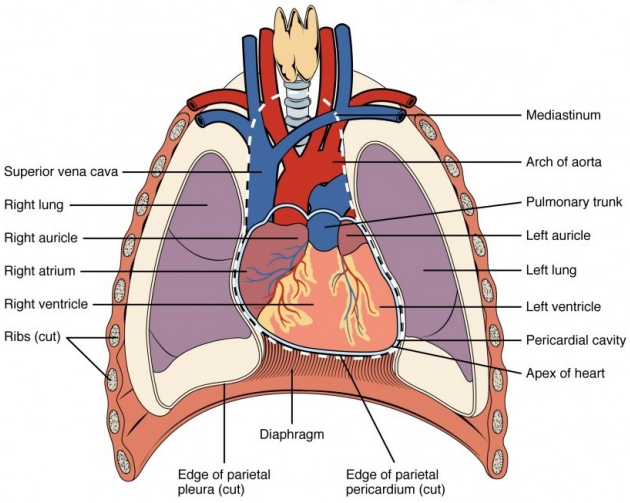
Direct conversion presents a higher conversion efficiency of incident energy to charged particles, and better resolution. However, technical limitations still make indirect conversion the preferred method [21].

2.2 Diagnostic Value

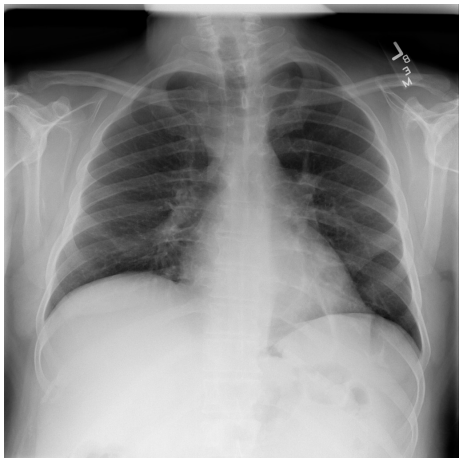
In addition to plain radiography, several other imaging techniques exist. Computerized Tomography (CT), Positron Emission Tomography and Fluoroscopy are examples of modalities involving radiation. Magnetic Resonance Imaging (MRI), on the other hand, uses magnetic fields. Even though some of these modalities have advantages regarding plain radiography - in CT, for instance, the acquired images are 3D and more informative - this is still the most commonly prescribed x-ray imaging method [23]. The main reasons are that it is quick to perform, inexpensive, there is a low dose of radiation involved, and it can be portable, making it ideal for different scenarios such as emergency departments.

Chest radiography, with the exception of dental x-rays, is the most frequent plain radiography procedure, composing around 56% of the total non-dental radiographs in Portugal (347 out of 621 per 1,000 population) [23]. This indicates the importance of studying pathologies of the thoracic cavity, and the diagnostic relevance of a chest radiograph. In Figure 2.2, the main structures of the thorax are shown alongside an example of a chest radiograph.

As it is possible to see, a CXR image covers several important anatomical parts of the human body, which are subjected to a myriad of conditions that require diagnosis. Below there is a description of several of those conditions that can be studied via chest radiography. Some of those are also depicted in Figure 2.3.



(a) Representation of the thorax [24].



(b) CXR image [9].

Figure 2.2: Anatomy of the thorax and example of a chest radiograph.

Mediastinum

The mediastinum is the central compartment of the thoracic cavity, located between the lungs. It contains, among other structures, the heart and major vessels. Some of the possible findings in these areas include abnormal mediastinal contour, consisting of changes in the outline of the mediastinum, or enlarged cardiac silhouette, characterized by an enlargement of the heart. The former can indicate the presence of a neoplasm, aortic aneurysm or enlarged lymph nodes [2]. The latter, on the other hand, is caused by heart problems such as Congestive Heart Failure (CHF). In CHF, a chronic condition, the heart is not able to pump enough blood and fluid builds up, leading to an enlargement of this organ.

Lungs

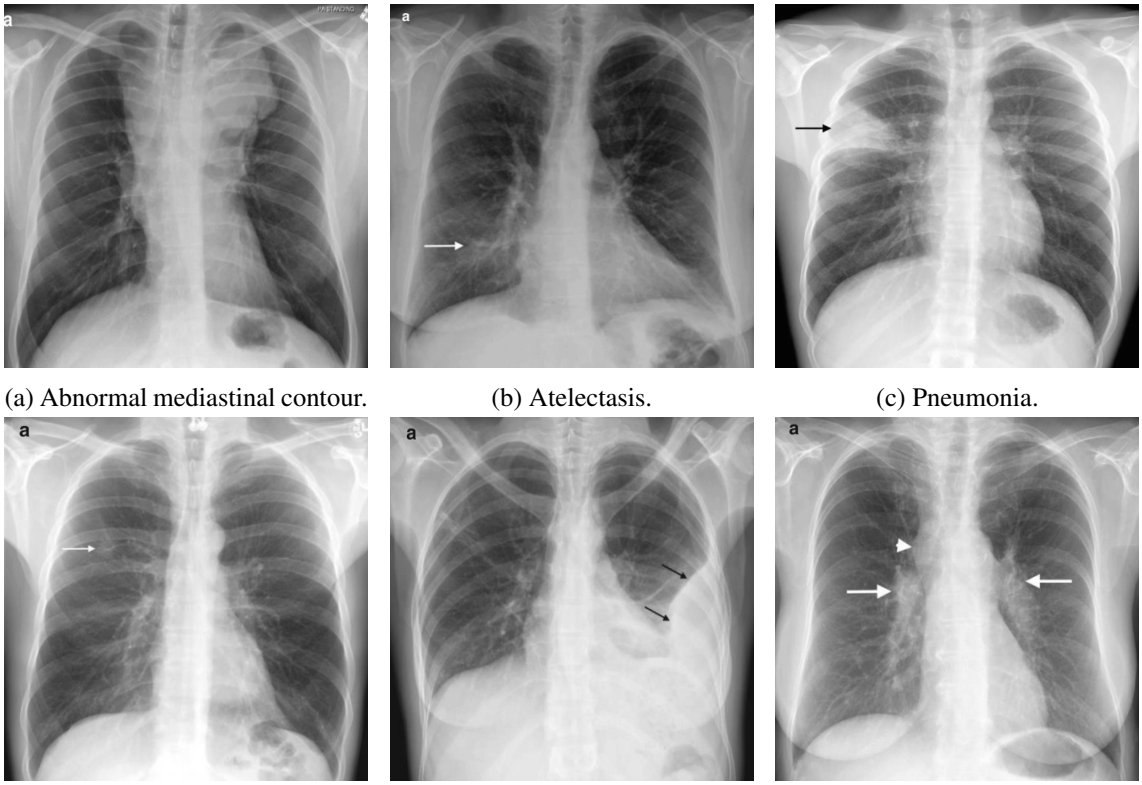
The lungs may contain many different pathologies. Atelectasis is the partial or total collapse of a lung, meaning the lung is uncapable of properly inflating. On the other hand, a high lung volume may indicate emphysema. It belongs to a group of lung diseases known as chronic obstructive pulmonary diseases, and it is characterized by the destruction of the alveoli walls, leading to obstructions that trap air inside the lungs. The presence of a consolidation in a CXR image - white patch located on the lungs - usually indicates the occurrence of pneumonia, an infection in the lungs. Groundglass opacities, diffuse areas in the lungs, constitute another type of finding and may suggest the existence of infections or interstitial diseases. Other possible findings are lung nodules or masses, which may indicate the presence of lung cancer. More recently, chest radiographs have also been used to diagnose COVID-19.

Pleura

The pleura is a thin membrane surrounding and protecting the lungs. Furthermore, it secretes a lubricant that aids the movement of the lungs during air intake. In terms of findings, the possibilities are pneumothorax (presence of gas in the pleura) or pleural abnormality. The latter can either constitute a pleural effusion (presence of fluid in the pleural space) or a pleural thickening (caused by plaques or neoplasms) [2].

Others

Besides the abovementioned findings, CXR images are also suitable to identify other pathologies. Despite not being mentioned thus far, fractures are usually diagnosed with plain radiography. Hiatal hernia is another condition occasionally present in CXR images and is characterized by the bulging of the stomach through the diaphragm. An enlarged hilum is another possible finding and it may occur due to a neoplasm, lymph node enlargement or pulmonary hypertension [2].



(a) Abnormal mediastinal contour. (b) Atelectasis. (c) Pneumonia.

(d) Lung nodule. (e) Pleural abnormality. (f) Enlarged hilum.

Figure 2.3: Examples of findings in a chest radiograph [2].

2.3 Automated Pathology Detection and Classification

Artificial Intelligence (AI) methods have been developed for a long time with the goal of assisting clinical practice. AI can have an important role in a wide variety of tasks, aiding medical practitioners to perform better decisions or simply by increasing the efficiency of their workflow. Examples of areas where AI can be introduced include risk modelling and stratification, patient prognosis, disease diagnosis, treatment planning, personalized screening, or clinical workflow optimization [25; 26]. Regarding disease diagnosis, many AI applications involve the use of medical images, either directly or through the use of extracted features. Several of these applications focus on CXR images to ease the burden of radiologists, caused by the massive amount of chest radiographs performed and by the difficulty in diagnosing certain conditions. Furthermore, there is a high inter-observer variability in CXR interpretation [27; 28], which advocates the need for a second opinion.

To this end, Machine Learning (ML) algorithms have been developed for many years. In summary, ML is a subfield of AI which focuses on autonomous learning from computers. Given some input data, the ML model is expected to learn from it and apply the acquired knowledge to new examples. Despite achieving good results for several applications, medical image analysis is a rather difficult task and the manual feature extraction required by ML imposes further problems. This motivated the appearence of DL in the medical imaging field, and more specifically of Convolutional Neural Networks (CNN). In DL, the models are composed of several layers which are capable of performing an automatic feature extraction followed by a decision. In the case of CNNs the layers are constituted by, among other components, convolution filters of small sizes that, by going through the image, are capable of extracting different features. After the process of feature extraction, a fully connected network composed of dense layers follows within the CNN architecture, in order for a prediction to be made. Other common components are pooling layers, which decrease the size of its input, or Rectified Linear Units (ReLU), that add another processing step usually after each convolutional layer. In Figure 2.4 there is a schematic representation of a typical CNN architecture.

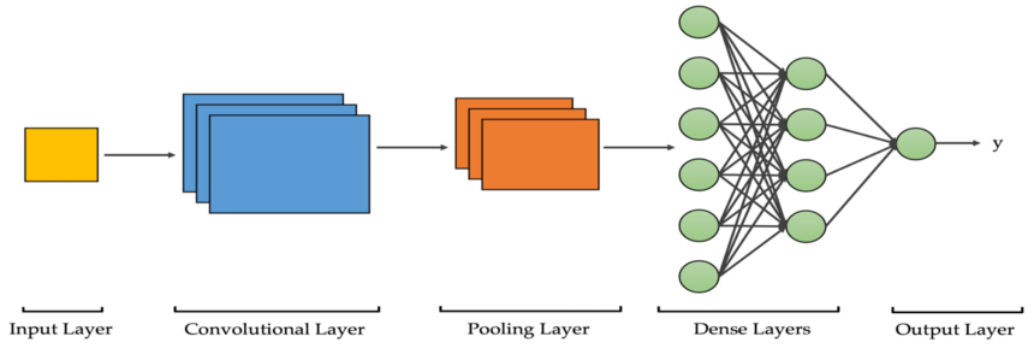


Figure 2.4: Simplified representation of a CNN architecture [3].

The first application of CNNs to medical images was in 1995 for lung nodule detection [29]. However, only more recently did this type of architecture gained momentum, mainly due to the successful application in the 2012 ImageNet challenge [30]. This, and further developments, were only possible because of the existence of more efficient training methods, availability of large datasets, and advances in parallel computer processing [31].

2.3.1 Datasets

In order for any DL model to perform well in CXR image analysis, it needs appropriate datasets for training. In Table 2.1, some of the most known public CXR datasets are described.

Table 2.1: Description of most common CXR datasets [17].

Dataset	Year	Patients	Images	Format	View	Labels	Annotation
PLCO [32]	2000	56,071	185,421	TIFF	frontal	12	manual
Open-I Indiana [33]	2015	3,996	8,121	DICOM	frontal	10	manual
ChestX-ray8 [34]	2017	32,717	108,948	DICOM	frontal	8	NLP
ChestX-ray14 [35]	2017	30,805	112,120	PNG	frontal	14	NLP
CheXpert [36]	2019	65,240	224,316	PNG	frontal/lateral	13	NLP
MIMIC-CXR [8]	2019	65,379	377,110	DICOM	frontal/lateral	13	NLP
PadChest [37]	2020	67,625	160,868	DICOM	frontal/lateral	193	manual/NLP
VinDr-CXR [38]	2020	-	18,000	DICOM	frontal	28	manual

Most of these datasets appeared in recent years motivated by the growing DL field and further reinforced this trend. The largest datasets contain hundreds of thousands of images, which also denotes the massive data requirements of current DL models. In order for a model to learn how to identify each class and to be able to generalize to new examples, it needs many samples. Only then can it learn relevant features that allow the model to distinguish different findings. Several distinct labels are also available in these datasets, which enables the models to screen each image for multiple pathologies. In most datasets the labels are global, but in some the findings are accompanied by bounding boxes denoting disease location. That is the case of VinDr-CXR dataset, and a portion of the ChestX-ray8 and PadChest datasets [38]. Another important information regarding these datasets is the method used to obtain the labels. Since the number of images is usually very large, manual annotation is cumbersome, and thus Natural Language Processing (NLP) algorithms are used for most datasets. These methods receive as input radiologist text reports regarding a CXR image and output the corresponding label(s). An example of such an algorithm is CheXbert, a transformer-based technique [39]. However, NLP algorithms are fallible and errors have been reported regarding image labels [40]. This means that manually annotated datasets offer an advantage since they contain more accurately labeled data.

2.3.2 Abnormality Detection

Abnormality detection refers to a more simple task where the goal of a model is simply to indicate if a medical image is normal or abnormal. These models can either be applied in scenarios where multiple pathologies might be present, or to detect the occurrence of a single disease. Since the task is less complex, DL models can usually show a quite good performance. However, there is a reduced applicability of such models. They can be used, for instance, in triage scenarios.

Most abnormality detection implementations simply use off-the-shelf models. In [41], several CNN architectures are compared, namely the AlexNet [42], VGG [43], GoogleLeNet [44], ResNet [11], and DenseNet [12]. These models are trained using the ChestX-ray14 dataset in a binary classification setting. All models achieved good results, especially when using transfer learning (TL), with DenseNet being slightly superior. Other approaches use more complex architectures. For instance, in [45], an autoencoder is used to perform CXR reconstruction and also to estimate the pixel-wise uncertainty relating to the reconstruction. The autoencoder is trained solely on normal images and, when applied to abnormal ones, differences occur in the pixel-wise uncertainty, allowing the detection of abnormalities. In [4], a similar approach is performed in the sense that only normal images are used during training. The model uses an autoencoder to reconstruct the images, it contains a discriminator to determine if the images were real or fake, and a second encoder that takes the generated image as input. The training functions similarly to that of a Generative Adversarial Network, and several losses are calculated: the reconstruction loss between the original and generated image, the discriminator loss, and loss terms related to the similarity between the feature maps/encoded features of the autoencoder and of the second encoder. Since the model only trains on normal images, it will perform poorly on the unseen abnormal class, making it possible to detect the presence of abnormalities. These models have the advantage of training exclusively with normal images, since abnormal images are only used upon inference, which discards the necessity of annotating pathologies. Figure 2.5 depicts the model described in [4].

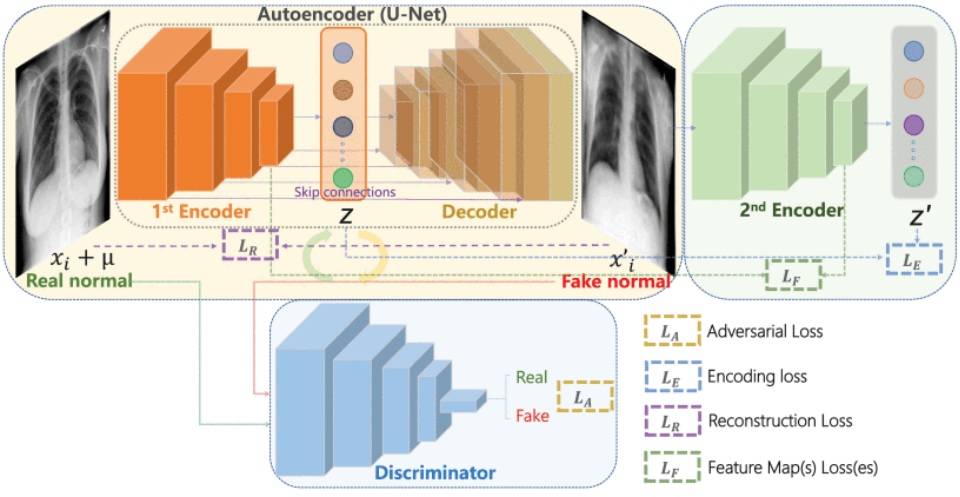


Figure 2.5: Architecture proposed in [4].

Contrarily to the previous methods, some do use abnormal images during model development. In [46], a multi-CNN framework is used to distinguish normal from abnormal images. The model is composed of three different CNNs, training either with the full image, the left half of the image, or the right half. Each CNN outputs a prediction, and these predictions are then combined to form a final evaluation of the input image. Finally, one other approach uses a subset of abnormal images to define the hyperparameters of the model, and then it trains only with normal images [47]. The model consists once more of an autoencoder followed by another encoder, but without a discriminator. Both the original and generated images are passed through the second encoder and the loss is computed by comparing the obtained feature maps. During inference on abnormal examples the loss is expected to be higher, allowing the detection of pathologies.

2.3.3 Pathology Classification

Automated pathology classification is more common than a simple abnormality detection task, since it is capable of predicting what disease(s) might be present in a CXR image. Initial approaches have consisted in more straightforward strategies, like applying a DenseNet using TL to the ChestX-ray14 dataset [48]. Even though the DenseNet is a relatively simple model, it was capable of outperforming a group of radiologists in the Pneumonia class. Other research efforts have focused on tuning certain aspects of the training process, like developing updated optimizers [49], or including weights in the loss function to deal with class imbalance [50]. Another possibility is to change the input format. In [51], the input image is divided into patches of fixed size and each patch is evaluated regarding the presence of pathologies. This allows the model not only to classify images but also to highlight disease location. In [52], the input does not consist in a single CXR image, but in a triplet of images. Two of the images contain the same pathology, while the third has a different label. By introducing this triple input, it is possible to force the model to produce similar feature maps between images containing the same findings, and different feature maps between images which do not. That way, the model is capable of learning more distinctive features regarding different pathologies. Another alternative method present in the literature consists in using more than one CNN architecture. The DualCheXNet [53], for instance, uses a DenseNet and a ResNet in parallel. Since the two models differ, the extracted features and predictions regarding the same input image will also differ. The authors argue that, with both architectures working together, it is possibly to obtain complementary information from each image, leading to a better outcome. A classifier then outputs a final prediction from the joined feature sets.

Different techniques also explore the concept of attention learning. This concept consists in shifting the focus of DL models towards relevant parts of the images, so that they are capable of performing more justified predictions. Examples of such techniques can be seen in [5], where a model with three branches was developed, as displayed in Figure 2.6. The global branch is responsible for analysing the entire image. From that branch, a heatmap is created regarding the part of the image in which the model was focusing more. This heatmap is used to create a mask that will select a patch of the original image to use as input to a local branch. In the end, a fusion branch concatenates the feature vectors of both the global and local branches and uses that to make

a prediction. By incorporating the local branch in the model, less attention is given to noisy and irrelevant parts of the CXR images, improving the performance.

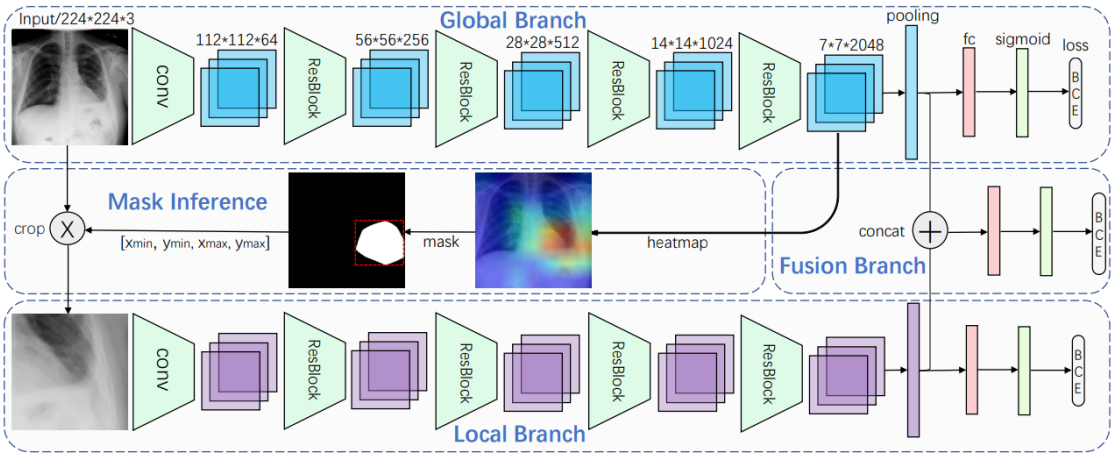


Figure 2.6: Architecture proposed in [5].

Finally, in [54], a combination of positive and negative samples for each class is used to guide the attention of the model. In this case, all images are previously adjusted in order to have the same scale and orientation. Then, both the positive and negative images pass through an encoder, and the respective feature vectors are subtracted. The resulting vector is then multiplied by the feature vector of a second encoder taking as input just the positive image. Consequently, only the features where a difference was registered between both examples are kept, forcing the model to make predictions with class-specific information. These are just a few cases where attention learning is used to guide the training process, a concept which is also explored in this work.

2.3.4 Explainability

The concept of explainability relates to the ability of a model to provide evidence to support its decision process. These explanations can be used in a training stage, to optimize the models, or in a prediction phase [55]. The latter type of explanation concentrates most of the attention in the DL field, since it is essential for successful applications to take place. In clinical practice, for instance, the implementation of DL models faces the challenge of making key decisions regarding human lives. Consequently, it is absolutely vital for the model to deliver insights, parallel to the predictions, that generate trust from doctors. Furthermore, these explanations need to be accessible, in order to be easily understood by non-experts.

Nowadays, many methods exist to produce explanations from DL models. These methods may be organized in different groups, that vary throughout the literature. In [55], explainability methods are separated in three main types: visualization, distillation, and intrinsic methods. Visualization methods highlight parts of the input that contribute the most for the prediction. This is performed either on a gradient-basis, or on an occlusion-basis. In the former, a backpropagation is performed and, through the analysis of the gradients, the importance of distinct features is computed. In the

latter, the input is altered, through the occlusion of specific regions, and its corresponding output is compared to the original one. The more the output changes, the more likely it is for that area to be important for the prediction. Distillation methods, on the other hand, consist in using separate models, that are inherently explainable, to translate the functioning of a DL model. They might work for small subsets of data, or attempt to mimic the global behaviour of the black-box model. Lastly, intrinsic methods consist in an integrated strategy where DL models are developed in a way that, alongside a given output, an explanation is also provided. This can be done through a joint training approach, in which a model trains simultaneously on two tasks. The primary one being the classification of a disease, for instance, while the secondary task renders some form of explanation.

Explainability methods may also be grouped according to their specificity. Model-agnostic methods are independent of the considered DL models, while model-specific methods are designed for a particular type of models [56]. Another possible categorization relates to the timing in which the explanation is obtained, if during or after running the model. The latter corresponds to the so called post-hoc methods.

Several of these methods can be applied to medical images. In most cases, saliency maps or heatmaps are obtained, which can be overlapped with the input images and highlight relevant areas.

2.4 Use of Eye-Tracking Data

So far, ETD has been used in the medical imaging field mainly to study the gaze patterns of experts, either towards an educational end or to assess screening efficiency [57]. However, with the introduction of DL, new applications using this type of data have been developed to study medical images. In [58], a CNN was used to model the search behaviour of radiologists when screening a mammogram for nodules. The model is capable of classifying each part of the mammogram as being likely, or not, the target of fixations from the radiologists and of predicting a confidence score relative to the presence of a malignancy. To that end, the model uses both the ETD and disease labels as Ground Truth (GT). This approach can also be used to screen regions of a mammogram typically unnoticed by radiologists, since nodule detection is usually imperfect [58].

Another application of ETD might be the automated segmentation of lesions. In [59], masks obtained from ET and hand annotated masks were both used as GT to train a CNN model with the goal of segmenting menangiomas in MRI images. In the test set, an average Dice Similarity Coefficient (DSC) of 0.85 was registered between the predictions of both CNNs, showing that similar results can be obtained with the two different GTs. Despite the positive outcome, this study is heavily limited. Only one radiologist participated in the work, and was responsible for both the ET and hand annotations, increasing the similarity between the two mask types. Furthermore, the physician was explicitly told to delineate the menangiomas with his eye sight, something which does not correlate to real ET patterns.

The studies mentioned previously used ETD as GT and images as input to the models, while other approach is to use ETD as input. In [6], a gaze analysis algorithm was used to cluster fixations and to eliminate isolated ones. The remaining data is then used to delineate regions of interest (ROI) in the corresponding 3D chest CT images. The CT images alongside the ROI are then passed to a CNN model that verifies if those regions are indeed pathological, and outputs a final segmentation. This method can be used to provide a second opinion in dubious areas focused by radiologists, or to perform an automatic segmentation of findings. A schematic of the proposed framework is depicted in Figure 2.7.

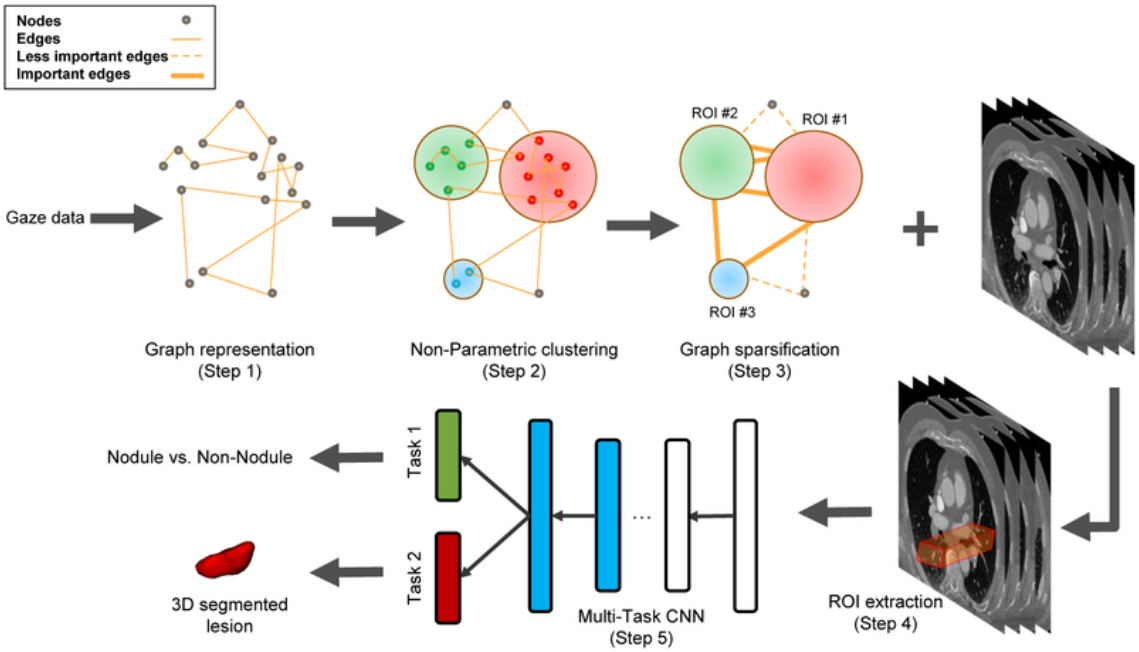


Figure 2.7: Framework proposed in [6].

In [60], another study was performed using ETD regarding chest CT images. In this case, a CNN architecture based on the YOLOv3 model [61] was developed and applied in the task of lung nodule segmentation. By using the ETD to select only regions where little attention was paid by radiologists, the model was capable of improving the detection sensitivity while not increasing the number of false positives.

Regarding CXR images, three applications of ETD can be highlighted. In [62], images with or without pneumothorax were used to collect ETD and to create a new dataset. Relevant features were then extracted from the ETD, namely patch-specific statistics (number of visited patches, maximum time dedicated to a patch, and maximum time spent on any local region, where each local region is an average of neighboring patches), and total time spent per image. These features were used as GT to train a CNN model, alongside the image labels, acting as a second source of supervision. Even though this strategy worked for the studied binary problem, cases including more pathologies will not be distinguishable based solely on a small number of features extracted from ETD.

The two other applications are presented in [7], which have also developed a new ET dataset, named here the EGD dataset. The first one is represented in Figure 2.8 and uses the ETD as input, in the form of temporal heatmaps (sequential heatmaps composed of single fixations). Each temporal heatmap is passed through an encoder, and the corresponding representations are combined using a 1-layer bidirectional Long Short-Term Memory with self-attention [63; 64]. The result is then concatenated with the feature vector derived from an encoder that receives the CXR image as input and used to classify the existing pathology.

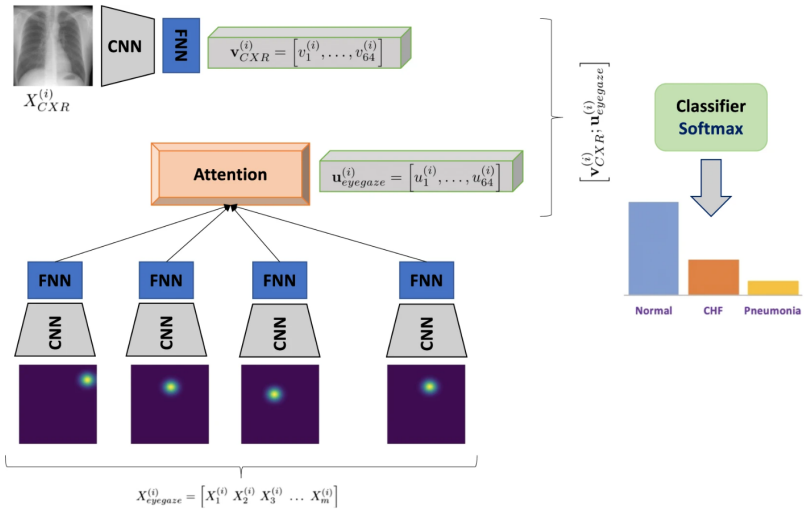


Figure 2.8: Model developed in [7], using CXR images and temporal heatmaps as input.

The second approach uses the ETD as GT, and instead consists in a UNet with a pretrained encoder that generates heatmaps from CXR images. The model also includes a classifier that performs a prediction based on the feature vectors derived from the encoder. The training of the UNet and of the classifier is integrated, using both the classification loss and the reconstruction loss. This architecture is depicted in Figure 2.9.

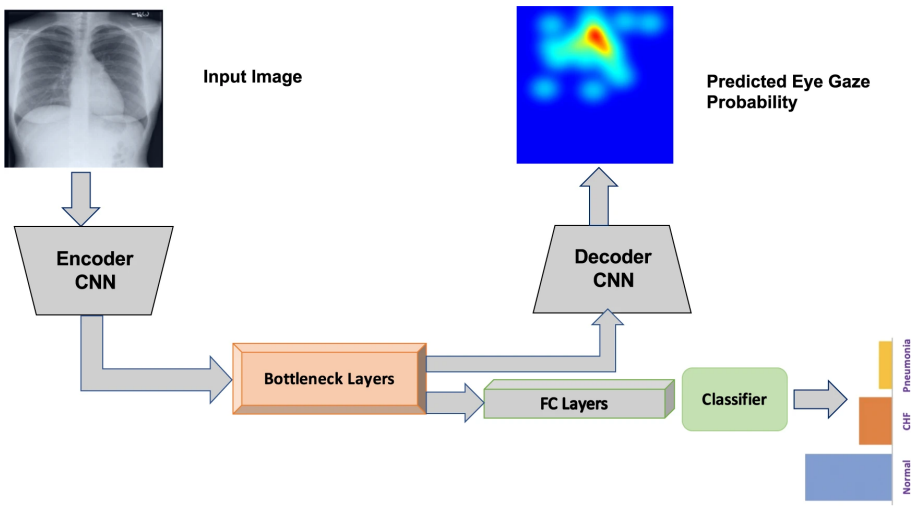


Figure 2.9: Second architecture developed in [7].

2.5 Advances and Contributions

This work aims to contribute towards new advances in the field by proposing two novel approaches. Both involve the use of a UNet with a pretrained encoder, responsible for performing the heatmap reconstruction, and a classifier, in charge of making the predictions. Figure 2.10 shows the representative backbone of both approaches.

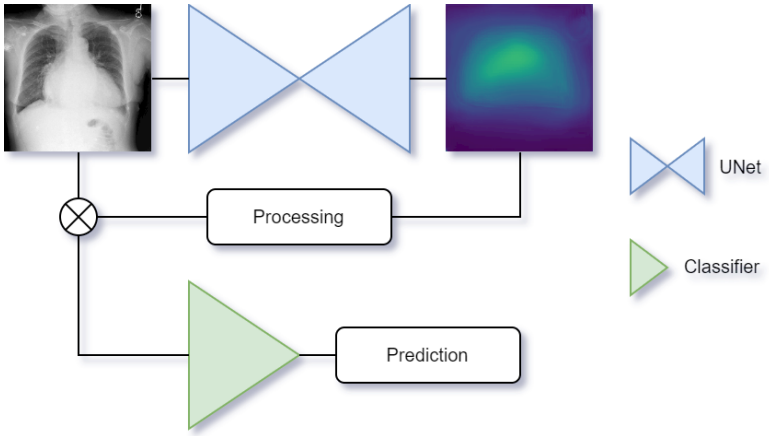


Figure 2.10: Proposed framework.

The developed models offer several advantages in comparison to previous works. First of all, the approaches here implemented strive to be independent of ETD during the inference stage. The main reason behind this imposed condition is the fact that acquiring ETD may be cumbersome. ET systems are not widely available, there is a constant need for calibration, zooming might not be allowed, and the restriction of movements can lead to increased fatigue [65]. Nevertheless, in case ETD becomes increasingly available through more practical data collection methods, for example with the use of virtual reality (VR) [66], these approaches also introduce new ways of directly incorporating ETD in the classification process, by bypassing the UNet.

The two approaches differ in the heatmap processing step, but both consist in using the reconstructed heatmaps to alter the input to the classifier. This offers an advantage in comparison to the UNet model proposed in [7], since it guarantees that the information contained in the reconstructed heatmaps is used to make predictions. The goal is to guide the classifier into focusing on important parts of the CXR images, while discarding irrelevant features such as common markings in the top corners. Ultimately, both approaches aim at offering a higher degree of explainability.

Chapter 3

Materials and Methods

This chapter starts by presenting a detailed description of the ET datasets used in this work. After that, the base architectures used to develop the models for heatmap reconstruction and for CXR pathology classification are depicted.

3.1 Datasets

In total, three datasets were found containing ETD regarding CXR images. Important details are presented in Table 3.1, followed by more extensive descriptions in the corresponding sections below.

Table 3.1: Datasets used and relevant informations.

Dataset	Year	Images	Labels	Radiologists	Masks	Image source
EGD [7]	2021	1,083	3	1	No	MIMIC-CXR
REFLACX [65]	2021	2,616	14	5	Yes	MIMIC-CXR
CXR-P [62]	2021	1,250	1	3	Yes	SIIM-ACR

As it is possible to see, all datasets contain ETD for a relatively small number of CXR images, since obtaining such data is time-consuming. Two of the datasets, EGD and CXR-P, present more simple tasks, while REFLACX contains a number of labels similar to the most common CXR datasets, as described in Table 2.1. Furthermore, the number of radiologists used in each study varies, and in EGD only one is used. This is undesirable since it might introduce some bias in the data. Regarding the existence of masks indicating the location of pathologies, they are present in two of the datasets. Concerning the source of the CXR images used to obtain ETD, they either come from MIMIC-CXR (already described in Table 2.1), or from the SIIM-ACR dataset. The latter was introduced in a 2019 pneumothorax segmentation challenge [9], and used in [62] to acquire ETD.

3.1.1 EGD Dataset

As shown above, this dataset includes ETD corresponding to 1,803 CXR images. The images were obtained from the MIMIC-CXR database, namely from a subset where other clinical observations were also available. This allowed a more rigorous selection process to ensure the exclusivity of each label. The three selected labels were Normal, CHF and Pneumonia, with 360, 363, and 360 examples, respectively. This denotes a very balanced dataset. In Figure 3.1, a CXR image for each of the labels is shown.

(a) Normal. (b) CHF. (c) Pneumonia.

Figure 3.1: Examples of CXR images used in the EGD dataset [8].

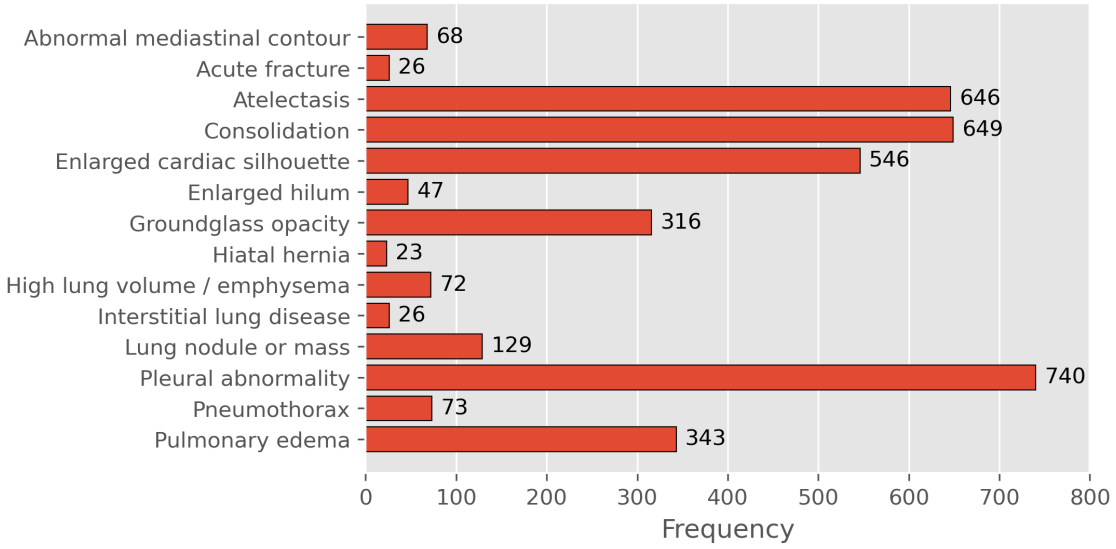
The ETD relative to these images was collected in several sessions, in order to avoid fatigue from the radiologist. The collection was performed with a Gazepoint GP3 Eye Tracker [67], which uses infrared light to detect eye fixation. The data consists essentially in fixation coordinates and respective time duration. The radiologist had only access to the images, and not to the associated medical reports. This created some disparity claimed by the authors between the labels of each image and the radiologist reports, since diagnosing diseases such as Pneumonia or CHF is difficult without further information. Instead, the radiologist would report the occurrence of findings such as lung opacity. This has the advantage of not biasing the radiologist, but has the drawback of possibly leading to gaze patterns not correspondent to image labels. Furthermore, to ensure the quality of the measurements, the radiologist was at a fixed position from the screen, and, in the beginning of each session, a calibration was performed.

Besides ETD, this dataset also contains audio reports and the respective transcripts. Additionally, it contains several bounding boxes relating to different anatomical structures, and segmentation masks for both lungs, for the heart, and for the aortic knob.

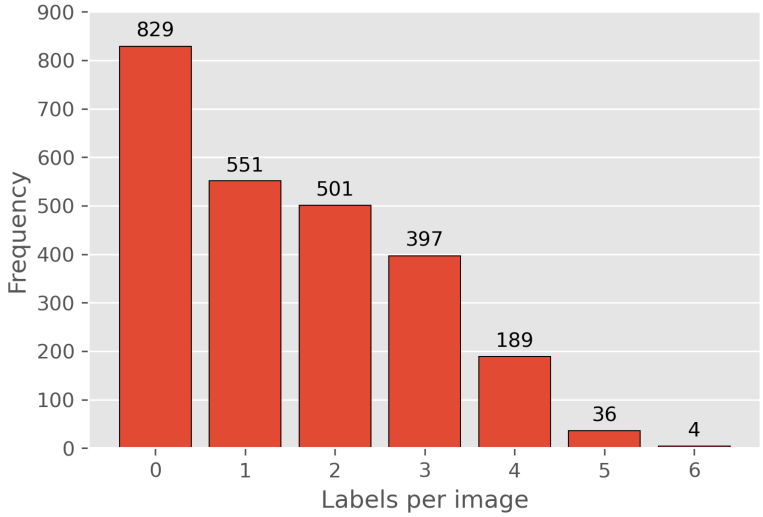
3.1.2 REFLACX Dataset

This dataset was developed with 2,616 images extracted also from the MIMIC-CXR database. The data collection process was divided into three stages, with the first two being used for optimization purposes. Therefore, only the images and corresponding ETD from the third stage were used, since the quality was higher [65]. This changed the number of images to 2,507. Regarding the number of labels, 14 exist and more than one label may be present in the same image, which is not the

case for the EGD dataset. Another difference is that the labels included in this dataset are obtained from the radiologists from which the ETD is collected. This has the obvious advantage of having gaze patterns corresponding to image labels. Furthermore, the labels are not binary, but instead are presented in a scale of 0 to 5. The first level on the scale indicates that the corresponding label was not selected by the radiologist. From 1 to 5, the levels correspond to unlikely, less likely, possibly, suspicious for/probably, and consistent with, respectively. To train the models, the labels were binarized, with a value of 3 or higher being considered a positive instance for the considered finding. In Figure 3.2, the number of images containing each label is shown, as well as the numbers of findings per image and respective frequencies.



(a) Frequency of each class.



(b) Number of labels per image.

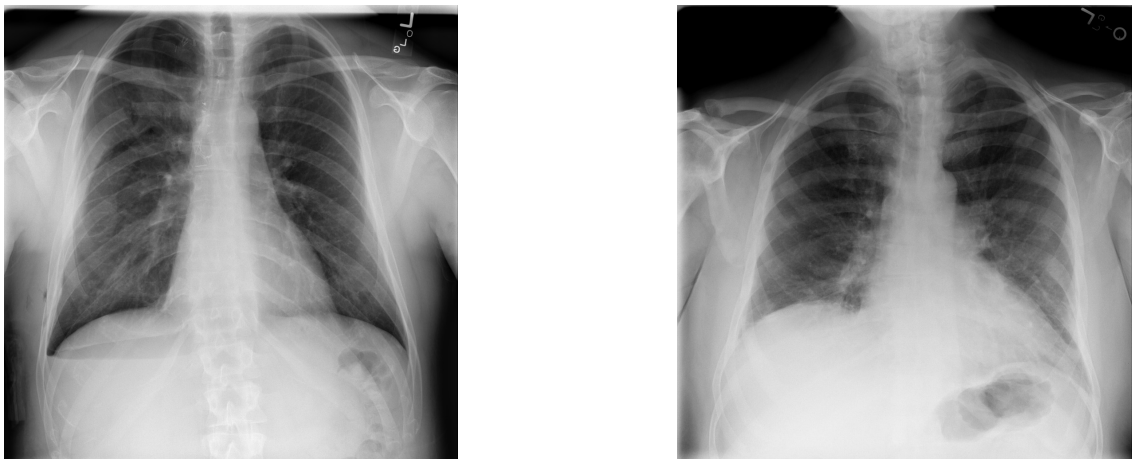
Figure 3.2: Class frequencies in the dataset and number of findings per image. Only labels with a certainty of 3 or higher were selected for this analysis.

To record the ETD, an Eyelink 1000 Plus system was used [68]. This system allows for some movements from the radiologists, although limited, but may suffer in case bifocals are used [65]. In the beginning of each session, a calibration was performed, in order to assure the accuracy of the measurements. During the session, the radiologist would dictate the report while ETD was collected. In the end, the labels were selected and ellipses highlighting the location of each finding were drawn by the radiologists. Zooming was allowed during the sessions, unlike in the other datasets, which makes the diameter of the fixations variable. To account for this, a value corresponding to the number of pixels per degree of visual angle is given for each fixation. The ETD consists of fixation coordinates in image space and respective duration.

Alongside the ETD, ellipses denoting anomaly locations, as well as bounding boxes showing the location of the thorax and report transcriptions, are also made available.

3.1.3 CXR-P Dataset

The CXR-P dataset was obtained using 1,250 CXR images collected from the SIIM-ACR dataset. It contains a single class, Pneumothorax, with 268 positive and 982 negative examples, denoting an imbalanced dataset. In Figure 3.3, two CXR images, with and without pneumothorax, are shown.



(a) With pneumothorax. (b) Without pneumothorax.

Figure 3.3: Examples of CXR images used in the CXR-P dataset [9].

The data collection was performed with a Tobii Pro Nano eye-tracker [69] and, at the beginning of each session, a calibration was performed. The data differs from the previous datasets, since the time duration of each fixation is not given in seconds. Adjacent gaze points are grouped together, and the means of the coordinates are used as the centers to new fixations. Then, the number of joined gaze points for each fixation is given as a time measurement.

This dataset also contains masks denoting the pneumothorax locations, as made available in the SIIM-ACR dataset.

3.2 Models

In order to perform the heatmap reconstruction and to classify different pathologies existent in CXR images, several models were tested. All the models, apart from the standard UNet, and pretrained weights are derived from [70].

3.2.1 UNet

The UNet architecture was proposed in 2015 with the goal of performing biomedical image segmentation [10]. Since then, it has been used in a myriad of medical imaging tasks with great success. In Figure 3.4, its architecture is depicted.

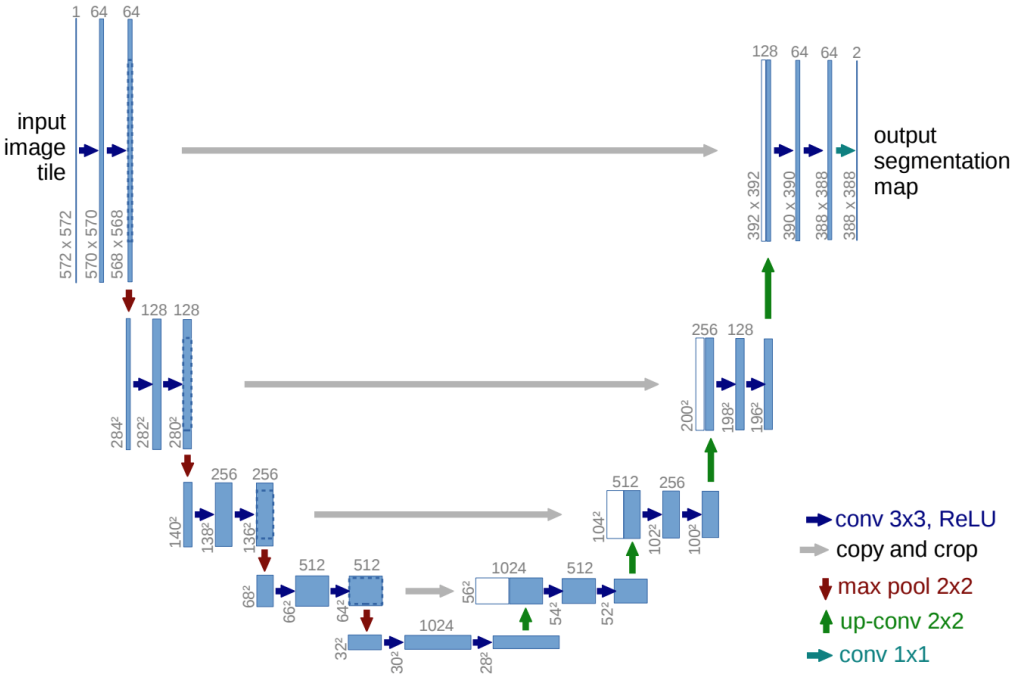


Figure 3.4: UNet architecture [10].

Its name derives from the U-shape, formed by a contracting and an expanding path. The contracting path applies a series of two 3x3 convolutions followed by the ReLU activation function, and a 2x2 pooling with stride of 2. The number of channels doubles at each block of the contracting path, while the width and height undergo an inverse trend. The opposite happens in the expanding path. Upsampling operations increase the width and height, and the number of channels gets reduced at each block until the final output. After every upsampling operation, there is a concatenation between the upsampled channels and features coming from skip connections starting in the contracting path. This helps the network to construct an output with a higher resolution.

Depending on whether padding is used in each convolution, the output might have a smaller or equal size in comparison to the input.

In this work, a smaller version of this network is used, with one block less and half the number of channels. This is due to limitations regarding the GPU memory. Also, padding is used in each convolution to assure that the final output has the same size as the input. Furthermore, variations of this network are also used to perform heatmap reconstruction, namely the inclusion of different pretrained encoders.

3.2.2 ResNet50

The ResNet architecture was proposed in 2015 [11] and since then it has been used widely in classification tasks. This network uses the concept of residual learning to make deep models easier to train. Instead of learning a given underlying mapping, each constituent building block has the task of learning the residual mapping, which the authors claim to be easier for the model. A simple representation of this process is shown in Figure 3.5.

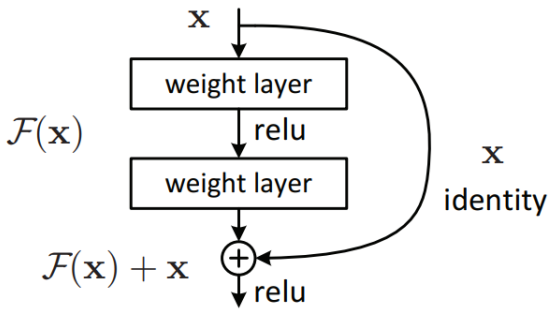


Figure 3.5: Building block of residual learning [11].

The input to the layers is added to the respective output through the existence of shortcut connections. The goal of each set of layers is then to learn the residuals that lead to an ideal mapping. The ResNet architecture may contain different numbers of layers, ranging from 18 to 152 layers.

In this work, a ResNet50 (the number 50 corresponds to the total number of layers) is used in the selection process of the models, since it is probably the most common ResNet architecture. It is employed either as a UNet encoder, or as a classifier. In both versions, pretrained weights are used.

3.2.3 DenseNet121

The DenseNet architecture was presented in 2016 [12], and since then became a commonly used model in DL. Its name derives from the densely connected blocks from which is composed of, in which the output of each layer is fed into the subsequent layers. Figure 3.6 shows a representation of such structure.

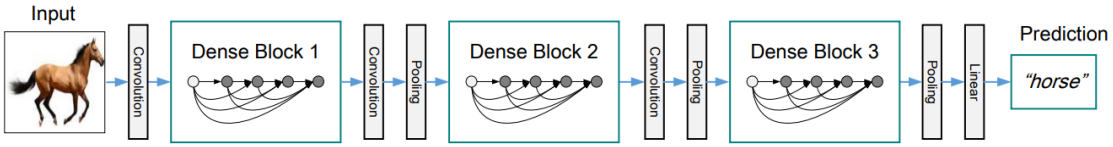


Figure 3.6: Schematic of a DenseNet architecture [12].

Multiple dense blocks are present in a DenseNet. Inside each one of them, the input to a layer is concatenated to the output of the previous layers, thus increasing the number of channels received by the layers. This results in increased complexity without the need to increase the number of parameters, by promoting feature propagation and reuse, and solves the vanishing gradient problem (this problem occurs in deep models, in which the gradients become very small and thus the model struggles to learn). Consequently, the number of connections increases dramatically. Between each dense block there is a transition layer, in which a sequence composed of a convolution followed by a pooling layer decreases the size of the feature maps.

DenseNets can contain different numbers of layers. In this work, the smallest DenseNet architecture, the DenseNet121, is used, which is composed of 121 layers. Again, it is tested either as a UNet encoder to perform heatmap reconstruction, or as a classifier. In both cases a pretrained version is used to speed up the training process by facilitating convergence.

3.2.4 EfficientNet-b0

More recently, a new type of network has been proposed, named EfficientNet [13]. The main goal of this type of network was to find a rational way of scaling the size of models. Until then, scaling methods consisted in trial and error experiments, in which either resolution, depth, or width were changed in isolation. In this work, however, the authors created the concept of compound scaling, in which all three dimensions of the network are simultaneously updated according to a set of rules. In Figure 3.7, the concept of compound scaling is elucidated.

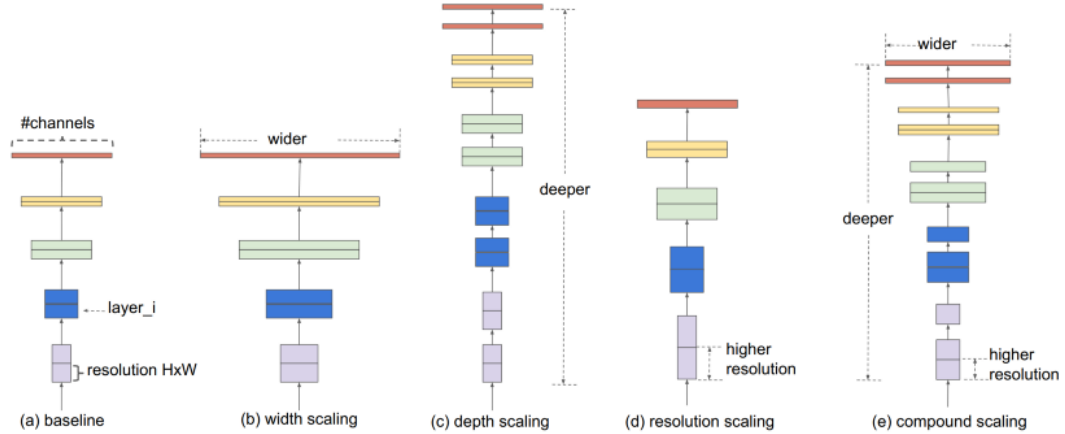


Figure 3.7: CNN scaling methods [13].

From a baseline architecture, the network is scaled according to this method. With this strategy, the authors were capable of surpassing state-of-the-art models in common benchmark datasets, while using smaller and more efficient networks [13]. Part of this is also due to the use of inverted residual blocks [71], which differ from normal residual blocks (like the ones used in the ResNet architecture). The difference is that normal residual blocks have a wide-narrow-wide configuration in terms of the number of channels, while inverted residual blocks have a narrow-wide-narrow structure. This greatly reduces the number of parameters.

Depending on the resources available, several EfficientNet versions are proposed which vary in size. It ranges from the EfficientNet-b0, with 5.3 million parameters, to the EfficientNet-b7, with 66 million parameters [13]. In this work, the smallest version is used, with pretrained weights, both as a UNet encoder or as a classifier. The EfficientNet-b0 architecture is depicted in Figure 3.8. It is composed of seven main blocks, each containing several inverted residual blocks (or MBConv blocks). The number of MBConv blocks varies depending on the main block, as well as the size of the convolution filters used.

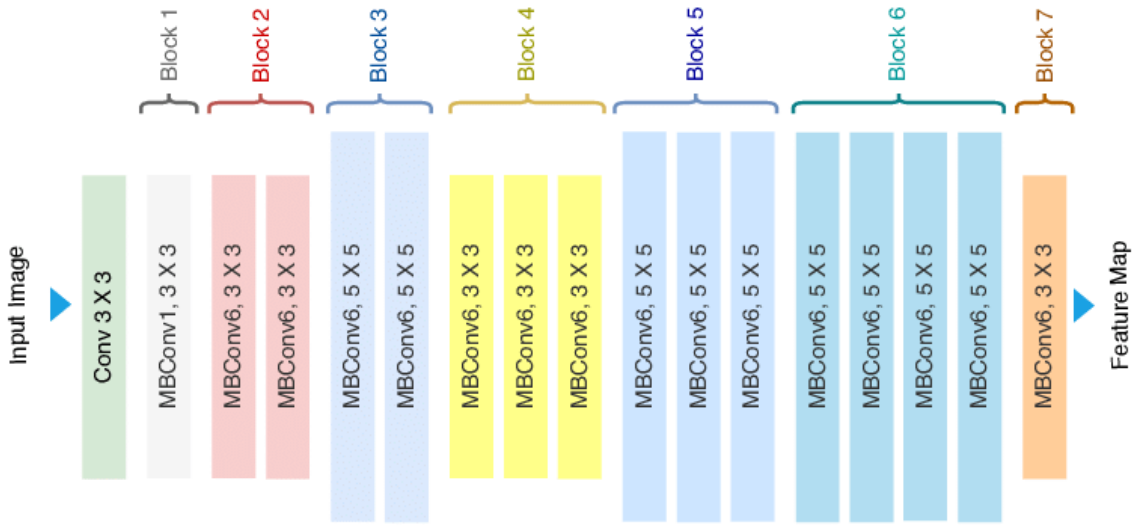


Figure 3.8: EfficientNet-b0 architecture [14].

3.2.5 Transfer Learning

TL, as the name implies, is the process of applying knowledge obtained in a given task into a different problem. This works especially well if both tasks are related, although unrelated tasks may also benefit from this strategy. The main advantages of using TL, regarding DL models, are the reduced training time and a possible improvement in the performance. It is used mainly when the dataset size relative to a given task is small. Therefore, first training a model in a different task with a large dataset and then reusing the weights for the second task might prove beneficial. Depending on the similarity between the two tasks, only the final layers might be retrained or otherwise the entire model.

In this work, since the three datasets used contain a relatively small number of examples, TL was applied with the goal of boosting performance. Another advantage is the notorious reduction in training time, since the models trained using TL converge much faster. Pretrained models are used both for heatmap reconstruction and for the classification of pathologies. In each case, the entire model is retrained in order to fine-tune the feature extraction part and adapt it to the task at hand.

Regarding the origin of the weights used, they all come from models pretrained on the ImageNet dataset [72]. This dataset is very broad, containing millions of images belonging to thousands of classes. Models trained from it learn generalized features, making them suitable for being applied in very diverse tasks. Also, it is a typical benchmark in the world of DL, which means that pretrained versions of the most common CNN architectures are easy to find. For the ResNet50 and DenseNet121, standard sets of weights are used here. For the EfficientNet-b0 architecture, both standard and noisy-student (NS) weight sets are used. The latter comes from training the model with the NS training scheme [15], depicted in Figure 3.9.

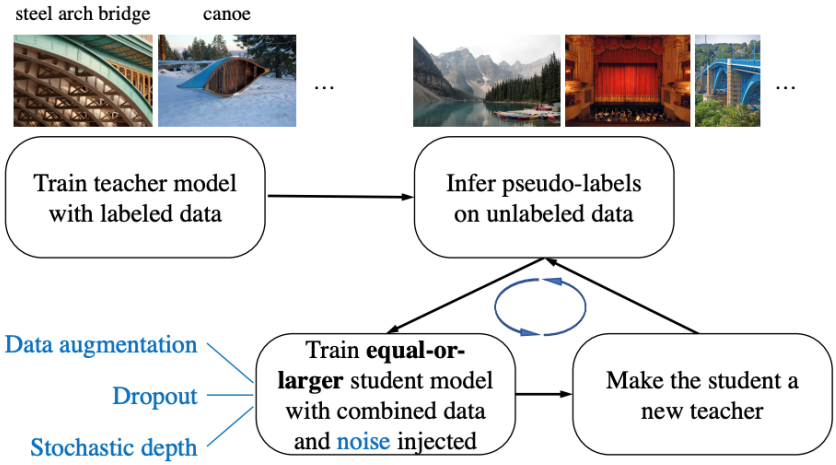


Figure 3.9: Noisy-student training [15].

In this training procedure, two models are used, which swap roles at each iteration. Initially, an EfficientNet is trained on ImageNet and then used to create pseudo labels. This will be the teacher model. The other model, the student, will then learn both from labeled and pseudo labeled images. At the end, the student becomes the teacher and vice versa. Furthermore, during the training of the student, noise is added in three forms, as seen in Figure 3.9. Data augmentation consists in changing the input so as to add variability, dropout [73] acts by switching off neurons within a network in order to decrease overfitting, and stochastic depth [74] is the process of randomly inactivating and skipping layers during training, making the network shorter. All these methods can result in better performances from the models. In addition, dropout and especially stochastic depth increase the training speed. As a result of the entire process, EfficientNets achieve better results and, at the time the work in [15] was released, managed to surpass the current state-of-the-art scores on the ImageNet dataset.

3.2.6 GradCAM

DL models, with their increased complexity, have the problem of not being easily interpreted, and thus are compared to a black-box. This is not necessarily bad, since more complex models achieve better performances, but the lack of interpretability demands for explanations in order to justify their predictions. Nowadays, several methods exist to study CNN architectures and to find class discriminative information that contributes to a given prediction. These methods allow, to some extent, the assessment of whether a model is trustworthy or not, regardless of how accurate it might be.

Gradient-weighted Class Activation Mapping (GradCAM) [75] creates coarse localization maps with respect to specific classes, allowing the visualization of which regions in an image are affecting a given prediction. GradCAM can be used with virtually any CNN architecture. It consists in firstly backpropagating the gradients, with respect to a specific class, up to the convolutional layer of interest, and global-average-pooling them for each channel in order to get an importance weight. Equation 3.1 describes the calculations performed to get the importance weights for each channel, where α_k^c is the weight for feature map k respective to class c, Z is the number of elements in each feature map, y^c is the class score, and A_{ij}^k is the k^{th} feature map value for coordinates i and j.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.1}$$

Once these values are calculated, a weighted mean of the channels belonging to the selected convolutional layer is performed, according to Equation 3.2. A ReLU activation function is used to discard negative values and to highlight only features with a positive influence on the class of interest [75].

$$L_{GradCAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{3.2}$$

The result, $L_{GradCAM}^c$, consists in a heatmap containing class discriminative information. Figure 3.10 shows an example of such heatmaps for two different classes.

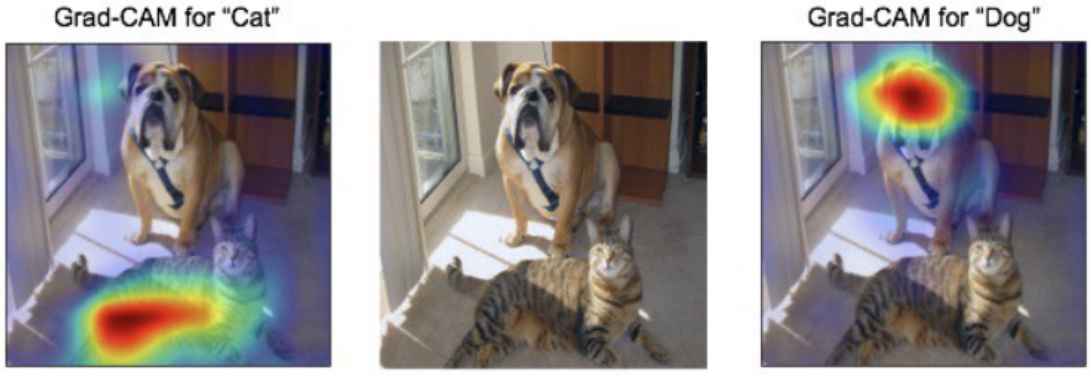


Figure 3.10: GradCAM example [16].

In both examples of Figure 3.10, the highlighted features correspond to the animals of interest. This shows that the predictions of that model were justified. However, this is a post-hoc method, and even after a thorough analysis for many examples no definite assessments can be made regarding the model, since incorrect explanations are still possible. Nevertheless, it provides a means of interpreting the underlying processes in a DL model on a case-by-case basis.

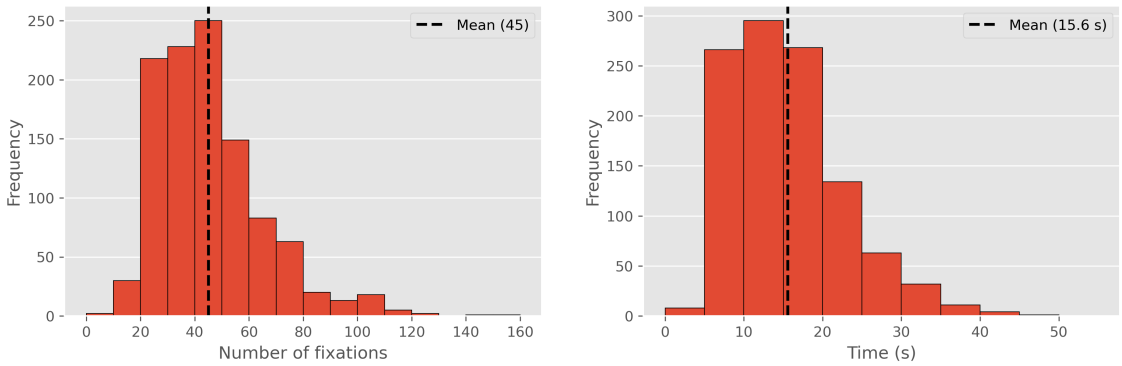
Chapter 4

Eye-Tracking Data Analysis

In this chapter, the results regarding ETD analysis for the three datasets used are shown. More specifically, the number of fixations and time per image are assessed, as well as the percentages of fixations that fall within the bounding boxes and/or anomaly masks. Class-specific values are also studied to see if any differences are present within the datasets.

4.1 EGD Dataset

Firstly, the number of fixations and time spent per image were evaluated. In order to achieve that, the number of fixations for each image was assessed, as well the time spent in the corresponding analysis. From the retrieved information, histograms showing the respective distributions were plotted, as depicted in Figure 4.1. On average, the radiologist performing the CXR screening would fix the gaze 45 times. This corresponds to a mean time of approximately 16 seconds per image.



(a) Number of fixations per image.

(b) Time per image.

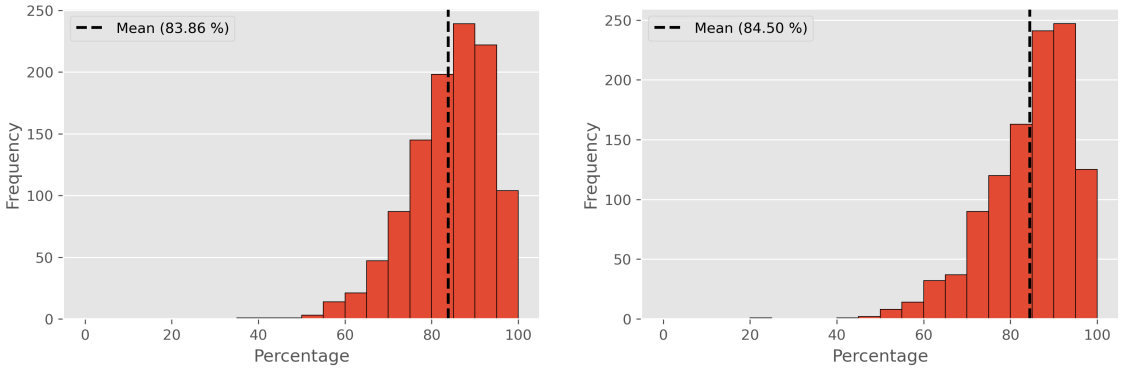
Figure 4.1: Histograms showing the distributions of fixations and time per image for EGD, with respective means.

Following this initial analysis, it was important to see how many of these fixations would fall into relevant parts of the CXR images, namely the lungs and the mediastinum. For that to happen, segmentation masks for the right and left lungs, and for the mediastinum, were joined. Figure 4.2 shows examples of the mentioned segmentation masks.

(a) Right lung [7]. (b) Mediastinum [7]. (c) Left lung [7]. (d) Thorax.

Figure 4.2: Segmentation masks for both lungs and for the mediastinum, and the resulting thorax mask.

Then, for each image, the percentage of fixations and time inside the thorax were calculated. For this analysis, the field of view was disregarded and only the center coordinates of each fixation were considered. The histograms in Figure 4.3 show the values obtained alongside the corresponding means. It is possible to see that most of the fixations were inside the thorax, as expected. In terms of time, the same pattern is observed.



(a) Percentage of fixations inside the thorax. (b) Percentage of time inside the thorax.

Figure 4.3: Histograms with percentages of fixations and time inside the thorax for EGD, and respective means. Thorax segmentation masks were created by joining segmentations of both lungs and the mediastinum.

Specific values for both lungs and mediastinum were also obtained, both on an overall basis and on a class-specific basis. These values are presented in Tables 4.1 and 4.2. As it is possible to observe, a higher mean number of fixations was recorded for the abnormal classes when compared to the Normal class. The same holds true regarding time. This indicates that the radiologist was capable of identifying pathological traits in those images, taking more time to analyse and report such findings. Furthermore, more fixations and more time were spent on the mediastinum for the CHF class as compared to the others. This is expectable since that was where the pathology was located. Interestingly, no apparent differences exist between CHF and Pneumonia regarding the

number of fixations targeting the lungs. However, this is most likely due to the overlap between the lungs and the mediastinum segmentation masks, with fixations aimed at the heart being accounted also in the lungs. Nevertheless, more fixations occur in the lung area for Pneumonia when compared to the Normal class.

Table 4.1: Class-specific and overall mean numbers of fixations for EGD.

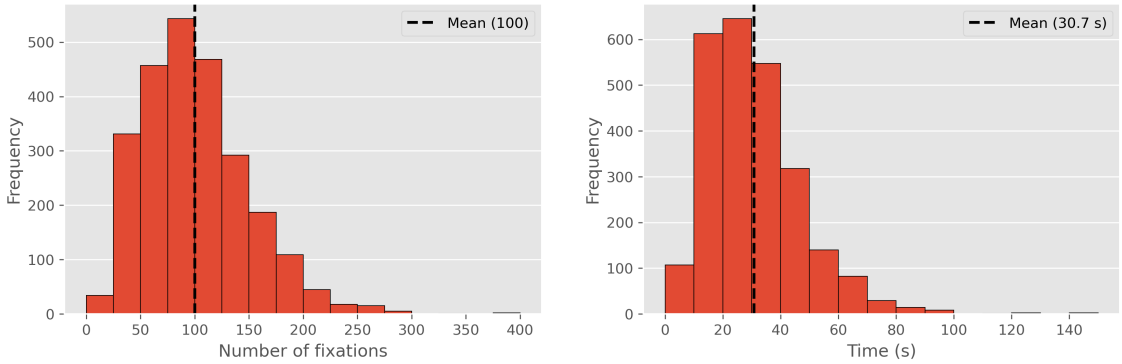
	Normal	CHF	Pneumonia	Overall
Right lung	11 ±6	19 ±9	20 ±12	16 ±10
Left lung	12 ±6	20 ±8	19 ±10	17 ±9
Mediastinum	12 ±7	18 ±8	15 ±7	15 ±8
Total	33 ±14	51 ±17	50 ±22	45 ±20

Table 4.2: Class-specific and overall mean time values in seconds for EGD.

	Normal	CHF	Pneumonia	Overall
Right lung	3 ±2	7 ±3	7 ±4	6 ±4
Left lung	4 ±2	7 ±3	7 ±4	6 ±4
Mediastinum	4 ±2	6 ±3	5 ±3	5 ±3
Total	11 ±5	19 ±6	18 ±8	16 ±7

4.2 REFLACX Dataset

As done in the previous section, the analysis for the REFLACX dataset started with an assessment of the number of fixations and time spent per image. Figure 4.4 shows the corresponding histograms. A striking difference exists between the REFLACX and EGD datasets, with the former having a mean number of fixations of 100 and a mean time value of 31 seconds, representing a 2-fold increase. This is probably due to the higher number of classes within the dataset, and also because of the simultaneous occurrence of several pathologies, as shown in Figure 3.3b. This increases the difficulty of the task and thus the time required to analyse such images.

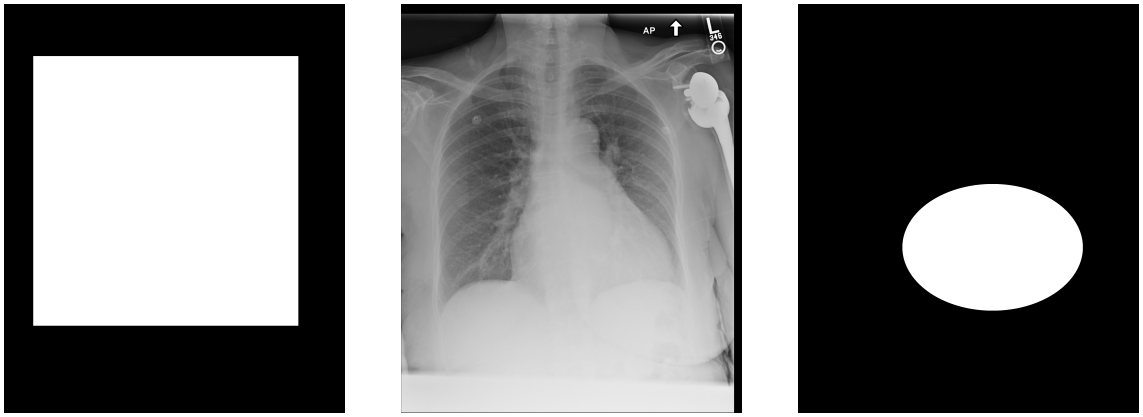


(a) Number of fixations per image.

(b) Time per image.

Figure 4.4: Histograms showing the distributions of fixations and time per image for REFLACX, with respective means.

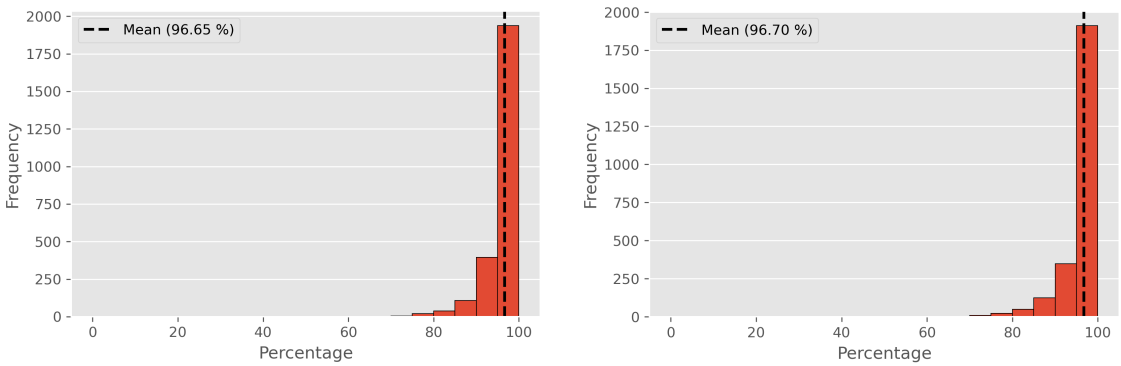
Since in this dataset both bounding boxes and ellipses highlighting disease locations are available, the next step was to study the extent to which fixations overlapped with these regions. In Figure 4.5, an example of a bounding box is shown, alongside the corresponding CXR image. Figure 4.5 also shows the anomaly ellipse from that image, clearly encircling an enlarged heart.



(a) Bounding box. (b) CXR image [8]. (c) Anomaly ellipse.

Figure 4.5: Example of bounding box, anomaly ellipse, and respective CXR image.

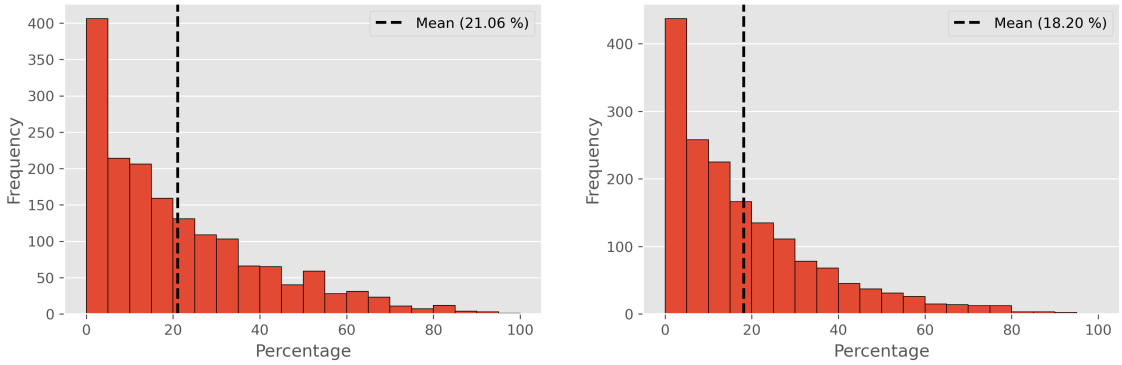
Firstly, the percentages relating to the number of fixations and time inside the bounding boxes were calculated. Figure 4.6 displays the histograms that represent the distributions of these percentages across the dataset. It is clear that for this dataset, the mean percentage values of number of fixations and time inside the bounding boxes are very high. This reveals that, as expected, the focus of the radiologists was in the lungs and mediastinum. These values are superior when compared to the EGD dataset, probably because bounding boxes were used instead of segmentation masks. The bounding box is a rectangle, and not a well delineated mask containing both lungs and the mediastinum. Consequently, it might capture more fixations.



(a) Percentage of fixations inside the bounding box. (b) Percentage of time inside the bounding box.

Figure 4.6: Histograms with percentages of fixations and time inside the bounding box for REFLACX, and respective means.

Afterwards, the same analysis was performed with the anomaly ellipses instead of the bounding boxes. Figure 4.7 shows the results obtained. Interestingly, only around a fifth of the fixations focus on anomaly ellipses with relatively high certainty. Furthermore, there is a huge gap to the results obtained for the bounding box analysis, that might be partially explained by the smaller size of the ellipses. Another obvious reason is that, in order to locate every finding, radiologists need to perform a complete screening of the thorax. While doing so, other dubious regions of the CXR image may catch the attention of radiologists, distributing the time spent around the image.



(a) Percentage of fixations inside the anomaly masks.

(b) Percentage of time inside the anomaly masks.

Figure 4.7: Histograms with percentages of fixations and time inside the anomaly masks for REFLACX, and respective means. Only images containing ellipses with a certainty level of 3 or higher were used in this analysis.

The mean number of fixations and mean time per image dependening on the pathology class were then computed. However, given that multiple pathologies can be present in the same image and that the same anomaly mask can contain more than one finding, the analysis was limited to a division between normal and abnormal images. Table 4.3 shows the mean values regarding number of fixations and time for both scenarios. As expected, more fixations occur and more time is spent, on average, in abnormal images.

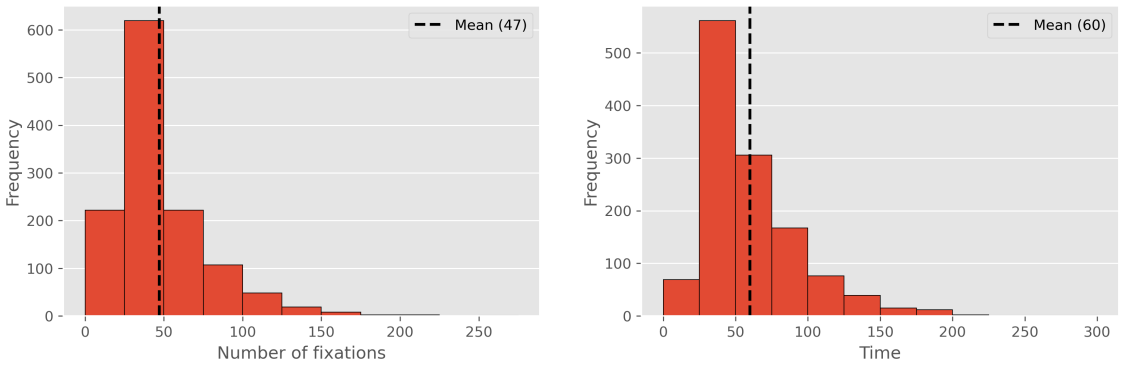
Table 4.3: Normal, abnormal and overall mean values regarding number of fixations and time per image for REFLACX.

	Normal	Abnormal	Overall
Number of fixations	78 $\pm$35	111 $\pm$50	100 $\pm$48
Time per image (s)	23 $\pm$11	35 $\pm$17	31 $\pm$16

4.3 CXR-P Dataset

For the CXR-P dataset, the analysis also started by assessing the number of fixations and time spent per image. However, there is a key difference in the values made available in this dataset. In the other datasets, angular velocity thresholds were used to detect saccades - rapid eye movements - and the data points between saccades were joined to form a single fixation (coordinates were averaged and time was added). For this dataset, however, the threshold used to select fixations relates to the distances between data points. If the distances between both coordinates of two data points were less than 5% of the image size, and if the data points were consecutive in time, they would be joined. Then, the time was given as the number of joined data points. Since every data point takes the same amount of time, corresponding to the sampling period, this will be a measure of how long each fixation was.

Figure 4.8 shows the histograms representing the distributions for the collected data. The average number of fixations per image is 47, and the associated mean time is 60. This means that most fixations are composed of single or very few data points. It is also possible to see that the mean number of fixations per image is very similar to the recorded in the EGD dataset. This is probably due to the similarities between both datasets, which contain very few classes. When compared to REFLACX, on the other hand, this value is only around half of its mean number of fixations.



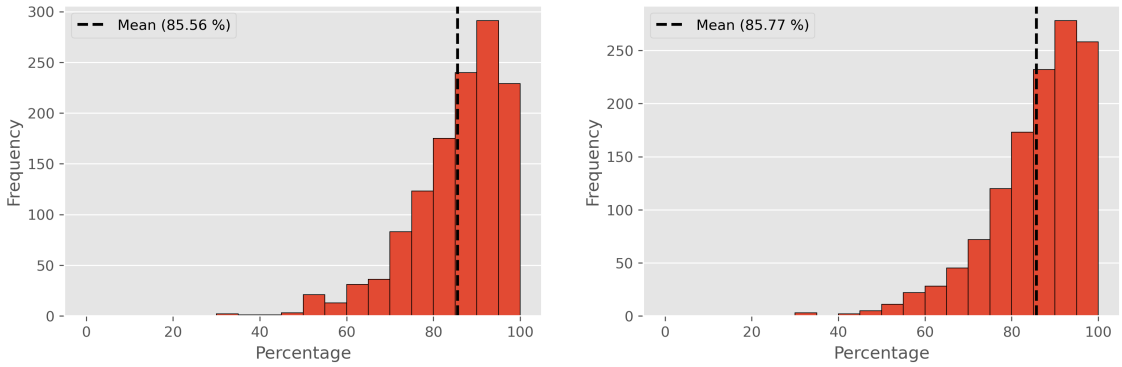
(a) Number of fixations per image.

(b) Time per image.

Figure 4.8: Histograms showing the distributions of fixations and time per image for CXR-P, with respective means. Time units are not given in seconds, but instead as the total number of data points per image.

Since no bounding boxes are available in this dataset, a thorax detection model developed in-house was used. It is based on the YOLOv5 architecture and was trained with nearly a thousand CXR images from three datasets, the JSRT [76] and the Montgomery and Shenzen datasets [77]. Furthermore, the GT bounding boxes used in the training process were obtained by drawing rectangles that fit the lung segmentation masks provided in those datasets.

After applying the thorax detection model to every image included in the CXR-P dataset, and selecting the bounding boxes with the highest confidence score for each case, it was possible to compute the percentages of fixations and time inside the bounding boxes. The results are shown in Figure 4.9. The obtained values are high, but inferior to the ones observed in REFLACX. This could be related to the origin of the bounding boxes used in the analysis, or to the fact that, since pneumothoraces occur preferentially in the contour of the lungs, some of the fixations get recorded as being slightly outside the bounding boxes.



(a) Percentage of fixations inside bounding boxes.

(b) Percentage of time inside bounding boxes.

Figure 4.9: Histograms showing the percentages of fixations and time inside the bounding boxes for CXR-P, alongside the respective means.

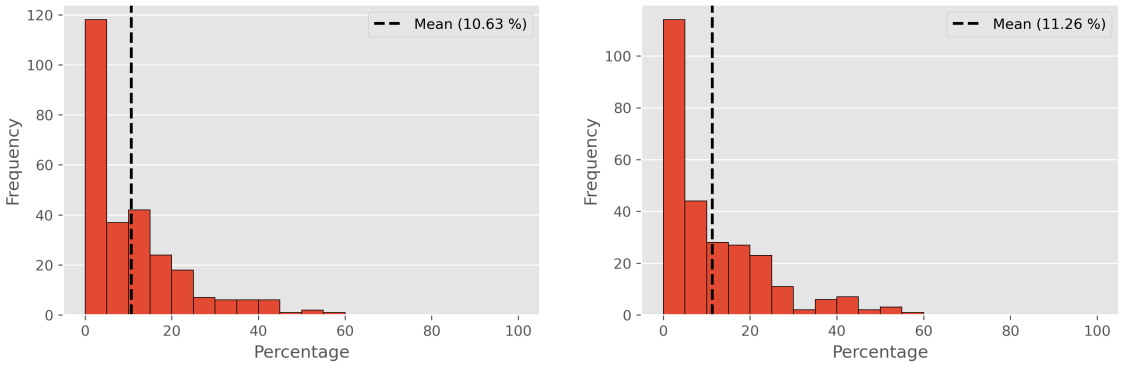
The following analysis consisted in looking at the percentages of fixations and time inside the pneumothorax masks. Figure 4.10 displays an example of a mask present in the dataset, alongside the corresponding CXR image. The collapse of the right lung is quite clearly visible in the location highlighted by the mask.

(a) Pneumothorax mask.

(b) CXR image [9].

Figure 4.10: Example of pneumothorax mask and corresponding CXR image.

The number of fixations and time spent in these masks were then calculated, as shown in Figure 4.11. It is possible to see that the mean values are relatively low. The most likely cause is that the pneumothorax masks are smaller and narrower than the masks highlighting other types of findings (such as for REFLACX). This will then make it harder for the fixation coordinates to fall within the masks. Other possible reasons include measurement errors, causing fixations to be outside the narrow masks, or perhaps the radiologists failed to spot the pneumothoraces. Since the labels were not given by the radiologists, but instead came from the SIIM-ACR dataset, it is impossible to know whether the opinion and gaze data from radiologists was in accordance with the class and mask of each image.



(a) Percentage of fixations inside masks.

(b) Percentage of time inside masks.

Figure 4.11: Histograms showing the percentages of fixations and time inside the pneumothorax masks for CXR-P, alongside the respective means.

Lastly, the mean values on an overall and class-specific basis were studied to see if differences existed between both classes. Table 4.4 shows the results obtained. It is quite clear that the values are higher for the abnormal class. This indicates that the radiologists, at least for most images, were capable of detecting the occurrence of pathologies.

Table 4.4: Normal, Pneumothorax and overall mean values regarding number of fixations and time per image for CXR-P.

	Normal	Pneumothorax	Overall
Number of fixations	43 $\pm$26	61 $\pm$36	47 $\pm$29
Time per image	55 $\pm$31	77 $\pm$41	60 $\pm$35

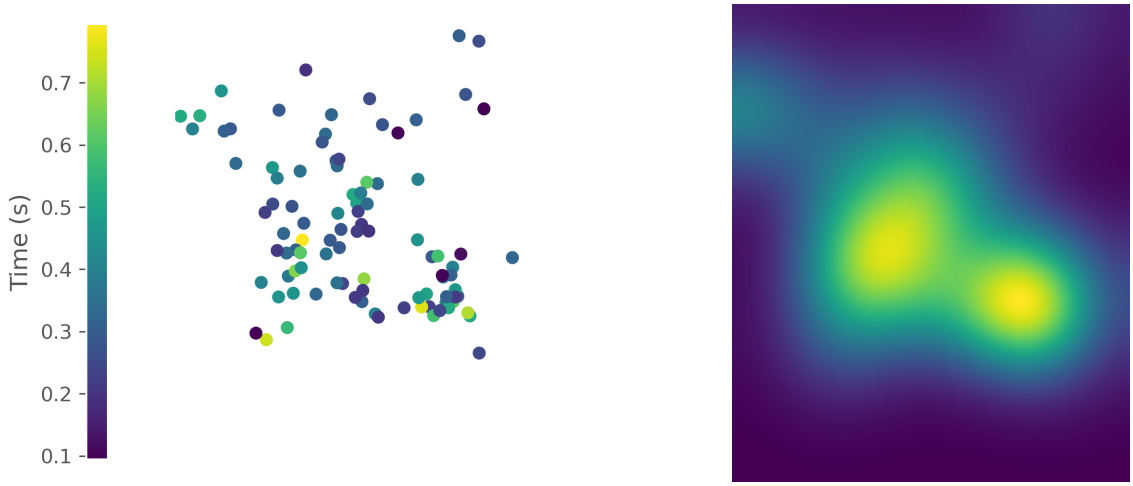
Chapter 5

Heatmap Generation

In order to train DL models, ETD had to be converted into images. This was done in the form of heatmaps, that model the distribution of fixations across each image. This chapter describes the methods used for each dataset and shows the corresponding results.

5.1 Methods

For all datasets, the heatmaps were generated by applying a gaussian kernel to the ETD. More specifically, the gaussian kernel was applied to arrays of the same size as the images, containing the time values for each fixation in the respective coordinates. Figure 5.1 elucidates that process.



(a) Scatter plot. (b) Heatmap.

Figure 5.1: Scatter plot showing the fixations and corresponding times, and the final heatmap after applying a gaussian kernel. The colormap applied to the heatmap is the same as the one used in the scatter plot, with yellow and dark blue corresponding to large and small values, respectively.

The rationale behind using a gaussian kernel was that it suits well the importance of the area around each fixation, with the center point being more relevant, and with the importance decreasing when moving away from it. This required, however, the selection of an appropriate standard deviation. Unlike in [7], where the standard deviation selected was the same for every image regardless of its size, here specific standard deviations were calculated. To obtain these values, several steps were performed.

1. Calculation of the angular resolution in the screen space;
2. Conversion of the angular resolution to the image space;
3. Computation of the visual diameter for each image;
4. Calculation of the standard deviations.

The first step consists in performing a simple trigonometric operation by using the distance from the radiologist to the screen, and the screen size. Afterwards, the obtained value in inches per degree is converted to pixels per degree by knowing the screen resolution. Then, the second step converts the angular resolution in the screen space to the image space with the Equations 5.1 and 5.2, where r is the angular resolution in the image/screen space, w and h are the width and height of the image/screen, and pad corresponds to the padding performed in the screen, either at the left, right, top, or bottom of the image. Two equations were required since this conversion step outputs slightly different values in terms of width and height.

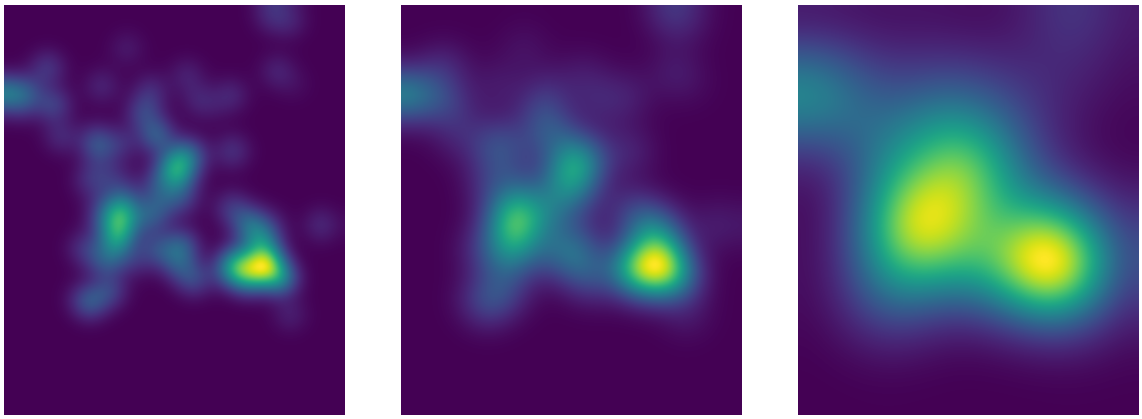
$$r_{image,w} = r_{screen} \frac{w_{image}}{w_{screen} - pad_{left} - pad_{right}} \tag{5.1}$$

$$r_{image,h} = r_{screen} \frac{h_{image}}{h_{screen} - pad_{top} - pad_{bottom}} \tag{5.2}$$

For the third step, first an amplitude regarding the visual angle of the radiologists had to be selected. The defined value was 5 degrees, the same as used in [60], since it corresponds to the amplitude of the foveal vision. The foveal vision corresponds to the central part of the field of view, and is used for tasks requiring high visual detail, like reading a CXR image. With the amplitude defined, and knowing the angular resolutions, it was possible to compute the visual diameters in the image space. Two values were obtained, one for width and one for height, which were then averaged.

With the image diameters calculated, it was then possible to define the standard deviations to apply in each case. In [60], the standard deviation is defined as a sixth of the diameter corresponding to the foveal vision. This means that the highlighted area by the gaussian will almost entirely

be contained within the corresponding diameter. However, this has the downside of neglecting the outer part of the foveal vision. Therefore, the standard deviations were defined as being half of the diameter for each image. In Figure 5.2, a comparison between heatmaps generated with distinct standard deviations is shown, in order to visualize the differences. As it it possible to see, with a standard deviation corresponding to half the size of the diameter, the highlighted regions are wider, and the heatmaps are smoother.



(a) 1/6. (b) 1/4. (c) 1/2.

Figure 5.2: Examples of heatmap generated with different standard deviations.

Since the datasets differ in terms of the amount and type of information complementing the ETD, the process of generating the heatmaps will also vary. In the sections below, the specificities for each dataset are described.

5.1.1 EGD Dataset

In order to select an adequate standard deviation for every image, the complementary data provided in the EGD dataset was used. It is stated in [7] that the eyes of the radiologist were approximately 28 inches away from the screen while the ETD was being collected. The screen brand and model are also disclosed, alongside the resolution used, 1920x1080 pixels. Altogether, it was possible to compute the angular resolution in the screen space, which corresponded to 40 pixels per degree. Through Equations 5.1 and 5.2 this value was converted to the image space, and the diameters were calculated. Figure 5.3 shows the results up to this point, with the vast majority of the diameters between 550 and 600 pixels. Nevertheless, for some images, the diameters were close to only 300 pixels. This was due to the variable image size, which supports the proposed method. Then, the standard deviations were obtained from the diameters, and a gaussian kernel was applied to each image.

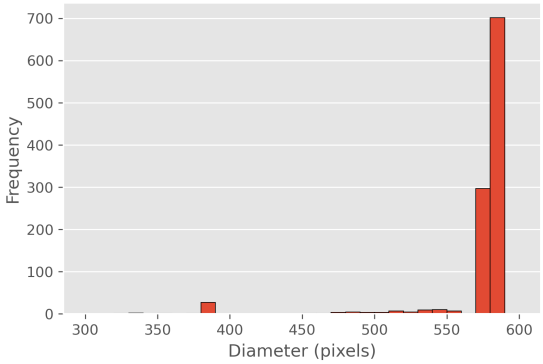


Figure 5.3: Histogram showing the different diameter sizes in the image space for EGD.

5.1.2 REFLACX Dataset

In this dataset, a key difference exists: the radiologists were allowed to zoom in on images during the ETD collection. This was an advantage for the radiologists, but created the issue of having different visual diameters for each fixation. Therefore, the authors of the dataset registered for each fixation the angular resolution in the image space, both horizontally and vertically, which allowed the bypass of the first two steps of the proposed method. With these values, the visual diameters were easily obtained. In Figure 5.4, the distribution of the calculated diameters for each fixation is displayed.

Given the wide range of diameters, a single gaussian kernel per image would be insufficient. However, the application of a gaussian with a different standard deviation for each fixation would necessarily imply a large computational demand. In order to avoid this, fixations were grouped according to the respective diameters, and a single value was used for each group. Five groups were created, since that number offers some resolution while not increasing the computation time significantly. Then, as previously, the standard deviations were defined as being half of the selected diameters. After obtaining these values, each heatmap was generated by using up to 5 different gaussian kernels, one for every existing group of fixations.

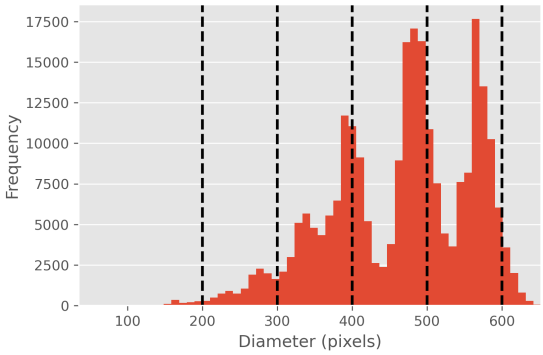


Figure 5.4: Histogram showing the diameter distribution across the REFLACX dataset for every fixation. The black dashed lines indicate the chosen values for the five groups. Each fixation is assigned to the nearest group based on the respective diameter.

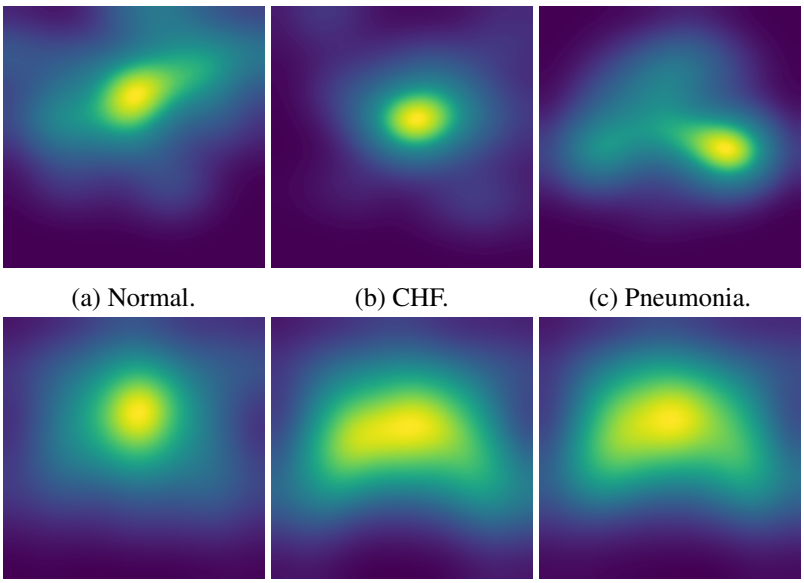
5.1.3 CXR-P Dataset

For this dataset, no information was given which could possibly be used to calculate the standard deviations. To circumvent that, it was assumed that the data collection conditions were identical to the EGD setup, namely the screen size and resolution, and the distance to the radiologist. Since the images provided by the SIIM-ACR dataset have relatively small dimensions, of 1024x1024 pixels, they would fit directly in the screen without any resizing operation. Consequently, the initial screen space angular resolution computed for EGD would be equivalent to the angular resolution in the image space. By doing the product between this value and the amplitude of the foveal vision, the image space visual diameter could be computed, which would be equivalent to 200 pixels and identical for every image since they all have the same size. The remainder of the process was identical to the one used in EGD, with standard deviations corresponding to half of the diameter and using one gaussian kernel per image.

5.2 Results

5.2.1 EGD Dataset

In Figure 5.5, examples of the resulting heatmaps for the EGD dataset are shown, alongside the mean heatmaps for each class. By comparing normal with abnormal cases, it can be seen that the heatmaps differ. In the former, the highlighted areas are more centered, while in the latter the lungs are also highlighted. This might be caused by an involuntary attraction of the gaze towards the center of the image that, when no abnormalities are detected, concentrates the majority of fixations. Curiously, no significant differences were observed between the mean CHF and mean pneumonia heatmaps.



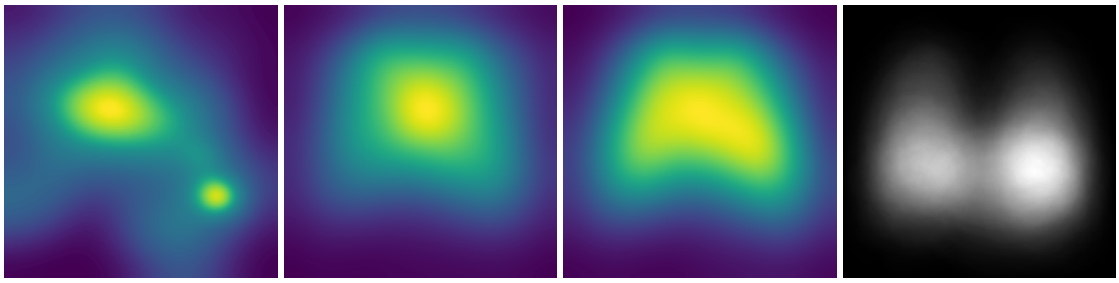
(a) Normal. (b) CHF. (c) Pneumonia.

(d) Mean Normal. (e) Mean CHF. (f) Mean Pneumonia.

Figure 5.5: Examples of heatmaps for each class in EGD and mean heatmaps.

5.2.2 REFLACX Dataset

In REFLACX the generated heatmaps differed slightly, since multiple gaussian kernels were used to create each heatmap. Figure 5.6 shows a representative example of a heatmap with different fixation diameters, alongside the mean heatmaps obtained for CXR images with no findings and for pathological ones. A similar aspect when compared to the EGD dataset was obtained. Again, normal heatmaps have the tendency to be more centered, while the abnormal ones highlight a vaster area, coincident with the thorax. In addition to this analysis, and since pathology masks are available in this dataset, the colocalization between heatmaps and anomaly masks was evaluated. By comparing Figure 5.6c with Figure 5.6d, it is possible to see that the masks are more incident in the inferior part of the lungs, while the heatmaps are less specific.



(a) Heatmap example. (b) Mean normal. (c) Mean abnormal. (d) Mean mask.

Figure 5.6: Example of a heatmap with different fixation diameters and mean heatmaps for normal and abnormal images. The mean of the anomaly masks is also displayed to assess the colocalization with the mean abnormal heatmap.

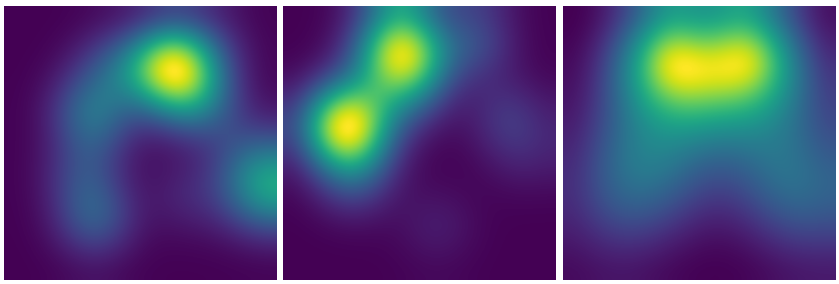
To complement this assessment, the heatmap mean intensity values within the masks were computed. These values were then compared to the mean intensities in the bounding boxes (excluding the masks), and to the mean intensities outside the bounding boxes. Table 5.1 contains the results obtained. The mean intensity is higher inside the masks when compared to other regions within the bounding boxes, showing that the generated heatmaps contain important information content regarding the occurrence of pathologies. Conversely, the standard deviation was quite high, which indicates that at least for some heatmaps this might not be the case.

Table 5.1: Mean intensity values for different regions in the REFLACX heatmaps, in the interval [0,1].

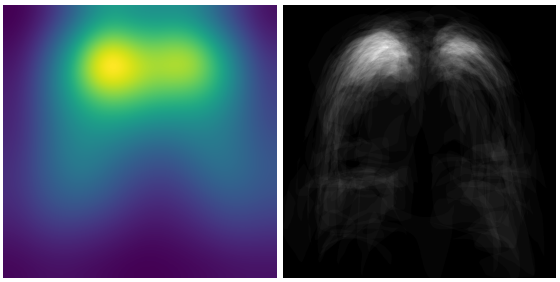
Mask	Bounding box	Outside
0.44 $\pm$0.17	0.26 $\pm$0.08	0.06 $\pm$0.04

5.2.3 CXR-P Dataset

In Figure 5.7, examples of heatmaps for the CXR-P dataset are shown alongside the mean heatmaps. It is possible to see that the generated heatmaps resemble the ones for previous datasets in terms of the fixation diameter, which ascertains the validity of the initial assumption that the conditions were similar to EGD. Regarding the mean pneumothorax heatmap, it is less centered and highlights a more specific area, namely the top of the lungs. This particular behaviour is most likely related to the pathology present in this dataset, which is located preferentially in the highlighted areas (as shown in Figure 5.7e). Curiously, the mean normal heatmap is similar and, unlike its previous counterparts, is not as centered. This might indicate some bias from the radiologists, which immediately search for pneumothoraces in the most likely locations.



(a) No pneumothorax. (b) With pneumothorax. (c) Mean normal.



(d) Mean pneumothorax. (e) Mean mask.

Figure 5.7: Examples of CXR-P heatmaps for images without and with pneumothorax and mean heatmaps, alongside the mean of the masks.

To further evaluate the information content of the heatmaps within the mask areas, a similar method to the previous section was used. The bounding boxes included in this analysis were the ones generated by the thorax detection model described in Chapter 4. Then, the different mean intensities were computed, as shown in Table 5.2. The mean value for the masks was even higher, denoting a significant intensity difference when compared to the bounding box values. Nevertheless, the existent bias in the dataset might exacerbate these differences, generating unrealistic cases that are not consistent with real world scenarios.

Table 5.2: Mean intensity values for different regions in the CXR-P heatmaps, in the interval [0,1].

Mask	Bounding box	Outside
0.61 $\pm$0.19	0.26 $\pm$0.09	0.11 $\pm$0.04

Chapter 6

Baseline Experiments

In order to perform the tasks at hand, namely the reconstruction of heatmaps and the classification of CXR images, several models were tested. In this chapter, those experiments are described, and a comparison is made regarding the performance of the models for both tasks. The best models are then selected to be used in combined approaches, where reconstructed heatmaps aid the process of CXR image classification. These combined approaches are explained later on, in Chapter 7.

6.1 Experiment Setup

6.1.1 Cross-Validation

To perform these comparisons, only one of the three datasets was used. The EGD dataset was the selected one, since it is the smallest of the three, which decreases the amount of time required to train and test all of the considered models. Furthermore, it was the dataset in which the model proposed in [7] was trained and tested, allowing for a more direct comparison with different methodologies. This dataset was split into five different folds, while maintaining a similar class distribution. Furthermore, images belonging to the same patient were kept in the same fold, in order to avoid leakage. The distribution of the classes for each fold is displayed in Figure 6.1.

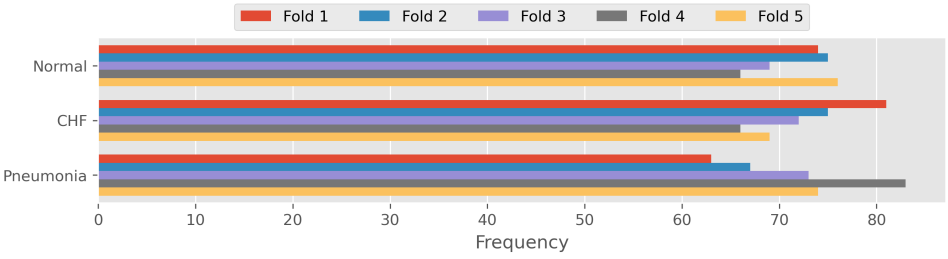


Figure 6.1: Bar plot with the class distribution for each fold of the EGD dataset.

The distribution is quite similar for every fold, either in terms of the total number of samples or in terms of the number of examples per class. Only small disparities are verified, due to the constraints applied during the splitting process.

During the training process for the models, three of the folds were used as a training set, one as a validation set, and one as a test set. Every model would run a total of five separate rounds. At each round, the sets would rotate so as to never repeat the same validation and test sets. Figure 6.2 offers a visual representation of how this process worked.

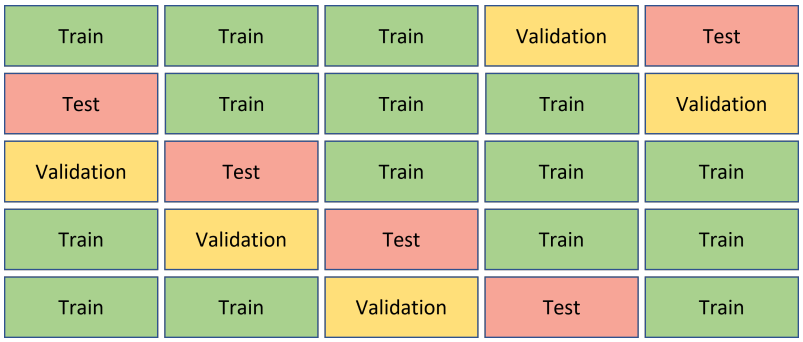


Figure 6.2: Distribution of train, validation, and test sets. Each row corresponds to a different split, and each column to a different fold.

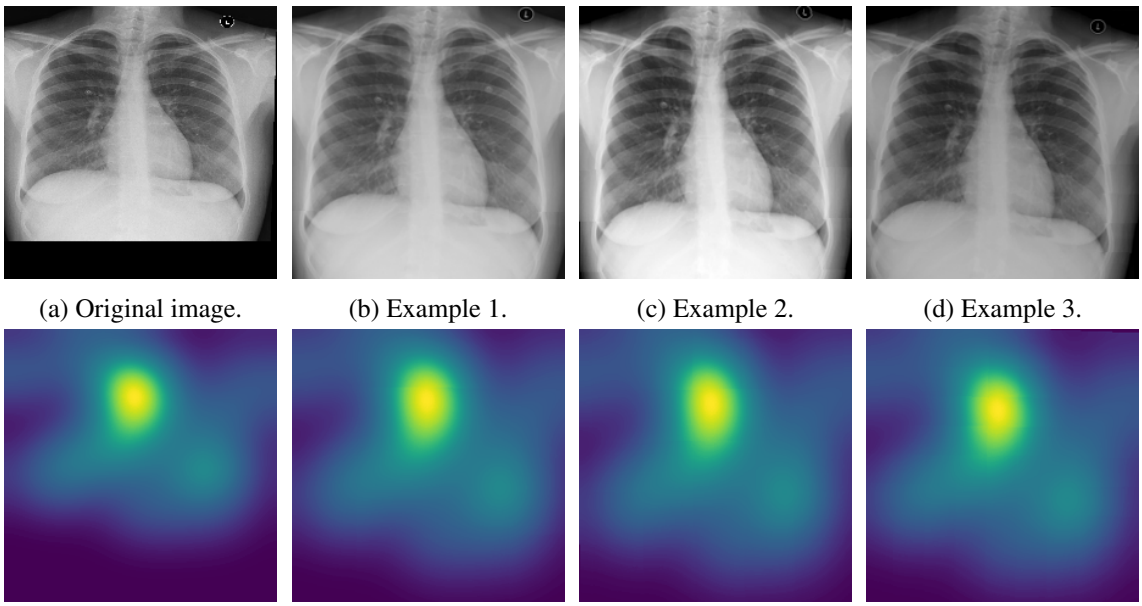
After every epoch, the loss was calculated for the validation set and, if no improvement was registered for ten epochs, the training process would stop. The model with the lowest validation loss would then be selected, and evaluated in the test set. The process was repeated for all five splits and, in the end, the obtained values for the task-specific metrics were averaged. Two main advantages exist from this training strategy. Firstly, the use of early stopping prevents the occurrence of overfitting, which happens when a model fits the training data too well and is not able to generalize to new samples. By using the validation set to stop the training and to select the best model, it is ensured that these decisions are made based on independent data samples. Secondly, by running the models for five different splits their intrinsic variability is taken into account, and more informed and rigorous analyses can be made.

6.1.2 Data Preprocessing

During the training process, data augmentation was used. This concept relates to the generation of modified examples from existing images or to the creation of synthetic data, with the goal of increasing the dataset size and its variability. In this setup, the CXR images contained in the datasets, before passing through the models, were transformed. A concise description of the transforms and respective order is given below.

1. Cropping of CXR image black borders (if existent);
2. Random rotation by α, in which $\alpha \in$ [-5º,5º];
3. Random changes in brightness and contrast up to 20%;
4. Random crop of borders by up to 10% of each dimension;
5. Resizing to a shape of 224x224 pixels;
6. Normalization to mean 0 and standard deviation 1.

Whenever heatmaps were used as GT, they were first concatenated with the CXR images and then submitted to the mentioned transforms, apart from the normalization step. This ensured the overlap between heatmaps and CXR images. Instead of the normalization, the heatmaps were scaled to an interval between 0 and 1. Furthermore, every parameter used in these transformations is randomly sampled from a uniform distribution. In Figure 6.3, three examples of the application of the mentioned transforms are shown for the same CXR image.



(a) Original image. (b) Example 1. (c) Example 2. (d) Example 3.

(e) Original heatmap. (f) Example 1. (g) Example 2. (h) Example 3.

Figure 6.3: Comparison between the original image and heatmap and three transformed versions.

In this sequence of CXR images the changes caused by the transforms are quite evident. The most visible ones are the cropping of the black borders and alterations in brightness and contrast. Additionally, some rotation and random cropping are also observed. More importantly, it can be seen that the transforms introduce variability without generating unrealistic images, which would defeat the purpose of using this strategy. Furthermore, it is possible to verify the overlap between the heatmaps and CXR images even after the modifications. Regarding the transforms applied

to the validation and test sets, it is important to note that only the cropping of the black borders, the resize, and the normalization operations occur. This adds no variability to these sets of data, meaning the models will be evaluated on the original samples. This is necessary since in order to compare models and experiments, the validation and test sets have to be identical.

6.1.3 Models

Regarding the tested models, for the heatmap reconstruction task they are as follows:

- Standard UNet with [32,64,128,256] channels at each level;
- UNet with pretrained ResNet50 encoder;
- UNet with pretrained DenseNet121 encoder;
- UNet with pretrained EfficientNet-b0 encoder.

From this point forward, for matters of simplicity, the models will be named UResNet, UDenseNet, and UEfficientNet, respectively.

Regarding the classification task, a similar set of models was tested:

- Pretrained ResNet50;
- Pretrained DenseNet121;
- Pretrained EfficientNet-b0.

Again, pretrained encoders were used, but this time they were not inserted into a UNet backbone. Instead, the encoders were attached to an untrained fully connected network responsible for the classification from the encoder extracted features. Since the encoders make up most of the classifier models, their names will be used interchangeably either relative to the encoders themselves or to the full models.

For every pretrained model, standard ImageNet weights were used, since they are the most common type. For the EfficientNet-b0, however, NS weights were also evaluated since they were available in [70].

6.1.4 Hyperparameters

To train these models, it was necessary to define values for the batch size and for the learning rate. Based on preliminary experiments, a value of 32 was selected for the batch size, and a value of 10^{-3} for the learning rate. Regarding the optimizer, the one used was Adam since it provides for a more efficient training process. For the classification task, several dropout values were also compared (0, 0.25 and 0.5). This is a commonly used strategy to avoid overfitting, where part of the neurons of the fully connected network are switched off during training.

6.1.5 Loss Function

The final step was to define a loss function. As in [7], the selected function both for classification and heatmap reconstruction was the Binary Cross Entropy (BCE) loss function. More specifically, the implementation used also applied a sigmoid function prior to calculating the loss. Equation 6.1 shows how the BCE loss is computed, where l_n is the loss for each element, y_n the GT, x_n the prediction, and σ stands for the sigmoid function.

$$l_n = y_n * log(\sigma(x_n)) + (1 - y_n) * log(1 - \sigma(x_n)) \tag{6.1}$$

This loss was calculated for every label prediction or for every pixel, depending on the considered task. Then, the values computed were averaged throughout the entire batch. The application of the sigmoid function shifts the range of the output to an interval between 0 and 1, which explains why the original heatmaps were also converted in the same way.

6.2 Results

6.2.1 Heatmap Reconstruction

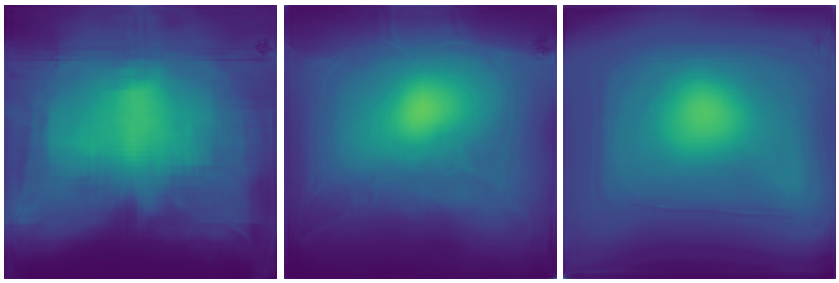
For the heatmap reconstruction task, two criteria were defined to compare the models: the reconstruction BCE and a visual analysis. For each model, the overall and class-specific BCE values were calculated for the EGD test sets, and an average was performed over the five splits. Table 6.1 shows the obtained mean values for the BCE, alongside the respective standard deviations. Besides these values, the mean number of training epochs for each model is also displayed. Even though the latter values were not used in the selection of the best model, they help to elucidate the advantages of using TL.

Table 6.1: Mean number of training epochs and mean BCE values for the EGD test sets.

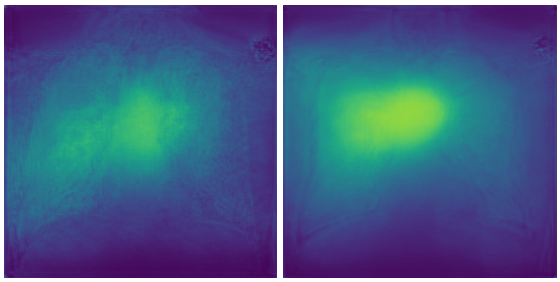
Model	Epoch number	Normal	CHF	Pneumonia	Overall
UNet	55 ± 16	0.479 ± 0.006	0.490 ± 0.007	0.467 ± 0.009	0.478 ± 0.003
UResNet	15 ± 1	$\mathbf{0.477 \pm 0.007}$	0.490 ± 0.006	0.466 ± 0.010	0.477 ± 0.005
UDenseNet	17 ± 1	0.478 ± 0.006	$\mathbf{0.486 \pm 0.007}$	$\mathbf{0.462 \pm 0.010}$	$\mathbf{0.475 \pm 0.004}$
UEfficientNet	14 ± 2	0.478 ± 0.004	0.488 ± 0.008	0.465 ± 0.007	0.477 ± 0.004
UEfficientNet (NS)	24 ± 7	0.481 ± 0.005	0.489 ± 0.008	0.463 ± 0.011	0.478 ± 0.006

The number of training epochs is much higher when the standard UNet is compared with the pretrained models. Furthermore, the standard UNet is one of the models with the highest overall BCE values. On the other hand, the UDenseNet is the model with the lowest overall, CHF, and Pneumonia BCE values. Even though the differences are marginal, it should be taken into account that these values are obtained through a sequence of averaging operations. First the pixel-wise mean, then the mean BCE over the test set, and then the mean over the five splits. Therefore, a difference of one thousandth can be quite significant.

For the second criterion, the reconstructed heatmaps were assessed visually and compared between models. Figure 6.4 shows one such case in which five examples regarding the same original heatmap, one for each model, were compared.



(a) UNet. (b) UResNet. (c) UDenseNet.



(d) UEfficientNet. (e) UEfficientNet (NS).

Figure 6.4: Comparison between heatmaps generated by different models.

This Figure is representative of the overall appearance of the heatmaps belonging to each model, with the most differentiating feature being the texture. For the standard UNet, a grid-like texture is present in the reconstructed heatmaps, while for the UEfficientNet models there is a coarse texture, especially for the ImageNet version. On the other hand, the UResNet and the UDenseNet output heatmaps with a smoother appearence. Another difference is the brightness, with the UEfficientNet (NS) model outputting heatmaps with higher pixel intensities. However, this feature will not be relevant given the reconstructed heatmap processing explained in Chapter 7.

Considering everything, the UDenseNet was selected, since it presents the best BCE values and smooth heatmaps.

6.2.2 Baseline Selection

For the classifier selection, the comparison was based on the average Area Under the Curve (AUC) values for the EGD test sets. Table 6.2 shows the obtained values for every model, as well as the mean AUC. It is clear that, for every model, there is a lower performance in the Pneumonia class, which might relate to a more difficult diagnosis of this pathology. Regarding the model comparison, the EfficientNet-b0 models presented a slightly higher mean AUC. EfficientNet-b0 (NS), more specifically, was the one with the best performance in the CHF and Pneumonia classes. Therefore, it was the selected classifier for subsequent experiments.

Table 6.2: Mean AUC values for different classifiers in the EGD dataset.

Model	Normal	CHF	Pneumonia	Mean
ResNet50	**0.88 ± 0.01**	0.86 ± 0.03	0.67 ± 0.03	0.80
DenseNet121	0.87 ± 0.03	0.87 ± 0.05	0.70 ± 0.06	0.81
EfficientNet-b0	0.87 ± 0.03	0.88 ± 0.03	0.70 ± 0.03	**0.82**
EfficientNet-b0 (NS)	0.86 ± 0.03	**0.89 ± 0.02**	**0.72 ± 0.03**	**0.82**

In order to optimize the performance of the chosen classifier, different dropout values were tested during training on the EGD dataset. The corresponding results are presented in Table 6.3. For 0.0 and 0.25, the mean AUC values were the same. However, for a dropout value of 0.5, the mean AUC was slightly higher. These results led to the choice of a 0.5 dropout to be used in the EfficientNet-b0 (NS).

Table 6.3: EfficientNet-b0 (NS) AUC results for different dropout values.

Dropout	Normal	CHF	Pneumonia	Mean
0.0	0.86 ± 0.03	**0.89 ± 0.02**	**0.72 ± 0.03**	0.82
0.25	**0.87 ± 0.02**	0.88 ± 0.03	0.70 ± 0.04	0.82
0.5	**0.87 ± 0.03**	**0.89 ± 0.03**	0.71 ± 0.03	**0.83**

Chapter 7

Heatmap-Aided Pathology Classification

This chapter depicts the two proposed frameworks joining the previously selected models. In these integrated strategies, reconstructed heatmaps are used to aid the final predictions. The first approach consists in performing thorax segmentations using the reconstructed heatmaps, which are then used to make predictions. On the other hand, the second approach does not binarize the heatmaps and multiplies the CXR images by the reconstructed heatmaps. The two approaches were evaluated in several datasets, both in terms of performance and explainability.

7.1 Experiment Setup

In these experiments, the EGD and REFLACX datasets were used to train and test the models. Additionally, the models trained on REFLACX were tested on the entirety of the other two datasets, since they contain equivalent classes. For EGD, the classes CHF and Pneumonia are similar to the Enlarged Cardiac Silhouette and Consolidation classes in REFLACX. For CXR-P, the Pneumothorax class is also present in REFLACX. No splitting was necessary for this testing phase.

As for the training process, it was similar to Chapter 6, but with extra steps to process the reconstructed heatmaps and CXR images. Other parameters were also included in the training process, which are described in the following sections alongside the respective models.

Furthermore, a baseline classifier - an EfficientNet-b0 (NS) with dropout of 0.5 - and the model proposed in [7], named here EGDmodel, were trained and tested in the same datasets. For the EGDmodel, every aspect of the training process described in [7] was replicated.

Every model was also trained without any optimization to perform a fairer comparison between the proposed approaches and the state-of-the-art. In the models developed throughout this work, no dropout was used. For the EGDmodel, the training process was simplified and the same parameters were used (no dropout, learning rate of 10^{-3}, and a γ value of 0.5). In these experiments, the heatmaps and transforms used in the proposed approaches were also applied to the EGDmodel.

7.1.1 Cross-Validation

The REFLACX dataset was split into 5 different folds, as previously done for EGD. The class distribution was maintained in every fold, and no images belonging to the same patient were present in different folds. The bar plot shown in Figure 7.1 represents the distribution of the number of examples per fold. The class distribution is quite even across the folds, with only small differences since finding a perfect balance is not feasible. On the other hand, some of the classes are not well represented due to the small number of examples present in the dataset.

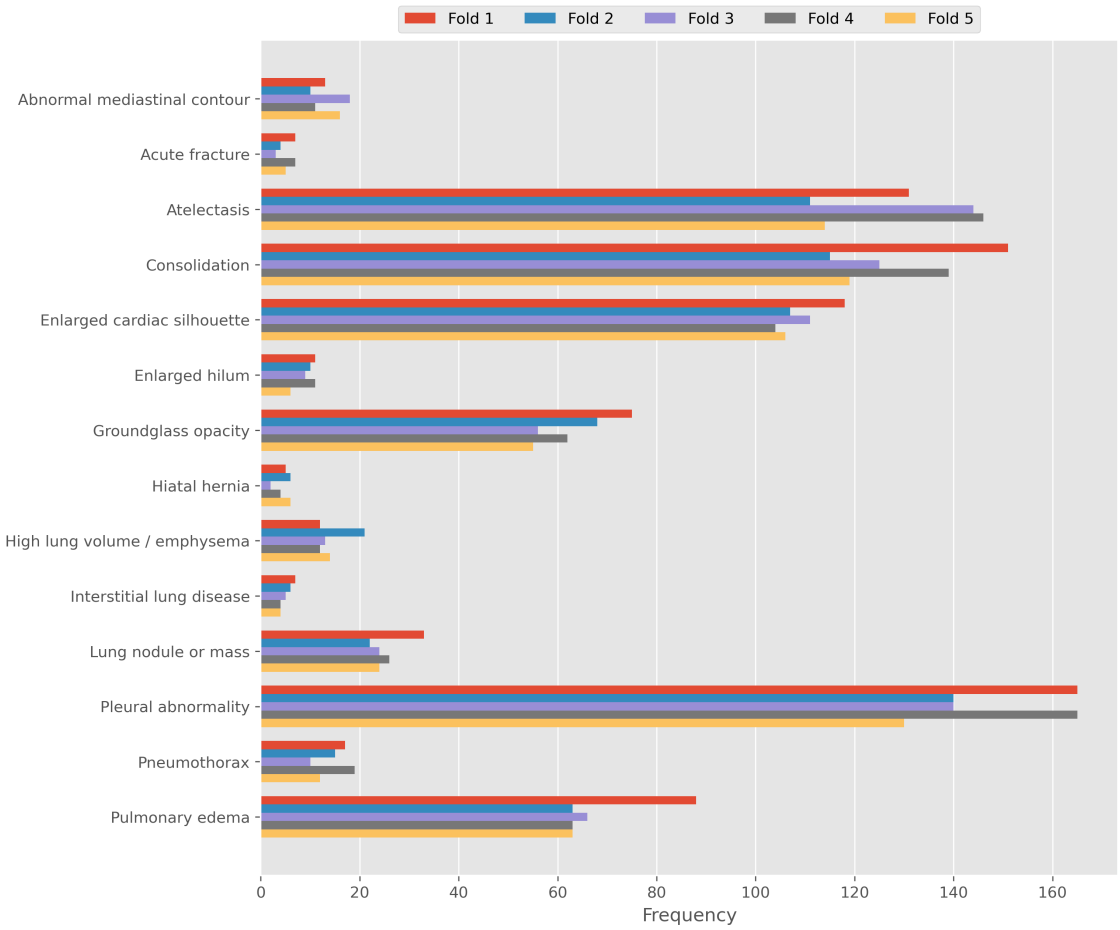


Figure 7.1: Bar plot with the class distribution for each fold of the REFLACX dataset.

7.1.2 Models

7.1.2.1 Thorax Segmentation

The first proposed framework consists in executing a heatmap-based thorax segmentation. For that to happen, a threshold is applied to the reconstructed heatmaps derived from the UDenseNet. The output is a thorax mask which can then be used to segment the corresponding CXR region. With this strategy, irrelevant areas of the input image, such as common markings in the top corners, are

eliminated. The resulting CXR images are then used as input to the selected classifier, which now only uses information located in the lungs and heart. Figure 7.2 shows a schematic of the proposed framework.

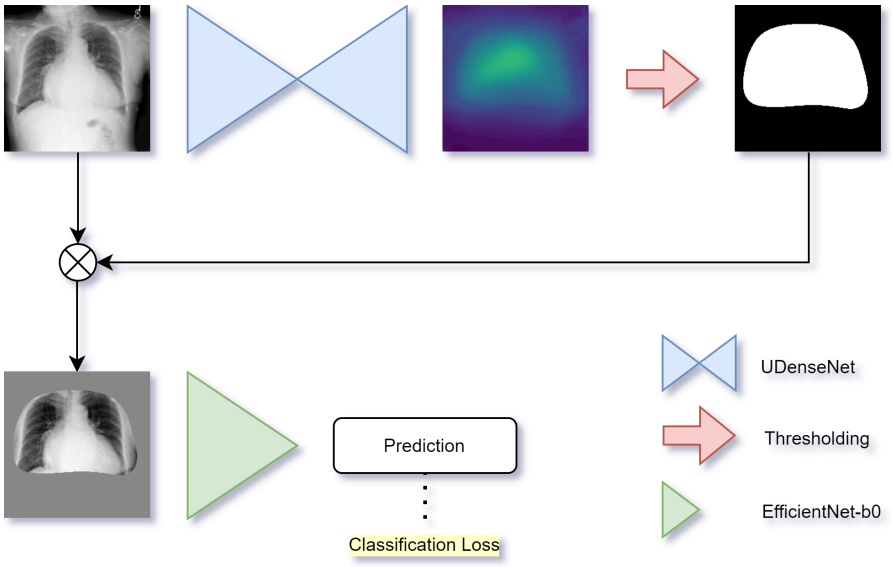


Figure 7.2: Schematic of the first approach.

More specifically, this approach encompasses the following steps:

1. Heatmap reconstruction using the UDenseNet (including the application of a sigmoid);
2. Application of a gaussian filter with a kernel size of 51 pixels;
3. Thresholding using the mean of each reconstructed heatmap;
4. Product between CXR image and obtained thorax mask;
5. Final prediction using an EfficientNet-b0 and the resulting image as input.

The UDenseNet is trained previously on the corresponding dataset and split and used in evaluation mode during this process. Then, the UDenseNet reconstructed heatmaps undergo a sigmoid function and a gaussian filter. The latter increases the smoothness of the heatmaps and consequently of the mask edges after the thresholding operation. The applied threshold corresponds to the mean of each heatmap, since it proved to be an appropriate choice in the experiments performed. The mean threshold constitutes an adaptive strategy, which results in a decreased sensitivity to changes in the reconstructed heatmap brightness. After the product between the CXR image and the mask, the thorax is preserved and the surrounding area acquires a gray color, which is explained by the normalization of the CXR images. An EfficientNet-b0, with NS weights and dropout of 0.5 as previously selected, is then trained on the resulting images using the classification loss.

The models present in this framework are not trained simultaneously since receiving inputs from an incompletely trained UDenseNet could pose a problem for the EfficientNet-b0. Therefore, a sequential training strategy is adopted.

To differentiate this approach from the baseline, it will be named UDenseEfficientNet-TS.

7.1.2.2 Differential Preservation

The second approach developed in this work differs from the previous one in two major aspects. Firstly, instead of a thresholding operation, a contrast adjustment is performed. This results in a final image with heatmap-dependent differential preservation instead of a thorax segmentation. Secondly, in a subsequent training stage, both models are trained simultaneously and in a joined process. This allows for the use of the classification loss not only to train the EfficientNet-b0 but also the UDenseNet. Figure 7.3 provides a visual guide on the functioning of the mentioned approach.

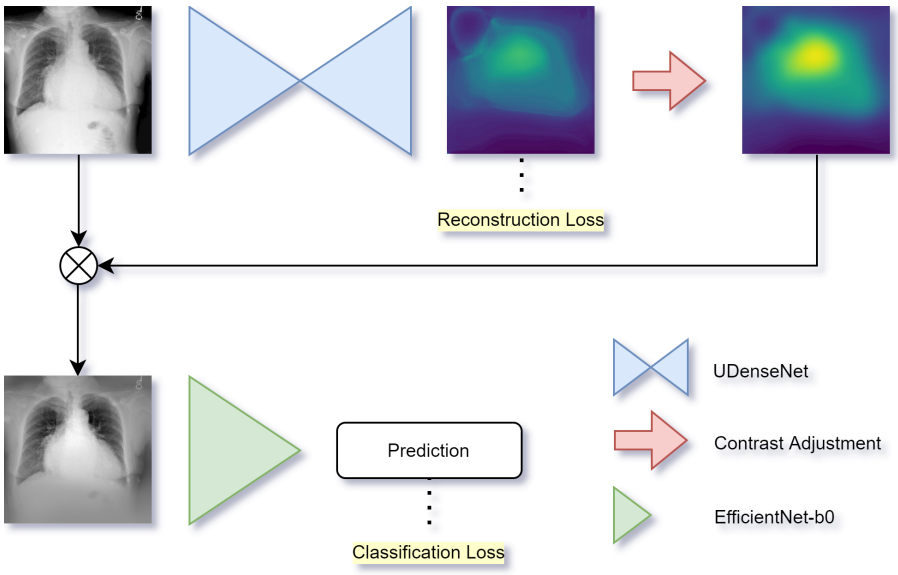


Figure 7.3: Schematic of the second approach.

As previously done, the UDenseNet is first trained on the corresponding dataset and split, in order to facilitate the convergence of the classifier. However, unlike the thorax segmentation approach, a second training stage occurs where both models are jointly trained. The purpose of this second stage is to train the EfficientNet-b0, but also to train the UDenseNet using a combination between the reconstruction loss and the classification loss. The former preserves the heatmap reconstruction ability of the model, while the latter updates the UDenseNet in order to output heatmaps that facilitate the classification process. This integrated strategy is only possible since no thresholding is performed, which would otherwise disrupt the gradient flow from one model to the other.

Regarding the heatmap processing steps, again a sigmoid function and a gaussian filter are applied. Then, a contrast adjustment is performed, which stretches the range of the reconstructed heatmaps to an interval between 0 and 1. This step leads to a higher preservation of the CXR images in the highlighted heatmap regions, and a lower preservation in the darker regions of the heatmaps. The process here described does not binarize the heatmaps, and instead takes advantage of its continuous values and differentially highlighted regions to guide the classifier.

Concerning the training process, certain alterations had to be performed. Based on empirical observations, training a UDenseNet in isolation with a learning rate of 10^{-3} proved to be adequate, but using this value for the joint training stage would not result in quality heatmaps. Therefore, a smaller learning rate, 10^{-4}, was selected. Since for the EfficientNet-b0 a learning rate of 10^{-3} was more suitable, a second optimizer had to be used during the training process to accommodate the two different learning rates.

Lastly, another parameter had to be included to combine the reconstruction and classification losses. This parameter, γ, controls the weight balance of the two losses, as described in Equation 7.1. A higher value of γ gives more importance to the reconstruction loss at the expense of the classification loss, while a lower value does the opposite. A γ of 0.5 was selected, constituting a balanced approach with equal weights for both losses.

$$CombinedLoss = ReconstructionLoss * \gamma + ClassificationLoss * (1 - \gamma) \tag{7.1}$$

To distinguish this approach from the previous one, it will be named UDenseEfficientNet-DP.

7.1.3 Evaluation Metrics

To study and compare the developed models, several evaluation metrics were used. Firstly, to assess the quality of the heatmap reconstruction process, the BCE between the original and reconstructed heatmaps was calculated for the different test sets on an overall and on a class basis. This analysis was not done for the EGDmodel since the heatmap generation process differs.

It was also important to verify the quality of the thorax segmentation process. For EGD, the obtained segmentations for the test sets were compared with the segmentation masks provided in the dataset using the DSC. For REFLACX, however, no segmentation masks were provided, only bounding boxes. To get an estimate of the segmentation quality for this dataset, the thorax segmentations were converted to bounding boxes and compared to the original ones using the DSC. Lastly, for CXR-P, the bounding boxes obtained from the thorax detection model described in Chapter 4 were used once more. Then, the evaluation was performed as in REFLACX.

Furthermore, to determine the information content of the reconstructed heatmaps, the mean intensity values were computed for the masks, and for the bounding box regions (excluding the masks). This process was repeated for the processed heatmaps.

Regarding the classification process, the performance of the models was compared by computing the mean AUC values for the test sets, both for the optimized and non-optimized versions.

7.1.4 Explainability

An important aspect of DL models, as mentioned previously, is their explainability. In this work, the GradCAM method was used to analyse the regions of each image that were more relevant to a given prediction. A Pytorch implementation was applied [78], which performed the gradient backpropagation up to a specific layer, and computed the importance weights and the mean feature map. The selected layer was the last convolutional layer of the classifier, which for every approach was an EfficientNet-b0. This method also performed an upsampling operation, which would result in a final GradCAM image with the same size as the input. For visualization purposes, the GradCAM images were overlapped with the input images to the classifier in RGB format. For the baseline used, the input images corresponded to the original, albeit transformed, CXR images. For the other two approaches, the images used were the result of the processing steps using the reconstructed heatmaps. Only the test sets for each dataset were evaluated.

Since CXR images may include different findings, or no findings, applying this method to the highest scoring class for each example is not suitable. Therefore, the test scores were binarized, according to specific thresholds. These thresholds were computed based on the ROC curves for the validation sets using the geometric mean, as demonstrated in Equation 7.2, where TPR corresponds to the True Positive Rate and FPR to the False Positive Rate.

$$Threshold = argmax\sqrt{TPR(1-FPR)} \tag{7.2}$$

Then, for each case in the test sets, GradCAM images were obtained for the predicted classes, i.e., with scores above the thresholds. This meant that, for the same example, none or multiple GradCAM images could be generated. Figure 7.4 shows an example for the EGD dataset of three ROC curves, one for each class, and the respective best thresholds.

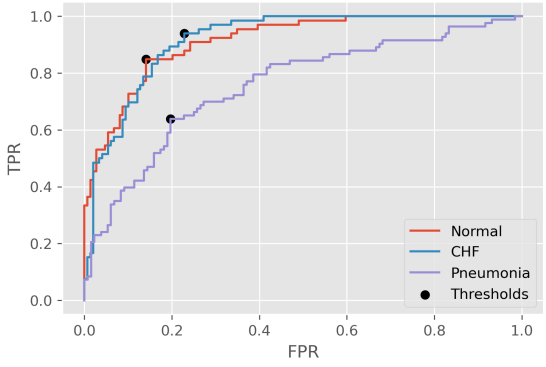


Figure 7.4: Example of ROC curves and respective thresholds for one EGD validation set.

7.2 Results

7.2.1 Heatmap Reconstruction

7.2.1.1 Binary Cross-Entropy

The first step in analysing the performance of the models was to assess the BCE related to the heatmap reconstruction process. This was performed for the models trained on EGD and REFLACX, in the respective test sets. For the latter models, the BCE was also evaluated for the other two datasets, in the follow up of the testing experiments. Each class was considered in separate, as to verify if significant differences occurred in terms of the class-wise BCE. The obtained values are shown in Tables 7.1 to 7.4.

Table 7.1: Mean BCE values for the two proposed approaches in the EGD dataset.

Model	Normal	CHF	Pneumonia	Overall
UDenseEfficientNet-TS	**0.478 ± 0.006**	**0.486 ± 0.007**	**0.462 ± 0.009**	**0.475 ± 0.003**
UDenseEfficientNet-DP	0.483 ± 0.008	0.492 ± 0.007	0.470 ± 0.011	0.482 ± 0.007

Table 7.2: Mean BCE values for the two proposed approaches in the REFLACX dataset.

Model	Normal	Abnormal	Overall
UDenseEfficientNet-TS	0.437 ± 0.004	**0.415 ± 0.002**	0.423 ± 0.003
UDenseEfficientNet-DP	0.437 ± 0.006	0.416 ± 0.002	0.423 ± 0.003

Table 7.3: Mean BCE values for the two proposed approaches trained on REFLACX and tested on EGD.

Model	Normal	CHF	Pneumonia	Overall
UDenseEfficientNet-TS	0.490 ± 0.004	0.498 ± 0.005	0.468 ± 0.003	0.485 ± 0.004
UDenseEfficientNet-DP	0.490 ± 0.003	**0.496 ± 0.003**	**0.467 ± 0.002**	**0.484 ± 0.002**

Table 7.4: Mean BCE values for the two proposed approaches trained on REFLACX and tested on CXR-P.

Model	Normal	Pneumothorax	Overall
UDenseEfficientNet-TS	**0.613 ± 0.004**	**0.591 ± 0.004**	**0.609 ± 0.004**
UDenseEfficientNet-DP	0.614 ± 0.003	0.594 ± 0.002	0.610 ± 0.003

Curiously, in most cases the BCE is smaller for the pathological classes (except for CHF). This is interesting, since the original heatmaps present more centered highlighted areas, which could provide for an easier task. Furthermore, Table 7.1 reveals that the UDenseEfficientNet-DP presents a slightly higher BCE, probably due to the use of the classification loss to train the reconstruction process. In Table 7.2, however, the BCE values are similar. Additionally, the values are smaller for REFLACX when compared to EGD, possibly as a result of the larger dataset size. When the REFLACX trained models are tested in EGD, the values are close to the ones obtained in Table 7.1, which may be due to the common CXR image source. For the testing phase in CXR-P, the BCE values are higher. This might be justified by the dataset characteristics, and by a distinct source of the CXR images.

Mean reconstructed heatmaps for the different models and datasets, for one of the splits, are available in Figures A.1 to A.4 from Appendix A. The range of the mean heatmaps was adjusted to increase the contrast of the images. As in the original mean heatmaps, the normal ones focus mainly on the center, while the abnormal mean heatmaps highlight more extensive regions. This difference is slightly more significant in the UDenseEfficientNet-DP case in comparison to the UDenseEfficientNet-TS.

7.2.1.2 Segmentation Evaluation

In the UDenseEfficientNet-TS, thorax segmentations were performed from the reconstructed heatmaps. The obtained masks were compared to the GT provided in the datasets (either segmentation masks or bounding boxes), or to the bounding boxes obtained from an algorithm with that purpose (in the case of CXR-P). The obtained DSC values are shown in Table 7.5. It is possible to see that the values are relatively high in all datasets, indicating an adequate thorax segmentation from the proposed approach. Curiously, the mean DSC for the model trained on REFLACX and tested on EGD is higher than the value computed for the model trained and tested on EGD. This reveals once again the effect of the dataset size in model performance.

Table 7.5: Mean DSC values for the thorax segmentations.

Train Dataset	Test Dataset	DSC
EGD	EGD	0.82 $\pm$0.01
REFLACX	REFLACX	0.86 $\pm$0.02
REFLACX	EGD	0.85 $\pm$0.02
REFLACX	CXR-P	0.82 $\pm$0.02

To further assess the quality of the thorax segmentations, several examples were collected in which the GT was compared with the segmentation masks obtained. Figure 7.5 shows five examples for the EGD dataset. It can be seen that for some examples the segmentation masks overlap quite well with the GT, with slight disparities at the edges. For the third example, however, a significant portion of the heart is not included in the segmentation mask. Even though for some cases these differences might not be necessarily bad, since the excluded regions could be irrelevant

for the prediction (and thus not highlighted by the reconstructed heatmaps), that is not the case for this particular example, since it belonged to the CHF class. For the REFLACX trained models these events were less frequent, which corroborates the DSC values in Table 7.5 and shows the importance of the dataset size. Examples for these models are shown in Figures A.5 to A.7 from Appendix A.

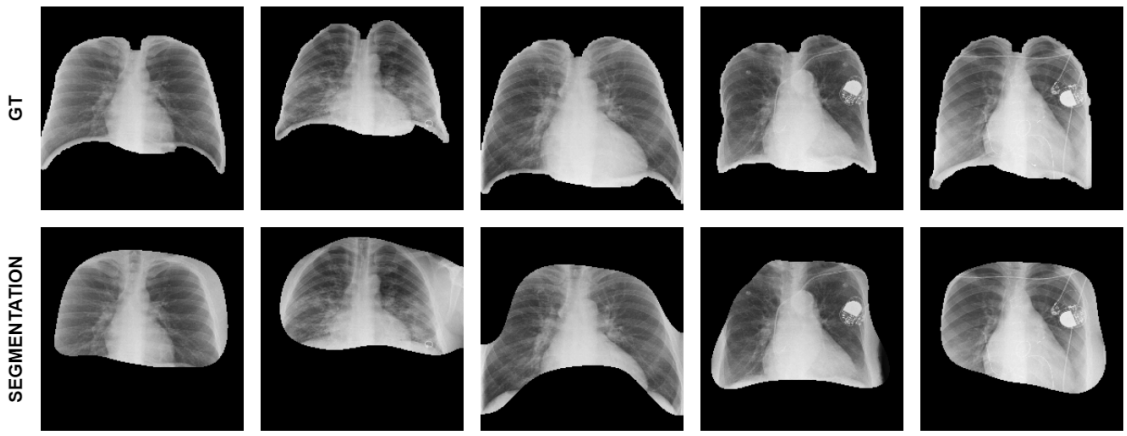


Figure 7.5: Examples of segmentation masks and corresponding GTs for one of the EGD test sets.

7.2.1.3 Intensity Measures

In the UDenseEfficientNet-DP, the contrast adjusted heatmaps were multiplied by the CXR images, to differentially highlight/darken certain regions. To assess if the adjusted heatmaps contained important information regarding the location of pathologies and relevant anatomical areas, the mean intensity values were computed for the masks, bounding boxes, and for the remainder of the image. Figures 7.6 and 7.7 show the results of this process for the datasets where masks were available, REFLACX and CXR-P. Furthermore, the UDenseEfficientNet-TS and the EGDmodel were also included in this analysis for comparison. Moreover, the mean intensities were computed both before and after the heatmap processing.

For REFLACX, the mean intensities within the masks were higher for every model, especially for the UDenseEfficientNet-TS and for the UDenseEfficientNet-DP. The latter also presents a bigger contrast between the masks and bounding boxes, which can be attributed to the second training stage. This revealed that these reconstructed heatmaps contain important information regarding the location of pathologies. Additionally, there was also a clear difference in intensity inside and outside the bounding box for every model. Conversely, for the CXR-P reconstructed heatmaps no significant differences were registered between the mask and bounding box mean intensities. This correlates to the higher BCE values, and shows that the REFLACX trained models might not be suitable to be applied to this dataset. Nevertheless, the bounding boxes are clearly highlighted.

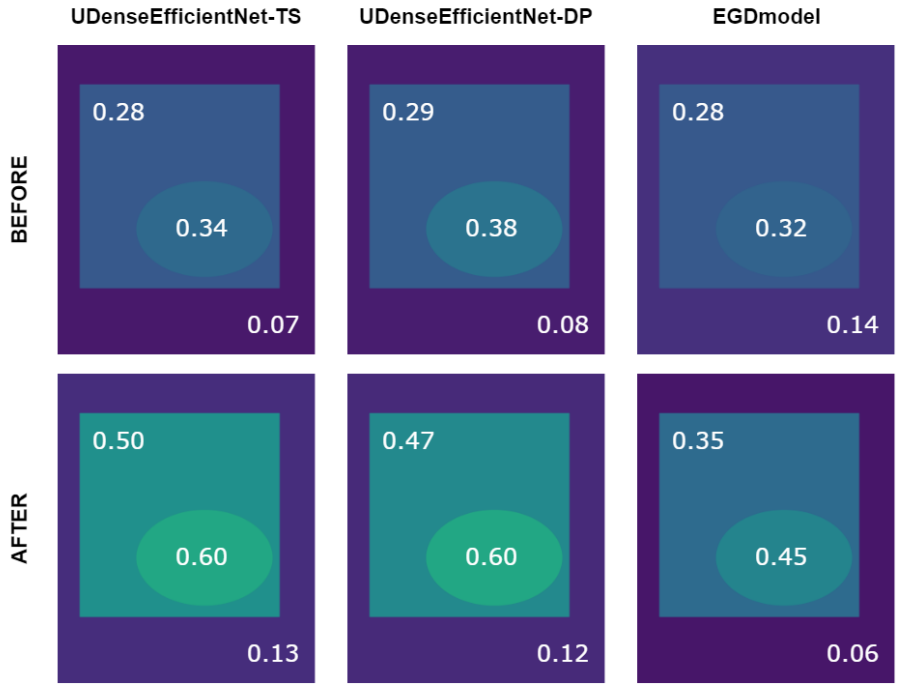


Figure 7.6: Mean intensity values within the masks, and inside and outside the bounding boxes (before and after processing). These values correspond to the average over the five splits for the REFLACX test sets, and are in the range [0,1].

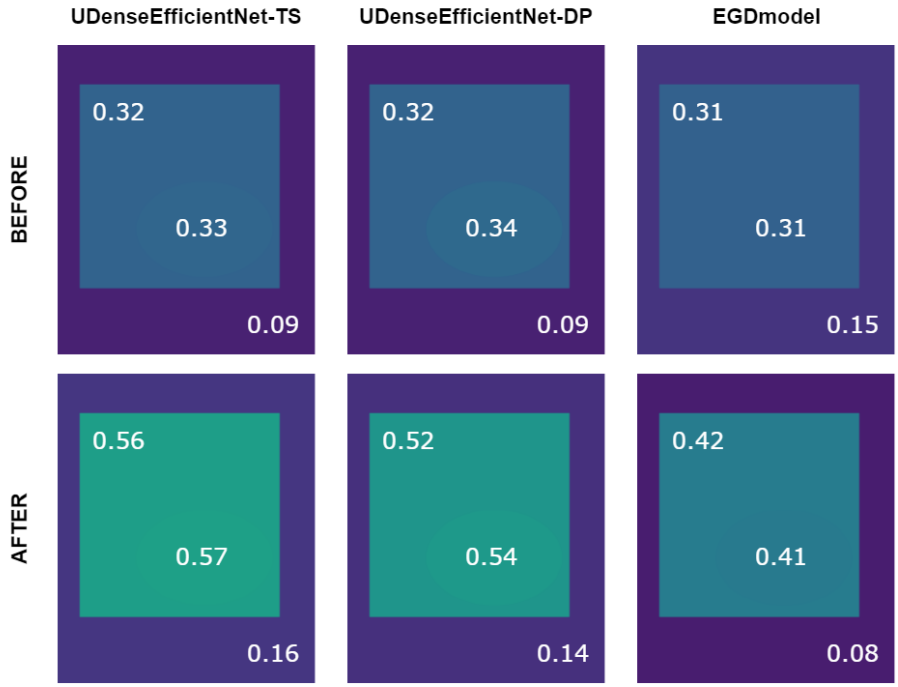


Figure 7.7: Mean intensity values within the masks, and inside and outside the bounding boxes (before and after processing). These values correspond to the average over the five models, one for each split, trained on REFLACX and tested on the CXR-P dataset. Values are in the range [0,1].

7.2.2 Pathology Classification

To assess the performance of the models in the classification task, AUC values were computed. The obtained values are shown in Tables 7.6 to 7.10. In Table 7.6, the mean values for the EGD test sets are displayed. For the optimized section, the EGDmodel outperforms the other models in every class. However, when only the non-optimized models are considered, this is not verified. Furthermore, no significant differences were registered between the UDenseEfficientNet-TS and the UDenseEfficientNet-DP. Regarding the baseline, it slightly outperforms the proposed approaches.

Table 7.6: Mean AUC values for different approaches in the EGD dataset.

	Model	Normal	CHF	Pneumonia	Mean
Optimized	EfficientNet-b0	0.87 ± 0.03	0.89 ± 0.03	0.71 ± 0.03	0.83
	UDenseEfficientNet-TS	0.88 ± 0.02	0.87 ± 0.03	0.70 ± 0.03	0.82
	UDenseEfficientNet-DP	0.88 ± 0.03	0.87 ± 0.05	0.69 ± 0.03	0.81
	EGDmodel	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.73 \pm 0.06}$	**0.84**
Non-optimized	UDenseEfficientNet-TS	0.85 ± 0.01	0.88 ± 0.03	$\mathbf{0.71 \pm 0.02}$	0.81
	UDenseEfficientNet-DP	$\mathbf{0.88 \pm 0.02}$	0.88 ± 0.02	0.68 ± 0.07	0.81
	EGDmodel	0.86 ± 0.03	0.88 ± 0.02	0.69 ± 0.02	0.81

For the REFLACX dataset, the AUC values were also computed for every class, alongside the respective overall means. Since some of these values correspond to less represented classes, in which the performance is lower and highly variable, another mean AUC value was computed excluding classes with less than 50 examples (Acute fracture, Enlarged hilum, Hiatal hernia, and Interstitial lung disease). In Table 7.7, only the optimized model values are displayed, with the EGDmodel presenting the best performance in terms of the overall mean. However, when considering the exclusive mean, the UDenseEfficientNet-DP performance is equivalent, matching also the mean value for the baseline classifier. The UDenseEfficientNet-TS, on the other hand, was one AUC point below in terms of the exclusive mean. In Table 7.8, in which the non-optimized models are compared, the two proposed approaches outperform the EGDmodel. This indicates once more that the extensive optimization performed in the EGDmodel is most likely responsible for the differences verified for the EGD dataset. Moreover, the UDenseEfficientNet-TS and the UDenseEfficientNet-DP present the same mean AUC values, revealing a similar performance in this scenario for both approaches.

Table 7.7: Mean AUC values obtained from the optimized models for every class in REFLACX.

Class	EfficientNet-b0	UDenseEfficientNet-TS	UDenseEfficientNet-DP	EGDmodel
Abnormal mediastinal contour	0.63 ± 0.05	0.63 ± 0.05	**0.69 ± 0.05**	0.67 ± 0.04
Acute fracture	0.64 ± 0.22	0.65 ± 0.05	**0.72 ± 0.07**	0.71 ± 0.09
Atelectasis	**0.78 ± 0.01**	0.77 ± 0.03	**0.78 ± 0.03**	**0.78 ± 0.02**
Consolidation	0.80 ± 0.02	0.80 ± 0.01	**0.81 ± 0.02**	0.80 ± 0.03
Enlarged cardiac silhouette	**0.84 ± 0.02**	**0.84 ± 0.01**	0.83 ± 0.02	0.83 ± 0.02
Enlarged hilum	0.57 ± 0.06	0.58 ± 0.07	0.57 ± 0.07	**0.65 ± 0.07**
Groundglass opacity	**0.70 ± 0.03**	0.67 ± 0.03	0.66 ± 0.02	**0.70 ± 0.02**
Hiatal hernia	0.51 ± 0.19	0.55 ± 0.14	**0.62 ± 0.07**	0.56 ± 0.10
High lung volume / Emphysema	**0.87 ± 0.08**	0.85 ± 0.07	0.84 ± 0.09	**0.87 ± 0.06**
Interstitial lung disease	0.79 ± 0.11	0.74 ± 0.15	0.78 ± 0.09	**0.83 ± 0.09**
Lung nodule or mass	0.63 ± 0.05	0.64 ± 0.05	**0.66 ± 0.05**	0.61 ± 0.05
Pleural abnormality	**0.85 ± 0.02**	0.84 ± 0.01	0.84 ± 0.03	**0.85 ± 0.01**
Pneumothorax	**0.74 ± 0.08**	0.71 ± 0.04	0.72 ± 0.08	0.71 ± 0.05
Pulmonary edema	0.86 ± 0.02	0.84 ± 0.02	0.84 ± 0.03	**0.87 ± 0.02**
Overall mean	0.73	0.72	0.74	**0.75**
Exclusive mean	**0.77**	0.76	**0.77**	**0.77**

Table 7.8: Mean AUC values obtained from the non-optimized models for every class in REFLACX.

Class	UDenseEfficientNet-TS	UDenseEfficientNet-DP	EGDmodel
Abnormal mediastinal contour	0.68 ± 0.03	**0.70 ± 0.07**	0.63 ± 0.08
Acute fracture	0.66 ± 0.12	**0.70 ± 0.17**	0.63 ± 0.17
Atelectasis	**0.77 ± 0.02**	**0.77 ± 0.03**	0.76 ± 0.04
Consolidation	**0.81 ± 0.02**	**0.81 ± 0.02**	0.80 ± 0.02
Enlarged cardiac silhouette	**0.84 ± 0.02**	0.83 ± 0.01	0.81 ± 0.02
Enlarged hilum	**0.64 ± 0.10**	0.57 ± 0.05	0.59 ± 0.10
Groundglass opacity	0.67 ± 0.01	0.66 ± 0.04	**0.69 ± 0.02**
Hiatal hernia	0.58 ± 0.08	**0.61 ± 0.13**	0.57 ± 0.11
High lung volume / Emphysema	0.85 ± 0.10	0.84 ± 0.10	**0.88 ± 0.06**
Interstitial lung disease	0.68 ± 0.12	**0.76 ± 0.08**	0.71 ± 0.10
Lung nodule or mass	**0.62 ± 0.06**	0.60 ± 0.05	0.58 ± 0.06
Pleural abnormality	0.83 ± 0.03	**0.84 ± 0.02**	**0.84 ± 0.03**
Pneumothorax	**0.73 ± 0.04**	**0.73 ± 0.05**	0.68 ± 0.11
Pulmonary edema	0.84 ± 0.02	0.84 ± 0.03	**0.85 ± 0.02**
Overall mean	**0.73**	**0.73**	0.71
Exclusive mean	**0.76**	**0.76**	0.75

The AUC values were also computed for the other two datasets using the REFLACX trained models. Table 7.9 contains the results for EGD, while Table 7.10 contains the results for CXR-P. In the first case, the baseline and the UDenseEfficientNet-TS present slightly higher mean values. For the non-optimized models, curiously, the EGDmodel presents the highest mean AUC. By comparing the values with the ones obtained in Table 7.6, it is possible to verify that they are similar for CHF. For Pneumonia, however, the AUC values were inferior. This is probably due to some disparities between the REFLACX Consolidation class and the EGD Pneumonia class.

Table 7.9: Mean AUC values for the models trained on REFLACX and tested on EGD.

	Model	CHF	Pneumonia	
Optimized	EfficientNet-b0	0.87 ± 0.01	$\mathbf{0.62 \pm 0.01}$	**0.74**
	UDenseEfficientNet-TS	$\mathbf{0.88 \pm 0.01}$	0.61 ± 0.02	**0.74**
	UDenseEfficientNet-DP	0.86 ± 0.02	0.61 ± 0.02	0.73
	EGDmodel	$\mathbf{0.88 \pm 0.01}$	0.58 ± 0.02	0.73
Non-optimized	UDenseEfficientNet-TS	0.86 ± 0.01	0.59 ± 0.04	0.73
	UDenseEfficientNet-DP	0.86 ± 0.02	0.60 ± 0.01	0.73
	EGDmodel	0.86 ± 0.01	$\mathbf{0.61 \pm 0.02}$	**0.74**

For the models trained on REFLACX and tested on CXR-P, the obtained results were similar to the ones achieved in the REFLACX test sets, and even higher in the UDenseEfficientNet-TS, UDenseEfficientNet-DP, and EGDmodel cases. Conversely, the value for the baseline classifier was 4 AUC points inferior. The EGDmodel presented the best value among the optimized versions, while the opposite was verified for the non-optimized models.

Table 7.10: Mean AUC values for the models trained on REFLACX and tested on CXR-P.

	Model	Pneumothorax
Optimized	EfficientNet-b0	0.70 ± 0.02
	UDenseEfficientNet-TS	0.73 ± 0.03
	UDenseEfficientNet-DP	0.74 ± 0.02
	EGDmodel	$\mathbf{0.76 \pm 0.01}$
Non-optimized	UDenseEfficientNet-TS	$\mathbf{0.72 \pm 0.02}$
	UDenseEfficientNet-DP	$\mathbf{0.72 \pm 0.01}$
	EGDmodel	0.68 ± 0.05

7.2.3 GradCAM Experiments

The GradCAM experiments allowed the analysis of several aspects. For the UDenseEfficientNet-TS, it was important to verify if the highlighted areas would focus on the segmented parts of the CXR images. Figure 7.8 shows four examples of CHF and Pneumonia cases where GradCAM images from the EfficientNet-b0 are compared to GradCAM images from the UDenseEfficientNet-TS. Even though the baseline examples highlight relevant regions, like the heart for the CHF cases, other unimportant areas are also highlighted. Surprisingly, the GradCAM method highlights areas outside of the thorax for the UDenseEfficientNet-TS, despite not containing any information. This raises concerns regarding the applicability of the GradCAM method for the proposed approaches.

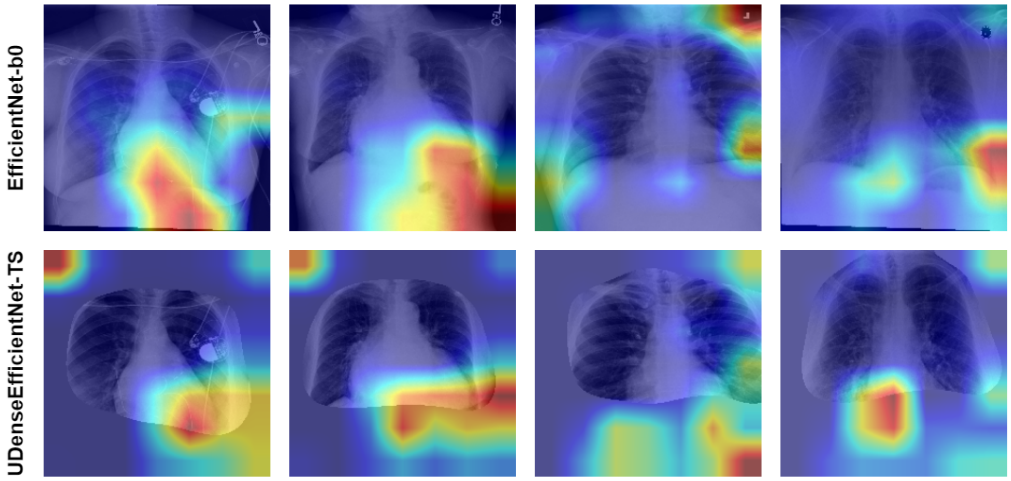


Figure 7.8: Four EGD examples of GradCAMs for the EfficientNet-b0 baseline and for the UDenseEfficientNet-TS. The two cases on the left correspond to the CHF class, and the two cases on the right to the Pneumonia class.

Concerning the UDenseEfficientNet-DP model, it was relevant to see if the reconstructed heatmaps, through the product with the CXR images, could guide the classifier. Therefore, the anomaly masks were compared with the reconstructed heatmaps and GradCAMs, to see if a correlation existed. Figure 7.9 shows five REFLACX examples of anomaly masks, corresponding to different pathologies, and the UDenseEfficientNet-DP reconstructed heatmaps, after the processing steps. It is possible to see that the heatmaps present a higher intensity in the pathological area. However, even though the classifier correctly predicts the existing classes, the GradCAM images were incongruent with the heatmaps. Given the previous results, it is hard to determine the cause of this phenomenon, whether it is related to the application of the GradCAM method to the EfficientNet-b0, or with the proposed framework.

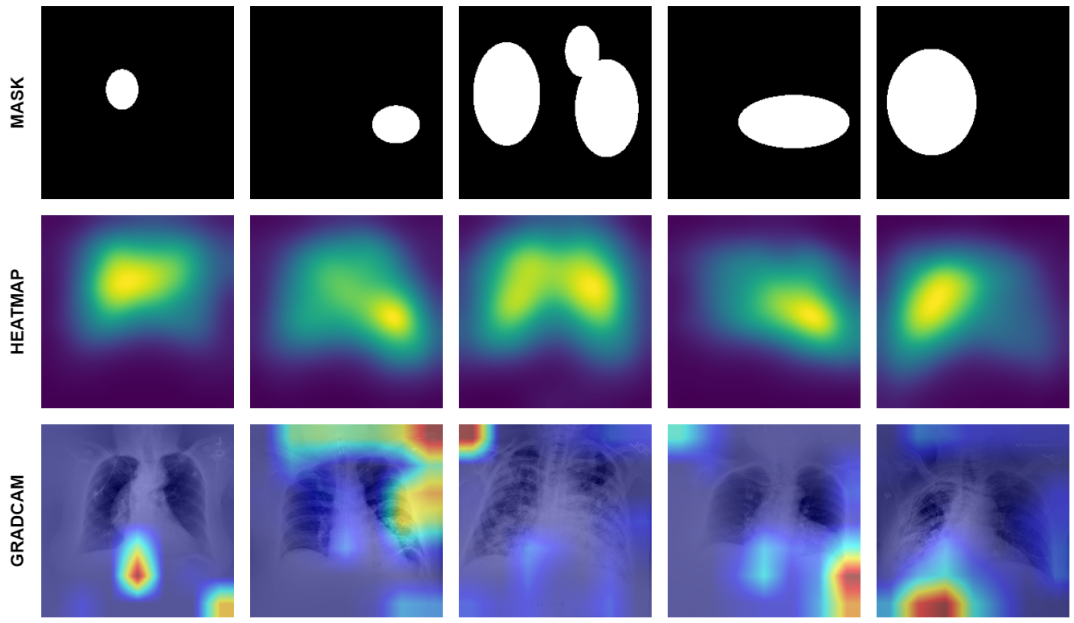


Figure 7.9: Examples of anomaly masks, processed heatmaps, and respective GradCAMs for the UDenseEfficientNet-DP. The corresponding findings are, from left to right, Abnormal mediastinal contour, Atelectasis, Consolidation, Enlarged cardiac silhouette, and Pleural abnormality.

In addition to the previous analyses, the mean GradCAMs were computed for every class of the datasets on which the models were tested. In Figure 7.10, the mean CHF and Pneumonia GradCAMs for one of the EGD test sets are displayed. The CHF mean GradCAMs highlight mainly the bottom right corner, while the Pneumonia ones are more diffuse. In some of the images, and in other computed examples, one of the corners is highlighted. This happens, for example, for the UDenseEfficientNet-TS, in which those corners do not contain any information. This indicates that the GradCAM method is not ideal to study the explainability of these models, and thus no rigorous comparison can be performed.

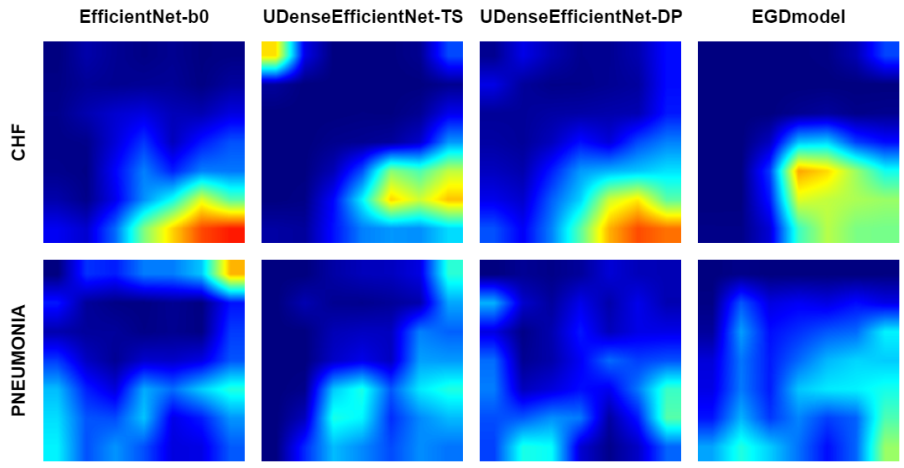


Figure 7.10: Mean GradCAM images for different models regarding the CHF and Pneumonia classes from one EGD test set.

Chapter 8

Discussion

This Chapter starts by discussing the data analyses performed, both in terms of the ETD and of the generated heatmaps. This is followed by an assessment of the heatmap reconstruction and classification performances of the models. Lastly, the limitations of this work are discussed, and the future path to continue improving the developed models is delineated.

8.1 Data Analysis

8.1.1 Eye-Tracking Data

The analysis performed regarding the ETD shows its relevance and utility for the development of DL models. As expected, most of the fixations are located within the thorax, indicating that this type of data can be used for segmenting that region of CXR images. ETD also contains important information relating to the occurrence and location of pathologies. For all the datasets, a higher number of fixations and time were registered in abnormal images when compared to images with no findings. For REFLACX and CXR-P, datasets where anomaly masks were available, it was also possible to verify that a considerable portion of the fixations were contained within the masks, more precisely around 21% and 11%, respectively. Regarding time, similar values were observed. For many cases, however, the percentages were much smaller. This is partially explained by the fact that the visual angle of the radiologists was not taken into account in the analyses performed, and only the center point of the fixations. Therefore, fixations in the vicinity of the masks might not be considered, even though they relate to those findings. Another reason is the fact that ETD is not specific for pathological areas of CXR images, since the radiologists focus on many areas in the search for multiple findings. In the case of CXR-P, there is also a possible scenario where the radiologists do not find existing pneumothoraces in the images. Obvious findings may also require less attention from the radiologists, which spend most of the time in such examples searching for other pathologies. Another dimension unexplored in this work is time, in which ETD may also contain important information.

8.1.2 Generated Heatmaps

The generated heatmaps through the proposed methods, concomitantly with the previous analyses, also contain important information. It is quite clear by observing the mean heatmaps that differences exist between normal and abnormal cases. Heatmaps corresponding to images without findings highlight more centered areas, whether heatmaps corresponding to pathological cases highlight more extensive areas, that stretch more towards the lungs. The exceptions are the heatmaps generated for CXR-P, due to the explained bias, with both normal and abnormal mean heatmaps highlighting the top of the lungs. For the latter dataset there is also a very high correlation between the mean heatmaps and the mean of the masks. For REFLACX, which presents a more realistic scenario, this overlap is not as obvious. To assess this correlation in a more quantitative way, mean intensities were computed on a case by case basis, within the masks, and inside and outside the bounding boxes. The average values clearly indicate that the heatmaps are brighter inside the masks, meaning that radiologists focus more on these areas. They also reveal the existing contrasts between the mask intensities and the intensities in the remainder of the bounding boxes, and between the inside and outside of the bounding boxes. However, the higher standard deviations for the mean intensities inside the masks indicate that some cases may be different. Altogether, the analysis of the generated heatmaps indicates that they successfully translate the information content present in the ETD, and that they highlight abnormal patches and also the thorax.

8.2 Model Performance

8.2.1 Heatmap Reconstruction

A noticeable difference exists between generated and reconstructed heatmaps regarding their brightness. While the generated heatmaps stretch through the entire [0,1] interval, the same does not happen for the reconstructed ones. Overall, the maximum intensity is smaller, which occurs both for the UDenseEfficientNet-TS and for the UDenseEfficientNet-DP, but also for the EGDmodel. This is probably the consequence of an attempt by the models to decrease the training loss, since outputting smaller values penalizes less the model when mistakes happen. To circumvent this issue, a contrast adjustment is included in the differential preservation approach, which results in a better CXR conservation in the heatmap highlighted regions. Furthermore, some isolated cases occur where anomalies are verified in the reconstructed heatmaps. To erase such artefacts, a gaussian kernel is applied before the thresholding or contrast adjustment operations. This also results in smoother edges and eliminates any noise that might be present.

Even though the maximum intensities of the reconstructed heatmaps are smaller, the same does not occur for the mean intensities inside the bounding boxes. Both for the REFLACX and CXR-P datasets, these mean intensities are slightly higher. Moreover, there is a quite significant contrast relative to the mean intensities outside the bounding boxes. Altogether, these results show that the reconstructed heatmaps are suitable to be applied with the goal of performing thorax segmentations. This is demonstrated by the high DSC values obtained. Nevertheless, as shown

in Chapter 7, cases occur where the segmentations do not cover the entire thorax, especially for EGD. Alongside dataset size, another important factor might be the number and diversity of the classes, which can affect the area highlighted by the reconstructed heatmaps.

Additionally, the REFLACX reconstructed heatmaps contain higher mean intensities within the anomaly masks, proving that the models are capable of more than just highlighting the thorax. Although these values are not as high as the ones for the original heatmaps, there is still a significant contrast relative to the bounding box values. Interestingly, the mean intensities within the masks for the processed heatmaps are higher than in the original heatmaps. The contrast relative to the bounding box also increases in the processed heatmaps. For the UDenseEfficientNet-DP, the contrast is slightly higher, which can be attributed to the second training stage using also the classification loss. In comparison to the EGDmodel results, the mean intensities within the masks are higher for the models developed in this work, and for the UDenseEfficientNet-DP the contrast is also higher. For the CXR-P test, on the other hand, no significant differences were observed between the mean intensities for the masks and for the bounding boxes, which happened for every model. The most probable cause is the low number of Pneumothorax examples in the REFLACX dataset, which hinders the capacity of the models to highlight such findings.

The reconstructed heatmaps were also averaged on a split and class-basis. The mean reconstructed heatmaps were similar to the mean original ones, with normal cases being more centered, and abnormal examples covering more of the lungs.

8.2.2 Pathology Classification

The two proposed approaches, despite being different in terms of the information provided to the classifiers, achieved similar results in the experiments performed. Therefore, it is not possible to determine which model is better. The UDenseEfficient-TS offers additional guarantees, in the sense that it restricts more the information passed to the classifier. On the other hand, the UDenseEfficient-DP has a greater potential to guide the classifier into focusing on pathological regions. Regardless of the model, the limiting effects of the dataset sizes/class distributions are quite clear. CXR pathology classification models in the literature are usually trained with hundreds of thousands of images. Therefore, it is not unexpected to observe that the AUC values are not very high, especially for the under-represented classes in REFLACX. In EGD, even though the class distribution is the same, the AUC values for the Pneumonia class are inferior. This is probably due to the fact that pneumonia is a harder disease to diagnose when compared to CHF, in which the signs are clear, something exacerbated by the low number of examples in the training set.

Concerning the comparison between the developed models and the state-of-the-art, it could be seen that the results are comparable. In the EGD dataset, the results are slightly inferior, which is also explained by the fact that the EGDmodel was developed and optimized in that dataset. For REFLACX, however, the UDenseEfficientNet-DP had a similar performance taking only into account the exclusive mean, something also verified for the EGD test using the REFLACX trained models. In this last case, the UDenseEfficientNet-TS had the best performance, 1 AUC point above its counterparts. Regarding the non-optimized versions of the models, a different trend is observed,

with the approaches developed in this work achieving betters results when the REFLACX trained models were tested on the same dataset or on CXR-P.

Interestingly, the results obtained for the models trained on REFLACX and tested on EGD were similar to the ones originally obtained in the EGD dataset, with the exception of the Pneumonia class (most likely due to differences in comparison to the REFLACX Consolidation class). For the test performed on CXR-P, the same thing was observed, with the Pneumothorax AUC values being similar to the ones registered for the REFLACX test sets. This reveals that the models did not overfit to the dataset used for training, supporting their applicability to other sources of data in which the same classes are present.

8.2.3 Explainability

The explainability of a model is of extreme importance, especially in models with healthcare applications. To guide the models into outputting more reliable explanations, different approaches were considered. The thorax segmentation approach aims at forcing the model to focus only on relevant parts of the CXR images, namely the lungs and heart. The differential preservation method, on the other hand, consists in conserving important areas of the CXR images, and dimming less relevant regions. To confirm the reliability of the models, explanations were generated using the GradCAM method, since it is one of the most commonly used. However, surprising results were obtained, especially for the UDenseEfficientNet-TS. The GradCAM method highlighted regions outside of the thorax masks, especially in the corners, in which no information existed. This creates doubts regarding the applicability of this method in the considered classifier, the EfficientNet-b0. The occurrence of highlighted corners has however been previously reported for EfficientNet models [7; 79]. This phenomenon can either be attributed to some characteristic of the model, or to the GradCAM method itself. Therefore, it is important to explore other explainability methods in order to get better insights regarding the developed models.

8.3 Limitations and Future Work

This work inevitably contains some limitations. Most of them derive from the datasets used, which present certain issues. The main concern relates to the representativeness of the data, which applies both in the standpoint of the data collection process and to the data itself. Since a small number of radiologists participates in each study (only one in EGD), the ETD may be biased towards the individual gaze patterns of specific radiologists. Moreover, since the chosen images contain a reduced number of possible findings, the developed models are unable to be applied in more complex scenarios. Even in the case where more classes are present, namely in the REFLACX dataset, a significant part includes only a limited number of examples. With an increase in complexity, DL models become more and more data-hungry. If the data quantity is insufficient, the performance of these models and their generalization capacity will be drastically affected. In the CXR-P dataset, there is also a clear bias from radiologists, which explains the similarities between mean normal and abnormal heatmaps. DL models are extremely apt at learning these

biases, since it constitutes an easy way to decrease the training loss. This explains the higher reconstruction losses verified in the CXR-P test, and why this dataset was not used for training throughout this work.

There is also a lack of complementary data, mainly for EGD and CXR-P, which hinders the heatmap generation process and/or the analyses performed. In EGD, the anomaly masks are not provided, which makes it impossible to assess the information content of both the original and reconstructed heatmaps in those areas. In CXR-P, the lack of complementary data was notorious, which demanded an adaptation in the heatmap generation process. Furthermore, the bounding boxes were not available, which required the use of a thorax detection model to extract them. An additional concern arised from the different ETD processing used in CXR-P. Most of the fixations contain a single data point, which drastically reduces the temporal resolution, since the time unit used is the number of joined data points per fixation.

Another important issue is the fact that, for EGD and CXR-P, the labels were not determined by the participating radiologists. This creates a possible scenario where the gaze data does not correspond to the pathology, or pathologies, labeled in the CXR image.

The abovementioned limitations in the datasets likely had a negative impact on the performance of the models, either regarding the information content in the reconstructed heatmaps, or the classification scores for the different experiments. Nevertheless, the developed models still presented adequate results, showing that the proposed approaches functioned as predicted. However, the link between the reconstructed heatmaps and the reasoning of the classifiers was not demonstrated, due to unexpected results from the explainability method used. This imposes a serious limitation in the developed work, since the goal was to use the heatmap reconstruction task to shift the attention of the subsequent models into providing more reliable explanations. It also reveals the fragility of common methods used to acquire such explanations, and how limited the provided evidence can be.

Regarding future work, several options can be explored. For instance, the development of new processing methods for the ETD, either in the spatial or temporal domain. The clustering technique present in [6], applied to CT ETD, could prove beneficial in the 2D space as well. Excluding fixations outside of the thorax could also have a positive impact, albeit small. Another possibility could be to exclude certain time periods of the ETD, such as the beginning or the end, intervals where the focus of the radiologist might not be directed to the detection of pathologies.

Another direction consists of further optimizing the proposed models, since for the EGDmodel this created a significant impact on the results. Parameters like the learning rate, batch size, dropout, and γ can be studied to a bigger extent. The inclusion of class weights to deal with class imbalance is also a possibility, especially for the models trained on the REFLACX dataset. Furthermore, the training process in DL is constantly evolving, providing different methods that could be applied here. Using a bigger EfficientNet could be an option as well, depending on the GPU limitations. Another future improvement could be related to the threshold selection for the UDenseEfficient-TS approach, through the inclusion of a more complex method.

Lastly, and more importantly, it is essential to provide evidence regarding the role of the reconstructed heatmaps in guiding the classifiers. The GradCAM method was unsuccessful, which demands the use of other alternatives. LIME [80] constitutes an interesting method, which also outputs visual information relative to important areas for a given prediction, but through a different approach. Moreover, explainability methods should be accompanied by the use of metrics to verify the reliability of the provided explanations [55]. Only then can solid conclusions be made concerning the quality of the explanations.

Chapter 9

Conclusion

The work here developed reveals how relevant ETD can be. Nevertheless, its practical implementation in DL models is not trivial, with many factors influencing the quality of the data, and consequently the performance of the models.

The proposed approaches have focused on using ETD to guide the classification process, in an attempt to provide reliable explanations regarding the underlying functioning of DL models. By including a component responsible for the reconstruction of heatmaps, these approaches also guarantee their independence of ETD during the inference stage, being less subjected to data constraints. The results obtained were similar to the state-of-the-art, without extensive optimization of the models, which shows potential. The thorax segmentations were satisfactory, and for the differential preservation strategy the reconstructed heatmaps contained pathology-related information. Even so, the explainability analysis of the models did not yield the expected results, and further experiments are still required to ascertain the quality of the rendered explanations.

The biggest limitation throughout this dissertation proved to be the dataset sizes. Current DL models train on enormous amounts of data in order to learn generalized attributes related to the task at hand, something which was not possible in this work. The scarcity of data had a negative impact on the models, both on the heatmap reconstruction and pathology classification tasks. More ETD is therefore needed to explore the full potential of the proposed approaches.

The greatest promise of ETD relates to the inexpensive acquisition of important data, while radiologists perform their daily routine of diagnosing CXR images. However, that promise is still far from being a reality. ETD collection suffers from several drawbacks, some related to the comfort and efficiency of radiologists, which hinders the widespread use of these technologies. As a matter of fact, all the datasets included in this work have collected their data from simulated scenarios, and not from real workplaces. This shows the limitations of current ET systems, which end up restricting the amount of collected data, hence the reduced number of DL applications. It is therefore essential to design new ETD collection methods, which do not disturb the normal workflow of radiologists. VR can introduce a change of paradigm in CXR pathology screening, allowing for a more practical gathering of ETD. If implemented, VR real-time collected data could be used to provide second opinions on the areas focused by the radiologists, or to detect unseen find-

ings. Another application could be the ETD-assisted annotation of CXR images, both regarding the pathology present and its location. Ultimately, successful DL models can only be developed if more ETD is available, which can only be achieved through the integration of practical ET systems in radiology departments.

Bibliography

[1] M. Berger, Q. Yang, and A. Maier, "X-ray imaging," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11111 LNCS, pp. 119–145, 8 2018.

[2] J. S. Klein and M. L. R. de Christenson, "A systematic approach to chest radiographic analysis," pp. 1–16, 2 2019.

[3] C. Ashley, "An overview on convolutional neural networks." `https://medium.com/swlh/an-overview-on-convolutional-neural-networks-ea48e76fb186`.

[4] Y. X. Tang, Y. B. Tang, M. Han, J. Xiao, and R. M. Summers, "Abnormal chest x-ray identification with generative adversarial one-class classifier," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, pp. 1358–1361, 4 2019.

[5] 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," 1 2018.

[6] N. Khosravan, H. Celik, B. Turkbey, E. C. Jones, B. Wood, and U. Bagci, "A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning," *Medical image analysis*, vol. 51, p. 101, 1 2019.

[7] A. Karargyris, S. Kashyap, I. Lourentzou, J. T. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, and M. Moradi, "Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development," *Scientific Data 2021 8:1*, vol. 8, pp. 1–18, 3 2021.

[8] A. E. 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 2019 6:1*, vol. 6, pp. 1–8, 12 2019.

[9] "Siim-acr pneumothorax segmentation | kaggle." `https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation`.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation,"

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 12 2015.

[12] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, 8 2016.

[13] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, 5 2019.

[14] T. Ahmed and N. H. N. Sabab, "Classification and understanding of cloud structures via satellite images with efficientunet," 9 2020.

[15] Q. Xie, M. T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 11 2019.

[16] M. Chetoui, "Gradient-weighted class activation mapping - gradcam." `https://medium.com/@mohamedchetoui/grad-cam-gradient-weighted-class-activation-mapping-ffd72742243a`.

[17] 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, pp. 16–32, 11 2021.

[18] *Forum of International Respiratory Societies. The Global Impact of Respiratory Disease - Second Edition*. Sheffield, European Respiratory Society, 2017.

[19] "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, pp. 1736–1788, 11 2018.

[20] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, pp. 394–424, 11 2018.

[21] D. R. Dance, S. Christofides, A. D. A. Maidment, I. D. Mclean, and K. H. Ng, "Diagnostic radiology physics: A handbook for teachers and students,"

[22] "Digital radiography. a comparison with modern conventional imaging," *Postgrad Med J*, vol. 82, pp. 425–428, 2006.

[23] E. Commission and D.-G. for Energy, *Medical radiation exposure of the European population*. Publications Office, 2015.

[24] "Heart anatomy | anatomy and physiology ii." `https://courses.lumenlearning.com/suny-ap2/chapter/heart-anatomy/`.

[25] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, pp. 1347–1358, 4 2019.

[26] I. Castiglioni, L. Rundo, M. Codari, G. D. Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N. C. D'Amico, and F. Sardanelli, "Ai applications to medical images: From machine learning to deep learning," *Physica Medica*, vol. 83, pp. 9–24, 3 2021.

[27] R. M. Hopstaken, T. Witbraad, J. M. van Engelshoven, and G. J. Dinant, "Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections," *Clinical Radiology*, vol. 59, pp. 743–752, 8 2004.

[28] B. Moifo, E. W. Pefura-Yone, G. Nguefack-Tsague, M. L. Gharingam, J. R. M. Tapouh, A.-P. Kengne, S. N. Amvene, B. Moifo, E. W. Pefura-Yone, G. Nguefack-Tsague, M. L. Gharingam, J. R. M. Tapouh, A.-P. Kengne, and S. N. Amvene, "Inter-observer variability in the detection and interpretation of chest x-ray anomalies in adults in an endemic tuberculosis area," *Open Journal of Medical Imaging*, vol. 5, pp. 143–149, 8 2015.

[29] S. C. B. Lo, S. L. A. Lou, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE Transactions on Medical Imaging*, vol. 14, pp. 711–718, 1995.

[30] A. Krizhevsky, I. S. A. in neural . . . , and undefined 2012, "Imagenet classification with deep convolutional neural networks," *proceedings.neurips.cc*.

[31] B. van Ginneken, "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning," *Radiological Physics and Technology*, vol. 10, p. 23, 3 2017.

[32] J. K. Gohagan, P. C. Prorok, R. B. Hayes, and B. S. Kramer, "The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: History, organization, and status," *Controlled Clinical Trials*, vol. 21, pp. 251S–272S, 12 2000.

[33] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, pp. 304–310, 3 2015.

[34] 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," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 3462–3471, 11 2017.

[35] "Nih chest x-rays | kaggle." `https://www.kaggle.com/datasets/nih-chest-xrays/data`.

[36] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, 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," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 7 2019.

[37] A. Bustos, A. Pertusa, J. M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101797, 12 2020.

[38] 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," 12 2020.

[39] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert,"

[40] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," *Academic Radiology*, vol. 27, pp. 106–112, 1 2020.

[41] 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 2020 3:1*, vol. 3, pp. 1–8, 5 2020.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 5 2017.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 9 2014.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1–9, 9 2014.

[45] Y. Mao, F. F. Xue, R. Wang, J. Zhang, W. S. Zheng, and H. Liu, "Abnormality detection in chest x-ray images using uncertainty prediction autoencoders," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12266 LNCS, pp. 529–538, 2020.

[46] P. N. Kieu, H. S. Tran, T. H. Le, T. Le, and T. T. Nguyen, "Applying multi-cnns model for detecting abnormal problem on chest x-ray images," *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018*, pp. 300–305, 12 2018.

[47] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," *IEEE Access*, vol. 9, pp. 118571–118583, 6 2020.

[48] 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," 11 2017.

[49] K. Urinbayev, Y. Orazbek, Y. Nurambek, A. Mirzakhmetov, and H. A. Varol, "End-to-end deep diagnosis of x-ray images," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, pp. 2182–2185, 3 2020.

[50] 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," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11401 LNCS, pp. 757–765, 2019.

[51] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L. J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8290–8299, 12 2018.

[52] C. Zhang, F. Chen, and Y.-Y. Chen, "Thoracic disease identification and localization using distance learning and region verification," 6 2020.

[53] 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.

[54] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu, "Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 10631–10640, 10 2019.

[55] G. Ras, N. Xie, M. V. Gerven, and D. Doran, "Explainable deep learning:a field guide for the uninitiated explainable deep learning: A field guide for the uninitiated,"

[56] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 7 2022.

[57] C. C. Wu and J. M. Wolfe, "Eye movements in medical image perception: A selective review of past, present and future," *Vision*, vol. 3, 6 2019.

[58] S. Mall, P. C. Brennan, and C. Mello-Thoms, "Modeling visual search behavior of breast radiologists using a deep convolution neural network," *Journal of Medical Imaging*, vol. 5, p. 1, 8 2018.

[59] J. N. Stember, H. Celik, E. Krupinski, P. D. Chang, S. Mutasa, B. J. Wood, A. Lignelli, G. Moonis, L. H. Schwartz, S. Jambawalikar, and U. Bagci, "Eye tracking for deep learning segmentation using convolutional neural networks," *Journal of Digital Imaging*, vol. 32, p. 597, 8 2019.

[60] G. Aresta, C. Ferreira, J. Pedrosa, T. Araujo, J. Rebelo, E. Negrao, M. Morgado, F. Alves, A. Cunha, I. Ramos, and A. Campilho, "Automatic lung nodule detection combined with gaze information improves radiologists' screening performance," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 2894–2901, 10 2020.

[61] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement,"

[62] K. Saab, S. M. Hooper, N. S. Sohoni, J. Parmar, B. Pogatchnik, S. Wu, J. A. Dunnmon, H. R. Zhang, D. Rubin, and C. Ré, "Observational supervision for medical image classification using gaze data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12902 LNCS, pp. 603–614, 2021.

[63] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 551–561, 1 2016.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, 6 2017.

[65] R. B. Lanfredi, M. Zhang, W. F. Auffermann, J. Chan, P.-A. T. Duong, V. Srikumar, T. Drew, J. D. Schroeder, and T. Tasdizen, "Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays,"

[66] C. Moreira, I. B. Nobre, S. C. Sousa, G. Lusíadas, P. J. ao Madeiras Pereira, and J. Jorge, "Improving x-ray diagnostics through eye-tracking and xr,"

[67] "Gp3 eye tracking device - affordable, research-grade equipment | gazepoint." https://www.gazept.com/product/gazepoint-gp3-eye-tracker/.

[68] "Eyelink 1000 plus - the most flexible eye tracker - sr research." https://www.sr-research.com/eyelink-1000-plus/.

[69] "Easy to use, small, portal eye tracker - tobii pro nano." https://www.tobiipro.com/product-listing/nano/.

[70] P. Yakubovskiy, "Segmentation models pytorch." https://github.com/qubvel/segmentation_models.pytorch, 2020.

[71] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 1 2018.

[72] "Imagenet." https://www.image-net.org/.

[73] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[74] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth,"

[75] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 618–626, 12 2017.

[76] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. I. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *AJR. American journal of roentgenology*, vol. 174, pp. 71–74, 2000.

[77] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, p. 475, 12 2014.

[78] J. Gildenblat and contributors, "Pytorch library for cam methods." https://github.com/jacobgil/pytorch-grad-cam, 2021.

[79] "Efficientnet gradcam comparison to other models | kaggle." https://www.kaggle.com/code/meaninglesslives/efficientnet-gradcam-comparison-to-other-models/notebook.

[80] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pp. 97–101, 2 2016.

Appendix A

Supplementary Figures

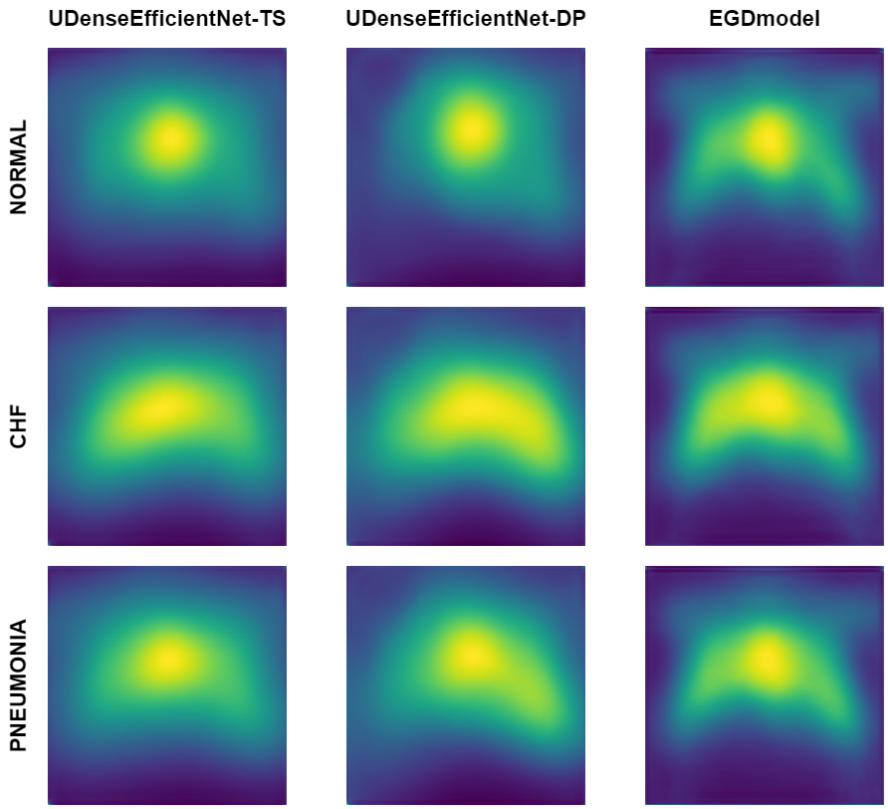


Figure A.1: Examples of mean reconstructed heatmaps for one of the EGD test sets.

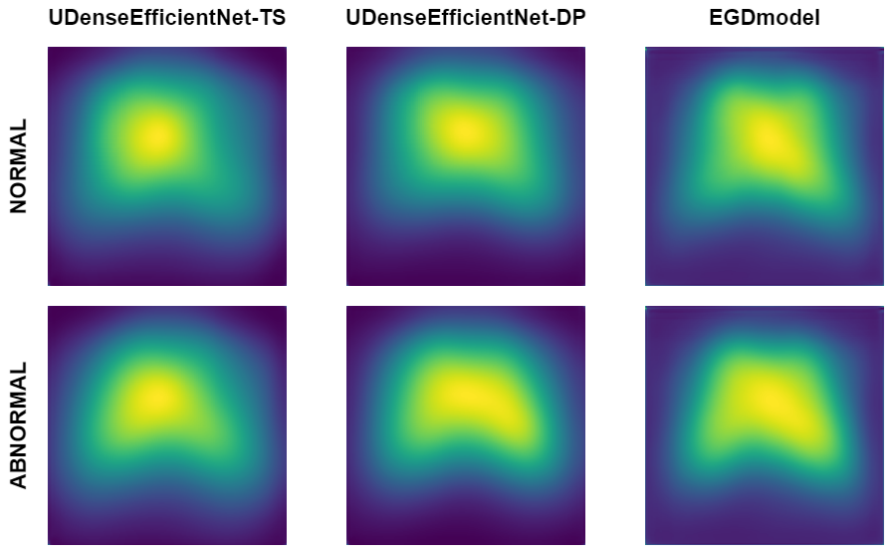


Figure A.2: Examples of mean reconstructed heatmaps for one of the REFLACX test sets.

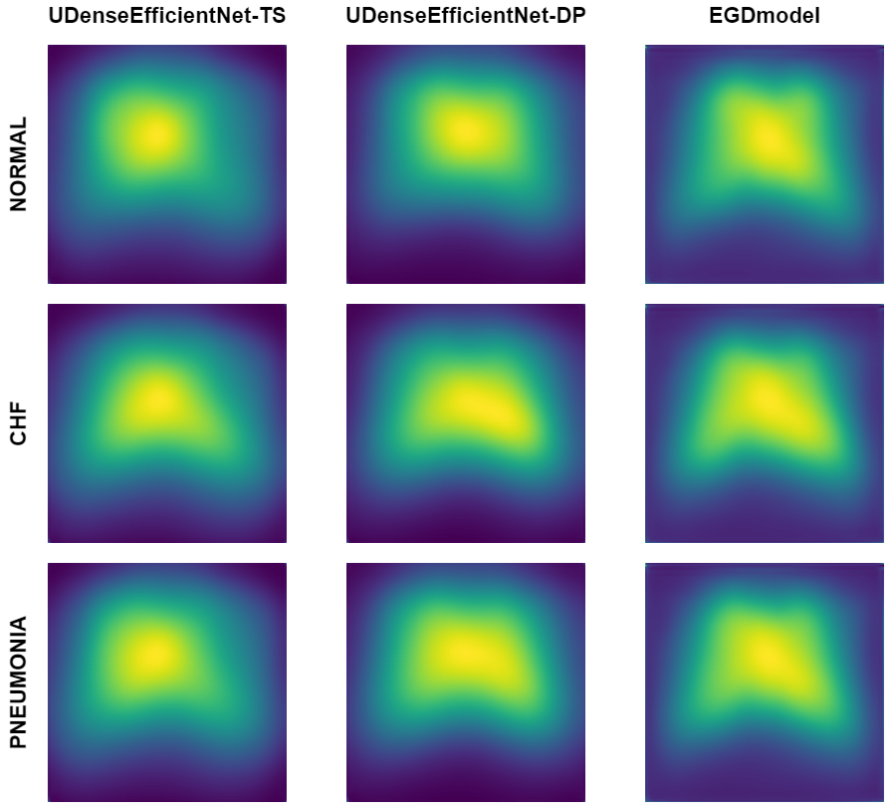


Figure A.3: Examples of mean reconstructed heatmaps for models trained on REFLACX and tested on EGD.

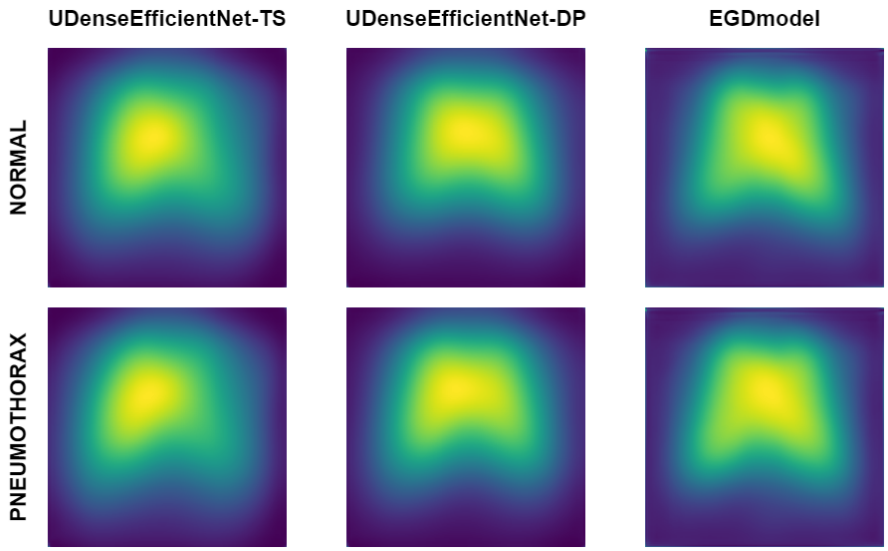


Figure A.4: Examples of mean reconstructed heatmaps for models trained on REFLACX and tested on CXR-P.

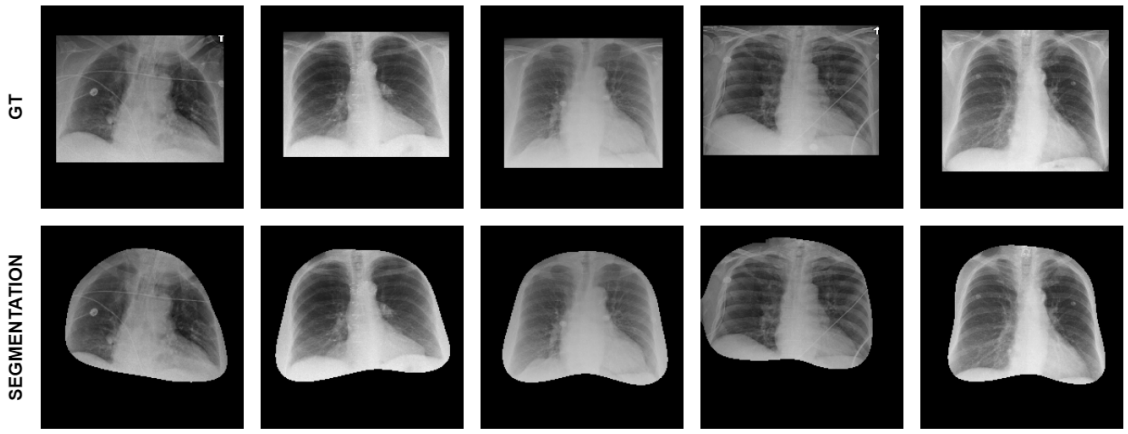


Figure A.5: Examples of segmentation masks and corresponding GTs for one of the REFLACX test sets.

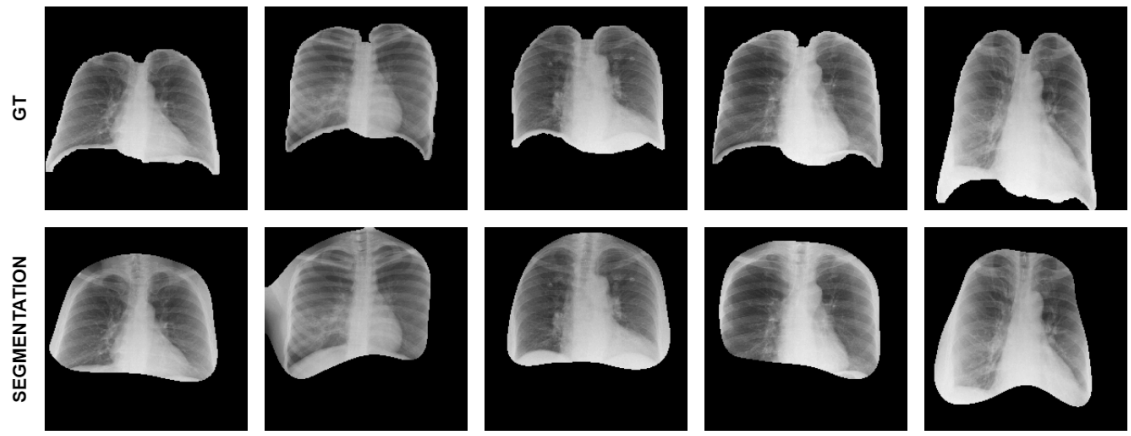


Figure A.6: Examples of segmentation masks and corresponding GTs for the UDenseEfficientNet-TS model trained on REFLACX and tested on EGD.

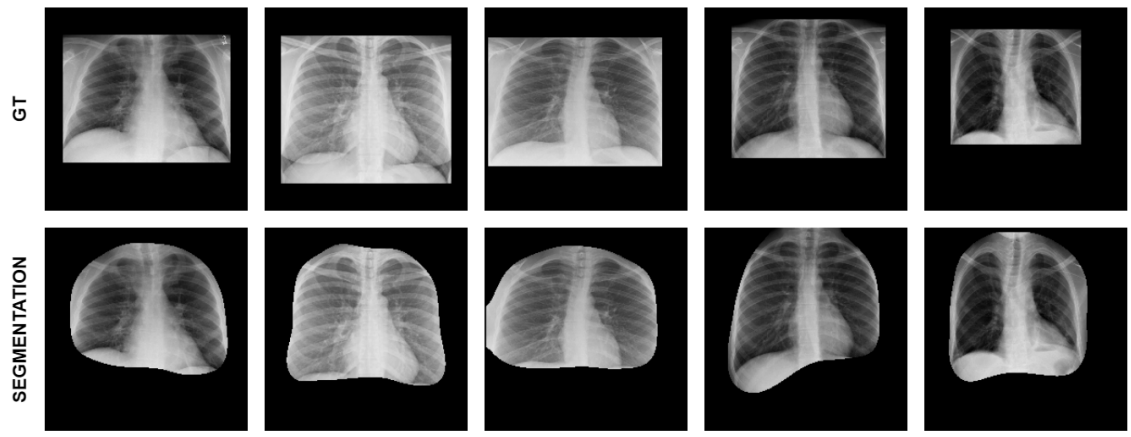


Figure A.7: Examples of segmentation masks and corresponding GTs for the UDenseEfficientNet-TS model trained on REFLACX and tested on CXR-P.