FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO


# Leveraging Longitudinal Data in Chest Radiography Pathology Detection

**Raquel Morais Belo**

Master in Bioengineering

Supervisor: João Manuel Pedrosa, PhD

Co-Supervisor: Joana Neves Rocha, MSc

July 4, 2023

# Leveraging Longitudinal Data in Chest Radiography Pathology Detection

**Raquel Morais Belo**

Master in Bioengineering

July 4, 2023

# Resumo

Raio-X torácico é um exame de imagem médica comum, usado para analisar a região torácica. Através de imagens de raio-X é possível distinguir diversas estruturas, permitindo a deteção de anormalidades. O uso extensivo de raio-X torácico, juntamente com o desenvolvimento tecnológico, levou à crescente necessidade de métodos automáticos para análise de imagem e relatórios médicos. Múltiplos métodos têm sido desenvolvidos, com diferentes objetivos, nomeadamente a deteção de anormalidades numa imagem e correspondente localização. Quando um raio-X torácico é analisado por um profissional, este é normalmente comparado por imagens adquiridas previamente. Este processo permite ter uma referência longitudinal, levando a um diagnóstico mais preciso. A análise automática de raio-X pode beneficiar da utilização de dados longitudinais, uma vez que estes podem levar à inclusão de informação relevante para a decisão efetuada, contudo, esta é uma área pouco estudada. Neste trabalho, a aplicação de informação longitudinal foi estudada, para deteção de anormalidades e deteção de mudança em pares de raio-X torácico.

Inicialmente, um método para alinhamento de raio-X torácico foi construído, com o objetivo de alinhar duas imagens. Sistemas automáticos tendem a beneficiar de um processo de alinhamento quando múltiplas imagens são usadas, uma vez que permitem efetuar uma melhor correspondência entre as características das imagens. O método de alinhamento desenvolvido utiliza segmentações de pulmões para alinhar uma imagem de acordo com uma referência, calculando parâmetros de rotação, translação e escalamento para aplicar transformações rígidas. Esta técnica permitiu alinhar pares consecutivos de imagens, cujas segmentações atingiram um DSC médio de $0.895 \pm 0.080$.

Múltiplas experiências foram efetuadas relativamente à deteção de uma patologia e deteção de mudança num par de imagens de um mesmo paciente. Relativamente aos algoritmos de classificação, vários modelos foram usados, nos quais informação longitudinal foi incluída a diferentes níveis. Um modelo treinado sem dados longitudinais foi usado como base para comparação. Nas restantes experiências, a inclusão de informação longitudinal foi feita ao nível das características e ao nível das imagens de entrada no modelo. Nesta última, o processo foi realizado com pares não alinhados e repetido com pares alinhados através do método desenvolvido. Mapas de explicabilidade foram gerados para estas configurações experimentais. Nas experiências iniciais foram usados pares de imagens consecutivas. A utilização de pares alinhados revelou um melhoramento das métricas finais, em comparação com pares não alinhados, quer para a deteção de uma patologia, quer para a deteção de mudança. O modelo que utiliza as características das imagens concatenadas superou o desempenho dos restantes na deteção de mudança, com uma AUC de 0.858, e apresenta uma AUC de 0.897 para a deteção de patologia, mostrando que as características associadas a patologia podem ser usadas na previsão de comparação entre as imagens.

De modo a melhorar os resultados dos métodos desenvolvidos, técnicas de aumento de dados foram estudadas. Estas técnicas provaram que aumentar a representação de classes minoritárias aumenta o ruído no conjunto de dados, levando, consequentemente, a resultados piores. O aumento do número de amostras de treino, mantendo a proporção de cada classe, mostrou ser uma

técnica de aumento de dados vantajosa em estudos longitudinais.

A possibilidade da existência de anotações incorretas no conjunto de dados levou à realização de outra experiência, em que um novo conjunto de dados retificado foi gerado, alterando as anotações do conjunto de dados original usando informações longitudinais. A alteração foi feita com o objetivo de eliminar os casos em que um mesmo paciente apresenta múltiplas alterações na presença da patologia num curto espaço de tempo. A retificação do conjunto de dados aumentou a sua consistência e os resultados mostram que as características aprendidas usando o conjunto de dados original levam a um alto desempenho no conjunto de dados retificado, provando que a retificação facilitou a tarefa de teste.

Concluindo, independentemente da escassez de informação temporal e comparativa nos conjuntos de dados de raio-X torácico mais comuns, o uso de imagens longitudinais provou fornecer informações relevantes que permitem a previsão de uma patologia e mudança entre duas imagens num par. O uso de características de patologia mostrou resultados promissores na previsão da mudança entre duas imagens, sem afetar a deteção de patologia. O uso de imagens alinhadas comprovou a importância dos métodos de alinhamento quando múltiplas imagens são usadas para previsão. A utilização de todos os dados, negligenciando a sua ordem temporal, pode ser usada como uma técnica de aumento de dados considerável, cujo estudo deve ser expandido. Da mesma forma, mais estudos devem ser realizados relativamente à retificação longitudinal. Os dados longitudinais podem ser usados como uma ferramenta poderosa para retificar um conjunto de dados, uma vez que fornecem informações temporais que podem ser usadas para verificar se uma anotação é coerente com os aspetos fisiológicos da anormalidade.

**Palavras-Chave:** Aprendizagem Profunda, Raio-X Torácico, Dados Longitudinais

# Abstract

Chest radiography is a common medical imaging exam that is used to analyze the thoracic area. Through X-ray images, it is possible to distinguish different structures, which allows the detection of abnormalities. The extensive use of chest radiography, along with the development of technology, led to an increasing need for automated methods for image and report analysis. Multiple methods have been developed, with different objectives, namely the detection of abnormalities in a scan, as well as their localization. When a chest scan is being analyzed by a medical professional, it is usually compared with previous scans, acquired at different time points. This is done in order to have a longitudinal reference, and provide a more accurate diagnosis. The automated analysis of scans might benefit from using longitudinal data, as it might provide relevant information for the presented decision, however, this field that has not been much studied. In this work, the application of longitudinal information for detection of abnormality and detection of change in pairs of CXR images was studied.

Initially, an alignment method was constructed, with the goal of aligning two images. When using multiple images, automated systems often benefit from such process, as it allows a better matching of the image features. The developed alignment method uses lung segmentation features to align one image according to a reference, computing rotation, translation, and scaling parameters for rigid transformation. This technique allowed the generation of aligned consecutive images pairs, whose segmentations reach an average Dice Similarity Coefficient (DSC) score of $0.895 \pm 0.080$.

Multiple experiments were performed regarding the detection of a pathology and the detection of change in an image pair, from the same patient. As for classification algorithms, various models were used, each integrating longitudinal information at a different level. A model trained without longitudinal information was used as a baseline. In the remaining experiments, the inclusion of this information was done at the features level and at the input level. The latter was done with the original images and also with images aligned with the developed method. Explainability maps were generated for all these experiments. Initial experiments underwent using consecutive pairs of images. The usage of aligned images revealed to improve the final metrics, in comparison with non-aligned pairs, for both the detection of a pathology and detection of change. The model that uses the concatenated image features outperformed the remaining in the detection of change, with an Area Under the Receiver Operating Characteristics Curve (AUC) of 0.858, and presenting an AUC of 0.897 for the detection of pathology, showing that pathology features can be used to predict comparison between images.

In order to further improve the developed methods, data augmentation techniques were studied. These techniques proved that increasing the representation of minority classes leads to higher noise in the dataset and consequently worse results. It also showed that increasing the number of training samples while maintaining the ratio of each class can be an advantageous augmentation technique in longitudinal studies.

The possibility of the existence of incorrect labels in the dataset let to another experiment,

where a new rectified dataset was generated, by altering the original dataset labels using longitudinal information. The alteration was done with the aim of eliminating cases where the same patient showed multiple changes in pathology presence in a short time span. The rectification of the dataset increased its consistency and the results show that the features learned from the original dataset have high performance in the rectified dataset, proving that the rectification facilitated the testing task.

In conclusion, regardless of the lack of temporal and comparative information in the most common CXR datasets, the usage of longitudinal scans proved to provide insightful information that allows the prediction of a pathology and change between two images in a pair. The usage of pathology features showed promising results for predicting a comparison label between two images, without affecting the detection of pathology. The usage of aligned CXR scans proved the importance of registration methods when multiple scans are used for prediction. Using all information, neglecting its temporal order, can be used as a considerable augmentation technique, which should be further explored in change studies. Similarly, further studies should be performed on longitudinal rectification. Longitudinal information can be used as a powerful tool to rectify a dataset, as it provides temporal information that can be used to verify if an annotation is coherent with the physiological aspects of the abnormality.

**Keywords:** Deep Learning, Longitudinal Radiography, Chest Radiography

# Agradecimentos

Estou imensamente grata por tudo o que vivi nos últimos cinco anos. Nos momentos bons e nos menos bons, tive ao meu lado várias pessoas com quem pude partilhar as minhas experiências, conhecimentos, e, acima de tudo, crescimento. Começo por agradecer ao meu orientador, João Pedrosa, por me guiar neste desafio e por toda a sua disponibilidade. Agradeço também à minha co-orientadora, Joana Rocha, por todas as vezes que me ajudou e impulsionou, e por ter sempre acreditado em mim.

Agradeço aos meus colegas, que estiveram presentes no meu percurso académico desde cedo. Obrigada ao Gonçalo Ferreira e Marco Teixeira por me acompanharem em todas as vivências que a Engenharia Biomédica nos proporcionou. Obrigada José Pedro Araújo e Rodrigo De Marco pelas aventuras vividas longe de casa. Obrigada Mafalda Cortez pela paciência e palavras simpáticas. Obrigada ainda Matilde Costa pelo companheirismo e presença.

Não posso deixar de agradecer à Andreia Gouveia, por todas as memórias partilhadas, e por ser sempre a melhor conselheira e companhia. Obrigada à Inês Martins por todos os momentos, mais ou menos atribulados, e por estar sempre presente. Obrigada Maria Beatriz Calçada por todo o apoio e todos os momentos felizes. Obrigada Mariana Pereira pelos novos desafios e animação. Obrigada ainda Maria Leonor de Aguiar, Ana Margarida Ferreira, Ana Francisca Vieira, Catarina Fernandes, Jéssica Nascimento, Adelaide Santos, Ângela Coelho, Inês Calmeiro, Inês Rodrigues e Ana Rita Marques por todo o apoio durante este percurso.

Quero também deixar um agradecimento especial para o Pedro Serrano, por estar sempre presente e ser um grande impulsionador do meu sucesso. Obrigada por todo o crescimento e por todos os momentos, pois sem eles nada seria o mesmo. Por fim, agradeço aos meus pais, e à minha família em geral, pelo apoio incondicional que me deram desde sempre, e que tornaram esta etapa da minha vida possível.

Raquel Belo

*"Those who cannot change their minds cannot change anything."*

George Bernard Shaw

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| AP | Anteroposterior |
| AUC | Area Under the Receiver Operating Characteristics Curve |
| BB | Bounding Box |
| BCE | Binary Cross Entropy |
| CAM | Class Activation Mapping |
| CNN | Convolutional Neural Network |
| CT | Computed Tomography |
| CXR | Chest X-Ray |
| DL | Deep Learning |
| DSC | Dice Similarity Coefficient |
| DenseNet | Densely Connected Convolutional Networks |
| DoG | Difference of Gaussian |
| GAN | Generative Adversarial Network |
| GAT | Graph Attention Network |
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| ICP | Iterative Closest Point |
| ICU | Intensive Care Unit |
| IMV | Intermittent Mandatory Ventilation |
| L-SVM | Linear Support Vector Machines |
| LL | latero-lateral |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| MSE | Mean Squared Error |
| NLP | Natural Language Processing |
| PA | posteroanterior |
| PCA | Principal Component Analysis |
| PD | Pixel Distance |
| r | Pearson correlation coefficient |
| RICORD | RSNA International COVID-19 Open Annotated Radiology Database |
| RSNA | Radiological Society of North America |
| ReLU | Rectified Linear Activation Unit |
| ResNet | Residual Neural Network |
| SENet | Squeeze and Excitation Network |
| SIFT | Scale-Invariant Feature Transform |
| VGG | Visual Geometry Group |
| XAI | Explainable Artificial Intelligence |
| YOLO | You Only Look Once |

# Chapter 1

# Introduction

## 1.1 Context

X-rays are a type of ionizing radiation that has been used for medical imaging ever since their discovery. Multiple imaging techniques have been developed based on X-rays, like plain radiography, Computed Tomography (CT) and fluoroscopy. Chest radiography falls into the category of plain radiography. Of all X-ray exams performed, 30-40% are Chest X-Ray (CXR) scans, regardless of the level of health-care delivery. This is because CXR are associated with fast acquisition times, with low costs and low radiation exposure [6].

CXR images allow the visualization of structures in the chest area, like the lungs and the heart, which are the main targets for abnormality detection when analyzing a scan. The analysis of such images is a laborious task, as there is a high volume of exams, displaying complex structures that may overlapped. This factor led to the development of automated methods for CXR analysis, which intend to facilitate the job of radiologists and other medical professionals by, for instance, detecting and localizing abnormalities, or automatically generating reports. With the growth of such methods came the creation of multiple CXR datasets, which can be used to train automatic abnormality detection systems.

## 1.2 Motivation and Objectives

Most of the automated methods for CXR analysis use one image to produce a desired output. However, when the analysis is performed by human professionals, it is normally done by comparing multiple scans from the same patient, allowing the visualization of the evolution in the scans. It is important to look at images acquired at different time points simultaneously, so that a diagnosis can be done. The study of automated methods that utilize longitudinal information to produce an output is, consequently, a field of high importance, as it allows a more realistic automation of the diagnosis process. However, The study of longitudinal information for automated analysis of CXR is still a developing area. With the rise of the COVID-19 pandemic, more studies that use sequential scans from the same patient arose. However, there is still much to uncover in this field,

as the most common CXR datasets do not contain information specifically for longitudinal comparison, focusing on the abnormalities or findings in each individual image, and providing only the acquisition date or the age of the patient at the acquisition time, as for temporal information.

Hence, the prediction of not only the presence of an abnormality but also the comparison with previous scans from the same patient is a relevant matter. The comparison can be done by predicting if the abnormality remains or if it is no longer present, or by predicting whether the pathological situation improved or worsened. The usage of longitudinal data may improve the performance of detection algorithms, by providing additional information. It may also increase the robustness and transparency of these methods, by providing a prediction with the reference of previous scans. The automated analysis of more than one image may also benefit from image registration, which allows the alignment of anatomical structures, in the case of CXR. Thus, image registration is a topic to be kept in mind in this field of study.

The objectives of this work are:

- The development of a method that uses anatomical lung features for CXR alignment.
- The development and experimentation with different methods for predicting the presence of an abnormality and change in consecutive scans.
- The exploration of data augmentation techniques to improve the developed methods, and the exploration of the usage of longitudinal data to rectify CXR datasets.

## 1.3 Structure

The remainder of this thesis is divided in the following chapters:

- **Chapter 2** provides an introduction to radiography and, more specifically, chest radiography. A description of state-of-the-art methods for automated analysis of CXR is also presented.
- **Chapter 3** focuses on longitudinal analysis of CXR, including the longitudinal analysis of datasets and the description of automated methods that use over time information in CXR.
- **Chapter 4** describes the developed CXR alignment method and its comparison to a state-of-the-art solution.
- **Chapter 5** presents the created experimental settings for detection of a pathology and change in a pair of sequential scans.
- **Chapter 6** focuses on data augmentation techniques to improve the detection of pathology and change methods developed in Chapter 5.
- **Chapter 7** describes a technique that can be used for dataset rectification using longitudinal data, as well as a replication of the previous results on the rectified dataset.
- **Chapter 8** aggregates the conclusions of this work.

# Chapter 2

# Radiography Acquisition and Analysis

## 2.1 Radiography

### 2.1.1 Fundamental Physical Concepts

When looking at the electromagnetic spectrum, different kinds of electromagnetic radiation can be seen. This radiation is composed by electric and magnetic waves that travel through space and time, and it can be categorized into ionizing or non-ionizing radiation. The spectrum is composed of different kinds of radiations, distinguished by their frequency range.

X-rays consist of ionizing electromagnetic radiation, with wavelength ranging from 0.01 to 10 nanometers. X-rays were discovered by Wilhelm Conrad Röntgen in Germany, in 1895. When experimenting with Crookes tubes, he realized that a fluorescent screen could pick up on radiation that was passing through an object between the tube and the screen [7], which initialized his study on this kind of radiation. In 1896, the medical usefulness of this radiation was demonstrated for the first time, which granted Röntgen the Nobel Prize in Physics in 1901. Crookes tubes are composed of two electrodes. When a high voltage is applied between them, the air in the tube is ionized, due to the acceleration of the electrons in the gas. These electrons hit the anode or the glass in the tube. The acceleration (or deceleration) of loaded particles results in electromagnetic waves, in this case, X-rays.

Nowadays, this radiation is generated using X-ray tubes, powered by generators. The generator provides the source of electrical voltage to energize the tube. In opposition to the Crookes tubes, in the X-ray tubes the electron beam emitter is a cathode filament. This filament is connected to a circuit that heats it up (as it works as an electrical resistance). When the filament is heated, it releases electrons by thermionic emission, which accumulate at the filament's surface. As a high voltage is applied to the anode and cathode, the electrodes are accelerated to the anode. The interaction between the anode (target) and the accelerated electrons leads to the conversion of its kinetic energy into electromagnetic radiation (X-rays) with equivalent energy. This radiation is called *bremsstrahlung*, or "braking radiation". A scheme is represented in Figure 2.1. Closer interactions with the anode's atoms' nucleus, lead to a greater deceleration and thus a greater radiation energy. The result of this effect is a continuous spectrum of many X-ray energies, which

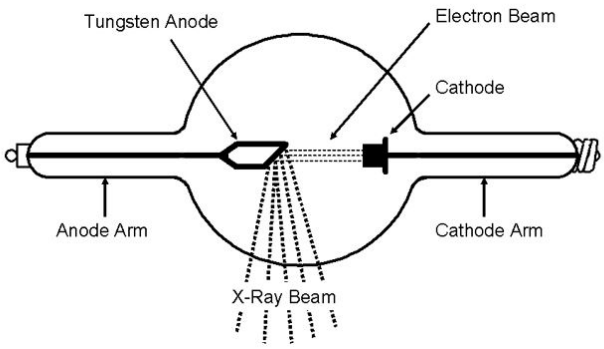



Figure 2.1: Scheme of hot cathode X-ray tube [1].

depend on the applied voltage. The average radiation energy in a typical X-ray spectrum is around one-third to one-half of the peak energy, depending on the beam filtration [8].

The number of released electrons by the filament can be increased by increasing the current on the circuit. The tube current is the number of electrons moving between the cathode and the anode and is expressed in milliampere, where 1 mA $= 6.24 \times 10^{15}$ electrons/s. The applied voltage typically ranges from 50 to 150 kV. Targets used in X-ray tubes are usually made of tungsten.

### 2.1.2 Radiography in Medical Imaging

X-rays are currently used in the medical field in a wide variety of applications, for both diagnostic and therapy. This radiation is used similarly in all imaging techniques: the X-ray beam is produced and directed to the patient, then, X-ray-sensitive plates or X-ray films are used to collect the radiation and produce the image. X-ray films consist of emulsions of silver halite crystals (commonly silver bromide or silver chloride), which are sensitive to X-rays. When exposed to light, some bromide ions are liberated and captured by the silver ions. When exposed to the developer (chemical solution), a reaction occurs, forming metallic silver. The silver is what generates the image.

With the development of technology came the digitalization of X-rays as an imaging technique. This was advantageous since digital radiography allowed a reduction of the radiation the patient was exposed to, while providing a high quality image, that can be easily processed, in opposition to a traditional X-ray image. In digital radiography, X-ray-sensitive plates are used to collect the radiation, after going through the patient's body. These detectors contain a combination of amorphous silicon detectors with cesium or gadolinium scintillators that convert X-rays to light, which is converted into a digital image by thin film transistors [9].

The exposure of the body tissues to X-ray is not risk-free since, as previously mentioned, it is ionizing radiation. This radiation has enough energy to damage DNA, which can lead to cancer. The risks of X-ray exposure are highly dependent on the radiation dose, as well as the patient's age, gender, and the body region that is exposed, since some body parts are more sensitive to X-rays than others. The exposure to X-rays should always be minimized, happening only

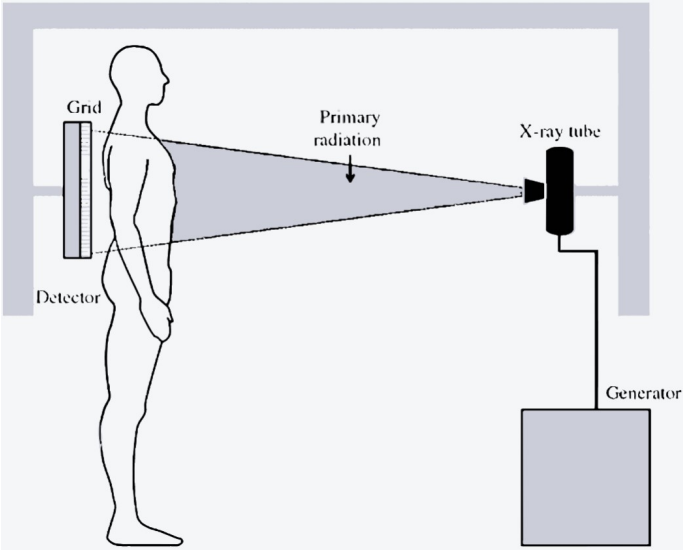



Figure 2.2: Plain radiography scheme [2].

when necessary, and applying the "As Low as Reasonably Achievable" (ALARA) principle, when choosing equipment settings to minimize exposure to the patient. The balance between the risk and the advantages should always be kept in mind. If the benefits of the radiation exposure overcome the risks, then the exam must be performed.

There are different applications of X-rays in medical image. In the following paragraphs, some of these applications are described.

#### 2.1.2.1 Plain Radiography

A plain radiography equipment can acquire X-ray images either vertically or horizontally, depending on the position of the X-ray emitter and the receptor. Grids for scatter radiation (collimators) are located immediately in front of the detectors, and protect them from scattered X-rays, which improves the quality of the final image (since only primary radiation is used to produce it). A scheme of a vertical radiography equipment is shown in Figure 2.2. The radiation goes through the body, and different tissues absorb it differently. It can easily go through fat and soft tissues, leading to a dark appearance in the final image. On the other hand, structures like bones that contain high levels of calcium, absorb the radiation, leading to light regions in the image. It is a useful technique to evaluate joints and bones, as well as for detecting pathologies in the lung area. Thus, chest X-ray is one of the imaging techniques included in plain radiography.

Mammography is a type of plain radiography, especially used to scan the breast. Just like a regular radiography equipment, a mammography unit contains a radiation source and a receptor, however, it also includes a compression paddle, which compresses the breast, making it less dense and widening the surface, which results in a better image. Mammography exams are done for breast cancer detection and diagnosis, as tumors tend to appear as masses that have higher cell density than the regular breast tissue.

#### 2.1.2.2 Computed tomography

CT provides many advantages in comparison with plain radiography. The most remarkable difference is the 3D imaging of the body structures, which overcomes the superimposition of different organs. It also has greater contrast, allowing for a general better visualization of the structures, at different angles.

In this technique, multiple projections of the same object are collected, obtaining its internal structure. Many of these measurements are acquired at different points during the translation motion of the tube and detector. A set of measured rays is designated a view, which is collected at many incremental angles in order to obtain all the information at the current translation point. The incremental translation creates different slices, with thickness corresponding to the thickness of the narrow beam. Each slice is a cross-sectional image of the structure, and the combination of all slices creates a three-dimensional X-ray image. Images are usually acquired in the axial plane, but sagittal and coronal images can be reconstructed.

There are variations of CT equipments, which provide different advantages depending on the situation where applied. Some of these variations include spiral or helical CT (allows a faster scanning time with less dosage, as well as few motion artifacts), multislice or multidetector CT (characterized by the acquisition of multiple slides simultaneously), dual source CT (which generates sharper images with less dosage and smaller acquisition time) and dual energy CT (permits the visualization of the same slice at different energies, which can be used to spot different materials and reduce artifacts).

#### 2.1.2.3 Fluoroscopy

Fluoroscopy is a type of medical imaging that consists in the creation of continuous X-ray images, forming a video in real time, where the movement of the body structures can be observed. It can be used for diagnosis, by following the path of a contrast agent inside the body. Following swallowed barium (esophagogram) and observing the blood with contrast agent flow through arteries and veins (angiograms) are examples. It can also be used for procedure guidance, for instance, for the placement of stents and catheter insertion and manipulation with a radiographic contrast agent. Even though this technique provides high advantages, it also exposes the patient to very high radiation doses in comparison with the previously mentioned techniques, thus, it should be used only in particular situations.

## 2.2 Chest Radiography

### 2.2.1 Definition and Advantages

Just a few years after the discovery of the X-ray, numerous medical applications emerged. The thoracic area was used as a target for radiography from early on, allowing the diagnosis of various chest diseases, like tuberculosis, pneumonia, and pneumothorax [10]. Despite the existence of equipment that allows the 3-dimensional visualization of the structures, chest radiography is

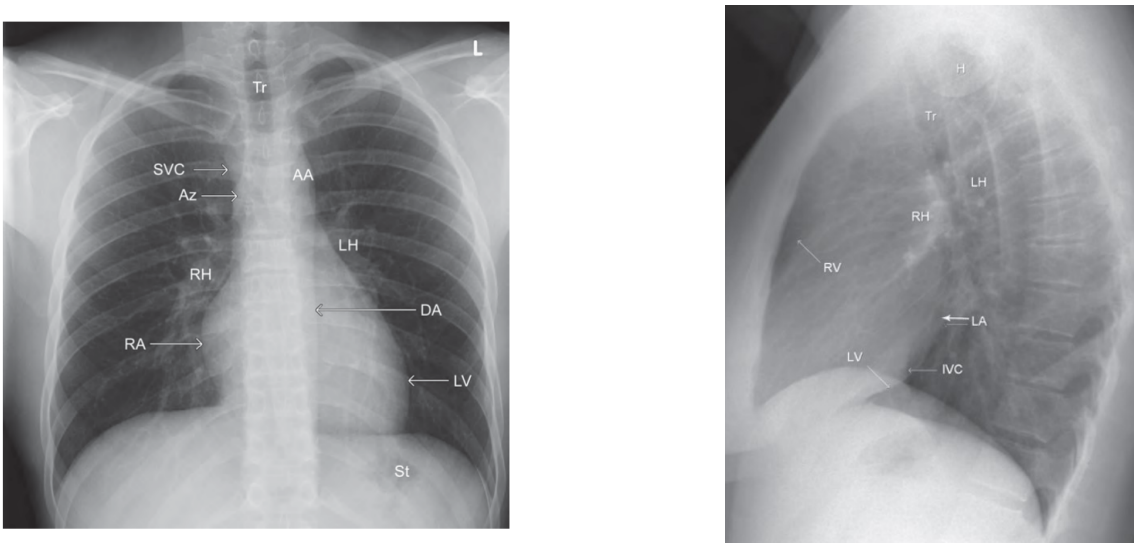



Figure 2.3: Normal PA and LL CXR. The following structures are noted: trachea (Tr), superior vena cava (SVC), azygos vein (Az), right hilum (RH), right atrium (RA), aortic arch (AA), right ventricle (RV), left atrium (LA), left hilum (LH), left ventricle (LV), inferior vena cava (IVC), humeral head (H), descending aorta (DA) and stomach (St) [3].

globally used as a first-line medical imaging technique for chest assessment. This is due to its advantages, which include fast acquisition and interpretation, low cost and low radiation exposure. It is estimated that in 2006, 129 million CXR images were acquired in the United States [11]. 30-40% of all X-ray exams performed are CXR scans, regardless of the level of health-care delivery [6].

During a chest radiography, the patient should be positioned with an erect posture and slightly extended chin, so it does not show up in the final image. The chest should be parallel (or perpendicular, depending on the view) to the beam source (and detector). In frontal views, the hands should be placed on the hips with palms facing out, and the shoulders should be rolled forward. In lateral views, the hands should be raised and crossed above the head. The image should be acquired in complete inspiration. Using both frontal and lateral views in CXR scans can be helpful in some situations, like localizing foreign bodies in the setting of aspiration or projectile injury [12]. The comparison of different views can also be useful in assessing hilar anatomy and lower lobe infiltrates, as well as providing a spatial mapping for intrapulmonary nodes [13]. Thus, since this imaging technique is bidimentional, the diagnostic value of using multiple views should not be undermined.

There are three types of views in chest radiography: anteroposterior (AP), posteroanterior (PA) and latero-lateral (LL). The view type is determined by the trajectory of the X-ray beam through the body. While in a scan with AP view, the beam passes firstly through the anterior anatomy and exits posteriorly (posterior structures are closer to the detector), in a scan with PA view, the opposite process happens (anterior structures are closer to the detector). In a LL view, the X-ray beam passes from a side of the patient to the other. Note that if the patient is confined to a bed, the detector is placed behind the back and thus only an AP view is possible (the cardiac magnification

has to be kept in mind in these cases). The mean radiation dose used in an adult for a CXR scan is around 0.02 mSv for a frontal view, and 0.08 mSv for a side view [14]. The radiation dosage is chosen for each patient according to the patient's size, age, and condition. The radiation dose used in chest radiography is dependent on the considered view. An AP view requires a higher radiation dose than a PA view due to the presence of the breasts [15]. In Figure 2.3 examples of normal PA and LL CXR can be seen.

Acquisition of an image where the patient is not correctly positioned can make it difficult to detect some subtle anatomical characteristics. Thus, there is a need to be aware of the patient position required to get a clear view of the structure that is being examined. The structures that are initially hit by the beam are magnified in comparison to those closer to the detector. Thus, in order to accurately measure a structure, it has to be placed closer to the detector. This is why CXR scans are preferably acquired with a PA view, which minimizes the magnification of the silhouette of the heart. In LL view scans the left side should be positioned against the receptor, for the same reason. As CXR scans are two-dimensional images, generated by different attenuation caused by the tissues, overlapping anatomical structures cannot be differentiated. Also, rotation of the thorax can cause anatomical distortions in the final image, and inadequate inspiration may lead to misdiagnosis of pulmonary opacity or collapse [3], due to the absence of air.

### 2.2.2 Diagnostic Value

The chest radiography technique allows the visualization of various anatomical structures in the thoracic and surrounding area. The trachea, lungs, hilum, scapula, diaphragm, heart, veins, arteries, ribs, clavicles, breasts, liver, and the stomach can all be observed. CXR can be used to identify a wide variety of abnormalities. This imaging exam is usually performed with the aim of visualizing the lungs and the heart, as the most common observed radiological findings are related to pathologies in these organs [3]. In Figure 2.4 a few of these common abnormalities can be observed, which are described in the following paragraphs.

- **Diffuse pulmonary shadowing:** One of the common findings in CXR is diffuse pulmonary shadowing. Radiologically (and anatomically) the lungs can be divided into the alveoli and the interstitium. Both these regions can be affected by disease, thus, both alveolar opacification and interstitial opacification can be present. Opacification refers to the attenuation of the X-ray beam that leads to a more opaque/ lighter appearance in the final image. It can be caused by edema, inflammatory fluid, blood, proteins, or cells.

- **Pulmonary consolidation:** Pulmonary consolidation is characterized by the loss of the usual boundaries of the lungs and heart in a CXR, and it is caused by the filling of the pulmonary alveoli with pus, blood, edema, proteins or cells. Consolidation can be observed in segments or lobes of the lungs, and it is often associated with pneumonia.

  Consolidation adjacent to pulmonary fissures, increased density of the lower thoracic spine on a lateral view and the loss of anatomical structure or border (in the CXR, as mentioned

previously, which is called silhouette sign) can be used to localize areas of pulmonary consolidation. The silhouette sign is visible when a part of the lung becomes non-aerated, and thus its density changes. In CXR images, there is a differentiation of the structures due to their contrasting densities. As the lungs are normally aerated structures, the X-ray beam attenuation is low in comparison with the neighboring structures. However, when an abnormality like lung collapse, consolidation, or mass is present, the lack of air in the lung originates an X-ray image where the mentioned differentiation is harder to observe. In PA view, the heart might limit the visualization of small areas of consolidation. In these situations, the lateral view is helpful at identifying lower lobe pneumonia, by detecting consolidation in either lower lobe.

- **Pulmonary collapse:** When air enters the pleural space (space between the visceral and parietal pleura – tissue that covers the lungs), a pulmonary collapse happens. The collapse can be focused, but it can also include segments and lobes. One of the causes of pulmonary lobar or segmental collapse is the passive collapse caused by external pressure on the lung, like a pneumothorax (total collapse caused by air pressure), pleural effusion (accumulation of fluid between the layers of the pleura) and diaphragm hernia (protrusion of abdominal organs into the thoracic cavity). In a CXR, an upper lobe collapse can be identified by occurring upwards and anteriorly, and a lower lobe collapse by occurring inferior and posteriorly.

- **Pleural disorders:** Pleural effusions and pneumothoraces are examples of pleural disorders. However, other disorders like pneumomediastinum (air leak into the soft tissues of the mediastinum) and pleural thickening (denser scar tissue development in the pleura) can be visible in a CXR.

- **Pulmonary nodules:** Pulmonary nodules are common findings in a CXR. A pulmonary nodule is a lung opacity that resembles a sphere with three centimeter diameter or less, and that is not associated with pulmonary collapse or lymphadenopathy. If the opacity is larger than a nodule, then it is called a mass. Masses have a higher tendency to be malignant.

  Pulmonary nodules are studied for malignancy to allow early diagnosis. Signs of a non-malignant nodule usually include calcification, a well-defined margin and a small size (without evidence of rapid growing). Opposite signs can mean a malignant nodule. Pulmonary nodules are usually considered incidental findings when alone (solitary nodule). However, when multiple nodules are present, the patient is usually symptomatic or with underlying pathology (malignancy, immunosuppression, etc.). The presence of small calcified granulomas is usually related to previous infections.

- **Mediastinal masses:** Mediastinal masses are agglomerations of cells that appear in the mediastinum (space between the lungs). They can be malignant or non-malignant, and can

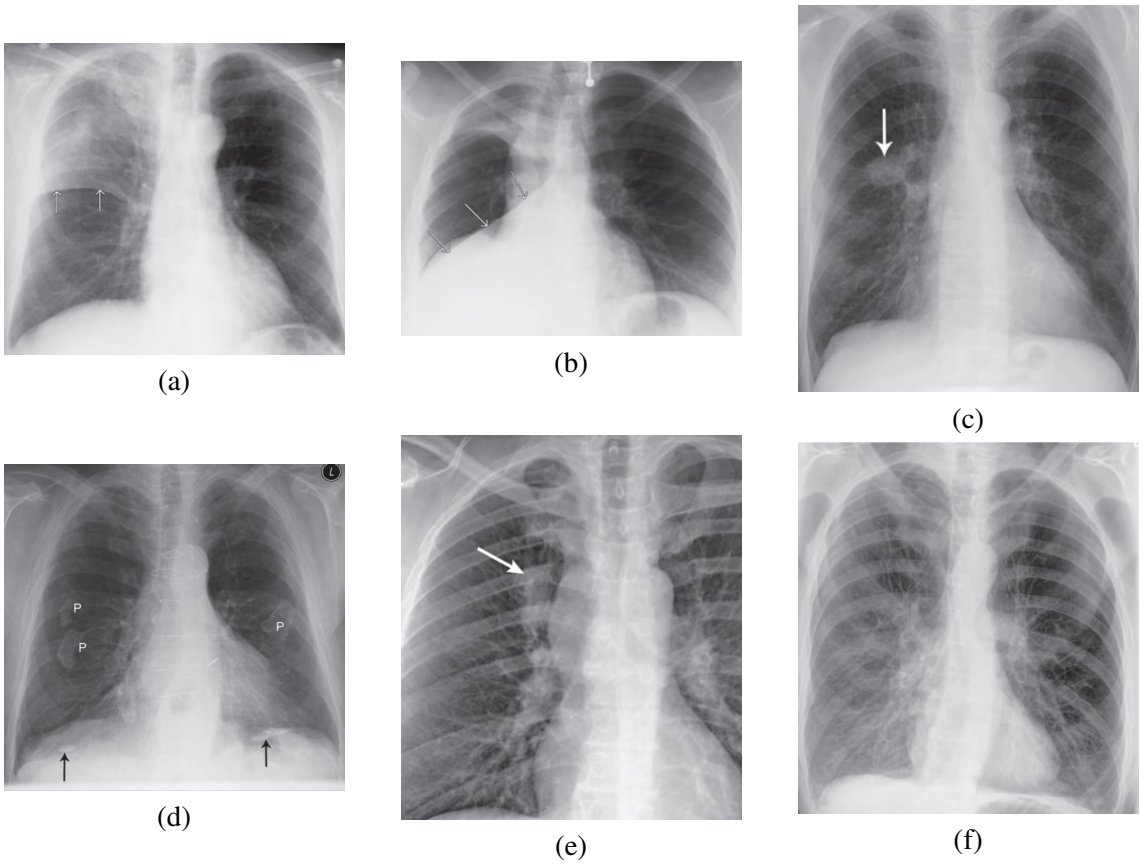



Figure 2.4: Scans with pathological representations of (a) right upper lobe consolidation, (b) right lower lung collapse, (c) solitary nodule, (d) pleural plaques, (e) mediastinal mass, (f) emphysema [3].

be spotted in a CXR due to their higher density. Their classification is based on the localization to the mediastinum: anterior, middle or posterior. Diagnosis includes thymomas, lymphomas, germ cell tumors and cysts, among others.

- **Hilar abnormalities:** Hilar abnormalities are also common findings in CXR. They can be spotted by analyzing the hilar complex components: pulmonary arteries, bronchus, pulmonary veins and lymph nodes. A change in position, size and/or density may be representative of these disorders. Hilar enlargement can be bilateral or unilateral.

- **Emphysema:** The presence of over-expanded lungs is one of the most evident signs of emphysema in a CXR. Emphysema is a condition characterized by the destruction of the alveolar walls, which leads to enlargement of air spaces over time. Centrilobular emphysema is the most common form of the disease, frequent in smokers [3].

- **Heart assessment:** CXR is also used to analyze the heart. Features like the position and size of the heart, valve calcification, outline of blood vessels, pulmonary edema and pulmonary vascular pattern can be used to assess cardiac failure.

The comparison between CXR scans from the same patient over time (longitudinal CXR) is commonly done in medical practice. Comparison with a previous image is useful to assess the stage of visible abnormalities. It allows the categorization of a pathology as acute or chronic, and it displays the possible growth or receding of structures. For instance, as mentioned before, pulmonary nodules are constantly being analyzed for growth, in order to identify malignancy promptly. CXR is also used to preclude further examinations in a nodule if no growth is observed in over two years and there is calcification [3]. The manual analysis and comparison of multiple scans is a time-consuming task, representing a major burden for radiologists, and it remains challenging, even for experienced radiologists. It is important to keep in mind that the position of the patient can differ between images, as well as some image quality parameters (like exposure and contrast). These factors make the process of comparison difficult, increasing the probability of incomplete or incorrect diagnoses.

## 2.3 Automated Chest X-Ray Analysis

The high number of exams, as well as the poor images that result from incorrect patient positioning or image acquisition, are some of the factors that lead to the difficulties in the manual analysis of radiographic scans. These characteristics made the automatization of such processes appealing and highly advantageous. Research in the chest radiography area dates back to the 1960s [16]. One of the first examples of automatization in CXR is the work done in [17], where quantitative measurements of the heart are automatically detected and extracted from PA scans. These features are then used to classify the heart projection using linear and quadratic discriminant functions. More recently, Machine Learning (ML) and, more specifically, Deep Learning (DL) have been widely employed in this field. These techniques demand high quantities of data, but they provide superior results in a short time, in comparison with traditional methods. ML techniques are usually based on initial feature detection and extraction, while DL approaches do this automatically.

Many types of problems can be distinguished in automated CXR analysis. Image-level predictions, based on the whole scan, are the most common task. Segmentation is also a frequent target, as well as localization, since both of these allow the identification of a region of interest, which can be used for further investigation. Other areas include image generation, report generation and image registration problems [16]. In this section, a small overview of automated methods for CXR analysis is presented. To do so, a few articles that target the previously mentioned types of problems were selected.

### Image-level prediction

Image prediction tasks can be separated into classification and regression. In [18], convolutional neural networks (CNN) are used to solve a classification problem, aiming to distinguish between PA scans and AP scans, which is useful for cleaning hospital data. The CNNs (Visual Geometry Group (VGG) variant [19] and Residual Neural Network (ResNet) [5]) are trained with the Radiological Society of North America (RSNA) dataset [20] and validated on a self-compiled dataset

labeled by a human expert, and Gradient-weighted Class Activation Mappings (Grad-CAM) [21] are generated in order to visualize and understand the model prediction. By ensembling the models, the F1-score attained was 0.958, and the Grad-CAMs showed that the anatomical structures used for the prediction were comparable to the ones used by a radiologist.

An example of a regression task is presented in [22]. Here, linear regression was performed to predict scores for extent of lung involvement and opacity in frontal CXR images, with the objective of identifying the severity of COVID-19 lung infection. The measurement of severity can be highly useful in the hospital, allowing for a fast triage of the patients, as well as monitoring of disease evolution. A public COVID-19 dataset was used, and it was annotated by three experts regarding the desired targets. A Densely Connected Convolutional Network (DenseNet) [23] model was pretrained on multiple non-COVID-19 CXR datasets (BIMCV-PadChest [24], MIMIC-CXR [25], CheXpert [26], ChestX-ray8 [27] and RSNA pneumonia dataset), and it was used to extract features from the COVID-19 images. The geographic extent score (range 0-8) and the lung opacity score (range 0-6) are predicted with a mean absolute error of 1.14 and 0.78 respectively. Saliency maps [28] were computed in order to better understand the relevant pixels for the prediction and, for most of the results, the model correctly looks at opaque regions of the lungs.

In [29] an end-to-end DL framework for X-ray image diagnosis is presented. Firstly, the image is classified as X-ray or not (first module), and then it is classified according to the type of X-ray (second module). The abnormality classification (third module) is then performed, based on CXR (14 different pathology labels). Multiple datasets were used, including Chest-Xray8, Musculoskeletal Radiographs (MURA) [30], Lower Extremity Radiographs (LERA) [31], Accurate Automated Spinal Curvature Estimation (AASCE) [32], Panoramic Teeth X-ray [33] and ImageNet [34]. The used classifiers are based on the DenseNet-121 network. The test set accuracy for the first module, second module, and third module are 0.987, 0.976, and 0.947, respectively.

One of the main disadvantages of DL algorithms is that most of them can be categorized as black box models, which means that their internal working process is hidden to the user. This is important because, as human beings, an explanation for certain decisions is needed, in order to trust, understand and interpret them. This is one of the main motivations for Explainable Artificial Intelligence (XAI) [35]. XAI methods aim for the construction of transparent artificial intelligence (AI), without affecting the existing performances. The mentioned Grad-CAMs and saliency maps are examples of XAI. Different kinds of groups can be used to categorize explainability methods. They can fit into the pre-model, in-model or post-model category. Pre-model interpretability is related to data analysis, using techniques for data visualization and description. In contrast, in-model interpretability concerns methods that have built-in interpretability, by applying constraints to its complexity. Post-model (post hoc) interpretability regards the analysis of the model after building it.

The type of explanation created can be used for XAI classification. Feature summary methods allow the interpretation of the model via summary statistics that can usually be visualized. Model internals methods are associated with intrinsically interpretable models, using internal characteristics for interpretability. Thus, these are model-specific methods. There are also methods that

output data points, which requires them to be interpretable themselves, thus, this is usually used for images and text. Another possible output for explainability methods is a surrogate intrinsically interpretable model. In these methods, another model (easily understandable) is used for interpretation, which is used to approximate local or global features [36].

### Segmentation

Segmentation tasks can be focused on the identification of anatomy, foreign objects or abnormalities [16]. As previously mentioned, segmentation masks can be used to improve efficiency, by limiting the image area used for further analysis. It can also be used to perform feature extraction, like shapes or area measurements.

In [37] a method for improving lung segmentation performance is proposed. An attention module was developed (X), as well as a variant (Y). The X-attention module is composed of channel and spatial attention (extracted from the input feature maps), enabling the effective extraction of global and local features. The Y-attention module is a variant of the aforementioned, which accommodates the global context from a deeper layer. The attention module was combined with a U-Net [38] in many different configurations, using ResNet-101 as the backbone network. Three public datasets were used, including the Montgomery dataset [39], the Japanese Society of Radiological Technology (JSRT) dataset [40] and the Shenzhen dataset [41]. The DSC score was used to evaluate the segmentation performance, as well as the sensitivity and Positive Predictive Value (PPV). The segmentation results were post-processed by keeping only the two objects with the bigger area, in order to improve segmentation performance. The DSC values attained for each dataset were $0.982 \pm 0.002$, $0.968 \pm 0.002$ and $0.954 \pm 0.002$, for the most favorable configuration. It also showed comparable performance to XLSor [42] (state-of-the-art DL model for lung segmentation). Nonetheless, the segmentation performance acquired is low for CXR with deformed lungs or ambiguous cardiac silhouette.

### Localization

Regarding localization algorithms, the objective is obtaining a Bounding Box (BB) or point coordinates that localize a certain structure. The most frequent target are pathologies, but there are also studies on localization of anatomic structures and objects like support devices [16]. As an example, in [43] a CNN with 23 convolutional layers was used to detect pneumothorax on CXR images after Percutaneous Transthoracic Needle Biopsy (PTNB) for pulmonary lesions, which is a method used for the diagnosis of pulmonary lesions with high diagnostic accuracy. The collected dataset comprises 1,596 CXR with pneumothorax at different levels of severity and 11,137 normal CXR. A set of 500 additional images were used for internal validation. Two radiologists manually drew regions of interest in the cases with pneumothorax. The network used for the inferences was fine-tuned using the You Only Look Once (YOLO) Darknet-19 [44] pretrained model. A temporal validation dataset was constructed using follow-up CXR at two different time points after PTNB (1,379 at 3 hour follow-up and 1,329 at 1 day follow-up). The model's AUC for pneumothorax

detection was 0.984. Regarding the 3 hour and 1 day follow-up, the AUC values were 0.898 and 0.905, respectively, meaning a good performance on post-PTNB follow-up CXR. Some structures or pathologies other than pneumothorax lead to false-positives, due to their similar nature, as pleural thickening, sclerotic rib margins, medial borders of scapula, and skin folds.

### Image Generation

The generation of images using DL technologies has been used in numerous fields. Some of the most common applications are data augmentation, visualization, abnormality detection through reconstruction, domain adaptation or image enhancement methods [16].

In [45], a method to classify a CXR as normal or abnormal is developed, using generative adversarial one-class learning. The proposed architecture is similar to Generative Adversarial Networks (GANs) [46], and it is composed of three main modules: a U-Net autoencoder, a CNN discriminator and an encoder. GANs are constituted by a generator and a discriminator. The generator takes a random noise vector as input and produces a sample in the data space. The discriminator identifies if a sample comes from the true data distribution or from the generator. The training process aims at getting the generator to construct samples that are not distinguishable from the discriminator, and both of them are trained alternately. In this work, the U-Net (autoencoder) functions as the generator and the CNN as the discriminator. The generator maps a first input image, and then a deconvolutional network (decoder) is used to inversely map the image, generating the reconstructed image. The used dataset was the Chest-Xray8 dataset. In the training phase, 4,479 normal CXR are used. The amount of normal and abnormal images used is 849 and 857, for validation, and 677 and 677, for testing. An abnormal CXR is considered to have at least one pathology. A CXR is distinguished as normal or abnormal by using the reconstruction result. If it is normal, the architecture can reconstruct the content. If it is abnormal, the model performs poorly in the reconstruction, since it has not seen pathology images during training. An anomaly score is also computed during testing. The final network achieved an average AUC of 0.841 on the testing set. The generated abnormal images contain blurry and messy regions, and the geometrical structures are distorted, proving that the classifier can be used to accurately distinguish the classes.

### Automated Report Generation

The automation in the CXR field is not only related to the scans itself, but it can also be related to the medical reports. In that sense, the automated generation of reports aims for a more efficient CXR analysis workflow.

In [47] a system called Vispi is proposed for classifying common thoracic diseases, as well as localization and generating a medical report. The first step in the algorithm is the classification and localization of pathologies. Then, the sentences that build the report are generated. The model used for classification has a DenseNet-121 backbone, pretrained with ChestX-ray8. In order to get the BB for the pathology, Grad-CAMs are used for the classification model. The generated heatmaps were used in a threshold based BB generation method, which builds a BB around the

regions of the heatmap with highest intensity. If no pathology is detected, a report is directly generated by an attentive Long Short-Term Memory (LSTM) network [48], using the full CXR. If a pathology is found, the generated BB is used to crop the image, and the subimage is used to build the description of abnormalities, while the original CXR is used to get the description of normality. The LSTM takes the image and the subimage as inputs, and generates a sequence of sentences for the entire report. For each disease class, a specific pair of LSTMs are trained. The IU Chest X-ray Collection dataset [49] is used for training, validating and testing. The classification module acquired an AUC of 0.804, and the report generation system got a CIDEr (metric for measuring the similarity of a generated sentence against a set of ground-truth sentences) of 0.553, outperforming all baseline models used for comparison.

### Image Registration

Finally, image registration is also a field of focus on automated CXR analysis. Image registration consists in aligning two images, allowing the comparison of the matching features. It is widely present in the medical imaging field, as multiple images from the same patient are commonly collected through time for diagnosis and therapy evolution, for example. Imaging techniques that acquire sets of images from multiple perspectives or slices, like CT, also benefit from image registration of the acquired images, allowing a smooth final result. This topic is further discussed in Section 3.2.2.

## 2.4 Longitudinal Chest X-ray Analysis

Most systems are designed to receive a single exam as input, whereas radiologists often compare multiple exams from the same patient in different time points in order to reach a conclusion on the analysis. Hence, the development of CXR analysis methods that take multiple input images must be taken into consideration. These methods could allow the analysis of evolution and comparison between exams, which would be a significant advantage for medical professionals, as it would make the manual process more efficient.

The majority of the developed automated methods do not perform the comparison between exams, aiming at the analysis of a single image, as it is the most common form of image input and datasets are usually not equipped with longitudinal data for comparison. The introduction of longitudinal comparison and analysis of scans in an automated manner can be highly useful. Using more information can improve the performance of already existing methods, and the comparison between scans can increase the transparency of the algorithm and increase its robustness. As such, the development of longitudinal systems is crucial for the successful integration of automatic system in the clinical practice.

# Chapter 3

# Automated Longitudinal Chest X-ray Analysis

Longitudinal data consists of information from the same patient throughout time. As previously mentioned, medical professionals usually compare multiple longitudinal studies from the same patient, but automated systems normally consider a single input image at the time for analysis, isolating temporal information, rather than taking advantage of previously collected scans. The inclusion of longitudinal data in automated CXR analysis is a topic that can improve current methods and facilitate comparison between scans. Thus, in this chapter, a longitudinal analysis of CXR datasets is done, as well as an overview of some algorithms that use longitudinal data.

## 3.1 Public Datasets

The challenges associated with the analysis of CXR led to an increasing development of automated strategies to facilitate the management of scans and the extraction of information from them. Such methods are usually dependent on high quantities of data. Thus, this led to the creation of large and various datasets for CXR. Regarding longitudinal information, some datasets include the date in which the scan was captured, while others might only contain information such as the age of the patient at the time of the collection or the follow-up number of the scan, which refers to a sequential numeration of the images from the same patient.

The original ChestX-ray8 [27] dataset contains 112,120 frontal CXR images (size of $1,024 \times 1,024$ pixels) from 32,717 patients, collected at the (US) National Institute of Health from 1992 to 2015. Eight common disease labels (atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax and consolidation) are generated from radiological reports using Natural Language Processing (NLP), and a small set of the images also contain hand labeled BB for these pathologies. The dataset was posteriorly updated, including six more diseases (edema, emphysema, fibrosis, pleural thickening and hernia). This update is called ChestX-ray14. Regarding this dataset, there are 13,302 patients with multiple images. Each image is linked with a patient

age and a follow-up number. In the majority of the cases, the age difference between sequential images is zero years.

The CheXpert dataset [26] consists of 224,315 CXR scans from 65,240 patients, collected at the Stanford Hospital between 2002 and 2017. The dataset is labeled for fourteen observations (which include twelve pathologies, a no findings label and the presence of support devices), decided on based on their prevalence in reports and clinical relevance. An automated rule-based labeler was used to extract the observations as labels from the medical reports. This dataset contains 37,161 patients with multiple studies. The only potentially useful parameter for sequential and temporal motives is the patient age. However, in the vast majority of the cases, the age difference between two studies is zero years.

The BIMCV-PadChest dataset [24] comprises 160,868 CXRs from 67,000 patients, collected at the San Juan Hospital (Spain) from 2009 to 2017. This dataset covers six different position views and information on patient demography. 27% of the reports were manually labeled by physicians, and a recurrent neural network with attention mechanisms was trained and used to label the rest of the dataset from the reports. The reports were used to extract 174 findings, 19 diagnoses, and 104 anatomic locations. In this dataset, there are 31,314 patients with more than one study (relevant patients for longitudinal purposes). Each study (might contain multiple images) is associated with a date, which allows the direct temporal comparison of the scans. Most commonly, the month difference between two studies is zero, however, there are studies fifty or more months apart.

The MIMIC-CXR dataset [25] is composed of 377,110 CXR of 65,379 patients, from studies performed at the Beth Israel Deaconess Medical Center in Boston, from 2011 to 2016. The images of the original version are in a DICOM format, but another version of the dataset with JPG format was also published (MIMIC-CXR-JPG). The dataset was automatically labeled from radiology reports using the same method and the same labels as CheXpert. The information about the date of the image is present in the dataset, however, to ensure anonymity, a date shift was assigned to each patient. There are 56,320 patients with more than one study. The majority of the studies have a zero-month time interval. The Chest ImaGenome dataset [50] contains 242,072 frontal MIMIC-CXRs, and it describes the relationships between images, including pathologies in common, the anatomical region associated with this comparison, the comparison label and BB information.

The RSNA International COVID-19 Open Annotated Radiology Database (RICORD) [51] contains not only CXR data, but also CT scans information. Regarding the CXR data, this dataset contains 998 images from 361 patients, collected by the Radiological Society of North America (RSNA) in four international institutions. Each CXR was classified by three radiologists as typical, indeterminate, atypical, or negative for findings of COVID-19 pneumonia. The regions with abnormal opacities were also classified as showing mild, moderate, or severe disease. In this COVID-19 dataset, there are 165 patients with multiple studies. The only temporal information available is the patient age.

Asides from the mentioned datasets, others are available. However, not every dataset contains temporal or sequential information that can be used for longitudinal studies. One example is the VinBigData dataset [52], which contains 18,000 PA view CXR scans, collected from two hospitals

Table 3.1: Longitudinal datasets' analysis.

| Dataset | % longit. patients | Number of longit. frontal images | Number of longit. lateral images | Min./max. number of longit. frontal images per patient | Median number of longit. frontal images per patient | Time measure | Abnormalities / longit. abnormalities |
|---|---|---|---|---|---|---|---|
| BIMCV-PadChest | 46.74 | 71,064 | 24,459 | 2/119 | 2 | Date | 191/194 |
| ChestX-ray14 | 40.66 | 94,617 | - | 2/184 | 4 | Age | 15/15 |
| CheXpert | 56.96 | 158,241 | 15,379 | 2/91 | 3 | Age | 13/13 |
| MIMIC-CXR | 86.14 | 269,031 | 53,984 | 2/172 | 4 | Date | 14/14 |
| RICORD | 45.71 | 802 | - | 2/32 | 3 | Date | - |

in Vietnam (the Hospital 108 and the Hanoi Medical University Hospital) between 2018 and 2020. The images are annotated by radiologists for the presence of 6 diagnoses and 22 critical findings. This dataset also provides the localization of critical findings.

In Table 3.1, an overview of the datasets is presented, where only the patients with more than one image were considered. Note that for the CheXpert dataset, the analysis is based solely on the author-defined train and validation sets. In Figure 3.1, histograms that represent the average time interval between images of the same patient are presented. The used time measurement is dependent on the dataset, as shown in Table 3.1. For the datasets in which the temporal data is uniquely the age (ChestX-ray14 and CheXpert), this time interval is defined in years, while in the other datasets (with a study date associated to the images), this interval is defined in days.

Using exclusively the patient age is not enough to get sequential information for a patient, since there are images with the same patient age, that could be from 1 day to 12 months apart. Thus, the usage of datasets where the only temporal clue is the patient age is challenging when considering longitudinal analysis.

## 3.2 State of the Art

### 3.2.1 Longitudinal Analysis

There are some studies regarding the utilization of longitudinal data for automated image analysis. Due to the COVID-19 pandemic, the tendency to study CXR evolved, and so, many studies addressing the pathology with temporal data appeared. In [53], the mortality and duration on Intermittent Mandatory Ventilation (IMV) are predicted for COVID-19 patients. A private dataset with 186 patients is used. Features are extracted from the images using a VGG-16, then, these are concatenated with longitudinal non-image data. The result is concatenated with non-image non-longitudinal data, and finally the model predicts the final outcomes. It is shown that using the

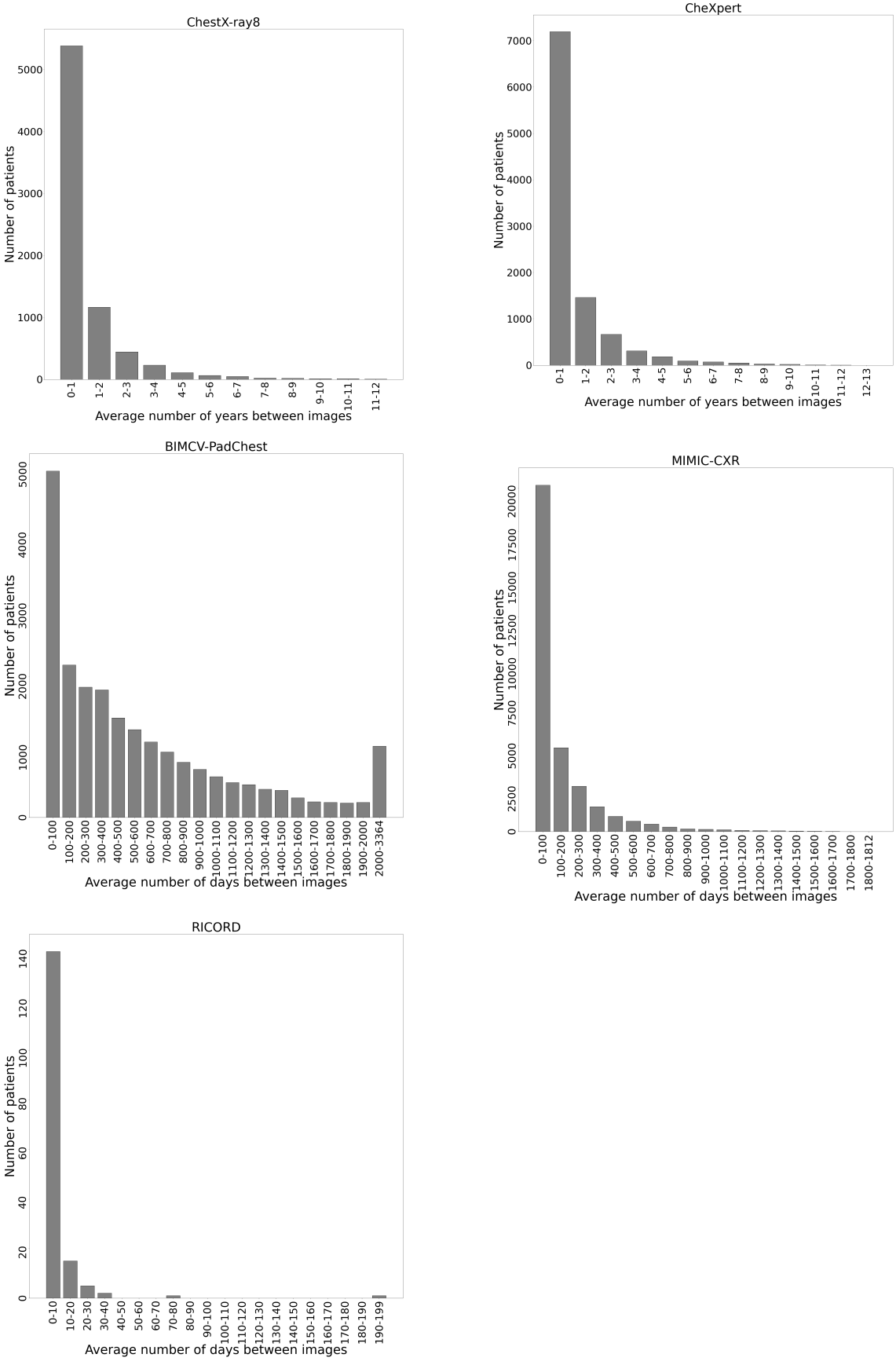



Figure 3.1: Histograms for average time between images from the same patients.

combination of all these data types outperforms the usage of each one independently. The best result was AUC of $0.870 \pm 0.050$ for mortality and a mean absolute error of $2.56 \pm 0.20$ days, for the predicted duration in IMV.

In [54], image and non-image data is also used, but to predict the probability of Intensive Care Unit (ICU) admission, ICU discharge, hospital admission, hospital discharge and death before a certain moment in time. The used dataset is private, including time dependent and time independent information for 1,894 COVID-19 patients. A baseline hazard is computed based solely on time. It allows the calculation of the hazard function and, consequently, the survival function, which is used to compute the probability of one of the events happening before a time point. The risk function is determined by the image and non-image data, which is processed using a convolutional LSTM and a LSTM. The computed concordance error outperformed all methods used for comparison, for all the time-to-event predictions. It is also shown that the usage of longitudinal (time-dependent) images significantly improves the predictions.

In [55] the aim is to predict disease severity (no disease, mild, severe or critical) and its outcome (worse, stable or improved), for COVID-19 patients. The CheXpert dataset is used to pretrain a DenseNet-121, which is used as a feature extractor. The last convolutional layer of the model is used to extract features for 10 random crops per image, using images from two COVID-19 datasets: the open-source MILA COVID-19 dataset [56], and a private COVID-19 ICU dataset. These are used to test whether the features from a first image can predict the outcome of its longitudinal pair. The used classifier and parameters were tuned, reaching a final model with a 0.810 AUC for the open-source dataset and 0.660 for the private dataset, regarding the outcome category. The disease severity prediction reached 52.3% accuracy.

A method for automated measurement of COVID-19 disease severity is proposed in [57], which is used for longitudinal disease tracking and outcome prediction. A Densenet-121 network is used. It is pre-trained with ImageNet, initially, and then with CheXpert, using a binary label on whether a pathology (lung opacity, lung lesion, consolidation, pneumonia, atelectasis, or edema) is present in the image or not. This model is used for transfer learning with an internal COVID-19 dataset, where the output is the mRALE score (measurement of severity of lung edema). A siamese structure is constructed using this model, taking as inputs the desired image and a pool of normal images from CheXpert, for comparison. The euclidean distance between the outputs is calculated in order to obtain the PXS score, which is a measurement of the severity. In the test set, the PXS score correlated with the mRALE score assigned (ground-truth) with a Pearson correlation coefficient (r) of 0.86. One of the used datasets is labeled for disease severity change, thus, the change in PXS score was evaluated between longitudinal image pairs. This change correlated with the change label with an r of 0.74. Patients with higher PXS scores were intubated or dead within 3 days of admission.

In [58], a modification of LSTM network [48] is presented, in order to consider different time intervals. LSTM are networks used for prediction of sequences. They are composed of memory blocks, which contain input gates, output gates and forget gates (which modulate how much information is used from the internal state of the previous time-step). The proposed alteration consists

in adding the time interval information, and testing whether the classification of sequential images improves. In order to predict the labels of a scan, the time-modulated LSTM (tLSMT) takes as an input the label of the previous image, the features of the scan and the time interval between the two images. In 3.1, the equations for this adaptation are presented. $h_t$ defines the internal state, while $f_t$, $i_t$ and $o_t$ refer to the forget, input and output gates, at time $t$. $X_i^t$ denotes the input image features at a time step, $l_i^{t-1}$ the labels that describe the images acquired at previous time points, $\delta_i^t$ the time difference between the image at time $t$ and the image at time $t-1$. The image labels for the last image in the sequence are computed, represented by $y_i^t$. The remaining parameters consist of learnable variables. A private dataset was used, containing longitudinal scans for 80,737 patients. Labels were extracted from medical reports using NLP methods. Image features are extracted using a pre-trained Inception-V3 network [59]. In comparison with a baseline CNN (predicts the labels from the image directly), the tLSTM showed an improvement of around 7% in F-measure, and around 8%, in comparison with a standard LSTM (with longitudinal data).

$$\begin{aligned} f_t &= \sigma(W_{fl} \times l^{t-1} + W_{fx} \times X^t + W_{fj} \times \delta^t + b_f) \\ i_t &= \sigma(W_{il} \times l^{t-1} + W_{ix} \times X^t + W_{ij} \times \delta^t + b_i) \\ o_t &= \sigma(W_{ol} \times l^{t-1} + W_{ox} \times X^t + W_{oj} \times \delta^t + b_o) \\ c_t &= tanh(W_{cl} \times l^{t-1} + W_{cx} \times X^t + W_{cj} \times \delta^t + b_c) \\ h_t &= f_t \times h_{t-1} + i_t \times c_t \\ y^t &= o_t \times tanh(h_t) \end{aligned} \tag{3.1}$$

In [60], the focus is the detection of abnormalities on CXR scans and the detection of change in pathologies over sequential images. The abnormality detection is done through Qure AI, which consists of a set of CNNs, trained to identify a certain disease on frontal CXR. The used dataset is ChestX-ray8, where 874 scans were selected from. These images were annotated by two radiologists for pulmonary opacities, pleural effusions, hilar prominence, enlarged cardiac silhouette or no findings. The image labels are compared between consecutive images, to assess the pathology change. The AUC reached by the method to detect change in the different pathologies ranges from 0.735 to 0.925, depending on the class. The results are compared with four test radiologists, which performed similarly or underperformed.

Another example can be found in [61], where the objective is to detect change in a lesion, when comparing two longitudinal images. A squeeze and excitation network (SENet) [62] is used to extract image features, which are used as local descriptors. A correlation score is computed for every possible local descriptor combination between the two images' feature maps, originating a geometric correlation map. Finally, a binary classifier is used to classify the sample as change or no-change. The used dataset is a private dataset of CXR from 5,472 patients. The image pairs have at least a 30-day time interval. This method showed an AUC of 0.890, outperforming all the comparison methods, including the aforementioned tLSTM [58], with an AUC of 0.780.

In [4] a model for tracking longitudinal relations between CXR (CheXRelNet) is proposed.

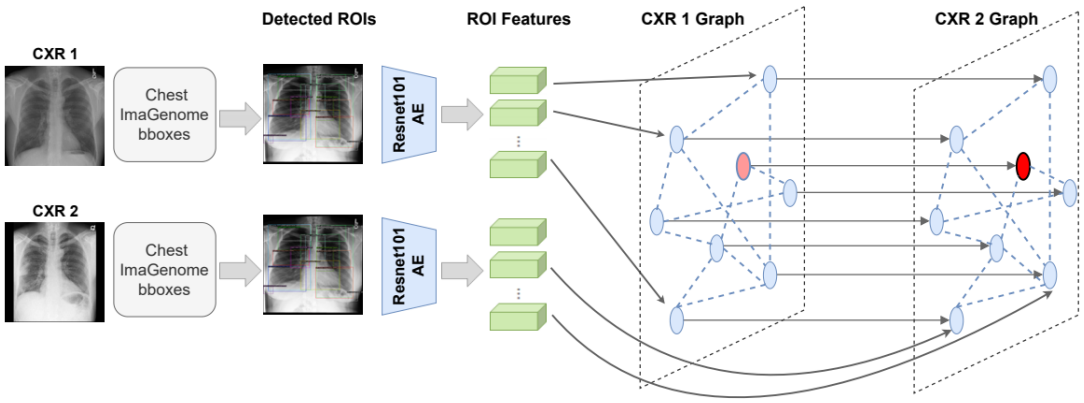



Figure 3.2: Graph construction scheme [4].

This model uses global information, but also local information. The used dataset is the Chest ImaGenome dataset. A pretrained Resnet-101 is used to extract global features, while a Graph Attention Network (GAT) [63] is used to extract local features, using the anatomical BB. GAT extracts inter and intra images features, using an adjacency matrix that expresses these relationships. A scheme of this construction is shown in Figure 3.2. The global and local features are concatenated and classification layers are added to provide the output. The output consists of the first image pathologies label, and on the second image "improved" or "worsened" label. The final model outperformed the baseline models, which use only global or local information, presenting an accuracy of 0.680 (average the pathologies test accuracy).

In Table 3.2 a summary of the mentioned methods is presented.

Along with the longitudinal analysis of images comes the necessity of correct image registration, as working with multiple images after registration simplifies their comparison. Thus, in the following chapters, work regarding image alignment is described.

### 3.2.2 Image Registration

As previously mentioned, image registration is one of the areas of active research in automated CXR analysis. The alignment of two CXR (which can be a longitudinal pair or not) may improve the management and analysis of the scans, either manually or by other automated methods.

Different alignment techniques have been explored in this field. Feature-based methods usually use features such as points, contours, curves, or other geometric references. Intensity-based methods directly use the image intensities for alignment. Frequency-based methods can also be used to align two images using their frequency domain. Hybrid methods use a mixture of the aforementioned techniques.

It is important to note that image registration techniques can be divided into rigid and non-rigid. Rigid transformations assume that the objects in the images are static, and, on the other hand, non-rigid transformations assume that the objects can be deformed by biological differences

Table 3.2: Summary of the described longitudinal studies.

| **Study** | **Year** | **Used dataset** | **Objective** | **Approach** | **Results** |
|---|---|---|---|---|---|
| Longitudinal Detection of Radiological Abnormalities with Time-Modulated LSTM [58] | 2018 | Private dataset with longitudinal scans for 80,737 patients | Testing whether a modification in the LSTM networks can improve prediction of sequences with different time intervals between images | Addition of the time interval between images as input information to the LSTM (tLSTM). Image features are extracted using a pre-trained Inception-V3 network. | In comparison with a baseline CNN and a stadatd LSTM, the tLSTM showed an improvement of around 7% and \8%, respectively, in F-measure |
| Deep learning in chest radiography: Detection of findings and presence of change [60] | 2018 | 874 manually annotated scans selected from ChestX-ray8 | Detection of abnormalities on CXR scans and of change in pathologies over sequential images | The abnormality detection is done through Qure AI, and the change detection is computed by the comparison of the pathology label in the pair. | The detection of change in the different pathologies ranges from 0.735 to 0.925. The comparison test radiologists showed similar or worse performances. |
| Longitudinal Change Detection on Chest X-rays Using Geometric Correlation Maps [61] | 2019 | Private dataset of CXR from 5,472 patients. | Detect change in a lesion, when comparing two longitudinal images | A SENet is used to extract image features, which are used as local descriptors to originate a geometric correlation map. A binary classifier uses this map to classify the sample regarding change. | AUC of 0.890, outperforming all the comparison methods, including the aforementioned tLSTM (AUC of 0.780). |
| Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks [57] | 2020 | ImageNet, CheXpert, an internal COVID-19 dataset | Longitudinal disease tracking and outcome prediction | A Densenet-121 is pre-trained and used to predict the presence of pathologies or not. This model is used in a siamese structure that recieves the input image and a pool of normal images. The euclidian distance between the outputs is used to compute a severity score. | The results correlated with the ground truth with a r of 0.86. The change in the severity score was evaluated between longitudinal image pairs and it correlated with the change label with an r of 0.74. |

| Study | Year | Used dataset | Objective | Approach | Results |
|---|---|---|---|---|---|
| Deep survival analysis with longitudinal X-rays for COVID-19 [54] | 2021 | Private dataset, including time dependent and time independent information for 1,894 COVID-19 patients. | Predict the probability of ICU admission, ICU discharge, hospital admission, hospital discharge and death before a certain moment in time. | A risk function is computed by processing image data (using a convolutional LSTM) and non-image data (using a LSTM). It is used to compute the final probability of each output. | The computed concordance error outperformed all methods used for comparison. The usage of longitudinal images significantly improved the predictions. |
| CheXRelNet: An Anatomy-Aware Model for Tracking Longitudinal Relationships between Chest X-Rays [4] | 2022 | Chest ImaGenome | Predicting the pathology label of the first image in a pair, and the the "improved" or "wornsened" label for the second image in a pair. | A pretrained Resnet-101 is used to extract global features, while a GAT is used to extract local features. These features are concatenated and classification layers are added to provide the output. | The final model outperformed the baseline models, which use only global or local information, presenting an accuracy of 0.680 (average the pathologies test accuracy). |
| Deep learning of longitudinal chest X-ray and clinical variables predicts duration on ventilator and mortality in COVID-19 patients [53] | 2022 | Private dataset with 186 patients | Mortality and duration on IMV prediction for COVID-19 patients | Image features are extracted using a VGG-16. They are concatenated with longitinal and non-longitudinal non-image data to predict the final outcomes | AUC of $0.870 \pm 0.050$ for mortality; mean absolute error of $2.56 \pm 0.20$ days for the predicted duration in IMV |
| Tracking and predicting COVID-19 radiological trajectory on chest X-rays using deep learning [55] | 2022 | CheXpert, the open-source MILA COVID-19 dataset, and a private COVID-19 ICU dataset | Predict disease severity and outcome for COVID-19 patients | A Densenet-121 is pre-trained and used as a feature extractor for image crops. The final model is used to test wheter the features from the first image can predict the outcome of the pair. | Outcome category reached an 0.81 AUC for the open-source dataset and 0.66 for the private dataset. The disease severity prediction reached 52.3% accuracy. |

(that lead to changes in shape or position over time), image acquisition and others [64]. Thus, rigid transformations, also referred to as affine transformations, include rotations, translations, scaling and shearing operations. These are transformations that preserve collinearity and ratios of distances. Non-rigid operations include local transformations in an image, modeling a transformation map that aligns it to its pair. Each type of transformation is associated with different types of tissues. Rigid transformations are applicable to bones, while non-rigid transformations are more applicable to soft tissues such as the abdominal organs [64].

In [65], keypoints based on lung features are extracted and used for registration. Intensity differences between lung and non-lung regions in a CXR scan are used to identify these keypoints, and different alignment techniques, like linear and polynomial transformations, are used to align the images. A small set of images was used for testing and defining the best type of transformation technique. As this method is quite simple and the acquisition of the feature points is widely dependent on image intensities, it may fail when aligning images with poor contrast or with pathologies that affect the lungs' opacity. Similarly, in [66], the intensities of the image columns are used to identify the limits of the lungs, as well as the spine. Intensity differences and mean density distribution, variance and density difference are used to obtain the upper and lower limits of the lungs, originating a BB. These features allow the acquisition of 9 control points that are used to align a pair of images. Thin-plate spline method is used to obtain the non-linear deformation matrix that aligns the pair. Quantitative results are not presented in this paper.

The application of rigid transformations is often not enough when dealing with CXR registration. However, the usage of non-rigid transformations can lead to unrealistic anatomical distortions of the image. In order to solve this problem, in [67] the anatomical information from the lungs and the heart is used to obtain a more realistic transformation of the image, avoiding harsh deformations of these organs. An encoder-decoder structure (similar to a U-Net) is used to predict the deformation field between a pair of images, and a differentiable warping module uses it to produce the deformed image. The lung and heart mask of the pair is also used, producing the warped image mask in the same manner. The used loss has three different components. The first one is the loss of the model that originates the deformation field itself. Then, there is the loss associated with the alignment between the target anatomical segmentation mask and the warped source segmentation mask, which allows the inclusion of both organs in the alignment, but does not guarantee a good global correspondence between the deformed image and the input image. In that sense, a final global loss is used. Denoising autoencoders are used to generate learned representations (contain information regarding relevant global anatomical features) of the masks. The global loss measures the euclidean distance between the learned representation of the deformed mask and the learned representation of the target segmentation mask. This global loss ensures that the deformed image is anatomically plausible. In this work, the usage of the heart and lung segmentations showed great advantage, and the presence of the three loss components was valuable at producing better results.

In [68] a hybrid Linear Support Vector Machine (L-SVM), composed of 6 models, is built based on Felzenszwalb Histograms of Oriented Gradients (FHOG) features and L-SVM models.

It is used to detect the lungs, ribs, and clavicles from a CXR scan. Initially, after the detection of the left and right lungs, the developed Spin Assisted Algorithm (SAA) is used to rotate the image to the correct orientation, by using the upper line of the lungs. This improved the detection of the ribs and clavicles by the model. Keypoints are selected from the detected regions. Then, Absolute Distance Matching Algorithm (ADMA) is used to match the landmarks from the input image and the landmarks from target image. Two categories of transformations are used to align the image pair, rigid and non-rigid. The method used for the rigid transformation is called Singular Value Decomposition (SVD), while the one used for non-rigid transformation is Elastix.

The work in [69] is adapted in [70], introducing image alignment and subtraction. The main objective of the work is to identify and locate pathologies. To align the input image pair, a ResNet-18 backbone is used. A target image is generated by averaging 500 images that are labeled as not containing pathologies. The model is trained to output the transformation parameters (translation, rotation, and scaling) that align the image to the target image. For this, a feature reconstruction loss is used, that encourages the image to have a similar feature representation to the target. The model for pathology detection and location contains two branches. One of them receives as an input a positive image and generates a feature map. The other one generates attention maps for a positive and for a negative image and then subtracts them. The resulting map is multiplied element-wise with the feature map from the first branch. Then, the paradigm used in [69] for the loss calculation and prediction generation for each class is employed. Images with and without BB information are used for training, using different losses for each case. The usage of more images with BB information during training showed to improve the location prediction. Even though this paper does not report results for alignment specifically and the used pair is not of longitudinal images, the reported results show that using an alignment method is advantageous.

Different types of metrics can be used to evaluate the performance of registration algorithms. In [67], the metrics used are the DSC, the Hausdorff Distance (HD) and the Average Symmetric Surface Distance (ASSD). DSC measures the overlap between objects. It is used to evaluate the CXR alignment algorithm by being applied to the lung segmentation masks. Thus, a DSC of 1 means complete overlap, while a DSC of 0 means no correspondence at all. HD corresponds to the maximum distance between the segmentation contours, thus, a lower HD value symbolizes better performance. ASSD is the average distance between the segmentation contours, so lower values imply better alignment. The DSC values for the JRST database, Montgomery County X-ray database and Shenzhen Hospital X-ray database were of 0.943, 0.953 and 0.931. In [68], for quantitative evaluation of the algorithm, the average registration error distance (MRED) is calculated for 15 pairs. This was done by manually annotating landmarks in 15 images. Comparison of this method with current benchmark methods that use only one category of transformations showed that using a combination of rigid and non-rigid methods the performance improved. The developed algorithm achieved a MRED of 24 pixels (for a 1,024 width image), while the two benchmark approaches that are used for comparison have a MRED of 41 and 473 pixels. In [65], 8 pairs of images are used for testing, by manually annotating points in them. The mean square error was calculated from the differences between the marked and transformed control points and

the corresponding marked points in the pair image. The best transformation acquired an error of 24 pixels (for an image of $2,000 \times 2,000$ pixels).

## 3.3 Final Considerations

A number of methods have been developed with the aim of utilizing sequential data in CXR analysis. Some of them focus on obtaining a better pathology detection or outcome prediction method by using longitudinal information, while others have the objective of predicting the difference between the images, which can be expressed as the presence of change or the presence of improvement or worsening in a pathology. In most of the methods, a DL model is used to extract features from the images. In general, the inclusion of longitudinal information seems to improve the performance of the algorithms.

Most of them aim at the prediction of labels that are not present in the most common CXR datasets, described previously. This might be a potential reason for the fact that a considerable amount of the methods use private datasets. The most common datasets do not present labels for comparison between longitudinal exams, thus, this can be considered a significant challenge in this field. Usually, the only reference to longitudinal information is either the age of the patient at the moment of acquisition or the date of the exam. The age of the patient is not a good factor to determine time between exams, thus, this can also be a problem if the time interval is relevant for the study.

Image registration is an important field regarding CXR analysis. Frequently, the acquired images show incorrect position of the patient, and deformations caused by the acquisition process might be present. Thus, the use of registration techniques is useful for introducing uniformity in the datasets. The analysis of simultaneous images also benefits from their alignment, when working with automated methods. Thus, alignment methods are closely related to automated longitudinal studies. The metrics that are used to validate the alignment methods are usually highly dependent on segmentations of the images, or manually annotated features. This factor might affect the robustness of method evaluation, as the manual annotation of features only allows the validation of a small amounts of images, and the usage of automated methods to get other features (as for segmentations or point coordinates) is always error dependent, leading to imperfect features.

# Chapter 4

# Chest X-Ray Image Pair Alignment

CXR scans often display various patient positions, despite the standardization of the image acquisition method. Rotation and tilt of the patient during acquisition, as well as wrong positioning, lead to images with different characteristics. When analyzing more than one image simultaneously, the comparison between scans might be difficult because of these factors. Thus, the development of an alignment algorithm to align pairs of longitudinal images might allow medical professionals or automated algorithms to pick on comparison features from the images easily.

An alignment algorithm was developed in this work, in order to align two CXR images. The alignment implies viewing the anatomical structures in the same regions of both images, thus helping to perceive any relevant differences between the pair and identify more easily the presence of a potential pathology. This developed method is mainly focused on the rigid alignment (explained in Section 3.2.2) of the lungs.

The features that are used to compute the transformations are lung segmentations and thoracic BB. These features are used to compute rotation, translation and scaling parameters, that align the two images. The developed method is described in this chapter, including different scaling techniques experiments and the comparison with a state-of-the-art solution for image alignment.

## 4.1 Methods

### 4.1.1 Datasets

To construct and evaluate this alignment algorithm, two different subsets were considered. These are constituted of images from the ChestX-ray14 dataset [27]. The first one is a subset generated by randomly picking 250 longitudinal image pairs from the original (**longitudinal subset**). The image pairs were constructed by picking scans with consecutive follow-up numbers, which refers to the sequential numeration of the images from the same patient, as previously mentioned in Section 3.1. The other used subset is a collection of 250 images that were randomly selected, to which a random, but known, rigid transformation was individually applied (**deformed subset**). This subset of deformed images allowed the evaluation of the alignment quality of the algorithm

with a ground truth reference (the original images), as it allows to check if the algorithm is able to reverse the artificially deformed images, making them as similar to the original ones as possible.

The applied deformations in the deformed subset images (rotation, horizontal and vertical translation and scaling) respect the following rules:

- the chosen angle is in the range $[-20, 20]$ degrees;
- the chosen horizontal and vertical translations are in the range $[-(\frac{\text{image size}}{100}), (\frac{\text{image size}}{100})]$ pixels;
- the chosen horizontal and vertical scaling factors are in the range $[0.75, 1.25]$;
- 85% of the lung area (after scaling) must be preserved after the rotation and translations are applied, otherwise a new random deformation is generated.

These deformations were chosen after experimentation, as they provide a reasonable transformation that keeps the lung area visible in the image.

### 4.1.2 Rigid CXR Alignment

In order to align the scans, features had to be extracted from the images, and posteriorly used as guides for the necessary transformations. In the developed method, the used features were the segmentations of the lungs, whose acquisition is described in Section 4.1.2.1. The segmentations provide information regarding the shape, size, and positioning of the lungs, thus, they can be used to compute alignment differences between two scans. The thoracic bounding boxes were also extracted from the images, with the aim of aiding the preprocessing and providing information for scaling, as described in Section 4.1.2.1. In the proposed solution, the segmentations are used to compute the parameters of the rigid transformations that might align one image with reference to another. Rigid transformations are used as they provide a simple alignment solution. The computed parameters take into consideration the presence of rotations, horizontal and vertical translations and different horizontal and vertical scaling factors.

In this algorithm, the lung masks are preprocessed and prepared for alignment. Then, vertical and horizontal scaling is performed, according to the reference, followed by the final transformations (rotation and translations). The final transformations to be applied are computed using Iterative Closest Point (ICP) algorithm [71], which is the main component of the alignment method. ICP is a classic technique for rigid image registration. This algorithm finds the rotation and translation parameters that better align two sets of points. It starts by associating the points from the different sets by the nearest neighbor criteria. Then, alignment parameters are computed based on a mean square cost function, aiming for the best possible overlap of the associated points. The points are transformed according to these parameters and a new iteration takes place. The algorithm converges when the alignment parameters stop changing. A convergence threshold is used for this matter, as well as a maximum number of iterations parameter.

A schematic of the full alignment method is represented in Figure 4.1. Different methods were experimented for scaling.

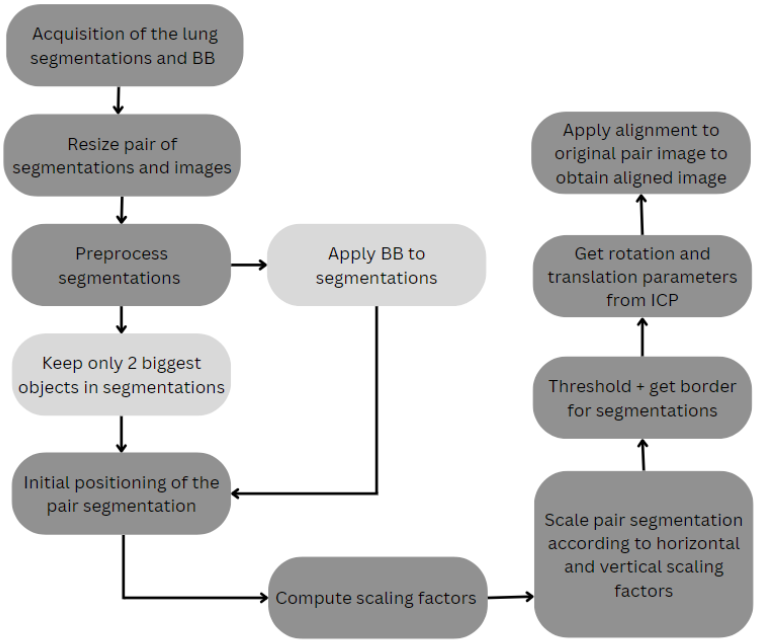



Figure 4.1: Developed method scheme.

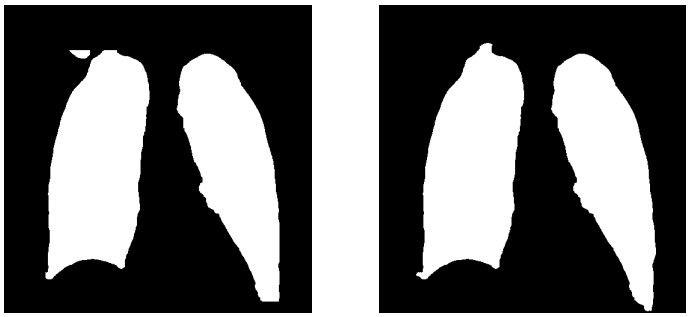
#### 4.1.2.1 Anatomical Segmentation and Thorax Localization

In the developed alignment method, the work in [72] was used in order to generate segmentations for the images. In the mentioned work, a U-Net architecture was used, and different experiments were performed, including the segmentation of the lungs, heart and clavicles. The model used to obtain the desired segmentations was trained solely on the JSRT dataset, using one image as input and outputting three segmentation masks, for the three mentioned anatomical structures. The performance of the model was evaluated using the DSC metric (described in Section 4.1.4). The lung, heart, and clavicles segmentations reached a DSC of $0.981 \pm 0.008$, $0.944 \pm 0.029$ and $0.927 \pm 0.027$, respectively.

For the thoracic BB generation, a localization algorithm in [73] was used. This model is based on a YOLO-V5, which was pre-trained on the COCO dataset [74] and trained with 956 CXR images. The corresponding ground truth BB were generated by drawing a BB around the manual lung segmentation masks provided in the JSRT, Montgomery, and Shenzhen datasets. This model reached an average precision of 99.84% at an Intersection over Union (IoU) superior to 0.5, in the validation dataset.

#### 4.1.2.2 Preprocessing and initial positioning

After generating the lung segmentations and the thoracic BB for the images, the input scans and the segmentations are resized to $512 \times 512$ pixels (as the usage of a large image size, like such, is common in CXR registration [67][68][70]). The segmentations are then preprocessed, in order to



(a) (b)

Figure 4.2: Application of the two possible cleaning methods in the same image. (a) Using the thoracic BB, and (b) keeping the two biggest objects.

eliminate small blobs and holes. This is done by performing the morphological operations opening followed by closing, using a circular structuring element of size 5 pixels. After preprocessing, the segmentations go through a cleaning process. There are two possible paths here, and the one that is followed depends on the scaling method that is used. The first consists in applying the thoracic BB as a mask, eliminating possible blobs and segmentation errors that lay outside that area by setting all pixels outside the BB to zero. The second consists in keeping only the two biggest objects in the segmentation, which ideally correspond to the lungs. In Figure 4.2 an example of the application of both processes is shown.

Note that YOLO-V5 can fail to detect any thoracic BB, particularly when the images are significantly rotated or contain large support devices. Thus, in these situations, the cleaning process that keeps only the two biggest objects in the image is used, independently of the scaling method.

As ICP is a gradient descent method, it should be used only when there is a good starting point. That is, if the original point sets are too different, the algorithm might not be able to converge to a good solution. If the initial point sets are originally similar, the result will probably be closer to the ideal. In order to guarantee that condition, an initial positioning was performed in the pair image. This consisted in getting the centroid of each segmentation, which is the central coordinate of the segmentation mask, and aligning the centroids of the two segmentations.

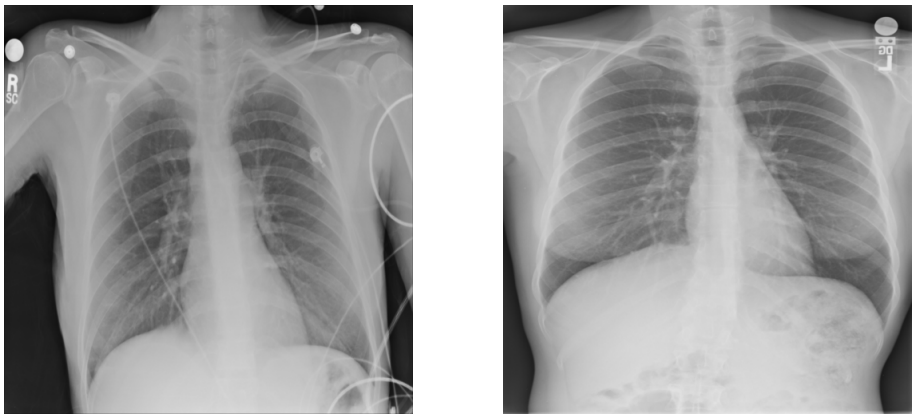


Figure 4.3: Two images from the same patient that show scaling deformation.

#### 4.1.2.3 Scaling

By looking at the scans, it is clear that different horizontal and vertical scaling between images from the same patient exist. These shape distortions that lead to the elongation or shortening of the target object can result from improper angulation of the image receptor or axis, from technical and/or structural errors of the X-ray tube [75], or from post acquisition errors. An example of an image pair with scale distortions is shown in Figure 4.3. Due to these distortions, and in order to align the image pair, horizontal and vertical scaling factors must be computed.

The used lung segmentations and BB present different characteristics, as the presence of different pathologies and the image exposure and contrast affect the segmentation and BB algorithms. Depending on the original image, their quality might vary and thus their utilization for the alignment algorithm must be adapted. Consequently, different scaling methods were developed, in order to experiment with each of them and study the best final scaling approach.

Initially, the comparison of the thoracic BB dimensions was used to scale the image pairs. However, when the lungs appear in a rotated position, or when the scan has poor contrast between the lungs and the surroundings, the BB fails to correctly delimit the lungs. Thus, a different approach was taken, where the dimensions used for scaling correspond to the BB generated by the limits of the segmentations (segmentation box). This works better for the cases where the thoracic BB failed, and, in opposition to the aforementioned method, the whole extension of the lungs is used to compute the scaling factor. An example is presented in Figure 4.4.

Although this method solves the scaling problem in some situations, the rotated lungs' scenario remains a problem, since they generate a higher or lower segmentation box dimension than the actual lungs dimension. This led to the necessity of obtaining the lungs dimensions more cautiously, using a rotation independent method.

The last scaling method that was implemented consists in using Principal Component Analysis (PCA) [76] to determine the size of the lungs in a segmentation. PCA is a commonly used dimensionality reduction method. It is usually employed in large multidimensional data, transforming them into a smaller set of variables that represent most of the original information. These new variables are the principal components, and they are linear combinations of the original variables that explain the variance in the original data. The generated set of principal components is such that the first principal component has the maximum variance explained, and the following will explain sequentially lower variances. Thus, the principal components (can also be called eigenvectors) represent the data in a new feature space, with a determined direction and the associated eigenvalues determine their magnitude, associated with the feature variance.

In the scaling method where PCA is used, the two principal axes of each lung in a segmentation are computed. After that, the extreme points of each lung are calculated, so that the dimensions of each lung can be computed. A scheme of this process is presented in Figure 4.5. The average of the dimensions of both lungs are used to determine the scaling factor in each direction, by doing the ratio of the lung dimensions for the input image and its pair.

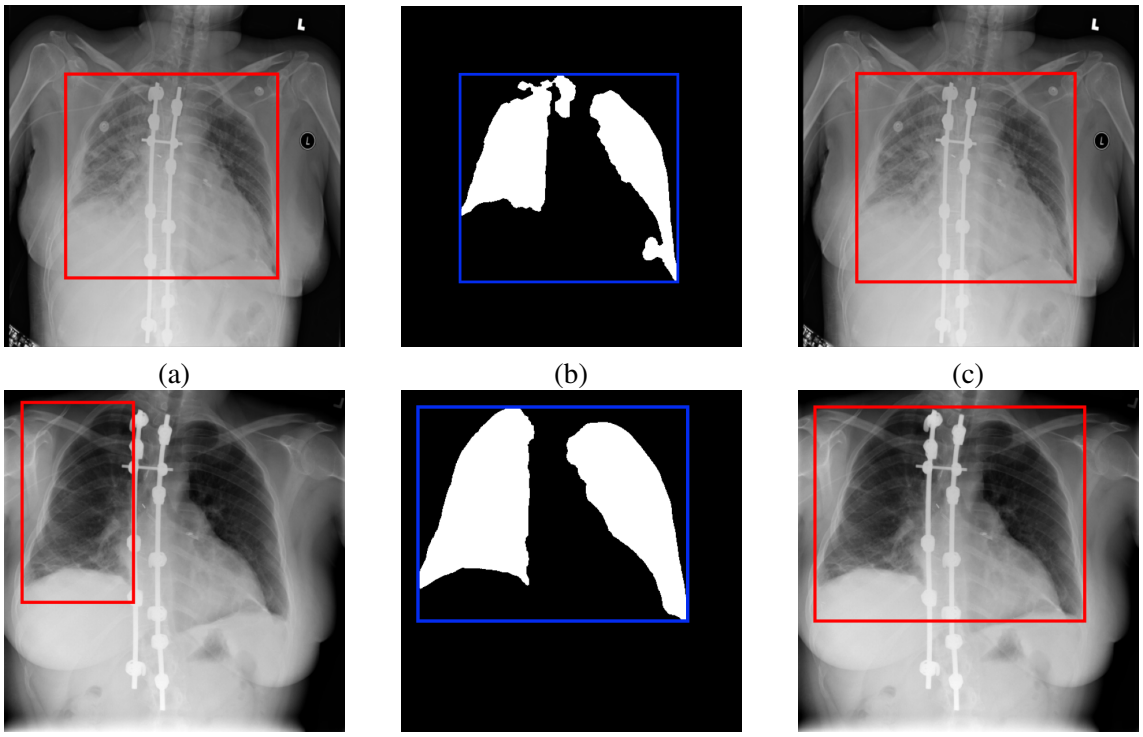


Figure 4.4: (a) Thoracic BB of the input image, (b) input image segmentation and corresponding segmentation box, (c) segmentation box of the input image, (d) thoracic BB of the image pair, (e) image pair segmentation and corresponding segmentation box, (f) segmentation box of the image pair.

To summarize, the used scaling approaches are: no scaling, scaling with the BB dimensions, scaling with the segmentation box dimensions, and scaling with the lung dimensions.

The diversity of the images and corresponding segmentations and BB makes it difficult to choose the best global scaling method. Thus, in order to obtain the best possible alignment for each pair, the images are aligned using all methods, but only the one that generates the best result is kept. This decision is made based on the DSC metric (explained in further detail in Section 4.1.4) and originates the results that are called **mixed results**. Note that if the image before alignment produces a better DSC metric than after alignment (for all scaling methods), then the image is not aligned.

As previously mentioned, the generation of the thoracic BB is not possible for all images. In order to provide an alignment solution for these cases, when one of the images in a pair does not contain a thoracic BB, all the scaling types are still used to align it, except for the one that uses the BB dimensions.

#### 4.1.2.4 Rigid Alignment

After scaling the segmentation pair according to the desired scaling method, it is thresholded and morphological operators (dilation and erosion) are used to obtain the border of the segmentations.

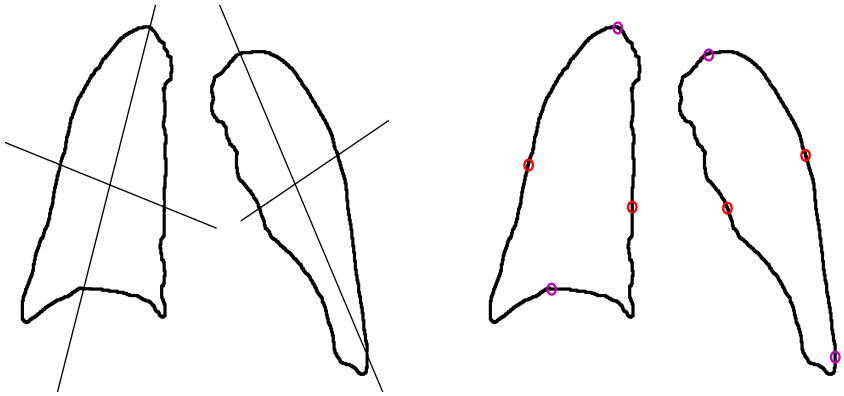


Figure 4.5: Segmentation border with PCA computed axis and extreme points in each lung.

The borders are used in the ICP algorithm, which returns the optimal rotation angle, horizontal translation and vertical translation. The convergence thresholds used are of 0.1, for both rotation and translation.

The final aligned image is generated by applying the initial positioning translations, followed by scaling, ICP rotation and ICP translations to the original pair image.

### 4.1.3 Scale-Invariant Feature Transform (SIFT)

In order to establish a comparison to a state-of-the-art solution, SIFT [77, 78] was used. This solution based on rigid transformations was adopted for feature alignment, using the lung segmentation borders.

This method obtains robust image feature points, which are invariant to scaling, rotation, limited affine distortion and changes in luminosity. The keypoints of two different images can be acquired and matched, and the affine transformation that converts one set of points into the other can be computed. This transformation can be applied to align an image according to its pair. SIFT starts by finding local maxima, using the Difference of Gaussian (DoG) algorithm at different image scales. More accurate points are chosen by using the Taylor series expansion of scale space. The neighborhood of the points is used to count the gradient direction of the neighboring pixels. Finally, a local coordinate is created with the main direction of each point [79]. After extracting the keypoints from both images, their descriptors are compared in pairs, using the L1-norm (sum of absolute value difference). The keypoints are matched by using the shortest L1-distance, and then used to compute the rigid transformation using singular value decomposition of the over-complete system of equation.

### 4.1.4 Evaluation and Metrics

The alignment algorithm was applied to the aforementioned deformed subset. Having the original image for each pair allows the computation of the distance between correspondent pixels in the two images. From a pixel's coordinates in the original image, the deformation applied to generate the deformed pair can be used to get the equivalent point coordinates in the deformed image and,

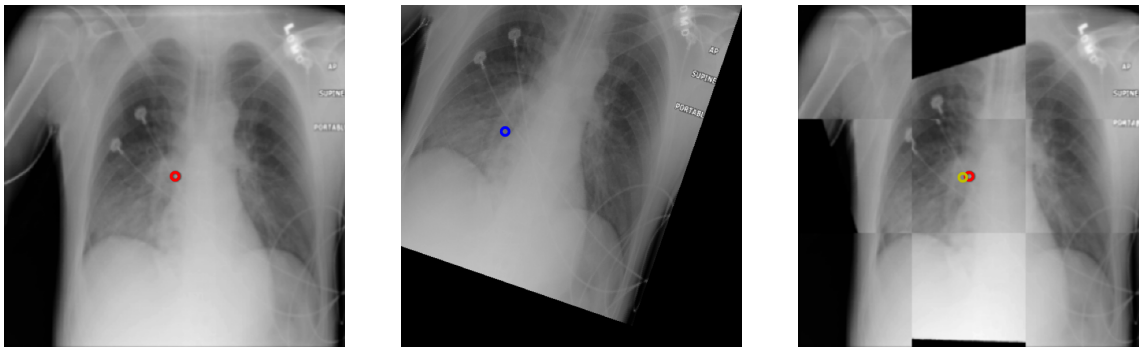


Figure 4.6: Representation of the same pixel after image deformation and alignment. The image order, from left to right, is: original, deformed, checkerboard (showing alternatively parts of the original and aligned images) with marked corresponding points (red is original and yellow is after alignment).

similarly, the final alignment parameters can be used to transform these coordinates into the new aligned image coordinates system. The average of all the pixel distances is used as a metric to evaluate the alignment performance – the Pixel Distance (PD). If the alignment is perfect, the PD should be zero. This distance is given in pixels. In Figure 4.6, the position of a certain pixel is shown in the original image. The position of the same pixel after the deformation (of the original image) and alignment (of the deformed image) is also represented. The measured distance corresponds to the distance between the position in the first image and the position in the aligned one.

The aforementioned metric can only be computed for the deformed subset, so, different metrics are necessary to evaluate the alignment performance on longitudinal pairs. If two images are similar, both in intensities and in orientation/scale, their difference should tend to 0. Thus, in order to compare longitudinal images before and after alignment, the Mean Squared Error (MSE) between the input image and its unaligned pair, as well as the MSE between the input image and its aligned pair were computed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2, \tag{4.1}$$

where $x_i$ represents the input image pixel intensities and $y_i$ represents the unaligned or aligned pair image pixel intensities.

If the MSE with the aligned image presents a lower value than the MSE with the unaligned image, then the alignment algorithm likely succeeded in making the longitudinal images pair more aligned. This metric was computed only in the union of the BB of the images, in order to avoid high values due to high or low intensity borders in an image, and making sure only the region of interest is considered. If one of the images in the pair does not contain a BB, then this metric is not computed. This metric, in opposition with the PD, is fully dependent on intensities. Thus, when the evaluated images have similar intensity maps, this metric is helpful in telling how aligned the images are. However, if that is not the case, this metric performs poorly as an objective

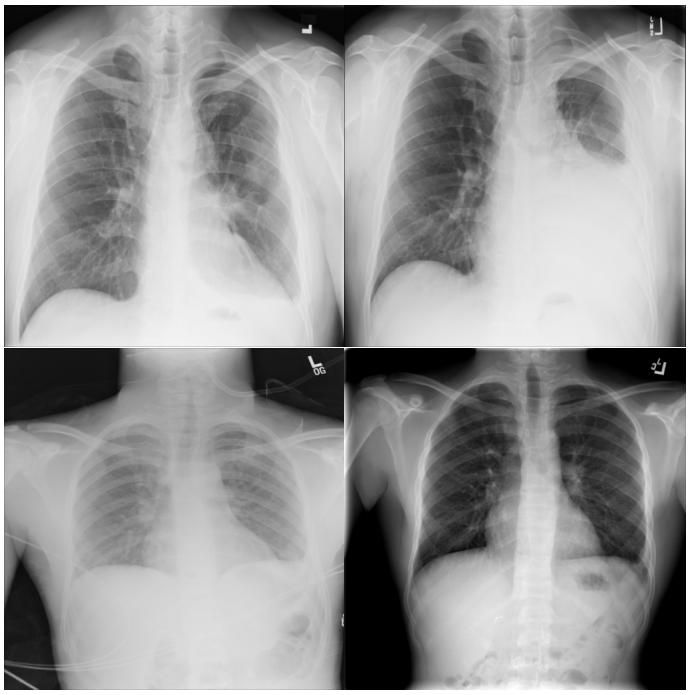


Figure 4.7: Example of two image pairs with different intensity maps. Each row presents a longitudinal pair of a unique patient

measurement of the images' alignment.

It is important to note that when comparing longitudinal images, pathologies of different natures may be present in the images. Some pathologies appear as darkened or lighter regions in the scan, due to their abnormal density. This might lead to different intensity maps between the images in the pair, leading to a higher MSE between images (whether the images are aligned or not). So, as the MSE metric might be enough to evaluate these type of situations, it requires manual observation of each case, as there are many factors that influence the image intensity. In Figure 4.7, examples of cases where the presence of pathologies affects the intensity maps are shown.

In order to obtain a more objective measurement of the image alignment, the DSC metric was employed as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{4.2}$$

where $X$ and $Y$ are the lung segmentations of a longitudinal pair. If the segmentations are perfectly overlapped, DSC tends to 1, as mentioned previously. The segmentation of the aligned image is generated by applying the same alignment transformations to the original pair segmentation. The segmentations used to compute this metric suffer from an initial preprocessing, including the application of the thoracic BB and the selection of the two biggest objects in the image. It is important to note that in spite of the fact that DSC is a more objective measurement of the segmentation overlap and, consequently, a more robust evaluation method of the alignment, it is affected by the error of the segmentations, as their generation is not ideal, and masks from

longitudinal scans are not equal.

## 4.2 Results and Discussion

The alignment algorithm results were generated individually for each scaling method in the developed algorithm, and for both subsets (longitudinal sample subset and deformed subset). The same images were used to generate the results for the SIFT algorithm.

The results for both subsets can be found in Tables 4.1 and 4.2.

Table 4.1: Deformed subset results. The images MSE parameter is the difference between the MSE of the original pair and the MSE of the aligned pair (aligned images MSE). A larger value should mean a higher impact in the alignment of the images. The DSC difference parameter refers to the subtraction of the DSC of the segmentations after alignment (DSC after) and before alignment.

| Scaling Method | Aligned images MSE | Images MSE | PD | DSC after | DSC difference |
|---|---|---|---|---|---|
| No scaling | 0.019 ± 0.020 | 0.032 ± 0.036 | 28 ± 18 | 0.832 ± 0.104 | 0.278 ± 0.211 |
| Thoracic BB | 0.012 ± 0.017 | 0.039 ± 0.031 | 22 ± 24 | 0.880 ± 0.106 | 0.324 ± 0.186 |
| Segment. box | 0.012 ± 0.014 | 0.039 ± 0.031 | 27 ± 39 | 0.876 ± 0.102 | 0.322 ± 0.195 |
| Lungs dim. | 0.011 ± 0.016 | 0.040 ± 0.034 | 25 ± 35 | 0.880 ± 0.123 | 0.326 ± 0.205 |
| Mixed results | 0.007 ± 0.009 | 0.044 ± 0.032 | 17 ± 21 | **0.908 ± 0.082** | **0.355 ± 0.195** |
| SIFT* | **0.006 ± 0.012** | **0.042 ± 0.036** | **12 ± 24** | 0.829 ± 0.149 | 0.230 ± 0.157 |

*could not align 4 of the 250 images.

Table 4.2: Longitudinal subset results. The images MSE parameter is the difference between the MSE of the original pair and the MSE of the aligned pair (aligned images MSE). A larger value should mean a higher impact in the alignment of the images. The DSC difference parameter refers to the subtraction of the DSC of the segmentations after alignment (DSC after) and before alignment.

| Scaling Method | Aligned images MSE | Images MSE | DSC after | DSC difference |
|---|---|---|---|---|
| No scaling | 0.025 ± 0.025 | 0.014 ± 0.020 | 0.859 ± 0.101 | 0.185 ± 0.167 |
| Thoracic BB | 0.021 ± 0.023 | 0.018 ± 0.020 | 0.894 ± 0.111 | 0.221 ± 0.173 |
| Segment. box | 0.024 ± 0.025 | 0.015 ± 0.022 | 0.884 ± 0.080 | 0.211 ± 0.166 |
| Lungs dim. | 0.024 ± 0.025 | 0.015 ± 0.021 | 0.872 ± 0.103 | 0.198 ± 0.171 |
| Mixed results | **0.021 ± 0.023** | **0.018 ± 0.020** | **0.910 ± 0.063** | **0.237 ± 0.161** |
| SIFT* | 0.029 ± 0.036 | 0.006 ± 0.035 | 0.798 ± 0.205 | 0.087 ± 0.199 |

*could not align 62 of the 250 images.

SIFT presents good metrics for both subsets, showing the lowest value for correspondent PD for the deformed subset (12 ± 24 pixels). However, it could not align all the image pairs. This is due to the fact that for these images, the algorithm couldn't find enough matching keypoints to compute the alignment parameters. In the case of the deformed subset, the detected keypoints

tend to be similar, as both images in the pair are, originally, the same. On the other hand, in the longitudinal subset, the images are different despite from being from the same patient. This leads to the detection of fewer corresponding keypoints, resulting in a higher feasibility of the SIFT algorithm in the deformed subset than in the longitudinal. An example of a longitudinal pair that could not be aligned is represented in Figure 4.8. The fact that SIFT performs more weakly in the longitudinal dataset presents an advantage of the proposed method as, in opposition to SIFT, it uses anatomical information to align the images.

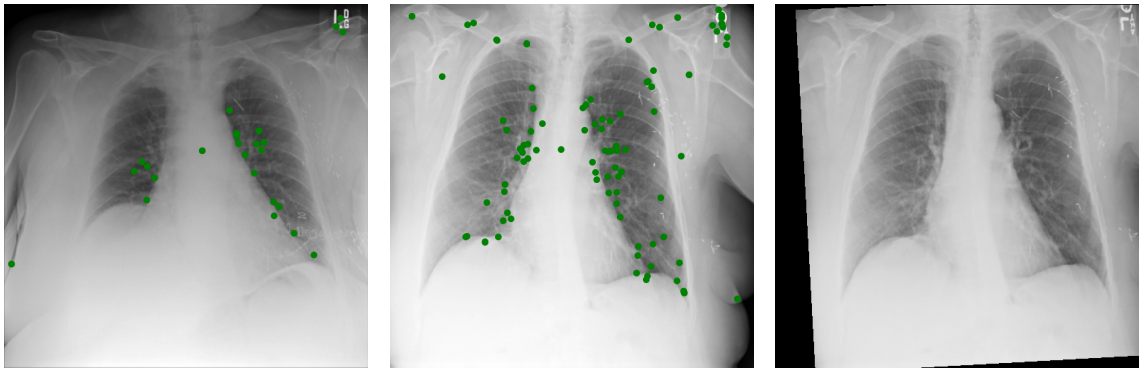


Figure 4.8: Example where SIFT failed to align the image pair. The first two images correspond to the computed keypoints, and the last one (right) represents the developed algorithm result, with a final DSC of 0.763.

In the deformed subset, the scaling method that provided the best results was the one based on the lungs dimensions. On the other hand, in the longitudinal subset, the best scaling method is the one based on the thoracic BB. Even though the MSE between images is not an objective metric to evaluate the alignment in all cases, the results show that it is coherent with the remaining metrics.

The mixed method results outperform the remaining, presenting a DSC of $0.908 \pm 0.082$ for the deformed subset and $0.910 \pm 0.063$ for the longitudinal subset. By looking at the image results, it is clear there is no ideal scaling method to align all images. Depending on the situation, there will be different problems. So, these confirm that the combination of the different methods provides a good solution. Table 4.3 shows the number of times a scaling method was selected as the best one, in the mixed methods results.

Table 4.3: Method Frequency on Mixed Results

| | Longitudinal subset | Deformed subset |
|---|---|---|
| No alignment | 0 | 0 |
| No scaling | 38 | 27 |
| Scaling thoracic BB | 124 | 71 |
| Scaling segmentation box | 54 | 54 |
| Scaling lungs dimensions | 34 | 98 |

Examples of good and bad DSC results are shown in Figure 4.11 and 4.12, respectively. The poor results are mostly caused by weak segmentations of the lungs. The DSC metric is computed based on the segmentations. Thus, it must be kept in mind that when the segmentations are weak,

this metric might not be indicative of the actual alignment. In Figure 4.10, an example of such a case is presented. Here, the alignment result seems visually good, but the metric leads to the opposite assumption. Nevertheless, the results show that all metrics are, in general, in agreement (both in the deformed and longitudinal subsets). Thus, the existence of such cases does not invalidate the fact that DSC is an adequate metric for the evaluation of this method.

Table 4.4: Alignment results for all consecutive image pairs in the used dataset.

| | **Aligned images MSE** | **Images MSE** | **DSC after** | **DSC difference** |
|---|---|---|---|---|
| Mixed results | 0.020 ± 0.021 | 0.016 ± 0.018 | 0.895 ± 0.080 | 0.212 ± 0.145 |

In Table 4.4 the final alignment results for all consecutive image pairs in the ChestX-ray14 dataset are presented. These results are similar to the ones obtained for the longitudinal subset (Table 4.2), which means the performance of the method is coherent throughout the data.

## 4.3 Conclusions

The developed alignment algorithm managed to align 250 longitudinal image pairs from the ChestX-ray14 dataset with a DSC of $0.910 \pm 0.063$. This result is satisfactory, as it means that there is a high overlap of the lung segmentations of the images in the aligned pair. The results shown for the deformed subset show that the method is able to realigned deformed images with a small error (PD of $17 \pm 21$ pixels for an image size of $512 \times 512$ pixels).

The metrics used to evaluate this algorithm (except for the PD, in the deformed subset) are dependent on the used segmentations or on the scans' intensity maps, which affects the objective evaluation of the method. The dependence on the pixel intensities does not allow the comparison of image pairs regarding structures alignment. That is because the scans might present abnormalities that lead to brighter or darker zones in the image, leading to an incorrect comparison regarding alignment. The dependence on the segmentations leads to lower performance when the segmentations are incorrect, which does not necessarily means that there is poor alignment of the structures. The metrics used to evaluate this sort of algorithm are usually dependent on features as the mentioned. Certain publications have used manually annotated keypoints in order to overcome this problem, however, this method is still not ideal as very few evaluations get to be performed.

The developed method was compared with a state-of-the-art approach for image alignment (SIFT). Although it outperforms the developed method in the deformed subset, it often fails at aligning images from the longitudinal subset. The proposed solution uses anatomical features extracted from the scans to perform the alignment, so, it manages to align all pairs (with a final DSC superior to the initial one), which is considered a relevant advantage.

Experiments where not only the lung, but also clavicles and heart segmentations were used were carried out. These points were integrated by being fed to ICP (adaptation for three point sets). The experiment showed that the presence of more points does not necessarily help the alignment, since poor segmentations worsen the ICP results and the presence of more point sets leads to a higher probability of having a bad segmentation. Thus, a suggestion for future work

consist in using these three segmentations in a more complex manner, allowing the extraction of relevant and more consistent features, that can be used to compute parameters for better alignment (that considers for example the localization of the heart and not only the lungs).

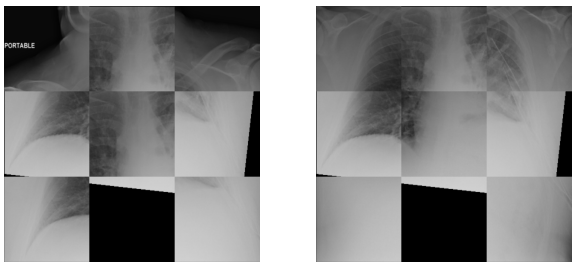



Figure 4.9: Example of a large DSC difference (0.704) between not aligned and aligned images. The left image is the checkerboard of the original pair (showing alternatively parts of the original images), and the image on the right is the checkerboard of the aligned pair (showing alternatively parts of the original and aligned images).

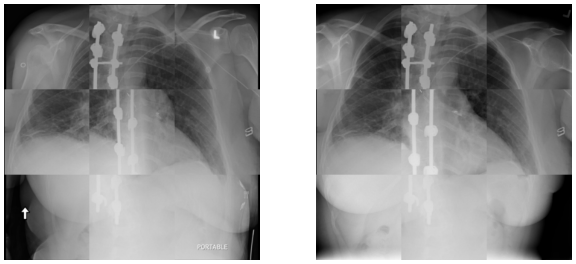
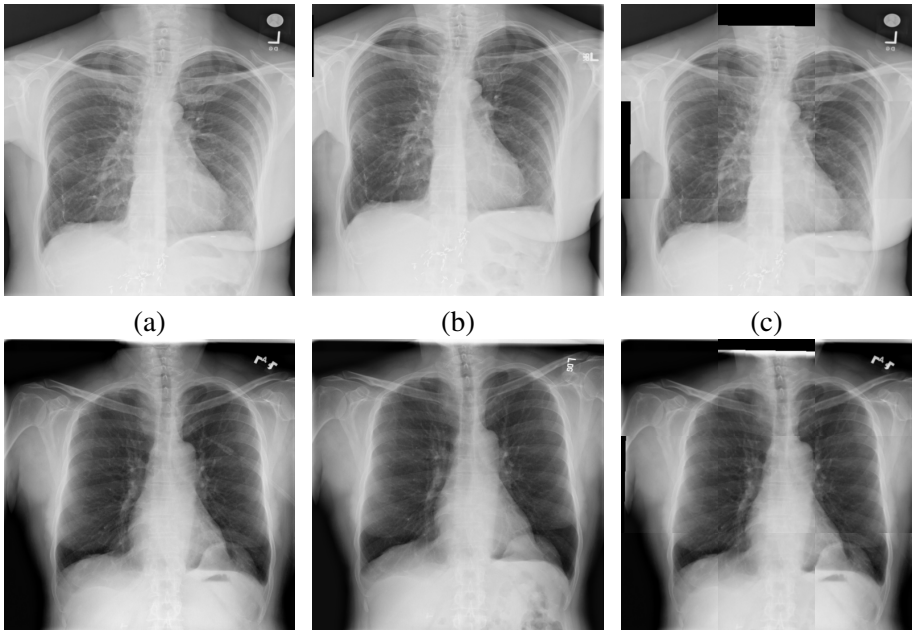


Figure 4.10: Example of an alignment result with a DSC (after alignment) of 0.609 and a more satisfying visual result. The checkerboard showing alternatively parts of the original image and it's longitudinal pair, and the checkerboard of the original and aligned images (showing alternatively parts of each) are presented, from left to right.



(a) (b) (c)

(d) (e) (f)

Figure 4.11: Examples of good DSC results. (a) and (b) represent two input image examples, (b) and (e) correspond to their pairs, and (c) and (f) represent the alignment results by a checkerboard (showing alternatively parts of the original and aligned images). The DSC value before and after the alignment is 0.770 and 0.976 (respectively), for the first example, and 0.883 and 0.976 or the bottom example.

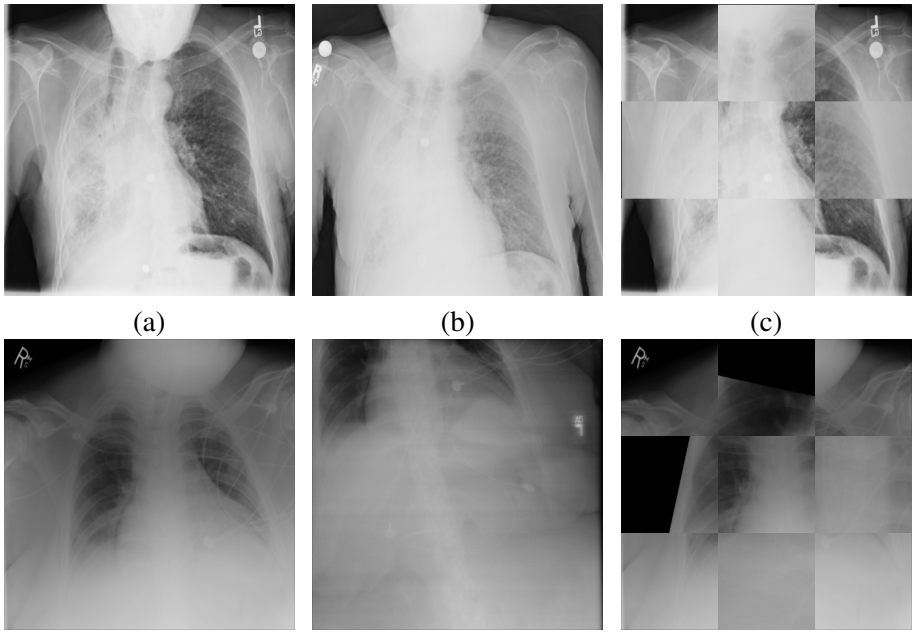


Figure 4.12: Examples of poor DSC results. (a) and (b) represent two input image examples, (b) and (e) correspond to their pairs (respectively) and (c) and (f) represent the alignment results by a checkerboard (showing alternatively parts of the original and aligned images). The DSC value before and after the alignment is 0.433 and 0.673 (respectively), for the first example, and 0.094 and 0.439 or the bottom example.

# Chapter 5

# Pathology and Change Detection

As previously mentioned, the usage of longitudinal CXR information in automated analysis is not common in state-of-the-art publications. However, as medical professionals normally look at multiple exams from the same patient, in order to compare them, the inclusion of this data in automated algorithms should be valuable. Thus, in this chapter, longitudinal information is used for studying the evolution of a pathology through a pair of consecutive scans from the same patient.

With the aim of predicting the pathologic differences between longitudinal images, that is, identifying the abnormalities in an image and whether these abnormalities remained in the follow-up image, different experiments were performed.

In these different experiments, various manners of integrating longitudinal information in the predictions were conducted. Initially, a multilabel model that predicts the presence of all abnormalities in ChestX-ray14 in a single scan was used, with the aim of establishing a comparison reference (model that does not use longitudinal information for training). The following experiments integrated longitudinal data at different levels, including at the image feature level and at the input level. The developed alignment algorithm (c.f. Chapter 4) was also used in the experiments, to test if aligned pairs provide an advantage in these methods.

In longitudinal experiments, image pairs were used. Each pair is associated with two labels: the pathology detection and the change detection.

- The **pathology detection** is related to the first image in the pair. Thus, an image was considered positive for a pathology if it has a positive label (1) for it, and negative (0) otherwise.

- The **change detection** is related to the comparison between the pathology labels in the two images in the pair. The presence of change is positive (1) if the pathology label is different between the two and negative (0) otherwise.

## 5.1 Methods

### 5.1.1 Datasets

The used dataset is the ChestX-ray14. In this dataset, 14 abnormalities are labeled, including atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening and hernia.

A longitudinal dataset, based on the original one, was created. Here, only images of patients that have at least two scans are included, which are grouped into pairs of longitudinally acquired images. Each sample of the longitudinal dataset thus corresponds to two images with consecutive follow-up numbers, being the first one (input image) the oldest scan, and the second one (image pair) the most recent one. Thus, the original dataset which contains 112,120 images was transformed into a longitudinal dataset composed of 81,315 samples.

In all the reported experiments, the images suffered the same pre-processing. The 3-channel images were resized to $256 \times 256$ pixels and normalized with the mean and standard deviation values of the ImageNet dataset [34], as it was used to pretrain the used models. Data augmentation was used during training, by applying random affine transformations, including rotations from -5 to 5 degrees and shear parallel to the x-axis, from -3 to 3 degrees. The used batch size was 8 in all training routines.

The images were split into five folds, maintaining the original class distribution in each one and ensuring no patient overlap between folds. This class and patient distribution was preserved both in the original and longitudinal dataset. Three folds were used for training, while one was used for validation and one for testing.

In order to facilitate the interpretation of the results, as well as allowing the development of multiple experiments and compare various methods, the performed experiments were based on an individual pathology: cardiomegaly. Cardiomegaly is an abnormality that can be found in CXR, and it refers to the enlargement of the heart.

In Figure 5.1 an example of an image pair is presented. This pair is part of the cardiomegaly longitudinal dataset.

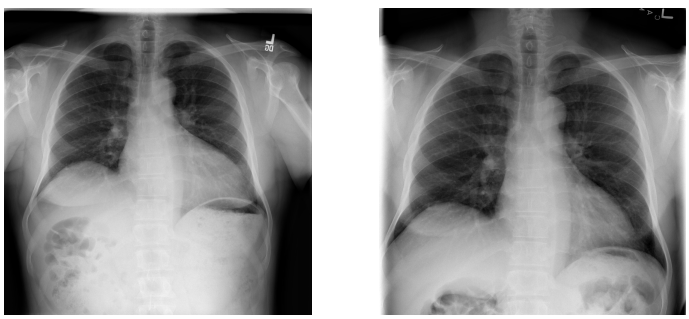


Figure 5.1: Example of a longitudinal pair. The first image has a positive label for cardiomegaly, while its pair has a negative label for the abnormality. Thus, the change label is positive.

To detect cardiomegaly, the Cardiothoracic Ratio (CTR) is measured (or visually assessed). The CTR is the ratio between the maximum transverse diameter of the heart and the maximum transverse diameter of the chest. If this value is higher than 0.5, then cardiac enlargement is present. In Figure 5.2 an example of the CTR measurements is displayed. It should be kept in mind that CTR alone is not reliable to provide the diagnostic, however, the usage of multiple CXR might allow the visualization of the evolution and thus help provide a more significant diagnostic [3].

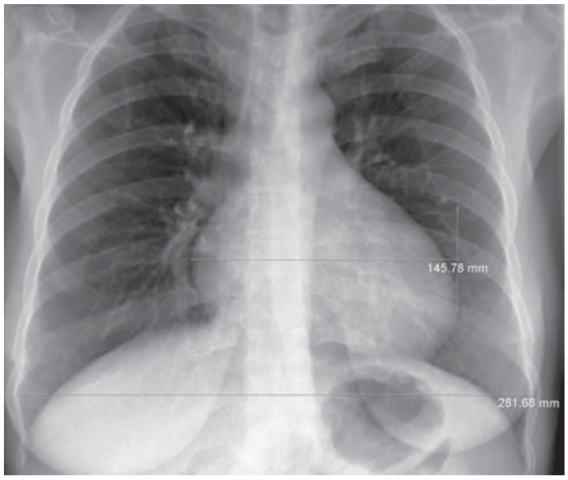



Figure 5.2: CTR measurements example [3].

For the original dataset, the percentage of images with a positive cardiomegaly label is 2.48%. When considering the longitudinal dataset, the distribution of the samples through the classes is presented in Table 5.1.

Table 5.1: Cardiomegaly and change cases numbers in the longitudinal dataset

| | **Positive Cardiomegaly** | **Negative Cardiomegaly** | **Total** |
|---|---|---|---|
| **Positive Change** | 1,419 (1.75%) | 1,375 (1.69%) | 2,043 |
| **Negative Change** | 624 (0.77%) | 77,897 (95.80%) | 79,272 |
| **Total** | 1,999 | 79,316 | 81,315 |

### 5.1.2 Pathology and Change Detection

The network that is used as a backbone for all the experiments is the ResNet-50. The ResNet-50 is a CNN, more specifically, a Residual Network. This kind of network includes skip connections, which consist of adding the original input to the output of the convolutional block. This happens only if the size of the input is the same as the size of the output, thus, transformations (like padding or convolutions) can be applied in order to make it possible. A building block example is shown in Figure 5.3. Here, $\mathcal{F}(x)$ represents the residual map to be learned and $x$ represents the input vector. The skip connections provide a solution to the vanishing gradient problem, that happens when,

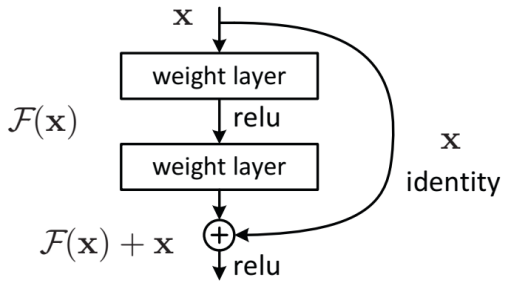



Figure 5.3: ResNet building block [5].

during backpropagation, certain activation functions map the input value into a smaller output space, leading to a small derivative. When the network has multiple layers, this causes a small gradient and, consequently, an inefficient update of the weights and poor training.

The ResNet architecture is constructed based on multiple convolutional layers. In the case of ResNet-50, it contains 48 convolutional layers, one MaxPool layer, and one average pool layer. The algorithms presented in this work were implemented using the *PyTorch* framework and an NVidia GeForce GTX 1080 GPU (8 GB).

In the following paragraphs, the various conducted experiments are summarized:

- **Experimental setting 1:** A multilabel model was used as a baseline for comparison with other experiments. This model was used to predict the presence of a single pathology in an image. These model predictions were used to compute the change class in longitudinal images;
- **Experimental setting 2:** The same multilabel model was used as a feature extractor, and the combination of image features were used to compute the presence of pathology in the input image, and the change class between the pair;
- **Experimental setting 3:** A model was trained using longitudinal image pairs. The images were concatenated and fed into the model, with the objective of predicting the presence of pathology in the first image, and the change class for the comparison between the two images;
- **Experimental setting 4:** The previous experiment was repeated with image pairs that were aligned, using the proposed alignment method (c.f. Chapter 4).

#### 5.1.2.1 Experimental Setting 1 - Baseline

A baseline model was established for comparisons with further experiments. This model focuses on a multilabel problem, predicting the 14 labeled pathologies in ChestX-ray14. A multilabel based model was used as it allows simple expansion of the performed experiment to other CXR abnormalities. As previously mentioned, a ResNet-50 backbone is used and the model was pre-trained with ImageNet [34].

The used loss function is the Binary Cross Entropy (BCE) loss, shown in equation 5.1, where $y$ (1 or 0) is the ground truth label, $p(y)$ is the predicted probability and $N$ is the number of samples. This loss function is commonly used in binary classification problems. The Adam optimizer [80] is used to compute the updated weights and biases. It is an algorithm that is adequate for noisy gradients, and requiring little tuning of hyperparameters. This optimizer was used with a defined learning rate of $10^{-4}$. A scheduler was used to reduce the learning rate throughout training, which allows the model to reach the best possible performance. The validation loss is used as the controlling parameter, and the learning rate is reduced by a factor of 0.1 when the validation loss does not decrease for over 3 epochs. Training was extended for a maximum of 10 epochs, saving the model weights on the epoch that provided the best validation loss.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} -(y_i \log(p(y_i)) + (1-y_i)\log(1-p(y_i))) \tag{5.1}$$

The described training conditions (including the model architecture, pretraining, loss function, scheduler, and optimizer) were maintained for all performed experiments, unless stated otherwise.

This baseline model was used to generate the probabilities of presence of each pathology, in each image of the test set. As in these experiments, individual labels were considered at a time, only the results relative to cardiomegaly were kept. The change class was also predicted for the longitudinal test set. This was done by computing the absolute value of the difference of probability of pathology in a longitudinal pair, as shown in equation 5.2. A scheme is represented in Figure 5.4

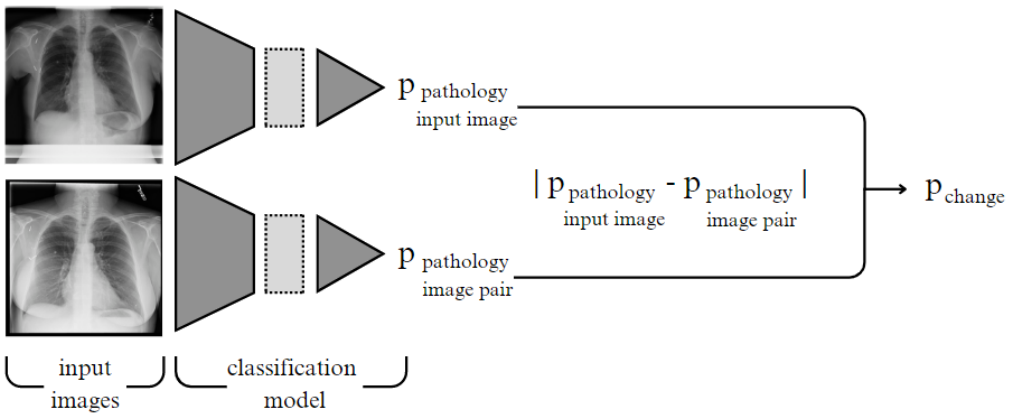



Figure 5.4: Baseline Model Scheme.

$$p_{change} = |p_{\text{pathology input image}} - p_{\text{pathology image pair}}| \tag{5.2}$$

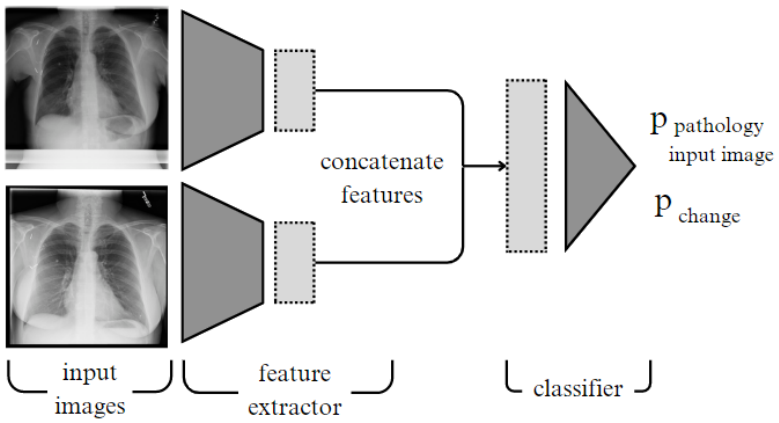



Figure 5.5: Features Model Scheme.

#### 5.1.2.2 Experimental setting 2 - Features Model

In this experiment, the encoder portion of the multilabel model previously described was used as a feature extractor. This was done by freezing all of its layers, and removing the final fully connected layer. The deeper layers are used to generate the features for each image in a pair. The features are then concatenated (forming a features vector with the double of the length), before being fed to a new trainable dense classifier layer, which outputs the presence of the pathology in the input image and the change class for its pair. In Figure 5.5, a scheme of the model is represented.

Training included 35 epochs, in this experimental setting, saving the model weights on the epoch that provided the best validation loss. The defined learning rate was $10^{-6}$.

#### 5.1.2.3 Experimental setting 3 – Longitudinal Model

In this experiment, both images from a longitudinal pair are used as the input to a model, aiming at the detection of both the presence of a pathology and the presence of change. The images in the pair are fed to the model after being concatenated, originating a 6-channel variable.

The used model (ResNet-50 [5] backbone) was adapted to include a 6-channel input image, and return the two desired outputs. A scheme of this model is represented in Figure 5.6.

Training included 15 epochs, saving the model that provided the best validation loss. The defined learning rate was $10^{-6}$.

#### 5.1.2.4 Experimental setting 4 - Longitudinal Aligned Model

The previously described experiment (longitudinal model) was repeated, but with image pairs that were aligned using the developed alignment method described in Chapter 4. Each image pair was aligned according to its reference, which is the first image in the pair. A scheme of this model is represented in Figure 5.7

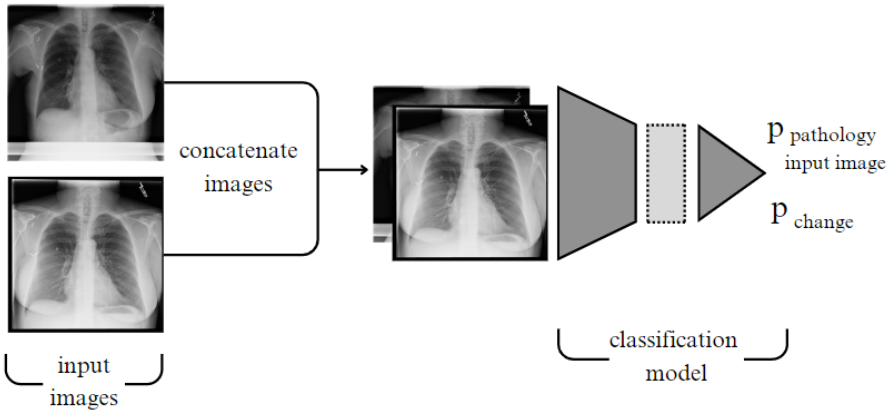



Figure 5.6: Longitudinal Model Scheme.

### 5.1.3 Explainability

When working with medical images, the most common form of XAI is saliency mapping [81]. Saliency maps are a visual explanation that highlight which parts of the image are important for a decision. The majority of these visual explanation techniques uses backpropagation, but there are also methods using perturbation-based or multiple instance learning-based approaches.

The simplest example of backpropagation techniques is image-specific class saliency maps [28]. In this technique, a forward pass is performed and the gradient values of the loss function are computed in respect to the input image, generating a final map. Grad-CAM [21] is also a backpropagation technique, that was introduced as a generalization of Class Activation Mapping (CAM) [82], which is a commonly used method that can be applied to any CNN.

In order to produce explainability maps for the constructed models, saliency maps were applied. This method was chosen since it allows the acquisition of maps for all experimental settings, independently of the input form: concatenated images (6-channel image), two images, or one image. Grad-CAM does not allow the separation of the images for its computation, since it uses the last convolutional layer of the network. Saliency maps were generated for both classes (pathology and change) and all experiments. This was done by acquiring the gradients for each image, normalizing them according to the maximum value, and applying a Gaussian filter with a sigma of 5 and a kernel with a quarter of the image size.

## 5.2 Results and Discussion

The final results for each experiment were computed by applying the corresponding model to the test fold, containing 16,283 samples in the longitudinal dataset. The explainability maps were generated for images in the test set. In Table 5.2, the results for the models can be seen.

Regarding the detection of cardiomegaly, the baseline model (experimental setting 1) outperforms the remaining, with an AUC of 0.897. This is probably due to the fact that this model was

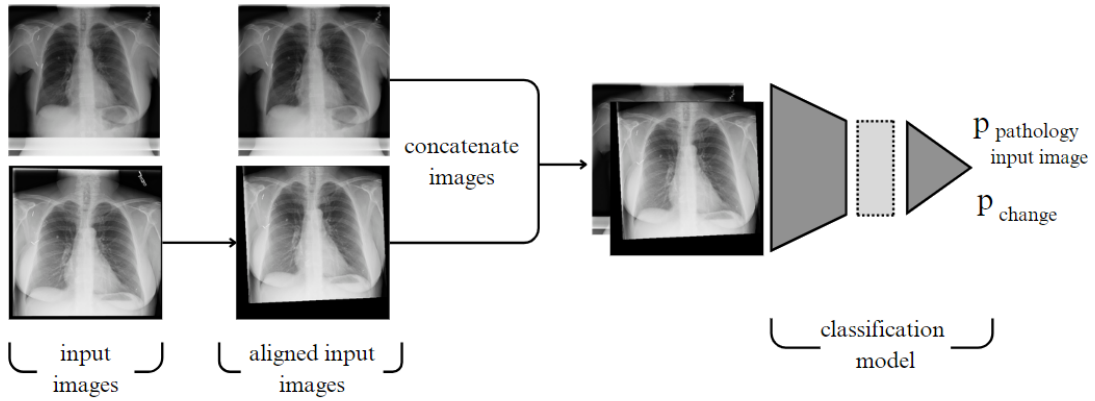



Figure 5.7: Longitudinal Model Scheme.

trained with the original dataset, which contains more samples than the longitudinal version, and might have provided a better cardiomegaly detection. Concerning the detection of change, the features model outperforms the rest of the experiments, with an AUC of 0.858.

The features model (experimental setting 2) performance in the pathology class is probably due to the usage of the same features as the baseline model (experimental setting 1) to reach the predictions. As the features from both images are used to compute the presence of cardiomegaly (and change), the similarity between these two cases was expected. Regarding the detection of change, the reached results show that the usage of features for cardiomegaly detection can be used to model the difference between two scans, predicting the label change when comparing the images in the pair.

In the longitudinal model, the images are concatenated before feature extraction. The usage of all longitudinal information as input was thought to be an advantage, since the feature extraction could capture the difference between the scans from the beginning. However, this factor can be the main reason for the weaker results shown by this approach, as it might make the feature extraction process more difficult. It is clear however that the usage of aligned images in the longitudinal model (experimental setting 4) improved the results. The alignment of the images leads to an alignment of the relevant structures in the scans and thus, it might facilitate the detection of features for comparison of the two images.

Table 5.2: Cardiomegaly and change detection results.

| | 1 - Baseline | | 2 - Features | | 3 - Longitudinal | | 4 - Longitudinal Aligned | |
|---|---|---|---|---|---|---|---|---|
| | Card. | Change | Card. | Change | Card. | Change | Card. | Change |
| AUC | **0.897** | 0.824 | 0.893 | **0.858** | 0.833 | 0.795 | 0.868 | 0.820 |
| Precision | 0.096 | **0.102** | **0.137** | 0.088 | 0.079 | 0.09 | 0.078 | 0.076 |
| Recall | **0.842** | 0.716 | 0.734 | **0.825** | 0.734 | 0.679 | 0.803 | 0.818 |
| Accuracy | 0.799 | **0.788** | **0.879** | 0.722 | 0.782 | 0.769 | 0.762 | 0.678 |

By looking at images from the same patient and corresponding ground truth labels, some

situations where images appear practically equal, but have a different ground truth label can be found. In Figure 5.8, an example of such a case is represented. It is important to acknowledge the existence of these cases, since in this dataset the labels were automatically generated from medical reports, and so, errors in the annotations can be present. Furthermore, given that the change label is computed by using the two pathology labels of an image pair, this label is associated with a higher error. This factor might introduce noise in the longitudinal dataset, that can lead to lower performance when this data is included for training.

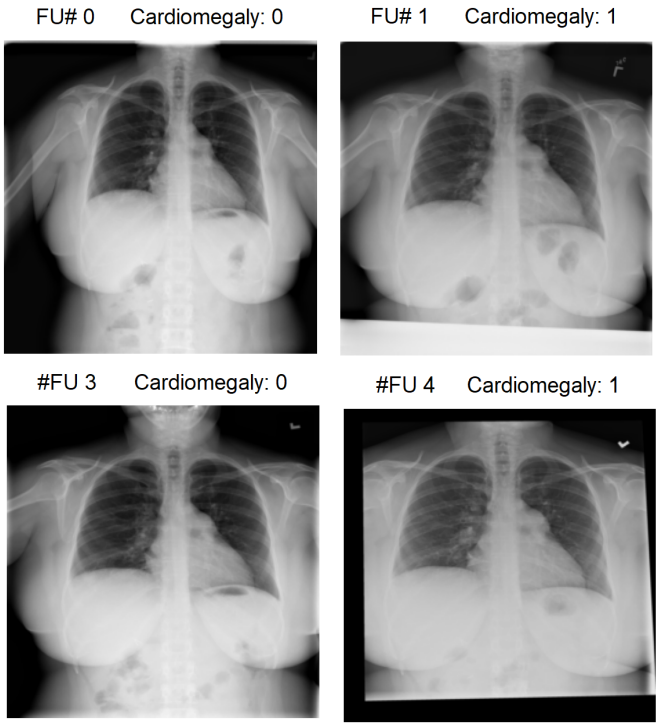



Figure 5.8: Example of a case where similar images from the same patient display different ground truth labels for cardiomegaly.

In Figure 5.9 and Figure 5.10, examples of cardiomegaly saliency maps are represented. The selected images are, in the first case, true positive and, in the second case, false positive cases, for the features model (experimental setting 3). In these figures, the same image pairs are represented, but overlaid with saliency maps generated by all different experimental settings. For each experimental setting, four pairs of images are presented, except for the baseline model, as it takes only one scan as input. Even though the cardiomegaly label is relative to the first image in the pair, saliency maps are shown for both images, as both of them were used as input.

The maps for the baseline model and the features model (experimental settings 1 and 2) are similar, which was expected since the features model is derived by adapting the baseline model. The maps for the longitudinal models (experimental settings 3 and 4) seem, in general, more disperse than the remaining experiments. However, the case that uses aligned pairs seems to present more focused maps, in comparison with the model that uses non-aligned pairs.

As the ground truth label is associated only with the first image, the saliency maps were expected to present low activations for the second image in the pair. However, by looking at the examples, it is clear that features from both images are used for the cardiomegaly prediction. This observation can be somewhat related to possible wrong annotations in the dataset, as previously mentioned, as they could lead the model to catch features from both images during training.

In Figure 5.11 and Figure 5.12, examples of change saliency maps are represented. Similarly to the pathology examples, the selected images are, in the first case, true positive and, in the second case, false positive cases, for the features model (experimental setting 3). In this case, four pairs are presented for each of the experimental settings where the change label is computed (2, 3 and 4).

In this scenario, as the change label concerns both images in the pair, it was expected that both would show activations that would highlight their differences. Similarly to the cardiomegaly saliency maps, the maps from the features model (experimental setting 2) are focused mainly in the heart area. The longitudinal models present saliency maps with more disperse activations, but still focused mainly on the logical anatomic region. By comparing the longitudinal models, the activations appear to be more focused on the same structures in the maps where aligned pairs are used.

In summary, for the baseline and features models (experimental settings 1 and 2), the metrics are concordant with the saliency maps. The features extracted by the baseline model seem to have a good performance when used to predict both the presence of cardiomegaly and change. This is concordant with the shown saliency maps, as, in these cases, they seem to be more focused on the anatomical logic area. The longitudinal models (experimental settings 3 and 4) show worse results in comparison, both in terms of metrics and saliency maps, providing more disperse activations. In the case of both classes, pathology and change, the generated saliency maps are similar in the two images used as input, which means that the models use mostly the same regions in both images to provide the predictions.

## 5.3 Conclusions

Multiple experiments were carried out with the aim of integrating longitudinal data for detection of pathology and change (comparison between the presence of pathology in two scans). This integration was performed at different levels. In the baseline model (experimental setting 1), no longitudinal data was integrated during training. In the features model (experimental setting 2), longitudinal information was included at the features level, by combining the features extracted from two images in a pair. In the longitudinal models (experimental settings 3 and 4), it was included at the input level, by concatenating the images in the pair in a 6-channel input.

It is important to note that, in the performed experiments, only one of the pathology labels from ChestX-ray14 (cardiomegaly) was used. In future work, it should be a priority to validate the carried out experiments in the remaining abnormalities.

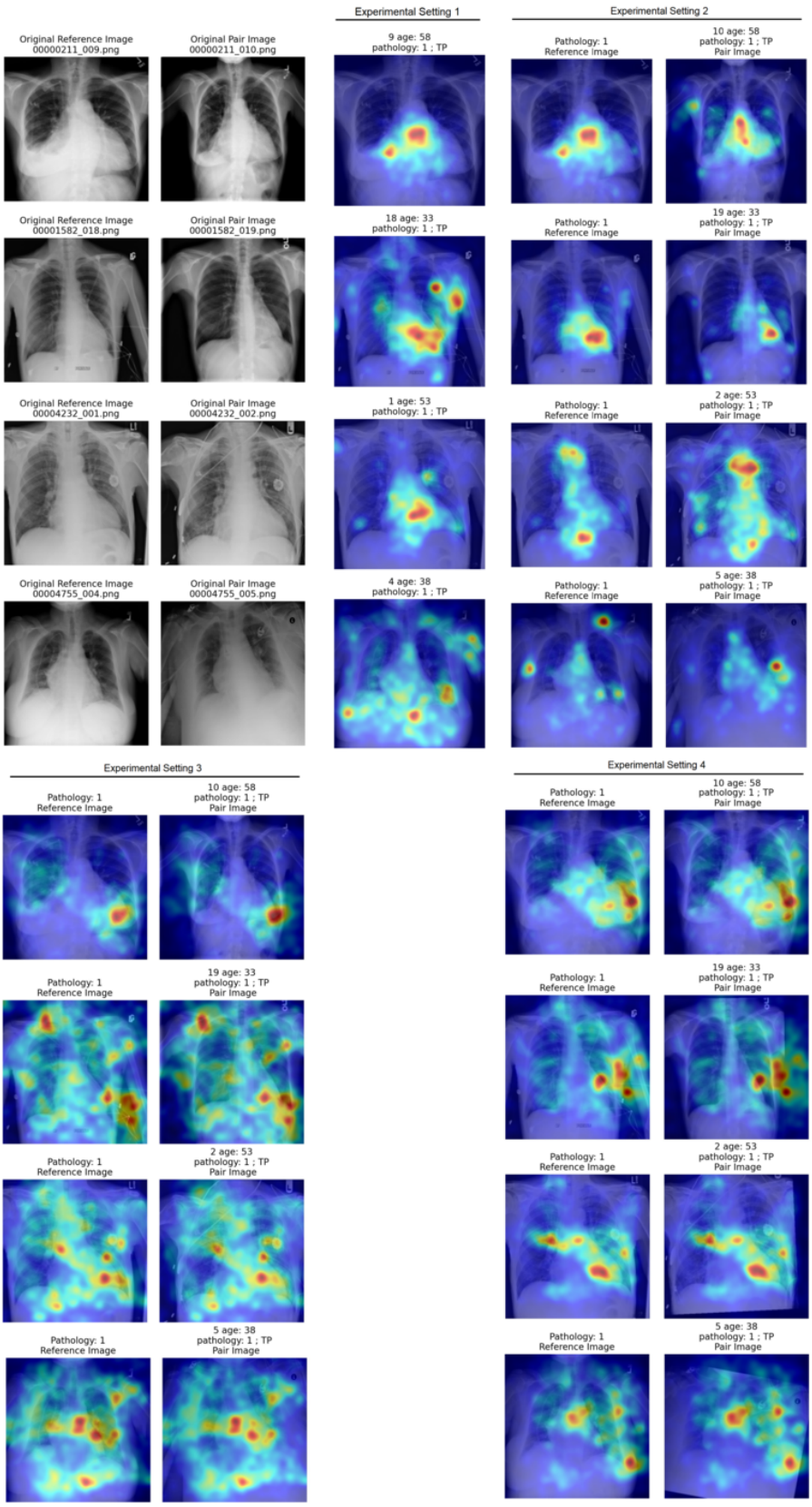



Figure 5.9: Comparison of saliency maps for all experimental settings on cardiomegaly detection. True positive cases.

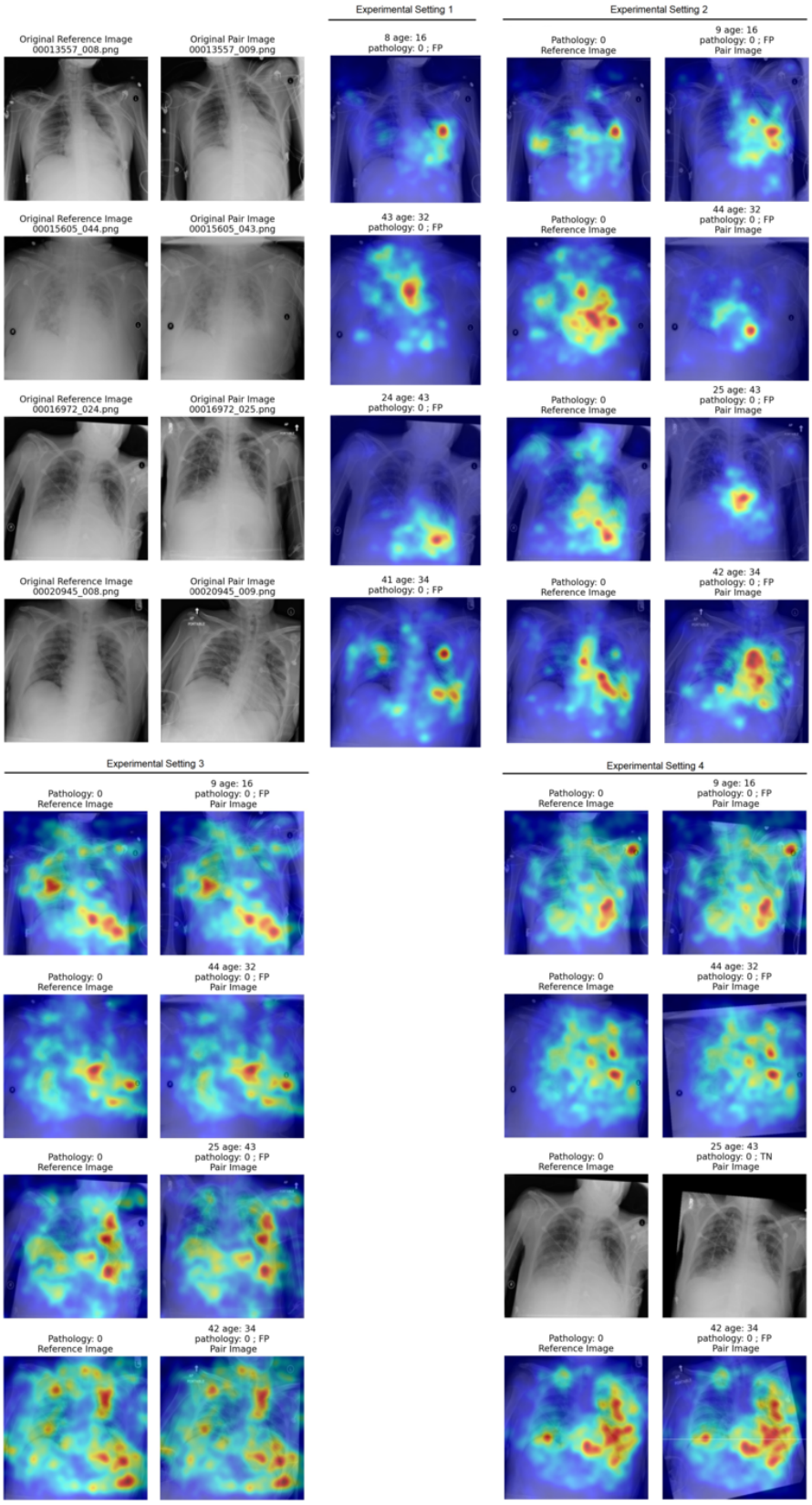



Figure 5.10: Comparison of saliency maps for all experimental settings on cardiomegaly detection. Mainly false positive cases.

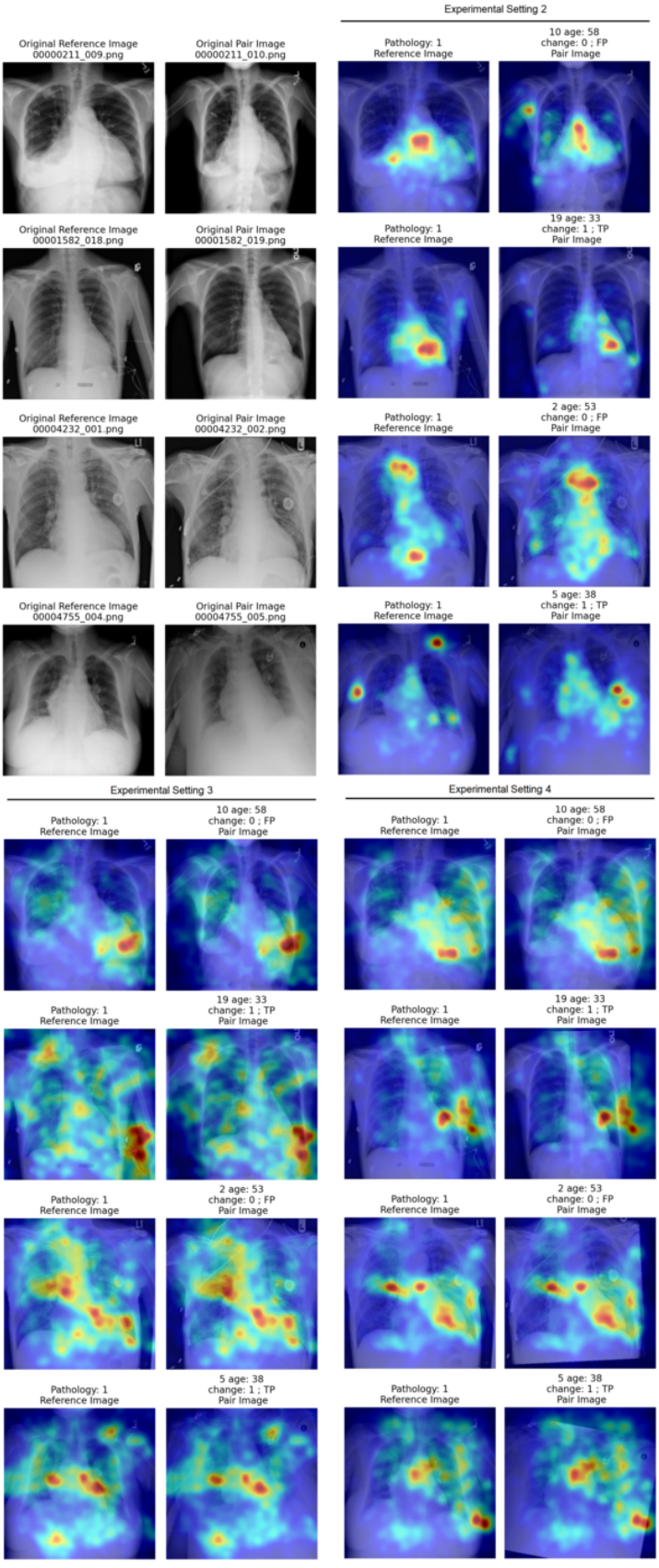



Figure 5.11: Comparison of saliency maps for all experimental settings on change detection. True positive cases.

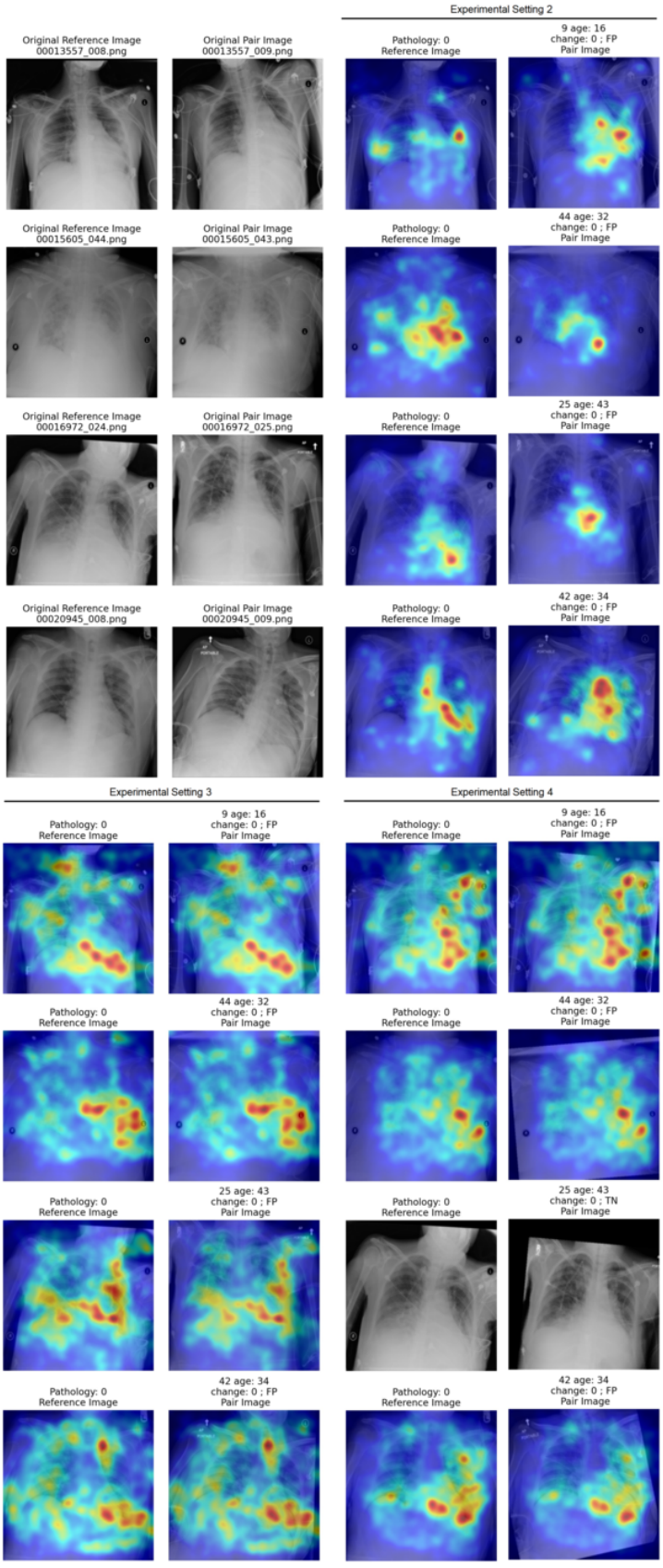



Figure 5.12: Comparison of saliency maps for all experimental settings on change detection. Mainly false positive cases.

The best performance obtained for cardiomegaly detection was reached by the baseline model (experimental setting 1), with an AUC of 0.897, even though the features model (experimental setting 2) reached a similar AUC of 0.893. Regarding the detection of change between the scans pair, the features model outperformed the remaining experiments, with an AUC of 0.858. The usage of the baseline model for feature extraction, allowing the computation of the presence of change, shows that longitudinal pathology data provides an advantage for automatic comparison of exams. On the other hand, the usage of 6-channel images for training (longitudinal models – experimental settings 3 and 4) showed that the concatenation of the input images affects the prediction of the pathology and change. The usage of aligned images improved the performance of the model that uses the concatenated images. The usage of aligned images improved the cardiomegaly detection AUC by 3.5%, and the change detection AUC by 2.5%. The alignment of the structures probably facilitates the extraction of features relative to both images, which shows that it can be advantageous for longitudinal problems. Information from the two images in the pair is usually used to predict both the presence of pathology in the first image and the change in the pair.

In the longitudinal dataset, there are similar images that contain different ground truth labels. This is important to note, as these situations might constitute errors in the dataset, that lead to a higher error. The presence of wrong labels is problematic mainly for the change class, as it is dependent on two pathology labels, leading to an augmentation of the error.

# Chapter 6

# Augmentation Techniques for Pathology and Change Detection

In the previous chapter, the usage of consecutive longitudinal information was studied, with the objective of predicting both the presence of cardiomegaly (in the first image of the pair) and the change class between the two scans of the pair. Here, the combination of the two exercises leads to four possible class combinations: [0,0], [1,0], [0,1] and [1,1], where the first label is relative to the presence of pathology, and the second one to the change class.

As previously shown in Table 5.1, the cases that contain positive labels for pathology or change are far less common than doubly negative pairs. This fact is thought to be one of the factors preventing the models from obtaining better performance, as the low representation of the minority cases ([0,1], [1,0] and [1,1]) might be hindering the learning of their representative features.

In order to improve the results established in Chapter 5, potential longitudinal data augmentation methods were explored. These methods have the objective of increasing the representation of minoritary classes during training, which might lead to learning more representative features and, consequently, better final prediction metrics. More specifically, in this chapter, different experiments were performed, in order to study data augmentation techniques for pathology and change detection. In these experiments, the longitudinal dataset was used in different alternative manners. These were used to train models, maintaining the same experimental settings presented in Chapter 5.

## 6.1 Methods

### 6.1.1 Datasets

Previously, the original dataset and the longitudinal dataset were described. The longitudinal dataset consists of pairs of consecutive scans from the same patient. However, as 9,189 patients have more than 2 images, multiple combinations can be done between the available scans. For instance, aside from combining two consecutive images in a pair, non-consecutive pairs can also be formed, either with a logical temporal order or an inverse temporal order. The datasets that can

be formed by creating all possible combinations of images (independently of the time order) are hereinafter referred to as the **pseudolongitudinal datasets**. Different versions of pseudolongitudinal datasets were used in the augmentation studies, since they allow the usage of more minoritary pairs for training.

The fully pseudolongitudinal dataset (all possible pairs for each patient) yields a very high number of combinations in comparison with the longitudinal dataset (Table 6.1). This fact would make the training process a lot more computationally expensive, so, it was not used in this chapter's study. In alternative, other datasets derived from it were explored as lighter options. These datasets, that also aim at reducing the effect of class imbalance and leading to better final performance, are described in the following paragraphs.

- The first formed dataset was called **pseudolongitudinal minoritary dataset**. It consists of the longitudinal dataset (consecutive image pairs), to which the pseudolongitudinal combinations are added, but only if the formed sample belongs to minority case ([1,1], [1,0] or [0,1]);
- Posteriorly, a similar dataset was generated, but where only the pseudolongitudinal samples with the [1,0] case were added to the longitudinal dataset (**pseudolongitudinal [1,0] dataset**). The objective of this experiment was introducing more minority samples in training, but without increasing the representation of the change class. As this class is by itself associated with a higher error (by being computed by the comparison between two pathology labels), maintaining its representation lower might prove to be advantageous;
- Finally, another dataset was formed by using the longitudinal dataset and adding to it the pseudolongitudinal combinations formed by the patients with N or fewer images. This was done in order to get an augmented dataset but without increasing too much number of samples. As for patients with more images, more combinations are possible, this restriction leads to the formation of multiple combinations only for patients with fewer images, while the remaining contribute with consecutive pairs only (longitudinal dataset). This dataset is called **pseudolongitudinal <N dataset**. Experiments were performed for N = 5 and N = 10.

Table 6.1: Number of training samples for each class case, for the pseudolongitudinal dataset versions.

| **Dataset** | **[1,1]** | **[1,0]** | **[0,1]** | **[0,0]** |
|---|---|---|---|---|
| Longitudinal | 883 (1.81%) | 363 (0.74%) | 842 (1.73%) | 46,673 (95.72%) |
| Pseudolongitudinal | 29,685 (2.57%) | 6,988 (0.61%) | 29,685 (2.57%) | 1,087,256 (94.25%) |
| Pseudolongitudinal minoritary | 29,685 (26.26%) | 6,988 (6.18%) | 29,685 (26.26%) | 46,673 (41.29%) |
| Pseudolongitudinal [1,0] | 883 (1.59%) | 6,988 (12.62%) | 842 (1.52%) | 46,673 (84.27%) |
| Pseudolongitudinal <5 | 1,024 (1.62%) | 455 (0.72%) | 1,117 (1.76%) | 60,775 (95.9%) |
| Pseudolongitudinal <10 | 2,070 (1.62%) | 645 (0.51%) | 2,072 (1.62%) | 122,895 (96.25%) |

The fold distribution scheme mentioned in Section 5.1.1 was maintained for training, validation, and test in this new scenario. The pseudolongitudinal combinations were implemented only for the train and validation datasets. The test dataset remains the same in all experiments, equal to the test dataset used in Chapter 5 (longitudinal test dataset). In Table 6.1, a comparison of all datasets generated for these experiments, as well as the longitudinal dataset, can be seen. Here, the number of samples of each case, in the train dataset, is displayed.

### 6.1.2 Pathology and Change Detection

In order to perform the initial experiments with the mentioned different datasets, the longitudinal model (experimental setting 3, explained in Section 5.1.2.3) was trained again to predict the presence of cardiomegaly in the first image, and the presence of change in the image pair. This model was chosen since it is trained with longitudinal information from the input level, in opposition to the baseline and features models (experimental setting 1 and 2), which might be an advantage at inferring longitudinal features from the images. As the aligned longitudinal model (experimental setting 4) depends on the previous alignment of the pairs, the longitudinal model was chosen over it, being used to test the performance obtained with each of the augmented datasets.

After determining the best longitudinal augmentation technique, the augmented dataset that provided the best final result for the longitudinal model was used for training in the remaining experimental settings (2, the features model, and 4, the aligned longitudinal model), maintaining the conditions described in Section 5.1.2. Similarly to the previous experiments, saliency maps were generated in order to provide explainability information of all models.

## 6.2 Results and Discussion

The results obtained for the pseudolongitudinal datasets, all applied to the longitudinal model (experimental setting 3) can be seen in Table 6.2. When comparing the model trained with the longitudinal dataset and the pseudolongitudinal minoritary dataset, the pathology detection metrics remain similar, and the change class metrics suffer from a small decrease. This fact is thought to be related to the noise associated with the change class. As previously explained, this class results from the combination of two cardiomegaly labels, thus, if one label is incorrect, this error will also be present in the change class. In the case of the longitudinal dataset, an error in a cardiomegaly label can impact two change labels (as the pairs are consecutive, so an image can be in a maximum of two pairs). However, in the pseudolongitudinal dataset, as all possible combinations of images (from the same patient) are used, the noise associated with the change label is augmented. In this scenario, an incorrect cardiomegaly label can affect as many change annotations as the number of pairs the corresponding image is included in.

For the pseudolongitudinal [1,0] dataset, as only the cases with negative change are considered, the noise associated with it is reduced. Thus, in this situation, both classes' metrics improve slightly, in comparison with both the longitudinal and pseudolongitudinal minoritary datasets.

Table 6.2: Results for the longitudinal model (experimental setting 3) and the pseudolongitudinal dataset versions.

| | Longitudinal | | Pseudolongit. minority | | Pseudolongit. [1,0] | | Pseudolongit. <5 | | Pseudolongit. <10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Card. | Change | Card. | Change | Card. | Change | Card. | Change | Card. | Change |
| AUC | 0.833 | 0.795 | 0.830 | 0.775 | 0.849 | 0.808 | 0.855 | 0.822 | **0.863** | **0.827** |
| Precision | 0.079 | 0.090 | 0.062 | 0.058 | 0.082 | 0.079 | **0.091** | 0.090 | 0.088 | **0.109** |
| Recall | 0.734 | 0.679 | **0.801** | **0.869** | 0.746 | 0.777 | 0.736 | 0.741 | 0.781 | 0.691 |
| Accuracy | 0.782 | 0.769 | 0.695 | 0.547 | 0.789 | 0.701 | **0.811** | 0.752 | 0.796 | **0.810** |

Regarding the pseudolongitudinal <N datasets (for N = 5 and N = 10), both of them display a higher performance than the remaining experiments, with the pseudolongitudinal <10 dataset presenting the best AUC (0.863 for cardiomegaly and 0.827 for change). In these experiments, the representation of each case ([0,0], [0,1], [1,0] and [1,1]) is similar to the longitudinal dataset, however, more samples are used for training. This fact, combined with the lack of noise caused by the augmentation of the errors in the cardiomegaly presence class, is probably the reason why these experiments provide the best results. The fact that the pseudolongitudinal <10 overperformed the pseudolongitudinal <5 dataset means that the usage of more combinations is probably beneficial for predicting both the presence of cardiomegaly and change.

As previously mentioned, the augmented pseudolongitudinal dataset experiments were also replicated on the remaining experimental settings (explained in Chapter 5). The chosen dataset was the pseudolongitudinal <5. This dataset was chosen over the pseudolongitudinal <10 dataset because there is only a slight performance improvement shown by the latter, and the longitudinal aligned model (experimental setting 4) requires the previous alignment of each pair (with the first image is used as a reference). Thus, by using the pseudolongitudinal <5 dataset there is no need for aligning as many pairs as if the pseudolongitudinal <10 was used.

In Table 6.3 the results using the pseudolongitudinal <5 dataset for all experimental settings are shown. The features model (experimental setting 2) outperformed the remaining approaches, with an AUC of 0.896 for cardiomegaly and 0.863 for change. These results are similar to the ones presented for longitudinal dataset and this experimental setting. Regarding the longitudinal model (experimental setting 3), the results for the pseudolongitudinal dataset are better than the ones presented for the longitudinal dataset, with an increase of 2.2% for cardiomegaly and of 2.7% for change AUC. In the case of the longitudinal aligned model (experimental setting 4), a very slight

Table 6.3: Results for all experimental settings, using the pseudolongitudinal <5 dataset.

| | 2 – Features | | 3 - Longitudinal | | 4 – Longitudinal Aligned | |
|---|---|---|---|---|---|---|
| | Card. | Change | Card. | Change | Card. | Change |
| AUC | **0.896** | **0.863** | 0.855 | 0.822 | 0.876 | 0.829 |
| Precision | **0.120** | 0.091 | **0.091** | 0.090 | 0.105 | 0.087 |
| Recall | **0.766** | **0.825** | 0.736 | 0.741 | 0.746 | 0.772 |
| Accuracy | **0.856** | 0.730 | 0.811 | **0.752** | 0.837 | 0.735 |

improvement is present, when comparing the pseudolongitudinal dataset with the longitudinal one.

One possible hypothesis for explaining these results, is the fact that training the models with more data (pseudolongitudinal augmentation) might facilitate the extraction of relevant features from the image pair. Also, as this augmentation leads to the combination of more temporally distant images, it is possible that the change between them is more noticeable. The usage of these relevant features might lead to a better classifier. Thus, as the features model (experimental setting 2) already seemed to be extracting relevant features (for cardiomegaly, when using the longitudinal dataset), it is possible that the addition of new pairs did not result in improvement. As in this case the change label is being predicted from the cardiomegaly features of both images, it is also possible that longitudinal features are not being further learned in this scenario.

In the case of the longitudinal model (experimental setting 3), the results using the longitudinal dataset suggest a weaker feature extraction. Thus, it is possible that the usage of more data (pseudolongitudal dataset) would leads to a more significant improvement in the final metrics. Regarding the longitudinal aligned model (experimental setting 4), the usage of more information only lead to a very slight improvement, thus, it is possible that the increase in information was overshadowed by the improvement provided by the alignment.

Examples of saliency maps, both for cardiomegaly and change, for all experimental settings are presented in Figure 6.1. It is clear that the features model (experimental setting 2) produces maps with more focused activations than the remaining models, for both cardiomegaly and change detection. Regarding the longitudinal models (experimental settings 3 and 4) saliency maps, when comparing them to the results from Chapter 5 (using the longitudinal dataset), they seem somewhat more focused. The maps produced for cardiomegaly and change, in an image pair, have the tendency to be similar, which means the models tend to use the same region in both input images to produce the predictions. The maps are, in general, focused on the heart region.

## 6.3 Conclusions

In this chapter, the use of non-consecutive longitudinal information as a data augmentation technique for the prediction of cardiomegaly and change in an image pair was explored. The integration of non-sequential longitudinal data was done by performing different combinations of images from the same patient, forming four versions of pseudolongitudinal datasets. These versions differ in the amount of class cases ([0,0], [0,1], [1,0] and [1,1]) that form each one.

These datasets were used to train the longitudinal model. The noise associated with the change class (due to the accumulation of error in the cardiomegaly label) proved to affect the performance of the model. When the used dataset contains a higher portion of samples with positive change, this error manifests itself, providing worse final results for change detection, in comparison with the model trained without augmentation. Experiments where the number of training samples was increased without affecting the ratio of each class seemed to improve the performance of the trained models. This shows that the usage of more image combinations during training can be used

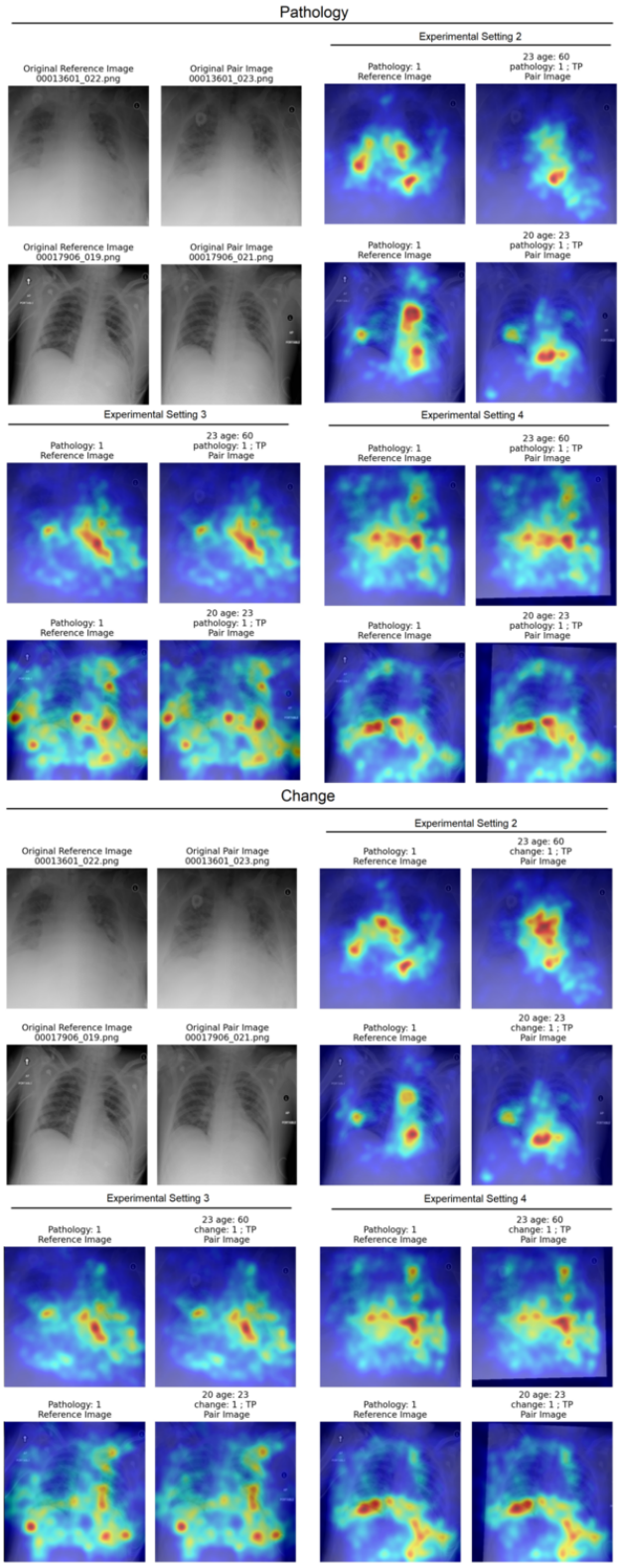



Figure 6.1: Comparison of saliency maps for all experimental settings using the pseudolongitudinal <5 dataset. True positive cases.

as an advantageous data augmentation technique, and it leads to the hypothesis that the higher the numbers of combinations used, the more notable the results will be.

The most favorable pseudolongitudinal dataset was used to train the remaining experimental settings (features and aligned longitudinal models). The features model (experimental setting 2) outperformed the remaining, with an AUC of 0.896 for cardiomegaly detection and 0.863 for change detection. The results show that the most notable improvement is present in the longitudinal model (experimental setting 3), while the features model (experimental setting 2) and the longitudinal aligned model (experimental setting 4) present similar results to the longitudinal dataset (no augmentation). As the features model seems to be using relevant cardiomegaly features from the scans, the augmentation was not as notable. Similarly, the effect of the alignment in the longitudinal aligned model probably overshadowed the augmentation advantage.

A possible suggestion for future work on data augmentation would be to use image combinations from different patients. This option could increase the differences between the images and possibly facilitate the extraction of features that compare the two images.

# Chapter 7

# Longitudinal Label Rectification

The experiments carried out in the previous chapters led to the conclusion that it is possible to predict change between two images, regarding the cardiomegaly findings. In these experiments, the change ground truth label was computed using the original cardiomegaly labels of both images, and both temporally sequential and non-sequential pairs of images were formed for training and testing.

As previously discussed in Section 5.2, the visual analysis of multiple images from the same patient, allowed the detection of cases where very similar looking images present different ground truth labels for cardiomegaly. Examples of such cases can be seen in Figure 5.8. As the original ChestX-ray14 dataset was labeled using NLP to automatically extracted annotations from medical reports, the possibility that errors exist in the dataset has to be kept in mind.

During the multiple pseudolongitudinal dataset experiments, the presence of more positive change labels proved to negatively affect the performance of the model, especially for the detection of change. On the other hand, datasets that preserved the distribution of the classes (similar to the longitudinal dataset) exhibited a performance improvement, when more samples were used for training. These results show that using a higher representation of positive change labels for training introduces noise to the training process, which results in poor final results.

With the aim of reducing the noise associated with the change class, the original dataset labels were altered using longitudinal data. This alteration led to the creation of a new dataset – referred to as the **rectified dataset** – which is described and studied in this chapter, being used to train the models for pathology and change detection and for pseudolongitudinal experiments.

## 7.1 Methods

### 7.1.1 Datasets

The created rectified dataset consists in an alteration of the original ChestX-ray14 labels. As cardiomegaly is related with heart enlargement, it is assumed that it cannot develop and fade rapidly, thus, in this alteration, it is presumed that the presence of two changes in the cardiomegaly label in a small period of time should be treated as an error in the dataset. For instance, if a patient

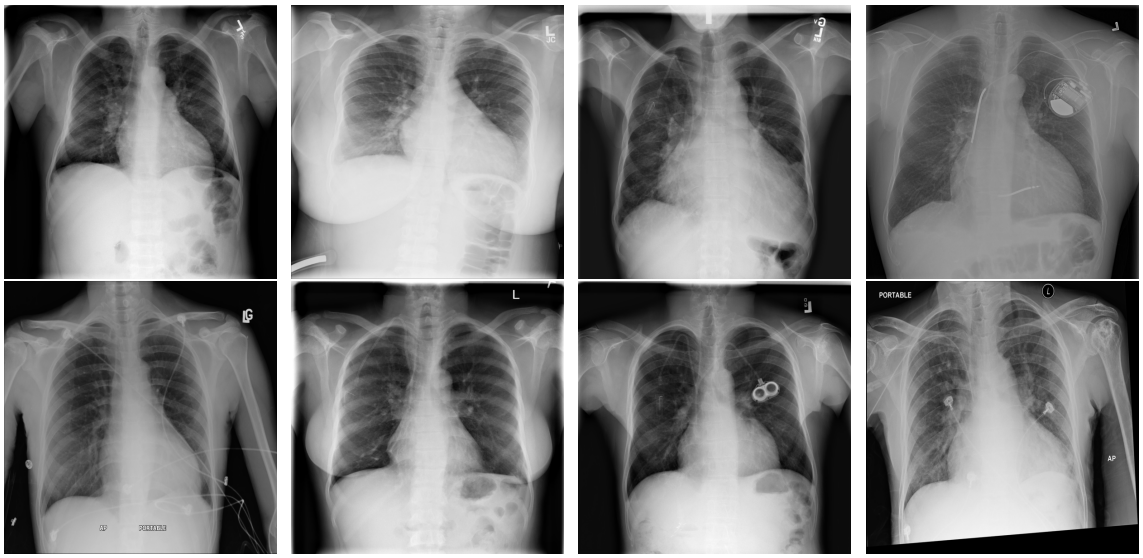



Figure 7.1: Examples of images whose cardiomegaly labels changed. In the top row, the label was rectified to positive, and in the bottom row to negative.

has 5 images acquired in less than 2 years with a ground truth label sequence of 0-0-0-1-0 (or 1-1-1-0-1), then it should be altered to 0-0-0-0-0 (or, similarly, 1-1-1-1-1).

In order to perform such alteration, the patients with multiple images were selected. For each patient, the scans were grouped by patient age, creating collections of images less than 2 years apart. The number of positive and negative labeled images were counted. If the negative images consist of more than 80% of the images in the group, then the labels of the remaining should be changed to negative. Similarly, if more than 80% of the images in the group are positive, then the remaining are changed to a positive label. It should be noted that, due to this condition, groups with less than five images remained unchanged.

This alteration led to a change of 949 cardiomegaly labels in the original dataset, and only 15 of these were changes from a negative label to positive. In Figure 7.1, examples of images whose labels were rectified can be seen.

The fold distribution mentioned in Section 5.1.1 was maintained, and the same folds were used for the train, validation and test datasets. After the creation of the rectified dataset, the longitudinal version of this dataset was also formed (longitudinal rectified dataset). The following versions of the rectified dataset were also generated, in order to conduct pseudolongitudinal experiments: pseudolongotudinal minoritary, pseudolongotudinal [1,0] and pseudolongotudinal <5. These rectified versions were considered for all train, validation and test datasets. Similar to the previous pseudolongitudinal experiments, the augmentation was performed solely on the train and validation datasets.

### 7.1.2 Pathology and Change Detection

The longitudinal rectified dataset was used to train the features model and the longitudinal models (experimental settings 2, 3 and 4), maintaining all training conditions previously described in

Section 5.1.2. Results were generated for the test set, for all experimental settings. The baseline model (experimental setting 1) results were obtained by inference on the rectified test dataset.

The rectified versions of the augmented pseudolongtitudinal datasets were used for experiments with the longitudinal model (experimental setting 3), similarly to the study done in Chapter 6. After that, the rectified pseudolongitudinal <5 dataset was used to train and generate results for all models (experimental settings 2, 3 and 4).

## 7.2 Results and Discussion

A rectified version of the ChestX-ray14 dataset was generated. It is important to note that the altered annotations were not verified, and thus it cannot be guaranteed that the alterations were beneficial, in terms of dataset accuracy. An example of a rectification that might not be correct is the last scan in Figure 7.1 (bottom right). In this case, the CTR measurement (and visual assessment) seem to identify this scan as positive for cardiomegaly, while the rectification altered the label to negative.

It also must be kept in mind that the method used for rectification is simple, and it might fail at identifying suspicious cases. This might happen principally in the cases that at the extremes of the formed collections (initial and final images), as they might indicate a real change that is postponed or advanced by the rectification. As an example, a ground truth full sequence of 1-0-0-0-0, or 0-0-0-0-1, will be rectified to 0-0-0-0-0. This alteration might not be improving the noise in the dataset, and it is weakly supported by the fundaments used for rectification, in comparison with alterations in samples in the middle of the sequence. Thus, the existence of such situations must be kept in mind.

### 7.2.1 Longitudinal Scans Experiments

In Table 7.1, the results for the longitudinal rectified dataset, in all experimental settings, can be seen. As this dataset consists in an alteration of ChestX-ray14, its results cannot be compared with the ones generated by the longitudinal and pseudolongitudinal datasets, as these are derivations of the original data.

Regarding the detection of cardiomegaly, the baseline model (experimental setting 1) outperforms the remaining, with an AUC of 0.914. However, the features model (experimental setting 2) and the aligned longitudinal model (experimental setting 4) present a similar performance. As the features model uses the same image features as the baseline model to provide the predictions, similar results regarding cardiomegaly detection were expected. Concerning the detection of change, the features model (experimental setting 2) present an AUC of 0.867, outperforming the remaining approaches.

These results should not be used to directly compare the models with the ones from Chapter 5 (longitudinal dataset). However, a comparative analysis to validate whether the performed rectifications helped the training and testing can be done. The fact that the baseline model inference

Table 7.1: Results for all experimental settings, using the longitudinal rectified dataset.

| | 1 - Baseline | | 2 - Features | | 3 - Longitudinal | | 4 - Longitudinal Aligned | |
|---|---|---|---|---|---|---|---|---|
| | Card. | Change | Card. | Change | Card. | Change | Card. | Change |
| AUC | **0.914** | 0.849 | 0.898 | **0.867** | 0.875 | 0.832 | 0.899 | 0.847 |
| Precision | **0.093** | 0.054 | 0.090 | **0.069** | 0.060 | 0.048 | 0.073 | 0.055 |
| Recall | 0.801 | **0.768** | 0.761 | 0.712 | 0.78 | 0.744 | **0.830** | 0.744 |
| Accuracy | 0.868 | 0.791 | **0.875** | **0.847** | 0.804 | 0.772 | 0.830 | 0.801 |

on the rectified longitudinal dataset outperformed the longitudinal results means that the parameters that were learned during training with the original dataset are more consistent at predicting cardiomegaly and change in the rectified longitudinal dataset than on the longitudinal dataset. Furthermore, the fact that the performance of all experimental settings improved for both cardiomegaly and change, means that the testing task became easier. Another hypothesis is that using the rectified version of the longitudinal dataset lead to a more efficient training, due to the usage of a more dependable dataset.

In Figure 7.2 some examples of images with rectified labels are presented. The ground truth label, as well as the predictions from the different experimental settings, are also displayed, for both the longitudinal dataset and the rectified longitudinal dataset.

### 7.2.2 Pseudolongitudinal Scans Experiments

In Table 7.2, the results for all versions of the rectified pseudolongitudinal datasets, used to train the longitudinal model, are shown. The results for the rectified longitudinal dataset are also displayed, for comparison.

The best performance for cardiomegaly detection was obtained when training with the rectified pseudolongitudonal minoritary dataset, reaching an AUC of 0.902. This value is similar to the ones obtained by the remaining pseudolongitudinal variants. Regarding the change class, the usage of the rectified pseudolongitudinal <5 dataset reached the highest AUC, with a value of 0.853. This value represents an increase of 2.1% in comparison with the rectified longitudinal dataset. The

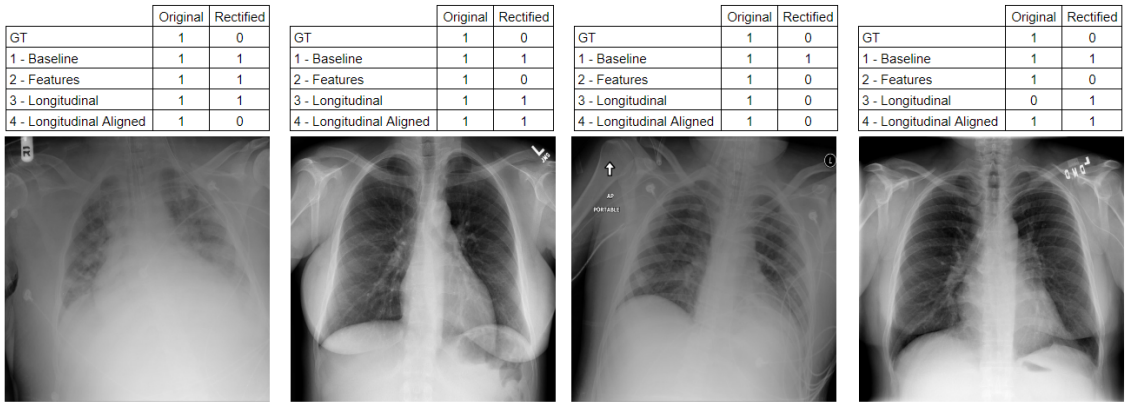



Figure 7.2: Examples of images whose cardiomegaly labels were rectified.

remaining pseudolongitudinal datasets also outperformed the longitudinal dataset results. As the usage of pseudolongitudinal information led to an improvement in the results in all cases, which show that, for both classes, the usage of non-consecutive combinations is helpful at bettering the model performance.

As previously mentioned, these results cannot be directly compared with the original labels pseudolongitudinal experiments, in terms of model performance, as different datasets were used. However, the effect of this different should be analyzed. Recalling the results shown in Section 6.2, the usage of the pseudolongitudinal minority augmentation led to a decrease in AUC in the change detection. In the original labels' scenario, the alteration of the ratio of the minority classes seemed to affect the performance of the models. Looking at the results generated by the rectified versions of the datasets, the usage of pseudolongitudinal augmentation proved to be an advantage in all situations, increasing the AUC. In this case, changing the ratio of minority classes did not provoke notable differences between the performances of the trained models. This shows that the rectification of the dataset succeeded at creating more consistent data.

Table 7.2: Results for the longitudinal model (experimental setting 3) and the pseudolongitudinal dataset versions.

| | Longitudinal | | Pseudolongitudinal minority | | Pseudolongitudinal [1,0] | | Pseudolongitudinal <5 | |
|---|---|---|---|---|---|---|---|---|
| | Card. | Change | Card. | Change | Card. | Change | Card. | Change |
| AUC | 0.875 | 0.832 | **0.902** | 0.845 | 0.893 | 0.843 | 0.893 | **0.853** |
| Precision | 0.060 | 0.048 | 0.078 | **0.068** | **0.082** | 0.059 | 0.073 | 0.062 |
| Recall | 0.780 | 0.744 | **0.819** | 0.700 | 0.768 | **0.748** | 0.799 | 0.728 |
| Accuracy | 0.804 | 0.772 | 0.844 | **0.847** | **0.860** | 0.815 | 0.835 | 0.825 |

Table 7.3: Results for all experimental settings, using the pseudolongitudinal rectified dataset.

| | 2 - Features | | 3 - Longitudinal | | 4 – Longitudinal Aligned | |
|---|---|---|---|---|---|---|
| | Card. | Change | Card. | Change | Card. | Change |
| AUC | **0.933** | **0.895** | 0.893 | 0.853 | 0.911 | 0.867 |
| Precision | **0.145** | **0.072** | 0.073 | 0.062 | 0.089 | 0.059 |
| Recall | 0.792 | **0.772** | 0.799 | 0.728 | **0.826** | 0.772 |
| Accuracy | **0.923** | **0.844** | 0.835 | 0.825 | 0.863 | 0.809 |

In Table 7.3 the results for the rectified pseudolongitudinal <5 dataset, for all experimental settings, are displayed. The model presenting the best performance on cardiomegaly detection is the features model (experimental setting 2), with an AUC of 0.933, meaning an increase of 3.5%, in comparison with the rectified longitudinal dataset. This model was the one presenting the highest improvement, however, both longitudinal models (experimental settings 3 and 4) also show an improvement when trained with pseudolongitudinal data. Regarding change prediction, the features model (experimental setting 2) outperformed the remaining, with an AUC of 0.895 (increase of 2.8% in comparison with the rectified longitudinal dataset). The longitudinal models (experimental settings 3 and 4) also present an increase in AUC. These results also show that

using the pseudolongitudinal augmentation is valuable, both for the detection of cardiomegaly and change between scans, and for all experimental settings.

## 7.3 Conclusions

The visualization of examples where the cardiomegaly labels seem to be possibly incorrect led to the exploration of a technique for dataset rectification. The creation of the rectified dataset aimed at the utilization of a more consistent data, with less noise associated with the change class. There was no manual verification to these alterations, so, the defined ground truth labels should not be trusted for comparisons with the original dataset.

The rectified dataset was used to train models in the experimental settings 2, 3 and 4, and inference results were produced for the baseline model (experimental setting 1). The best AUC for cardiomegaly detection (0.914) was achieved by the baseline model. Regarding change detection, the best performance was reached by the features model (experimental setting 3), presenting an AUC of 0.867. The results from these experiments cannot be used to compare the models with the previous ones, as a different dataset (with different labels) was used for training. However, the fact that the baseline model presents higher metrics for the rectified longitudinal dataset means that the features learned while training with the longitudinal dataset are more concordant with the rectifications.

The creation of rectified versions of the pseudolongitudinal datasets showed that the usage of data augmentation techniques can improve the results for all experimental settings. In this case, augmentations in the ratio of the minority classes did not affect the performance of the model, and all rectified pseudolongitudinal dataset versions improved the performance of the model. This also proves that the rectification of the dataset improved its consistency.

The usage of longitudinal data to clean-up noise in datasets with automatically generated annotations can be a powerful technique. To do so, it is important to keep in mind the time intervals which are considered reasonable for the evolution of a CXR abnormality, so that suspect annotations can be spotted. Despite the fact that it can not be used for all abnormalities, due to their spontaneous character, this rectification technique should be further studied, as it could provide an efficient solution for dataset correction. A deeper study should also be done as the developed technique is simple, and it is based on an arbitrary abnormality evolution time period, which is likely not ideal.

# Chapter 8

# Conclusion

The comparison of multiple CXR exams is a common practice by radiologists and medical professionals to analyze the most relevant differences and conclude on the represented findings. However, automated systems developed to aid this process and published as state-of-the-art are usually based on a single scan. The implementation of longitudinal data in automated CXR analysis is a field in development, but of high importance, as it provides solutions that take into consideration a more realistic problem.

In this work, an algorithm for alignment of CXR scans is proposed. Alignment methods usually accompany automatic solutions that use multiple images, as they might establish an improvement by facilitating the comparison. The developed method succeeds at aligning a pair of images, by using anatomical features extracted from lung segmentations. The usage of these features is an advantage, as it allows the alignment of all CXR pairs according to relevant marks common in all scans. The role of alignment methods should not be undermined in longitudinal problems. The state-of-the-art CXR registration techniques tend to use features as pixel intensity, anatomical keypoints or segmentations (similarly to the developed method), which cannot be fully relied on for computation of the transformations and evaluation of the method, when automatically acquired. The developed method constitutes a simple rigid transformation approach that does not overcome this general problem, however, it can be helpful especially to align images that are used in automatic comparisons, however, human professionals may also benefit from analyzing aligned pairs, instead of unaligned ones.

Experiments were performed with the objective of including longitudinal information in predictive models. Four experimental settings were designed, receiving a pair of CXR as input and predicting both the presence of pathology (in the first image) and the change in the pathology label (when comparing the images in the pair). The experimental settings varied the level at which the longitudinal data were integrated. The baseline model was trained with no longitudinal information, and subsequent models integrated data at the feature level and at the input level. The results showed that integrating longitudinal data at the features level succeeds at improving the prediction of change in the pair, while the integration at the input level makes it difficult to extract relevant features. This proves that pathology features can be used to predict relevant comparison cues be-

tween two images in a pair. The usage of aligned CXR pairs facilitated the feature extraction at the image level, exposing the advantage of using registration techniques in this field. Augmentation techniques were developed, where non-sequential pairs of images were used for training. The experiments showed the change class is associated with higher noise, and thus, an increase in its representation leads to poor results. However, this augmentation technique proved to be advantageous when the class ratios are maintained similar to original, and the addition of more training combinations seems to improve the results, thus, this method should be further explored. These methods should be replicated in other common CXR abnormalities, in order to assess the validity of the obtained results in other findings. The inclusion of longitudinal information proved to be useful at predicting the presence of change in a CXR scans pair, with little effect in the prediction of pathology, in comparison with a single input image approach. The prediction of change in a pair of CXR is important, as it can provide relevant information to a radiologist or other medical professional, mimicking the human analysis of CXR. Predicting this change allows a faster analysis of multiple exams from the same patient, reaching a possible diagnosis with more ease. It also allows an efficient comparison of a scan with the previous ones from the same patient, encouraging a diagnosis based on comparison with a previous state, instead of a standard reference. The explainability associated with automated methods is also a matter of high relevance for the human users. A model that predicts the presence of pathologies for one image can produce explainability maps to support such predictions. However, a model whose predictions are based on multiple images could allow the visualization of the changes throughout the scans, which facilitates the comprehension of the predicted output, for a human observer. Even though this topic was not further explored in this work, it should be kept in mind, as it is a valuable advantage present in longitudinal systems.

The presence of possibly wrong annotations in the dataset led to the creation of a technique to rectify them using longitudinal data. Thus, the dataset ground truth annotations were altered in order to make it more consistent and perhaps reduce the noise. This dataset was used to train models in the previously used experimental settings. The results showed that the rectified dataset seems to be more consistent and include less noise than the original dataset. Consequently, longitudinal rectification techniques should be further explored. Longitudinal information can be used to develop powerful techniques to rectify CXR datasets. This concept could not be found in state-of-the-art works, thus, the introduction of this topic is thought to have high relevance and possibly a high future impact after further studies. The presence of noise in a dataset can be prejudicial in DL applications, and it is especially common when the annotations are automatically generated. Thus, a valid correction of the labels could improve the performance of various developed techniques.

Altogether, different experiments were carried out with the objective of studying the inclusion of longitudinal data for automated comparison of scans. Asides from inspecting how to integrate such data in automated algorithms, a CXR pair alignment algorithm was developed, and data augmentation techniques and possible techniques for dataset rectification were analyzed. These studies demonstrate the positive impact that the inclusion of longitudinal information can have in

already existing technologies, as well as the possible benefits of using this type of data for new applications.

# References


[1] History of x-rays - 125 years in the making (pt 2) - excillum. URL: https://www.excillum.com/history-of-x-rays-x-ray-tubes/.

[2] Projection x-ray imaging | radiology key. URL: https://radiologykey.com/projection-x-ray-imaging/.

[3] David A Lisle. *Imaging for Students*. CRC Press, January 2012. URL: https://doi.org/10.1201/b13297, doi:10.1201/b13297.

[4] Gaurang Karwande, Amarachi Mbakawe, Joy T. Wu, Leo A. Celi, Mehdi Moradi, and Ismini Lourentzou. Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays, 2022. arXiv:2208.03873.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[6] Cornelia Schaefer-Prokop, Ulrich Neitzel, Henk W. Venema, Martin Uffmann, and Mathias Prokop. Digital chest radiography: An update on modern technology, dose containment and control of image quality. *European Radiology*, 18:1818–1830, 4 2008. URL: https://link.springer.com/article/10.1007/s00330-008-0948-3, doi:10.1007/S00330-008-0948-3/FIGURES/10.

[7] Martin Berger, Qiao Yang, and Andreas Maier. X-ray imaging. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11111 LNCS:119–145, 8 2018. URL: https://www.ncbi.nlm.nih.gov/books/NBK546155/, doi:10.1007/978-3-319-96520-8_7.

[8] J. Anthony Seibert. X-ray imaging physics for nuclear medicine technologists. part 1: Basic principles of x-ray production. *Journal of Nuclear Medicine Technology*, 32(3):139–147, 2004. URL: https://tech.snmjournals.org/content/32/3/139, arXiv:https://tech.snmjournals.org/content/32/3/139.full.pdf.

[9] Ian D. McLean and Jan Martensen. Chapter 2 - specialized imaging. In Dennis M. Marchiori, editor, *Clinical Imaging (Third Edition)*, pages 44–78. Mosby, Saint Louis, third edition edition, 2014. URL: https://www.sciencedirect.com/science/article/pii/B9780323084956000026, doi:https://doi.org/10.1016/B978-0-323-08495-6.00002-6.

[10] R. F. Mould. The early history of x-ray diagnosis with emphasis on the contributions of physics 1895-1915. *Physics in medicine and biology*, 40:1741–1787, 1995. URL: https://pubmed.ncbi.nlm.nih.gov/8587931/, doi:10.1088/0031-9155/40/11/001.

[11] Fred A. Mettler, Mythreyi Bhargavan, Keith Faulkner, Debbie B. Gilley, Joel E. Gray, Geoffrey S. Ibbott, Jill A. Lipoti, Mahadevappa Mahesh, John L. McCrohan, Michael G. Stabin, Bruce R. Thomadsen, and Terry T. Yoshizumi. Radiologic and nuclear medicine studies in the united states and worldwide: frequency, radiation dose, and comparison with other radiation sources–1950-2007. *Radiology*, 253:520–531, 11 2009. URL: https://pubmed.ncbi.nlm.nih.gov/19789227/, doi:10.1148/RADIOL.2532082010.

[12] Amin Tafti and Doug W. Byerly. X-ray radiographic patient positioning. *StatPearls*, 12 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK565865/.

[13] Robin Alexander Kluthke, Ralph Kickuth, Paul Martin Bansmann, Carolin Tüshaus, Stephan Adams, Dieter Liermann, and Johannes Kirchner. The additional value of the lateral chest radiograph for the detection of small pulmonary nodules—a roc analysis. *The British Journal of Radiology*, 89, 2016. URL: /pmc/articles/PMC5124842//pmc/articles/PMC5124842/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124842/, doi:10.1259/BJR.20160394.

[14] Fred A. Mettler, Walter Huda, Terry T. Yoshizumi, and Mahadevappa Mahesh. Effective doses in radiology and diagnostic nuclear medicine: A catalog1. *https://doi.org/10.1148/radiol.2481071451*, 248:254–263, 7 2008. URL: https://pubs.rsna.org/doi/10.1148/radiol.2481071451, doi:10.1148/RADIOL.2481071451.

[15] Aparna Tompe and Kiran Sargar. X-ray image quality assurance. *StatPearls*, 10 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK564362/.

[16] Erdi Calli, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1361841521001717, doi:https://doi.org/10.1016/j.media.2021.102125.

[17] Richard P. Kruger, James R. Townes, David Lee Hall, Samuel J. Dwyer, and Gwilym S. Lodwick. Automated radiographic diagnosis via feature extraction and classification of cardiac size and shape descriptors. *IEEE Transactions on Biomedical Engineering*, BME-19:174–186, 1972. doi:10.1109/TBME.1972.324115.

[18] René Hosch, Lennard Kroll, Felix Nensa, and Sven Koitka. Differentiation between anteroposterior and posteroanterior chest x-ray view position with convolutional neural networks. *RoFo Fortschritte auf dem Gebiet der Rontgenstrahlen und der Bildgebenden Verfahren*, 193:168–176, 2 2021. URL: http://www.thieme-connect.de/products/ejournals/html/10.1055/a-1183-5227http://www.thieme-connect.de/DOI/DOI?10.1055/a-1183-5227, doi:10.1055/A-1183-5227/ID/OR304-3.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. arXiv:1409.1556.

[20] George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein.

Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1, 1 2019. URL: https://pubs.rsna.org/doi/10.1148/ryai.2019180041, doi:10.1148/RYAI.2019180041/ASSET/IMAGES/LARGE/RYAI.2019180041.FIG3.JPEG.

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. URL: https://doi.org/10.1007%2Fs11263-019-01228-7, doi:10.1007/s11263-019-01228-7.

[22] Joseph Paul Cohen, Lan Dao, Karsten Roth, Paul Morrison, Yoshua Bengio, Almas F. Abbasi, Beiyi Shen, Hoshmand Kochi Mahsa, Marzyeh Ghassemi, Haifang Li, Tim Q. Duong, Joseph Paul Cohen, Lan Dao, Karsten Roth, Paul Morrison, Yoshua Bengio, Almas Abbasi, Beiyi Shen, Hoshmand Kochi Mahsa, Marzyeh Ghassemi, Haifang Li, and Tim Duong. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *Cureus*, 12, 7 2020. URL: https://www.cureus.com/articles/35692-predicting-covid-19-pneumonia-severity-on-chest-x-ray-with-deep-learning, doi:10.7759/CUREUS.9448.

[23] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. arXiv:1608.06993.

[24] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports, 2019. arXiv:1901.07441.

[25] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 12 2019. doi:10.1038/S41597-019-0322-0.

[26] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. arXiv:1901.07031.

[27] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. URL: https://doi.org/10.1109%2Fcvpr.2017.369, doi:10.1109/cvpr.2017.369.

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv:1312.6034.

[29] Kudaibergen Urinbayev, Yerassyl Orazbek, Yernur Nurambek, Almas Mirzakhmetov, and Huseyin Atakan Varol. End-to-end deep diagnosis of x-ray images, 2020. arXiv:2003.08605.

[30] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2018. arXiv:1712.06957.

[31] Lera- lower extremity radiographs | center for artificial intelligence in medicine imaging. URL: https://aimi.stanford.edu/lera-lower-extremity-radiographs.

[32] Aasce | aasce - miccai 2019 challenge: Accurate automated spinal curvature estimation. URL: https://aasce19.github.io/#challenge-dataset.

[33] Karen Panetta, Rahul Rajendran, Aruna Ramesh, Shishir Paramathma Rao, and Sos Agaian. Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1650–1659, 2022. doi:10.1109/JBHI.2021.3117575.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

[35] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 1 2021. URL: https://jair.org/index.php/jair/article/view/12228, doi:10.1613/JAIR.1.12228.

[36] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019. URL: https://www.mdpi.com/2079-9292/8/8/832, doi:10.3390/electronics8080832.

[37] Minki Kim and Byoung-Dai Lee. Automatic lung segmentation on chest x-rays using self-attention deep neural network. *Sensors*, 21(2), 2021. URL: https://www.mdpi.com/1424-8220/21/2/369, doi:10.3390/s21020369.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. arXiv:1505.04597.

[39] Montgomery county x-ray set. URL: https://ceb.nlm.nih.gov/repositories/tuberculosis-chest-x-ray-image-data-sets/.

[40] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule. *American Journal of Roentgenology*, 174(1):71–74, 2000. PMID: 10628457. URL: https://doi.org/10.2214/ajr.174.1.1740071, arXiv:https://doi.org/10.2214/ajr.174.1.1740071, doi:10.2214/ajr.174.1.1740071.

[41] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J. Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4:475, 12 2014. URL: /pmc/articles/PMC4256233//pmc/articles/PMC4256233/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233/, doi:10.3978/J.ISSN.2223-4292.2014.11.20.

[42] Youbao Tang, Yuxing Tang, Jing Xiao, and Ronald M. Summers. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation, 2019. arXiv:1904.09229.

[43] Sohee Park, Sang Min Lee, Namkug Kim, Jooae Choe, Yongwon Cho, Kyung Hyun Do, and Joon Beom Seo. Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *European Radiology*, 29:5341–5348, 10 2019. URL: https://link.springer.com/article/10.1007/s00330-019-06130-x, doi:10.1007/S00330-019-06130-X/FIGURES/5.

[44] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, 2016. arXiv:1612.08242.

[45] Yu-Xing Tang, You-Bao Tang, Mei Han, Jing Xiao, and Ronald M. Summers. Abnormal chest x-ray identification with generative adversarial one-class classifier. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1358–1361, 2019. doi:10.1109/ISBI.2019.8759442.

[46] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. arXiv:1406.2661.

[47] Xin Li, Rui Cao, and Dongxiao Zhu. Vispi: Automatic visual perception and interpretation of chest x-rays, 2020. arXiv:1906.05190.

[48] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi:10.1162/neco.1997.9.8.1735.

[49] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23:304, 3 2016. URL: /pmc/articles/PMC5009925//pmc/articles/PMC5009925/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5009925/, doi:10.1093/JAMIA/OCV080.

[50] Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest imagenome dataset, 2021. URL: https://physionet.org/content/chest-imagenome/1.0.0/, doi:10.13026/WV01-Y230.

[51] Emily B. Tsai, Scott Simpson, Matthew P. Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J. Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, Jody Shen, Mona Hafez, Susan John, Prabhakar Rajiah, Brian P. Pogatchnik, John Mongan, Emre Altinmakas, Erik R. Ranschaert, Felipe C. Kitamura, Laurens Topff, Linda Moy, Jeffrey P. Kanne, and Carol C. Wu. The rsna international covid-19 open radiology database (ricord). *Radiology*, 299:E204–E213, 4 2021. URL: https://pubs.rsna.org/doi/10.1148/radiol.2021203957, doi:10.1148/RADIOL.2021203957/ASSET/IMAGES/LARGE/RADIOL.2021203957.TBL3.JPEG.

[52] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022. arXiv:2012.15029.

[53] Hongyi Duanmu, Thomas Ren, Haifang Li, Neil Mehta, Adam J. Singer, Jeffrey M. Levsky, Michael L. Lipton, and Tim Q. Duong. Deep learning of longitudinal chest x-ray and clinical variables predicts duration on ventilator and mortality in covid-19 patients. *Biomedical engineering online*, 21, 12 2022. URL: https://pubmed.ncbi.nlm.nih.gov/36242040/, doi:10.1186/S12938-022-01045-Z.

[54] Michelle Shu, Richard Strong Bowen, Charles Herrmann, Gengmo Qi, Michele Santacatterina, and Ramin Zabih. Deep survival analysis with longitudinal x-rays for covid-19, 2021. arXiv:2108.09641.

[55] Daniel Gourdeau, Olivier Potvin, Patrick Archambault, Carl Chartrand-Lefebvre, Louis Dieumegarde, Reza Forghani, Christian Gagné, Alexandre Hains, David Hornstein, Huy Le, Simon Lemieux, Marie Hélène Lévesque, Diego Martin, Lorne Rosenbloom, An Tang, Fabrizio Vecchio, Issac Yang, Nathalie Duchesne, and Simon Duchesne. Tracking and predicting covid-19 radiological trajectory on chest x-rays using deep learning. *Scientific Reports 2022 12:1*, 12:1–14, 4 2022. URL: https://www.nature.com/articles/s41598-022-09356-w, doi:10.1038/s41598-022-09356-w.

[56] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection, 2020. arXiv:2003.11597.

[57] Matthew D. Li, Nishanth Thumbavanam Arun, Mishka Gidwani, Ken Chang, Francis Deng, Brent P. Little, Dexter P. Mendoza, Min Lang, Susanna I. Lee, Aileen O'Shea, Anushri Parakh, Praveer Singh, and Jayashree Kalpathy-Cramer. Automated assessment and tracking of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence*, 2:1–39, 7 2020. URL: https://pubs.rsna.org/doi/10.1148/ryai.2020200079, doi:10.1148/RYAI.2020200079/ASSET/IMAGES/LARGE/RYAI.2020200079.TBL4.JPEG.

[58] Ruggiero Santeramo, Samuel Withey, and Giovanni Montana. Longitudinal detection of radiological abnormalities with time-modulated LSTM. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 326–333. Springer International Publishing, 2018. URL: https://doi.org/10.1007%2F978-3-030-00889-5_37, doi:10.1007/978-3-030-00889-5_37.

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. arXiv:1409.4842.

[60] Ramandeep Singh, Mannudeep K. Kalra, Chayanin Nitiwarangkul, John A. Patti, Fatemeh Homayounieh, Atul Padole, Pooja Rao, Preetham Putha, Victorine V. Muse, Amita Sharma, and Subba R. Digumarthy. Deep learning in chest radiography: Detection of findings and presence of change. *PLOS ONE*, 13:e0204155, 10 2018. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204155, doi:10.1371/JOURNAL.PONE.0204155.

[61] Dong Yul Oh, Jihang Kim, and Kyong Joon Lee. Longitudinal change detection on chest x-rays using geometric correlation maps. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 748–756, Cham, 2019. Springer International Publishing.

[62] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. arXiv:1709.01507.

[63] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. arXiv:1710.10903.

[64] W. R. Crum, T. Hartkens, and Derek L.G. Hill. Non-rigid image registration: theory and practice. *http://dx.doi.org/10.1259/bjr/25329214*, 77, 1 2014. URL: https://www.birpublications.org/doi/10.1259/bjr/25329214, doi:10.1259/BJR/25329214.

[65] J. Csorba, B. Kormanyos, and B. Pataki. Registration of chest x-rays. In Panagiotis D. Bamidis and Nicolas Pallikarakis, editors, *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, pages 402–405, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[66] Shengwen Guo, Xiaoming Wu, and Zhaohui Luo. Automatic extraction of control points for chest x-ray image and elastic registration. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pages 2651–2654, 2008. doi:10.1109/ICBBE.2008.996.

[67] Lucas Mansilla, Diego H. Milone, and Enzo Ferrante. Learning deformable registration of medical images with anatomical constraints. *Neural Networks*, 124:269–279, 2020. URL: https://www.sciencedirect.com/science/article/pii/S0893608020300253, doi:https://doi.org/10.1016/j.neunet.2020.01.023.

[68] Yu Ching Lee, Muhammad Adil Khalil, Jui Huan Lee, Abdan Syakura, Yi Fang Ding, and Ching Wei Wang. Fully automatic registration methods for chest x-ray images. *Journal of Medical and Biological Engineering*, 41:826–843, 12 2021. URL: https://link.springer.com/article/10.1007/s40846-021-00666-4, doi:10.1007/S40846-021-00666-4/FIGURES/10.

[69] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision, 2018. arXiv:1711.06373.

[70] Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[71] E Recherche, Et Automatique, Sophia Antipolis, and Zhengyou Zhang. Iterative point matching for registration of free-form curves. *Int. J. Comput. Vision*, 13, 07 1992.

[72] Emanuel Ricardo Coimbra, Quintas Brioso, João Manuel, and Patrício Pedrosa. Anatomical segmentation in automated chest radiography screening. 8 2022. URL: https://repositorio-aberto.up.pt/handle/10216/143015.

[73] Joana Rocha, Sofia Cardoso Pereira, João Pedrosa, Aurélio Campilho, and Ana Maria Mendonça. Attention-driven spatial transformer network for abnormality detection in chest x-ray images. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 252–257, 2022. doi:10.1109/CBMS55023.2022.00051.

[74] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. arXiv:1405.0312.

[75] E Gronenschild. Correction for geometric image distortion in the x-ray imaging chain: local technique versus global technique. *Med. Phys.*, 26(12):2602–2616, December 1999.

[76] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi:10.1080/14786440109462720.

[77] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999. doi:10.1109/ICCV.1999.790410.

[78] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–, 11 2004. doi:10.1023/B:VISI.0000029664.99615.94.

[79] Satendra Pal Singh and Gaurav Bhatnagar. Chapter 1 - perceptual hashing-based novel security framework for medical images. In Amit Kumar Singh and Mohamed Elhoseny, editors, *Intelligent Data Security Solutions for e-Health Applications*, Intelligent Data-Centric Systems, pages 1–20. Academic Press, 2020. URL: https://www.sciencedirect.com/science/article/pii/B9780128195116000017, doi:https://doi.org/10.1016/B978-0-12-819511-6.00001-7.

[80] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.

[81] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001177, doi:https://doi.org/10.1016/j.media.2022.102470.

[82] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015. arXiv:1512.04150.