Ricardo Gil Salgado Evangelista

Representation learning methods for early detection of pathological lesions in chest X-rays images.

Biomedical Engineering Master's Dissertation Report

ORIENTADOR
Professor Doutor Miguel Lopéz
Professor Doutor Catarina Santos

Dezembro, 2024

Ricardo Gil Salgado Evangelista

Representation learning methods for early detection of pathological lesions in chest X-rays images.

JÚRI

Presidente: Professor Doutor Célio Pina, Instituto Politécnico de Setúbal (IPS/ESTS)
Orientador: Professor Doutor Miguel López, Instituto Politécnico de Setúbal (IPS/ESTS)
Vogal: Professor Doutor Luís Magalhães, Universidade do Minho, Escola de Engenharia

Dezembro, 2024

Acknowledgments

First of all, I would like to thank my supervisors, Professor Miguel López, for teaching me practically everything I know about Machine Learning and Artificial Intelligence and for continuously fostering my interest in this field of programming, and Professor Catarina Santos, for not letting me give up on writing this thesis and for encouraging me throughout the process. I would also like to thank the Master's coordinator, Professor Célio Pina, for giving me the opportunity to pursue this Master's degree, as well as all the professors for the knowledge they shared during this academic journey.

I want to thank my parents for giving me all the opportunities they have given me, for motivating me to continue my studies, and especially for allowing me to stay an extra year so that I could complete my Master's thesis.

Finally, I would like to thank all my friends for accompanying me throughout this academic journey, especially my Master's partner, André Oliveira, for his companionship throughout the Master's program and during the writing of this thesis.

Resumo

Atualmente existe uma dificuldade em interpretar imagens de Raio-X da zona do tórax, principalmente para médicos que não têm especialidade em radiologia, pois é uma tarefa complexa e apesar de estarem treinados para fazer essa análise, existe um grande grupo de doenças/patologias que podem ser detetadas radiologicamente na zona torácica. É aqui que entram as técnicas emergentes de Inteligência Artificial, como a "Computer Vision" e o "Machine (Deep) Learning", pois com a ajuda de ambas as técnicas é possível criar mecanismos de avaliação automática com elevado grau de certeza. Atualmente, já tenham sido publicados vários trabalhos, algoritmos, métodos e até soluções prontas para a avaliação automática dessas imagens, elas ainda não atingem o nível de precisão necessário.Por conseguinte, este é considerado um problema sem solução.

Esta dissertação é uma nova tentativa de automatizar e melhorar o processo de avaliação de imagens de raios X do tórax. A principal contribuição deste trabalho visa resolver o problema de desequilíbrio dos conjuntos de dados de domínio público, implementando um procedimento de "Aumento de dados", que permite melhorar o desempenho/precisão dos modelos de classificação de Deep Learning desenvolvidos anteriormente. Com isso, é possível aumentar o número de imagens de patologias sub-representadas. Testámos o nosso método utilizando o conjunto de dados CheXpert, que será descrito em pormenor mais adiante. Outra contribuição é o facto de dividir o conjunto de dados em vários subconjuntos binários e treiná-los isoladamente. Neste sentido, como mencionado, foi criado um subconjunto para cada patologia em vez de avaliar todas as patologias em conjunto. Selecionámos e afinámos dois modelos de classificação de aprendizagem profunda de alto desempenho desenvolvidos anteriormente: VGG19 e DenseNet121. No final, foi obtida uma tabela com todos os valores de AUC antes e depois do Data Augmentation, bem como um gráfico para cada patologia e para cada modelo. Estes passos resultaram em valores médios de AUC de 0,68 e 0,74 antes do Data Augmentation e de 0,96 e 0,97 após o Data Augmentation, para o modelo VGG19 e o modelo DenseNet121 respetivamente.

Palavras-chave: Radiografias do tórax, Doenças pulmonares, Machine (deep) learning, Aumento de dados, Inteligência artificial.

Abstract

There is currently difficulty in interpreting X-Ray images of the chest area, especially for physicians who do not have a specialty in radiology, because it is a complex task and although physicians are trained to do this analysis, there is a large group of diseases/pathologies that manifest themselves radiologically in the thoracic area. This is where emergent Artificial Intelligence techniques, such as Computer Vision and Machine (Deep) Learning come in, because with the help of both techniques, it is possible to create mechanisms that automatically assess with a high degree of certainty those X-Ray images. Although, at present, several papers have been published and various algorithms and methods exist, as well as some off-the-shelf solutions to evaluate these images automatically, they still do not have the required level of accuracy. Therefore, this is considered an unsolved problem.

This thesis is a further attempt to automate and improve the process of evaluating chest X-ray images. The main contribution of this work aims to try the imbalanced problem of public domain datasets by implementing a "Data Augmentation" procedure, which allows enhancing the performance / accuracy of previous developed deep learning classification models. With this, it is possible to increase the number of images of pathologies underrepresented. We test our method using the CheXpert Dataset, which will be described in detail afterwards. Another contribution is the fact of dividing the dataset into several binary subsets and train these alone. In this sense, as mentioned, a subset was created for each pathology instead of evaluating all the pathologies together. We selected and fine tuning two previously developed high-performance deep learning classification models: VGG19 and DenseNet121. In the end, it was obtained a table with all the AUC values before and after Data Augmentation, as well as a graph for each pathology, for each model. These steps resulted in average AUC values of 0.68 and 0.74 before Data Augmentation and 0.96 and 0.97 after Data Augmentation, for the VGG19 model and the DenseNet121 model respectively.

Keywords: Chest X-rays, Lung diseases, Machine (deep) learning, Data augmentation, Artificial intelligence.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI - Artificial Intelligence

AP - Anteroposterior

AUC - Area Under the Curve

AUROC - Area Under the Receiver Operating Characteristics

BR - Binary Relevance

CADe - Computer-Aided Detection

CADx - Computer-Aided Diagnosis

CC - Classifier Chain

CDSS - Clinical Decision Support Systems

CNN - Convolutional Neural Networks

CT - Computed Tomography

DEXA - Dual-Energy X-Ray Absorptiometry

FDA - Food and Drug Administration

GANs - Generative Adversarial Networks

GPU - Graphics Processing Unit

ICU - Intensive Care Unit

LP - Label Powerset

MRI - Magnetic Resonance Imaging

PA - Posteroanterior

PET - Positron Emission Tomography

RL - Representation Learning

RNNs - Recurrent Neural Networks

ROC - Receiver Operating Characteristic

SVM - Support Vector Machines

VAEs - Variational Autoencoders

1. Introduction

1.1. Background and Motivation

Chest X-Ray images have long been used to diagnose various lung diseases as well as other lesions of the organs located in the chest. The need to obtain high quality and non-invasive images of the human body, in particular, the thorax region has revolutionized the medicine field because it allows medical professionals to detect a wide range of diseases like cancers and pulmonary infections (Rubin et al., 2020). The early and precise diagnosis of these diseases is of significant importance because it can have a significative impact on patient diagnosis and treatment strategies (Wiener et al., 2015).

Pulmonary diseases like pneumonia, tuberculosis, and lung cancer are still one of the main conditions of global health problems, leading to thousands of deaths annually (Moon et al., 2023; World Health Organization, 2021). An early and effective detection of these diseases can lead to better results for the patients' health and can increase the survival rate and can reduce the healthcare costs. Radiography of the thoracic region, as it is usually the first step in the diagnosis of lung diseases, plays a crucial role in the early detection of these diseases.

Chest X-Rays has a significant role and "Physicians who are not specialists in radiology sometimes have difficulty reading and understanding the more complex patients' cases (images)". This is where emergent Artificial Intelligence techniques, such as Computer Vision and Machine (Deep) Learning come in, because with the help of both it is possible to create mechanisms that automatically assess with a high degree of certainty those X-Ray images (Esteva et al., 2019). Although, at present, several works have been published and various algorithms and methods exist, as well as some commercial (off-the-shelf) solutions to evaluate these images automatically, they still do not have the required level of accuracy. Therefore, this is considered an unsolved problem nowadays.

Due to these needs, this thesis aims to explore and improve the application of AI methods in the analysis of X-Ray images of the thoracic region for the early detection of some (more common) lung diseases, as the use of these technologies could lead to a significant advance in the diagnosis and treatment of these conditions, providing better results for patients and optimizing available health resources.

1.2. Problem Statement

The interpretation of Chest X-ray images, within the context of diagnosing lung diseases, is a complex and critical task. Although professionals in the field have improved their techniques, there aren't enough radiologists compared to the number of patient cases to evaluate. Inconsistencies in interpretation and the possibility of human error have led to a need for automated tools to help in the diagnostic process (Rubin et al., 2020). To this problem can be added the fact that traditional diagnostic

methods fail to achieve the necessary precision, and this is where AI-assisted diagnostic methods come in (Esteva et al., 2017).

The problem at hand is to improve the accuracy and fidelity of Machine (Deep) Learning models in detecting lung lesions in X-ray images of the thoracic region. In this regard, several (annotated) benchmarking datasets have been developed and made available in the public domain. One example is the CheXpert dataset (Irvin et al., 2019).

1.3. Objectives

This thesis has two overall objectives:

(1) Developing a method to correct the imbalance of class instances identified in chest X-Rays datasets, in particular, we will use the CheXpert (a reference database), which despite containing information on more than 200.000 patient cases, i.e., images (chest X-rays) and associated metadata shows a large imbalance in the number of patient cases for each of the fourteen pathologies/diseases collected. Therefore, we will explore emerging data augmentation techniques to balance the number of patient cases (images) in each of the fourteen (14) pathologies/diseases identified.

(2) Improving the robustness and accuracy of previously developed Machine (Deep) Learning models for the early detection and/or diagnosis of lung diseases.

1.4. Thesis Contributions

The main contribution of this thesis will be the improvement of the early detection/classification techniques using computer vision, deep learning and AI algorithms and methods, more specifically the introduction of data augmentation techniques to improve already well-known models. This data augmentation allows to enrich / fit the datasets imbalanced problem thus, improving the accuracy, robustness, and reproducibility of deep learning models.

1.5. Thesis Structure

This thesis is structured into five chapters; the first chapter discusses the background, the problem, the objectives, and the contributions of this thesis to its solution.

The second chapter presents a literature review and a state-of-the-art analysis that covers core concepts, such as X-ray imaging of the thoracic region, advances of Artificial Intelligence, in particular computer vision and machine (deep) learning in the field of medical imaging and novel existing approaches to analyzing X-ray images of the chest.

The third chapter presents and describes the dataset, and methods used / developed to analyze the dataset, detailing all the steps included, like the pre-processing, the data augmentation, the models used / evaluated and the training and testing processes.

The fourth chapter will present and discuss the main results obtained, showing the results of the models implemented with and without the use of data augmentation techniques. The results will then be compared with each other, as well as with the results of other studies that have already been published.

The last chapter summarizes the main conclusions and contributions of the research. It will also discuss some limitations of the work, as well as potential future work in this field.

2. Literature Review

Chapter 2 reviews the literature, analyzing the principal ideas / contributions and findings in the connection between Thorax (Chest X-Rays) images and AI algorithms and methods for the detection/classification of diseases located in the thorax region. This chapter presents the progress of AI technologies in medical imaging from the early used techniques to the more advanced deep learning models / frameworks by resuming the existence knowledge and the investigation results. Also, it is establishing the bases for the following discussions of the utilization of AI algorithms and methods to improve precision and efficiency in analyzing thorax radiographies.

2.1. Overview of Chest X-Ray Imaging and Pathological Lesions

Thorax X-Ray imaging started at the end of the 19th century when Wilhelm Conrad Roentgen discovered X-Rays in 1895 (Mould, 1995). Thorax radiography and fluoroscopy were rapidly used in medical diagnosis, including the detection of pneumonia, pneumothorax, and tuberculosis. The anatomic and pathological signals were identified and validated as the interpretation of thorax radiography was developed over time. With the rapid advancement of technologies, X-Ray imageology is a diagnosis tool that is fast, accessible, economical, uses low radiation and is utilized in 30-40% of all X-Ray investigations [ref]. Chest radiography is still a pilar of modern medicine, with its utility to overcome some medical specialities. Due to his bidimensional representation of a tridimensional anatomic structure, its interpretation can be hard (Jones et al., 2021). This can be even harder based on the patient's position, inspiration level, medical equipment, or object in the vision sight, as well as subtill density alterations between the pathology and the normal surrounding structures.

Even with these difficulties, thorax radiography is still an essential technique in medical diagnosis, especially in emergency (ICU – intensive care units) scenarios, where it is frequently used as the primary method of image evaluation of the chest area. The health professional that uses these systems must know the thorax radiography techniques and potential artefacts/problems, as it can alter the final image’s appearance and hide significant anomalies (Broder, 2011).

Due to longer acquisition times and limited availability, Magnetic Resonance Imaging (MRI) is less used than Computer Tomography (CT) in thorax imaging. However, it is preferred for certain indications, such as evaluation of cardiac function, mediastinal masses, and vascular anomalies, because it provides superior soft tissue contrast (Yamasaki et al., 2024). Besides, the latest modalities, positron emission tomography (PET) and Dual-energy X-Ray absorptiometry (DEXA) help obtain a more complete of the thoracic pathology. While PET exams provide useful information about metabolic activity, DEXA helps evaluate bone density, and both helps to evaluate and monitor cancers (Broder, 2011). Despite the expansion on the various numbers of imaging modalities, knowing which one to use when evaluating chest images, is based on the clinical indication, availability, patient factors and the

diagnostic in question, sometimes being required multiple modalities to optimize the performance and patient care.

A variety of anomalies can be seen on a chest X-Ray, however, a few prevalent illnesses stand out, like pneumonia, which is an inflammatory lung illness, that typically shows up on a chest X-Ray as segmental or lobar distribution. It can also manifest as localized or multifocal airspace opacities. It can result in bacterial, fungal, or viral infections (Cillóniz et al., 2021). Pleural Effusion is marked by the buildup of liquid in the pleural space, and it is characterized by the blunting of the meniscus sign or costophrenic angles in lateral views. In more extreme cases, it can occur a displacement of the mediastinum (Ferreiro et al., 2020). The pulmonary nodes are isolated growths that can be either benign or cancerous, and that is why they must be evaluated carefully, focusing on the size, shape, or growth pattern to select the best treatment approach. Malign nodules are often associated with lung cancer, but the benign nodules can be produced by malign ones, infections, or inflammation (Mazzone & Lam, 2022). Atelectasis is the loss of volume and the displacement of neighboring structures due to the collapse of the lung and it usually manifests as an opacification in the damaged lung lobe. Multiple diseases, such as airway obstruction, pneumonia, or surgical interventions, can cause atelectasis (Zeng et al., 2022).

Lung cancer is one of the most common pathologies detected/found in thorax radiographies, and it is one of the main causes of death in the world. It can differ, based on size, location, and histological subtype of the tumor, from solitary nodules to massive lesions. Small cell lung cancer and non-small lung cancer are two of the several types of lung cancer that differ in clinical manifestations and treatment methods (Nooreldeen & Bach, 2021).

Due to the difficulty of reading two-dimensional images of a three-dimensional anatomical structure studying thorax radiographies is full of difficulties. Difficulties include overlapping structures, little irregularities and problems with the depths and the projection of the images. The subjective interpretation of the radiologists and the variability between the observers happens because the diversity of the image, which is affected by the patient's anatomy, by the disease and by the imaging procedures. This leads to difficulty in creating a uniform diagnosis criterion. The analysis of chest X-Ray is more based on the visual examination and qualitative assessment, which leads to subjectivity and interpretation errors. This emphasizes the need for more objective techniques, like automated algorithms and computer-aided diagnosis systems (Dreyer et al., 2023).

Technological developments in AI, especially in the areas of computer vision and machine (deep) learning, are promising to increase the precision and effectiveness of chest radiographies interpretations. By utilizing large datasets and high-end machine (deep) learning algorithms AI systems can assist radiologists to identify anomalies, to evaluate the seriousness of a disease and to predict the patients results. However, for AI to be utilized successfully in medical decision making it is necessary to solve questions, such as large quality annotated datasets, developing and validate algorithms and

methods, establishing verified clinical workflows, and no less important to consider ethical considerations. Overall, to address those difficulties and increase the caliber and availability of chest X-Ray analysis to improve patient outcomes and healthcare delivery, it is necessary to keep researching and innovating this field of medical imaging (Majkowska et al., 2020a).

2.2. Evolution of AI in Medical Imaging

In the last decades, there were many breakthroughs in the application of AI in medical imaging, which revolutionized the way healthcare professionals read and utilize imagological data for diagnosis, treatment planning and patient management. These accomplishments show how AI technologies have developed, they also show how large datasets containing medical images are becoming more accessible and how AI-based solutions are growing increasingly in clinical contexts.

One of the first turning points in the use of AI in medical imaging was the introduction of computer aided detection / diagnosis (CADe/CADx) systems in the 1990s. The main objective of the CADe system was to help radiologists identify lesions in medical images, like thorax radiographies to detect pulmonary nodules or for breast cancer detection. Although these CAD systems had some precision problems and false positive rates, they set the terrain for the later developments in AI based medical imaging techniques (Barragán-Montero et al., 2021).

Machine learning techniques, including support vector machines (SVM) and neural networks, became popular at the beginning of the decade of 2000 for the analysis of medical images (Erickson et al., 2017). Thanks to these methods scientists managed to create AI models more complex capable of recognizing patterns and specific characteristics inside medical images. Convolutional neural networks (CNN) that transformed / enhanced the computer vision techniques and thus, increased the precision of images identification of locations including medical images were important discoveries for diseases detection in chest X-Rays (Ayalew et al., 2024).

The performance of image analysis and other tasks using AI improved significantly at the end of the decade of 2000 and in the early of the 2010 due to the growth of the availability of large datasets, increasing computer power and storage capacity and advances of deep learning techniques powered by a graphics processing unit (GPU). CNNs and other deep learning models have shown to be superior to human specialists in multiple tasks of medical image analysis such as predicting outcomes, images segmentation and disease detection (Barragán-Montero et al., 2021; Najjar, 2023).

A significant development in recent times has been the incorporation of AI based solutions in clinical practice. Images interpretation, optimization of workflow and Clinical Decision Support Systems (CDSS) through AI-based technologies can now be obtained by radiologists to help in certain activities. This AI intelligence system demonstrated the capacity to increase the healthcare provided to patients, reduce the reporting times and improve diagnostic accuracy (Najjar, 2023).

AI revolutionized the way healthcare professionals can utilize and understand imaging data in clinical practice, having a significant impact on workflow efficiency and diagnostic precision.

Among the main ways in which AI has improved diagnostic accuracy is the fact that it can give radiologists more sophisticated tools to interpret and analyze images. With the capacity to analyze large amounts of medical imaging data, the ability of pattern detection and the ability to detect features that may be invisible to the human eye, AI algorithms can help radiologists diagnose patients in a quicker and more precise way helping them detecting anomalies with her sensibility and specificity.

AI also offers the potential to reduce diagnosis errors and radiologists' variability, offering impartial and uniform evaluations of medical images. AI based solutions can help to reduce the variability and partiality of the observers by utilizing standardized algorithms and quantitative measurements, producing more accurate and precise diagnosis results (Najjar, 2023; Oren et al., 2020). Besides diagnostic precision improvement, AI has also been fundamental to improve workflow efficiency in the departments of medical imageology. Radiologists can focus on harder diagnosis tasks and patient healthcare activities, and use AI based solutions to automate repetitive tasks like the preparation, segmentation, and annotation of medical images. AI, can help radiology departments increase efficiency, reduce response times, and satisfy the growing demand for imaging services by optimizing routine workflows and eliminating some human labor (Jin et al., 2022; Juluru et al., 2021).

CDSS using AI can also classify imaging studies according to clinical urgency and complexity, which can help radiologists assign resources more effectively and do triage cases. AI technologies can help to speed up the diagnosis and treatment of patients with fatal diseases, improving the results and reducing healthcare costs through the recognition of crucial conclusions and highlighting the more complex cases for an immediate evaluation (Jin et al., 2022; Kotter & Ranschaert, 2021). Overall, AI has a significative effect on medical imaging workflow efficiency and diagnosis precision. AI has the potential to drastically change radiology as a profession and the way healthcare providers deliver medical imaging services as they develop and grow. Healthcare organizations can improve patient care, increase operational efficiency, and save more lives using AI based solutions to complement human knowledge and optimize clinical rutinary operations. To ensure a safer and more responsible utilization in clinical practice, AI implementation requires the resolution of a series of questions and ethical preoccupations. To effectively manage these difficulties which cover technological, regulatory, ethical, and social challenges, it is required a multidisciplinary teamwork and careful study. The first challenges are the technological ones, where it is required a strong validation and evaluation of performance. It is necessary to train and validate AI models in large scale datasets to develop models that will perform good evaluating and analyzing a wide range of patient groups, imaging modalities and clinical contacts. It is essential to evaluate safety, accuracy, and reliability to reduce diagnosis errors and unfavorable outcomes (Brady & Neri, 2020; Geis et al., 2019). Moving to the regulatory challenges, where it is necessary a strict regulatory accreditation. To maintain compliance with regulations like Food and Drug

Administration (FDA) clearance procedure, medical devices, and AI fed software must go through multiple strict regulatory review processes. Regulatory abbreviation requires clinical efficacy, safety, and demonstration of performance which can require a lot of resources and time from the developers (Naik et al., 2022).

Ethical considerations can be divided into two main ones. The first is the algorithm bias and fairness, where it is essential to reduce bias in algorithms and ensure impartial decision-making based on AI. Continuous discrepancies in healthcare can be caused by datasets that are biased which requires the need off openness, verification, and supervision to ensure the quality of healthcare applications (Geis et al., 2019; Naik et al., 2022). Additionally, the use of sensitive patient privacy and data security, since AI algorithms use sensitive patient data, questions regarding consent, privacy and data security are raised. Ensuring the anonymity, encryption and secure storage of patient data is key to maintaining patient privacy and meeting data protection regulations (Farhud & Zokaei, 2021). Finally, workforce displacement and professional autonomy, are significant social challenges, because it is required education, formation, and development programs in workforce education to address concerns that healthcare professionals will lose their jobs and become less qualified because of AI. It is required clear standards and oversight procedures to preserve the autonomy and clinical judgment of healthcare professionals when using AI systems (Farhud & Zokaei, 2021).

In summary, is required a thorough methodological approach to manage the difficulties and moral dilemmas associated with integrating AI in the medical imaging field. Healthcare professionals can take advantage of AI as it promises the improvement of patient care, healthcare delivery, and ethical and responsible innovation in medical imaging.

2.3. Representation Learning and its Importance

In the medical imaging sector, representation learning (RL) is essential because it allows to identify and extract automatically significant representations data of unprocessed images. This is especially relevant in the context of medical imaging where it is necessary to identify subtle and diverse characteristics associated with multiple diseases and anatomic structures due to the complexity and variety involved in medical images.

CNNs are a class of deep learning models that can automatically learn hierarchical representations of image data through successive layers of convolution and pooling operations, and they are a popular method used to learn the representation of medical images. In tasks like image classification, segmentation, and anomaly identification, CNN's have demonstrated an impressive performance (Getty et al., 2021; M. Kim et al., 2019). For example, CNN's can be trained to recognize distinctive patterns of nodules / masses and infiltrations that suggest the presence of malign lesions in Chest X-Ray or CT images, improving the speed and accuracy of the diagnosis of lung cancer. Unsupervised Learning is another type of RL technique used in medical imaging that utilizes unlabeled

data to acquire significative representations without the need of explicit supervision. When using autoencoders and generative adversarial networks (GAN), noise removal, image reconstruction and synthesis can be easily done to identify patterns and connections in medical imaging (Getty et al., 2021). RL is also critical for the integration of multimodal and multi-scale data to combine multiple sources of information, like imaging, clinical, genetical and demographic data for improving clinic decision-making and diagnosis precision. These methods make it easier to joint analysis tasks on precision medicine settings, by fusing different types of data in a single feature space (Tiwari et al., 2011). RL methods allow to automatically extract relevant characteristics from unprocessed imaging data, which makes it a powerful tool for medical imaging analysis. These algorithms increase the diagnosis effectiveness, efficiency, and interpretability of medical images systems, by learning directly from hierarchical and discriminative features from images, which ends up improving patient healthcare and clinic practice results.

In medical imaging, RL approaches, especially deep learning ones, like CNNs have been gaining importance in tasks that involve the extraction and categorization of features. These methods have been successful in automatically deriving discriminative representations from unprocessed image data, which makes it possible to analyze medical images with precision and quickness for multiple applications of diagnosis and prognosis (S.-C. Huang et al., 2021). To represent anatomical underlying structures and clinical abnormalities, characteristics extraction, the process of extracting main features from raw image data, is a fundamental step in medical imaging analyses. By automatically learning hierarchical image data representations though successive layers of convection and pooling processes, RL systems are excellent in feature extraction. These features allow an efficient differentiation between various types of tissues, lesions, and diseases by capturing local and global patterns, textures and structures observed when analyzing medical images (Yan et al., 2023). RL techniques are essential to recognize and classify images according to their attributes and content in a medical imaging context. CNNs in particular have been widely used for images classification tasks because they can be trained, with the help of imaging data, to differentiate between regular and unregular results, classify multiple types of diseases and predict patient outcomes (Milecki et al., 2023).

RL is a powerful tactic that uses models, pre-trained in large datasets to address the lack of data and the difficulties of the medical imaging field. To adapt to determined imaging modalities, diseases or clinical situations, CNN models can be pre-trained on large imaging datasets, such as ImageNet and can be refined on smaller medical imaging datasets. With the help of this transfer learning strategy, medical image analysis tasks with insufficient labeled data can be performed more efficiently by researchers and clinicians, using information obtained in other domains (Zheng et al., 2023). However, RL in medical image analysis offers several benefits and some limitations. One advantage is automatic feature extraction, where techniques like CNN's can eliminate the need for handcrafted feature engineering, by automatically extract pertinent characteristics from unprocessed medical image data (Chan et al., 2020; Shen et al., 2017). This leads to an improved performance, since RL algorithms are

often superior to manual techniques, when extracting characteristics, due to learning directly discriminative representations from the (pixels) data (Chan et al., 2020; Shen et al., 2017). Additionally, transfer learning is made possible by RL, which reduces the effects of a reduced number of data, by allowing pre-trained models to be improved by other medical images datasets (Chan et al., 2020; Kumari & Singh, 2023). These transfer learning methods also support multimodal integration which helps to integrate multimodal image data in order to conduct a comprehensive analysis and interpretation (Chan et al., 2020). The last advantage of RL is its high adaptability making them versatile in a wide range of medical imaging tasks because they can be adapted to different modalities, anatomical locations, and clinical circumstances (Chan et al., 2020). However, RL methods also present some disadvantages, like generalization, since most parts of research use small training datasets and lack a strict validation with extensive data from the real world, it is currently unknown if deep learning models can be applied to new patients or multiple clinical contexts (Shen et al., 2017). Another issue is data shortage, as it is expensive to gather a high number of anomalous cases to cover all the variabilities of images characteristics, and it's a challenge to collect enough medical images that englobe all patients' cases (distribution) with reliable annotations (Shen et al., 2017). Ethics and privacy are also common critical issues, as these concerns impede the progress/improvement of deep learning methods by prohibiting the medical imaging data community from openly distributing/sharing patient cases. (Dhar et al., 2023).

2.4. Existing Approaches for Chest X-Ray Analysis

Traditional image processing techniques have been essential when interpreting thorax radiographies, and they are useful in processes to improve image quality, segment anatomical structures, organs and tissues, extract characteristics, identify patterns and perform quantitative analyses. Radiologists have been using these methods to help identify anomalies, to assess the severity of diseases and to monitor the effectiveness of treatments (Giełczyk et al., 2022).

There are multiple conventional digital image processing algorithms and methods to easy the analysis of chest X-rays, like image pre-processing filters, which involve improving their quality, by eliminating artefacts, noise, and unnecessary data. In order to increase the anatomical structures and anomalies visibility, the most common procedures are noise reduction, contrast enhancement and histogram equalization. Also, the most elaborate, and popular methods for image segmentation tasks, where chest X-ray images are segmented using region extraction techniques for organs segmentation, like the heart, the ribs and lung fields. Other tasks, including lesion detection, measurement and classification can benefit from this procedure. Common techniques include active contour modelling, region growing and thresholding. Feature extraction is also commonly used to analyze chest images, the extraction of features quantifies essential attributes to represent anatomical structures and anomalies, which can be linked / associated with clinical variables. There are multiple techniques for capturing

spatial patterns and statistical characteristics of pixel intensities, like texture analysis, shape descriptors and intensity-based features (Meedeniya et al., 2022a; Mustafa & Nsour, 2023).

With the help of techniques, like pattern recognition / machine learning, that utilize extracted features from images' pixels of the Chest X-ray to classify theses into different diagnostic categories (e.g., pneumonia, atelectasis, cardiomegaly). Decision trees and support vector machines are two examples of machine learning algorithms, which are used to distinguish between normal and non-normal images, as well as different disease situations (Meedeniya et al., 2022a; Mustafa & Nsour, 2023). To assess the seriousness, progression, and response to the treatment of a disease, the radiographic parameters of chest X-ray images are quantitatively measured. Some examples of the parameters used in quantitative analysis are Lung capacity, airway measurements, the size of the cardiac silhouette and the evaluation of the density of lung opacities (Meedeniya et al., 2022a; Mustafa & Nsour, 2023). Although traditional image processing methods are useful, there are also multiple disadvantages, like the susceptibility to artefacts, handcrafted characteristics dependences, difficulties with complicated patterns and heterogeneity between observers. A more automated, dependable, and accurate analysis of Chest X-rays images is now possible thanks to the development of emergent machine (deep) learning and other AI based methods, which improve and complement the more established methods in clinical contexts (Subramaniam et al., 2023).

The use of machine (deep) learning techniques has considerably improved the interpretation of chest X-ray images, accompanying the development of the quantitative paradigm in radiology. These automated methods include data-driven methods for analyzing chest X-ray images, which help radiologists identify abnormalities, make medical diagnoses and predict patient outcomes.

When it comes to machine learning techniques, can be divided into different techniques, the first one, supervised training, is when machine learning models are trained on labelled datasets using supervised learning methods, and then each image from the dataset gets a corresponding label, like normal, pneumonia, tuberculosis, or other pathology. The most common supervised learning techniques used are support vector machines, logistic regression, decision trees, random forests and ANNs /DNNs (Ahmad et al., 2023; Tiu et al., 2022).

The second technique, feature engineering, is used to extract relevant features from chest images to represent anatomical structures and anomalies. In this technique, machine learning models receive features like texture descriptors, shape features and intensity histograms to help categorize and predict. Another interesting technique is the ensemble methods, which allow to combine of multiple machine learning models increasing the strength of the prediction. The most used techniques are bagging, boosting, and stacking. They are used to combine multiple data sources and reduce overfitting (Ahmad et al., 2023; Tiu et al., 2022).

There are also various deep learning methods like Convolutional Neural Networks (CNNs). CNNs use consecutive convolutional layers and pooling processes and can automatically learn

hierarchical representations of image data, which allows them to detect accurately and characterize anomalies in chest X-ray images (Meedeniya et al., 2022b). Another powerful technique often applied in conjunction with CNNs is transfer learning, which uses pre-trained CNN models on large image datasets like ImageNet and then refines them on smaller medical image datasets. While mostly applied to CNNs, transfer learning can also be used with other types of neural networks. With less labelled data, deep learning models can adapt to chest X-ray interpretation tasks accordingly, which improves generalization performance and reduces the need for extended training (Ait Nasser & Akhloufi, 2023; Meedeniya et al., 2022b).

Another interesting and very explored method is the Recurrent Neural Networks (RNNs) that have the function of making sequential data processing when interpreting chest radiographs. RNNs are useful for applications like modelling disease progression and predicting treatment response because they can extract temporal dependencies and sequential patterns from time series data (Meedeniya et al., 2022b). The last method studied on this work is the Generative Adversarial Networks (GANs) which are used to increase data (i.e., creating new artificial images). They can produce realistic synthetic images of chest X-rays, which can be used to complement the training data and improve the strength and generalization capabilities of deep learning models (Ait Nasser & Akhloufi, 2023; Meedeniya et al., 2022b).

Machine (deep) learning (MDL) techniques have improved the interpretation of chest X-rays images, providing a more scalable, accurate and automated medical image analysis. These methods are still in development but have the potential to improve medical imaging, by providing an earlier identification, giving more accurate diagnosis, and facilitating the decision to choose individualized treatment methods for individuals with diseases related to the chest area.

The methods described before work in different ways according to accuracy, speed, and interpretability. In terms of accuracy, MDL methods show a high accuracy in the detection, localization, and classification of diseases, they just vary in the type of techniques used. Deep Learning techniques use mainly CNNs, achieving a top performance in the interpretation of chest X-rays. Machine Learning techniques use logistic regression, decision trees and SVMs (Wu et al., 2020).

Concerning speed, Deep Learning methods are used in Bigdata scenarios (when analyzing more complex / data Interpretability intensive problems), but they can perform a fast chest X-ray image analysis, which makes them appropriate for real-time applications in clinical practice. Machine Learning methods are simpler, use less computer power resources and have a reduced training time compared to the Deep Learning ones. They are ideal for applications such as emergency triage and diagnostics at the point of care (Majkowska et al., 2020b). Comparing the machine learning techniques, as they are based on manually created features (handcrafted), are easier for clinicians to understand / justify medical decisions. Deep Learning techniques, despite improving the interpretation are more difficult to understand (Majkowska et al., 2020b).

When analyzing chest X-ray images there are multiple possible approaches available and most of them are still being developed nowadays. These developments have the objective of increasing diagnostic accuracy and workflow and providing an individual strategy to care for each patient.

In deep learning-based approaches, attention mechanisms have become a more effective tool. This tool allows models to ignore information that is not important and only focus on the more relevant areas in the images, improving interpretability and diagnostic performance (Meedeniya et al., 2022b). Self-supervised Learning techniques use the data from the dataset that is not labelled and pre-train them on auxiliary tasks, like image painting or rotation prediction. This increases the amount of data provided to the model, making better use of the data that is available. Additionally, multi-modal integration new developments make it possible to combine data from different modalities, like merging images from X-rays and CT scans. This helps on a deeper analysis and interpretation of images data, which leads to increased accuracy (Ait Nasser & Akhloufi, 2023).

Explainable AI methods have also been improved to justify machine (deep) learning model decisions. Methods like interpretability techniques, work by highlighting parts of an image showing some pathology or have patterns that can help diagnose a disease. This is made by techniques like gradient-based visualization, saliency maps, and occlusion sensitivity analysis. Another explainable AI technique is uncertainty estimation, this technique, quantifies all the uncertainties that are related to a model prediction and uses procedures like ensemble approaches, dropout regularization and Bayesian neural networks to help doctors to achieve better accuracy when treating a patient (Sun et al., 2023).

Clinical Decision Support Systems (CDSS) have been included in the workflow, giving radiologists an almost instant diagnostic. They use features like lesion identification / recognition, disease classification and therapy suggestions and, like many techniques, help improve the accuracy of diagnostics. Radiologists can also now make annotations on images, by placing markers in any region of interest and use AI tools to help them make a correct annotation and a good diagnosis, this technology, called semi-automated annotation speeds up the process of annotation and evaluation (D. Kim et al., 2022).

Collaborative research initiatives are probably the biggest improvement in recent AI in the field of medical imaging. There are available large datasets, like the CheXpert dataset that has over 200.000 chest images, where each image is associated with a radiology report and is labelled with one of the 14 different pathologies that exist / represented in the datasets (Irvin et al., 2019). There is also the MIMIC-CXR dataset that provides a larger number of images, with over 350.000 images that correspond to 227,835 radiographic studies (Johnson, Pollard, Berkowitz, et al., 2019a). Both these datasets help improve different areas, like decision support, natural language processing and image understanding and help radiologists and researchers a huge help for the detection of chest diseases. In addition to these datasets, nowadays, researchers, developers and radiologists who work with chest X-ray images share each other's knowledge with the help of open-source frameworks and libraries like TensorFlow,

PyTorch and FastAI. By sharing all different findings and conclusions, and by sharing codes, and pre-trained models, the newest researchers and professionals already have a steady working base and can improve even more the classification score (Cohen et al., 2022).

2.5. Previous Studies of the CheXpert Dataset

The CheXpert dataset is one of the most used benchmarking datasets in the world used to train and validate machine (deep) learning models for classifying pathological lesions in Chest X-rays. It consists of a collection of multiple chest images (instances) created by researchers from the University of Stanford (Irvin et al., 2019).

There are multiple studies available that have applied different deep-learning techniques in order to analyze the CheXpert images. In the study by Rahimiaghdam & Alemdar, 2024 the CheXpert dataset was used to train different CNN architectures like InceptionV3, DenseNet121, VGG16, ResNet50. MobileNetV2, and InceptionResNetV2, in 7 different pathologies included in the CheXpert dataset, and accomplished a mean AUC of 0.89, 0.88, 0.86, 0.88, 0.86 and 0.88 respectively, as can be seen in Table 1.

Table 1 - AUC values of different trained classifiers (Rahimiaghdam & Alemdar, 2024)

Class name	IRNetV2	MNetV2	VGG16	ResNet50	DenseNet121	InceptionV3
Cardiomegaly	79	84	79	79	82	82
Pleural effusion	90	88	91	91	91	92
Lung opacity	90	90	88	89	90	91
Consolidation	90	90	90	89	90	90
Atelectasis	75	72	75	76	78	77
Pneumothorax	98	85	81	86	82	95
No finding	94	94	95	94	93	92

In another study, developed by (Zhang et al., 2022), it was developed a two-stage deep learning model combining region proposal, feature pyramid, and multi-instance learning, achieving a mean AUROC of 0.912 on the CheXpert validation set and 0.784 on an external test set for classification. The work carried out by (Pham et al., 2020) used different CNNs, along with a label smoothing technique and were trained on over 200.000 images of the CheXpert dataset and achieved a mean AUC of 0.940 in predicting five selected pathologies, Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion from the validation set. Lastly, in the study by (Allaouzi & Ben Ahmed, 2019) the aim was to compare models created by the authors to the DenseNet-121 model in terms of AUROC, and they concluded that the differences were minimal. The three used models can be seen in Table 2 and achieved an average AUC of 0.812 for the Binary Relevance (BR) classifier, an AUC of 0.808 on the Classifier Chain (CC) classifier and a 0.785 AUC on the Label Powerset (LP) classifier.

Table 2 - AUC values of different classifiers (Allaouzi & Ben Ahmed, 2019)

Classifier	Training time	Metrics		
		Micro-F1	Hamming Loss	Averaged AUC
BR(LR)	**18min47s**	0. 622	**0.116**	**0.812**
LP(LR)	7h15min53s	0.619	0.126	0,808
CC(LR)	22min33s	**0.631**	0.118	0.785

2.6. Benchmark Dataset

The CheXpert dataset is considered a benchmark dataset in the field of medical imaging research, and because of its large scale, clinically relevant findings, and high-quality labels, it has become the best choice for evaluating and comparing the performance of machine learning and deep learning models in chest X-ray analysis.

CheXpert has been widely used in research to train, validate and evaluate models, including Convolutional Neural Networks (CNN) and other deep learning architectures. The application of CheXpert has resulted in a comprehensive system of comparative results, which enables researchers to compare their models with the effectiveness of current methods.

Another justification for CheXpert being a great benchmark is its balance between quality and scalability. The dataset integrates structured, high-quality labelling with a considerable number of images, enabling researchers to train models that can generalize effectively. Unlike smaller or less structured datasets, CheXpert offers enough statistical power to support a robust training and evaluation, making it a perfect dataset for creating models that can handle real clinical variations.

Although there are several other datasets of chest X-rays, these have some limitations that make them less suitable for this study than CheXpert. The ChestX-ray14 dataset, released by the National Institutes of Health (NIH), has 112,120 chest X-ray images identified with 14 pathology labels. Even though it is a large dataset, the labeling process uses basic NLP techniques, which can result in inaccuracies. Unlike CheXpert, the NIH dataset does not deal with label uncertainty, which can lead to uncertainty during training. In addition, the metadata and quality of the labels in CheXpert are higher, making it a more reliable dataset for model development (Kufel et al., 2023).

The MIMIC-CXR dataset contains chest X-ray images and the corresponding reports, mainly of patients in the intensive care unit (ICU). Although it offers detailed clinical information, it is more focused on ICU situations, which may not apply adequately to more common clinical contexts. In addition, preparing the reports to obtain results requires considerable effort and knowledge. Compared to CheXpert, MIMIC-CXR offers fewer labeled images and lacks the same degree of scalability (Johnson, Pollard, Berkowitz, et al., 2019b).

PadChest is another extensive thoracic X-ray dataset; however, it presents a unique challenge: the labels and reports are predominantly written in Spanish. This adds an extra pre-processing step for non-Spanish NLP utilities. Although PadChest has potential for specific investigations, its use in global research has been restricted compared to CheXpert (Bustos et al., 2019).

3. Materials and Methods

Chapter 3 presents the material and methods explored / developed to improve the classification of chest X-ray images, trying to achieve the highest possible accuracy in the lowest possible time. In this chapter, the CheXpert will be introduced, followed by an explanation of the processes involved in addressing the problem, such as data augmentation techniques, RL techniques, and the model's development.

3.1. Dataset

The CheXpert dataset is known for its large diversity of chest pathologies (14), mainly lung pathologies, which are the most common ones found in clinical practice, like pleural effusion, pneumonia, cardiomegaly, pneumothorax, and pulmonary edema. By having a large number of different pathologies, the CheXpert dataset allows researchers to develop, manipulate and evaluate different machine learning algorithms to improve the classification of those pathologies. Besides the large difference in pathologies CheXpert dataset also offers a large-scale annotation, since all the images available in the dataset (more then 200.000) were carefully labelled by a team of professionals, being classified with binary labels that indicate if a pathology is present, with the label "1", and if the pathology is not present, with the label "0". In addition, the images also can have a label "-1", which means that the pathology indicated has some level of uncertainty present. The images presented on the dataset also have the particularity of being present with different qualities, and with the pathologies presented in different regions of the chest, which helps to have a more robust classification module that is as close as possible to what radiologists face when analyzing these pathologies on these regions. The CheXpert, due to its large community also facilitates the communication and the exchange of already known/developed models, or algorithms, which allows researchers to compare with their own models improving the large state of the art already available (Irvin et al., 2019).

The CheXpert dataset also presents a very advanced data collection process, its main goal is to simulate the actual clinical practice when analyzing the chest X-ray images. Those images were collected from different healthcare facilities and images centers with the objective to include the most variety of patient demographics, imaging processes and clinical contexts (Bressem et al., 2020; Irvin et al., 2019). There are protocols that are very strict when it comes to patient privacy and data security. Those protocols are implemented to comply with data security law and to ensure that patient personal data is correctly anonymized or even deleted from the image (Nguyen et al., 2022). In the process of images collection, it is used different imaging instruments and procedures to obtain the chest images. The dataset includes scans of posteroanterior (PA) and anteroposterior (AP) directions and there are different resolutions, exposure settings and quality of the radiographies because of the different facilities that obtained these images. To finish the collection process, all the images were evaluated by professional physicians, to make sure that they all met the requirements to be added to the dataset. All the images that had problems, like placement errors or artefacts that could affect interpretation, were

either deleted of put to the side to be evaluated again and then added back to the dataset (D. Kim et al., 2022; Majkowska et al., 2020b).

Withing the number of available datasets, CheXpert is one of the biggest and most used because of its strengths, like the large scale and diversity and the presence of a binary labeling system paired with an uncertainty label, which measures the level of confidence for each label on the images (Tiu et al., 2022).

On the contrary, the dataset also presents some limitations, one of them is the unbalanced number of cases for each label, varying from under 10.000 patients' cases for some pathologies to almost 100.000 patients' cases for other pathologies. This affects the performance when trying to train a model, being that this may affect the accuracy to classify pathologies with fewer instances (patients' cases), as those with more cases will be more biased. Another problem is the noise on the images, as some, despite the effort, have noise that can affect the labelling of the images. Additionally, the dataset doesn't include much information on the associated medical reports of the patients, like the medical history and some clinical exams (lab results) that can help a machine learning model improve its classification. Also, the dataset has a big problem, that is the lack of images for each pathology, meaning that most of the images are single-view images which can lead to a non-capture of the whole extension of the affected area (Jang et al., 2020).

The radiographies in the dataset were collected between October 2002 and July 2017 and they were collected by a group of researchers at Stanford University (Irvin et al., 2019). The dataset has radiographies from two different perspectives, frontal and lateral, the frontal images have two different views, anterior-posterior (AP) and posterior-anterior (PA), and the lateral images only have one view. Figure 1 shows an example of both perspectives, selected randomly from the dataset.

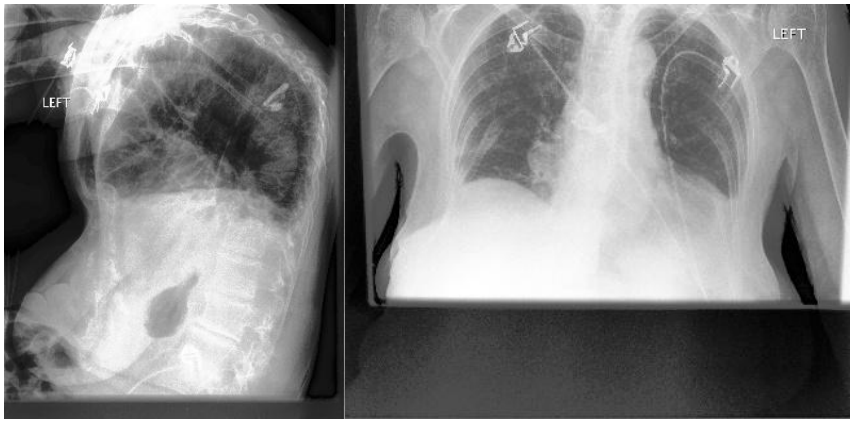

Figure 1 - Example of chest X-ray images from the CheXpert dataset. (A) – Lateral view. (B) – Frontal View (Irvin et al., 2019).

The dataset also has labels for different pathologies, consisting of fourteen different ones. There are fourteen different radiological observations, with twelve different pathologies, Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarge Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia and Pneumothorax. There is also one observation for the presence of Support Devices, and one named No Finding that indicates that there is no present pathology. Figure 2 presents a graphic of the classes' distribution. In the figure, it can be seen that "Lung Opacity" is the pathology in most cases, followed by "No Finding" which is the case where it was impossible to determine / associated a pathology, and that the ones with fewer cases are Pneumonia and Pleural Other. By analyzing the graphic, it is clear that it is a class imbalance dataset, which is one of the main limitations of this dataset. Dataset imbalance has a high / direct impact on the development and performance of a classification model.

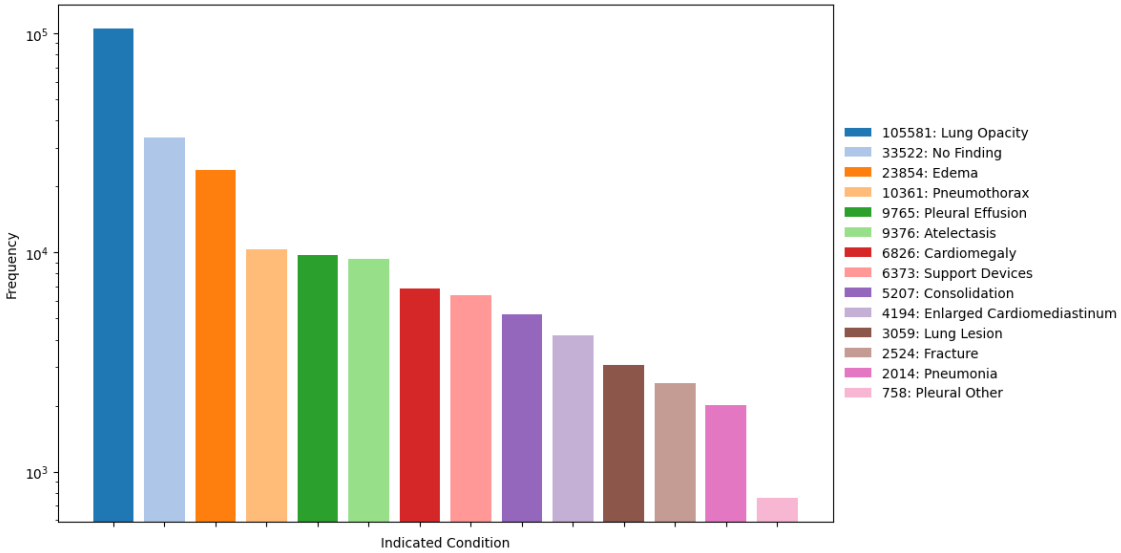

Figure 2 - Bar chart of the frequency of indicated conditions.

To help understand the correlation between the different pathologies, Figure 3 shows the correlation matrix of the pathologies and the "No Finding" class to provide a perception of how frequently different pathologies occur at the same time. By analyzing the matrix, there is a moderate positive correlation between "Consolidation" and "Atelectasis," "Pleural Effusion" and "Lung Opacity" and "Cardiomegaly" and "Edema" indicating that these conditions can occur with some frequency at the same time / together. There is also a negative correlation between "No Finding" and most pathologies, as expected, that indicates that when an X-ray image is classified as No Finding it can be classified as another pathology. Excluding the ones stated, almost all conditions show a week correlation between each other which indicates that most of the conditions happen alone in each radiography.

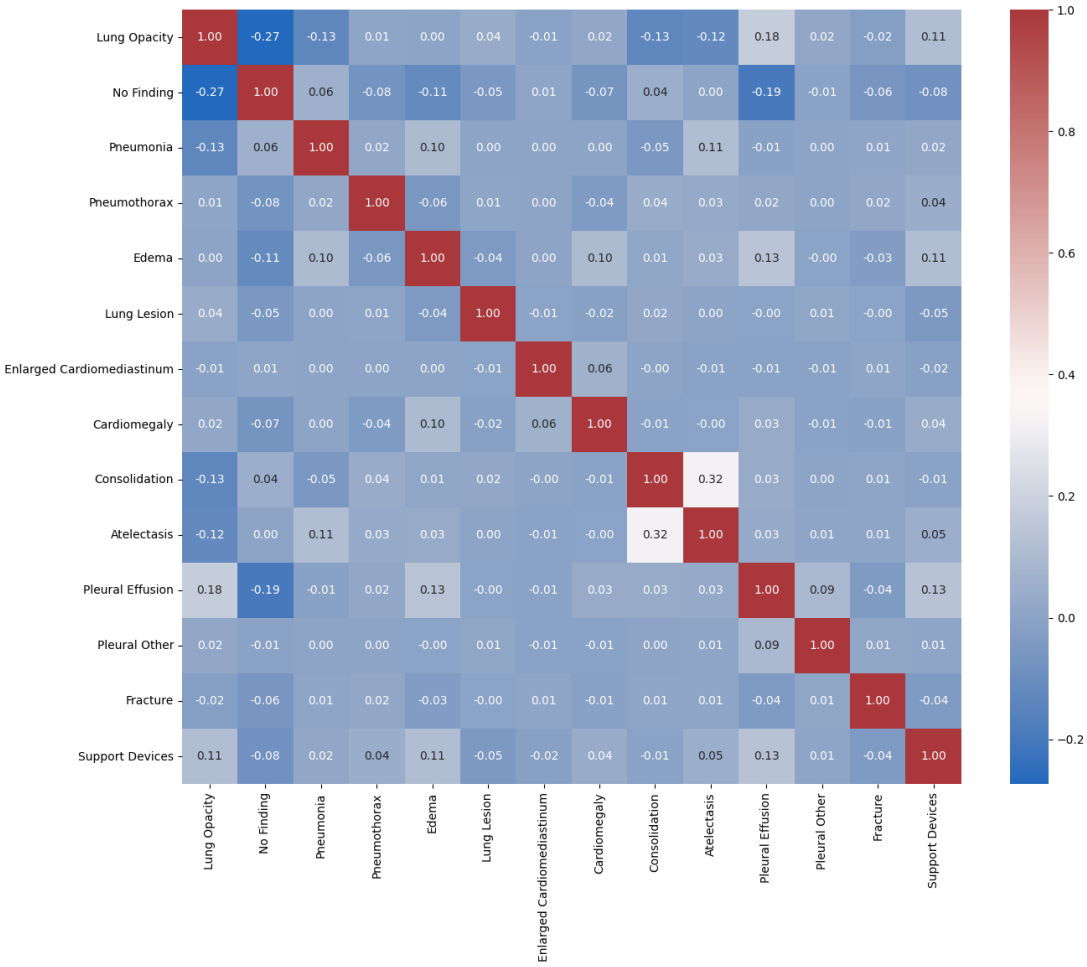

Figure 3 - Correlation Matrix of CheXpert Dataset Features

3.2. Proposed Method

In this part of the work, all the processes and developed methods used to achieve the desired result will be detailed, Figure 4 shows a flowchart of the sequential work done.

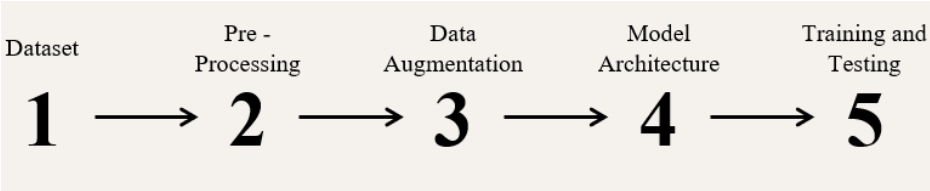

Figure 4 - Flowchart of the proposed method.

3.2.1. Pre-processing

In the area of medical imaging, it is often required to do some tasks to ensure that the dataset that will be used has the best quality possible, and that is where the pre-processing section enters. The first step is loading the dataset from a csv file that contains all the information relative to the medical images that will be analyzed, then the path to those images is added to the file and associates all images that are stored to each related line (specific patient case) on the dataset.

The following step aims to filter the dataset to include only images that were in the frontal view, because the number of images classified as “Lateral” was small, with only 32,387 cases out of 224,316. Next, all the images that had radiologic observations classified as “No Finding” or “Support Devices” were also removed from the dataset, leaving the dataset with 174,053 images.

While most models analyze all conditions / pathologies together, and it is a contribution of this work, we decided to separate each condition to train a binary classifier able to recognize/identify each class vs. the rest of the pathologies, so instead of having one dataset that englobes all pathologies, twelve (12) different datasets were created, one for each of the pathologies. In this step, a condition map algorithm was created that changes the dataset to have only one condition being analyzed and marking all other ones as no condition found, for example, in the case of Atelectasis, all images that had the pathologies didn’t change and all the other that didn’t have were classified as No Atelectasis. This piece of code is the only thing that differs from all datasets, which means that all the following processes were made for all twelve datasets. In the next step, we implement a procedure to remove all the columns that were not relevant, keeping only 3 columns, 'Path', 'Indicated_Condition', and 'full_path', and discard all the others to have a cleaner dataset with only pertinent information to be analyzed. Upon completion of the last steps, the dataset was split into two different subsets, training, and testing, this is a crucial step to achieve good training and hereafter, the evaluation of the produced machine learning model. This was made using a test size of 0.2, which means that 80% of the dataset will be used for training and the other 20% will be used for testing. In this part, it was guaranteed that in both subsets of the dataset, there was a balanced number of conditions and non-conditions. The patient's id was also considered because some patients have more than one image and so was created an algorithm to ensure that a patient did not have images in both subsets (training set and testing set).

To enhance images quality, there were preformed more pre-processing steps, such as background removal, noise removal, diaphragm region removal, and contrast enhancement that were adapted from a public code on Kaggle (Quý, 2023). The background removal process was made to eliminate any part of the image that was not relative to the chest area, in order to improve the image quality and the surrounding noise that can occur on X-ray images. The noise removal process is similar to background removal, but it focuses more on the parts inside of the image that can interfere with the classification. The noise removal process can help the clarity and the quality of the analyzed images. The diaphragm region removal is also a good step to improve the image quality, by excluding things that are not required in the analysis process, in this part, the diaphragm region, as it can interfere with the classification is removed from the image. The contrast enhancement, the last part of the pre-processing process, aims to enhance the visual contrast of the chest X-ray radiography. With this, the intensity levels of the pixels are adjusted to improve the contrast between the pathology and the anatomical structures.

3.2.2. Data Augmentation Techniques

The main goal of this work is to improve the accuracy of deep learning models, by training classification models using benchmarking the CheXpert dataset. As mentioned before, to fulfill that goal, instead of training models that tries to classify all conditions at the same time, in this thesis, we propose splitting all conditions and trying to classify them one by one making / developing a binary classification system.

The first part of the process was made in the pre-processing step, where the patients' cases data was split into twelve different binary datasets, one for each pathology. After that process twelve datasets were obtained, containing a different number of images for both condition and no condition (i.e., all other conditions). As shown in Table 3, it is possible to notice that there is an unbalanced number of images when comparing the ones with and without conditions. To balance the dataset, it is applied a data augmentation process, i.e., a set of small rotations to each image that belongs to the positive condition for increasing / obtaining new data without having to collect new images.

Table 3 - Number of images with and without condition for each pathology.

Condition	Atelectasis	Consolidation	Edema	Enlarge Cardiomediastinum	Fracture
Positive	29 720	12 983	49 675	9 187	7 436
Negative	144 333	161 070	124 378	164 866	166 617

Condition	Lung Lesion	Pleural Effusion	Pleural Other	Pneumonia	Pneumothorax
Positive	7 040	76 899	2 505	4 675	17 693
Negative	167 013	97 154	171 548	169 378	156 360

To ensure that the data augmentation process is performed correctly, and to avoid an image, after rotation, was present in both the training and test subsets, the data augmentation process was performed after the splitting/creation of the training and test datasets. Table 4 shows the number of rotations applied to each pathology, as some, such as Edema, only required 3 rotations while Pleural Other, due to the reduced number of images with the condition, required 68 rotations. Two cases require more attention, Lung Opacity and Pleural Effusion, despite having a similar number of images with and without the condition, it was required to also rotate the images with the pathologies because, without any rotation, the classification results for Lung Opacity and Pleural Effusion were not as high as the others.

Table 4 - Number of rotations per condition.

Condition	Atelectasis	Consolidation	Edema	Enlarge Cardiomediastinum	Fracture
No. Rotation Angles	4	12	3	18	22

Condition	Lung Lesion	Pleural Effusion	Pleural Other	Pneumonia	Pneumothorax
No. Rotation Angles	24	3	68	36	9

At the end of the data augmentation process, the datasets were populated with “new artificially created” images from images of the patients with pathologies and, due to all the care taken during the process, datasets were obtained that are overall twice as large as the initial ones, which leads to better prepare datasets for training. Table 5 shows the number of instances of each dataset, before and after the data augmentation process.

Table 5 - Final Sizes of Augmented Training and Test Sets

Condition	Train Dataset Before	Test Dataset Before	Train Dataset After	Test Dataset After
Atelectasis	139 242	34 811	234 346	58 587
Cardiomegaly			251 490	62 873
Consolidation			253 488	63 378
Edema			258 462	64 616
Enlarge Cardiomediastinum			271 542	67 877
Fracture			270 120	67 525
Lung Lesion			274 410	68 603
Lung Opacity			365 349	91 337
Pleural Effusion			323 799	80 951
Pleural Other			275 514	68 879
Pneumonia			273 882	68 471
Pneumothorax			266 628	66 662

3.2.3. Model Architecture

In order to classify and evaluate a large dataset, it is also necessary to have models that can keep up with the large number of images provided, and in this work, it was decided to use two models that are already well-known and worked on, which are DenseNet121 and VGG19. These models were selected because they have already been proved to work well with large datasets, especially ones with large number of images. Before running the models on the CheXpert dataset, both models were pre-trained in another well-known dataset, ImageNet and then they were tuned to the CheXpert dataset, for each condition both models were used to classify the presence or the absence of each condition.

The VGG19 model is a deep convolutional neural network that consists of 19 layers and 5 convolutional blocks, where each of them is followed by a max pooling layer and it was created by the

Visual Geometry Group at the University of Oxford. Each convolutional block contains several convolution layers, and each layer is followed by a batch normalization layer and a rectified linear unit (ReLU) activation function. The model is usually pre-trained on the ImageNet dataset that has over 1 million images and more than 1000 classes / categories of objects. Since the model is trained in many categories and has many image examples, makes the model very good for any kind of task related to image classification. (Bansal et al., 2023).

The model works with the principle of transfer learning, which is when the model has pre-trained CNN's and then uses those pre-trained layers on the following model layers. This functions by taking advantage of a pre-trained network that extracts useful features from the images and then uses custom layers to classify those features. When analyzing the model, Figure 5, it is worth noting the second layer where the feature extractions are made, which are taken from the pre-trained model in ImageNet, which removes certain patterns and details for use on CheXpert images. Another important layer, the Global Average Pooling Layer, reduces the dimensional space of the features, which reduces the number of parameters without losing information and the last layer is the one that takes out the final output, with the desired number of labels, that in this case were 2. In this layer, it is utilized the sigmoid activation function that maps all values to values between 0 and 1 producing the binary classifier.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_4 (InputLayer)        [(None, 128, 128, 3)]     0

 vgg19 (Functional)          (None, None, None, 512)   20024384

 batch_normalization (BatchN  (None, 4, 4, 512)        2048
 ormalization)

 global_average_pooling2d (G  (None, 512)              0
 lobalAveragePooling2D)

 dense_1 (Dense)             (None, 256)               131328

 dense_2 (Dense)             (None, 2)                 514

=================================================================
Total params: 20,158,274
Trainable params: 20,157,250
Non-trainable params: 1,024
_________________________________________________________________
```

Figure 5 - VGG19 model summary

The second used model, the DenseNet121 was created in 2016 by Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Weinberger, is a CNN model of the DenseNet family, which is characterized by their dense connectivity, meaning that each layer is connected to all other layers that are deeper in the network. Similar to the previous model, the DenseNet121 was also pre-trained on the ImageNet dataset. The model has 121 layers and consists of four dense blocks, which are followed by

transition layers. The difference between DenseNet121 and VGG19 is that all the layers on DenseNet121 are connected to all layers that are deeper in the network and the VGG19 the layers are only connected to the previous one (G. Huang et al., 2017).

The DenseNet121 can be seen in Figure 6 and consists only of an input layer which is where the model accepts the input of images, followed by the model layer, which consists, as the name states, in 121 layers. The first layer is a 7x7 Convolution layer, followed by 58 3x3 Convolution layers, the next is 61 1x1 Convolution layers and ends with 4 AvgPool layers and 1 Fully Connected Layer. In this model layer, the pre-trained ImageNet weights are loaded and finished with the output layer that has also the sigmoid function that maps all values to values between 0 and 1.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_2 (InputLayer)        [(None, 128, 128, 3)]     0

 densenet121 (Functional)    (None, 1024)              7037504

 dense (Dense)               (None, 2)                 2050

=================================================================
Total params: 7,039,554
Trainable params: 6,955,906
Non-trainable params: 83,648
_________________________________________________________________
```

Figure 6 - DenseNet121 model summary

3.2.4. Training and Testing

In this section it will be detailed the training and testing processes of both models detailed earlier, the DenseNet121 and the VGG19. These models were trained on the CheXpert dataset to identify cases of 12 different conditions, Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarge Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia and Pneumothorax, and all the steps that will be described next will be performed on the 12 different datasets of the 12 different conditions. The training of both models was performed with the help of TensorFlow and Keras frameworks and they were trained on Machine B (a machine learning based server), which has enough power to sustain training a large number of images, and it was performed a hyperparameter tunning to ensure the best conditions when preforming the training.

During the training process, and after multiple combinations of parameters, heuristically it reached the following numbers (values of the parameters). The number of steps per epoch, which sets the number of batches that get processed on each epoch was selected as 500. this number allows the model to learn in relatively small chunks, which facilitates an efficient training approach. The number of epochs used was 5, and this makes it so that the whole dataset will be trained 5 times which provides better results in a short number of epochs, it was noted that after the 5^{th} epoch, the "lost and accuracy"

values remained constant and they did not improve. After each model, for each chest-specific condition, had finished training, the model's architecture, weights, and training configuration were saved, in order to be used later to perform the testing part.

The test dataset, which was not used in the training processes, was used only for the evaluation of the models, in this work, it was used the Area Under the Curve (AUC), which seems to be one of the most appropriate (if not the best) metric for the developed work. This metric is popular to evaluate the performance of binary classification models, and it is used together with the plot of the Receiver Operating Characteristic (ROC) curves. These metrics are important to understand the model ability to detect between positive and negative conditions. These two metrics were calculated for the dataset before and after the process of data augmentation, in order to be able to compare the results before and after, and to check if there was any improvement in the AUC and the curves values.

In this work, all the codes and algorithms used were implemented using Python 3.10.0. programing language. To help in the creation of this work, there were used multiple libraries and frameworks that are already built in python. NumPy was used because it has a collection of mathematical functions to work on large datasets. Pandas was used to work with the dataset, and Matplotlib and Seaborn were used to create and show graphics. TensorFlow and Keras were used to create and train Deep Neural Networks.

All the work was divided into two different machines, machine A, with one GPU and machine B, which has two nodes of 8 GPUs each. Machine A was the main working tool, handling all the pre-processing, data augmentation and image generation processes, while machine B, as it is a shared resource provided by VISTA Lab Research Center of the University of Évora, Portugal, handled all the models training process that required the most intensive GPU work.

4. Results and Discussion

This chapter presents and discusses the main results reached in this work when applying the developed methods and models to classify medical conditions in the CheXpert dataset. To provide an easier way to present the achieved results, first, it will be presented the pre-processing results and then the final results will be presented by model, showing first the results for the VGG19 model and hereafter for the Dense121 model. For each model, the conditions will be grouped according to the improvement (from lowest to highest) of the AUC values.

4.1. Pre-processing Results

The pre-processing steps described in the methods section, including background removal, noise removal, diaphragm region removal, and contrast enhancement, were applied to the CheXpert dataset images and below are the results of each step. The background removal process, as can be seen in Figure 7, demonstrates the benefits of this pre-processing step as the image after the background removal shows a cleaner and more focused image to work with.

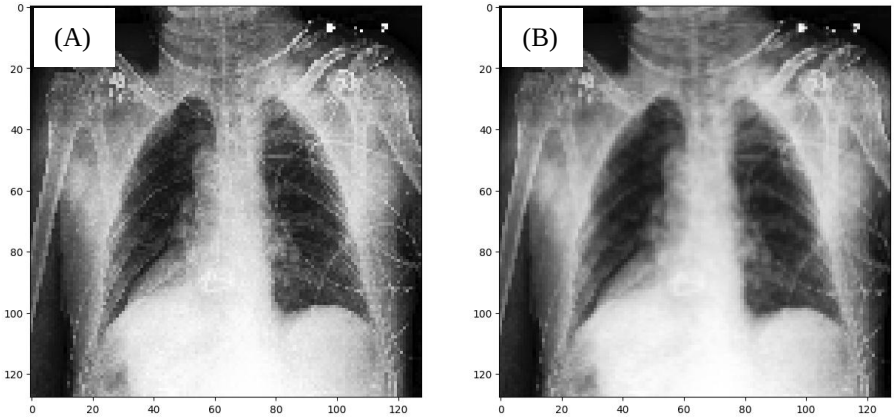

Figure 7 - Image before (A) and after (B) background removal.

Figure 8 demonstrates the effectiveness of the noise removal process as the image before the process shows some noise and artefacts in the chest region and after applying the noise removal technique, the noise on the image was reduced, resulting in a clearer image.

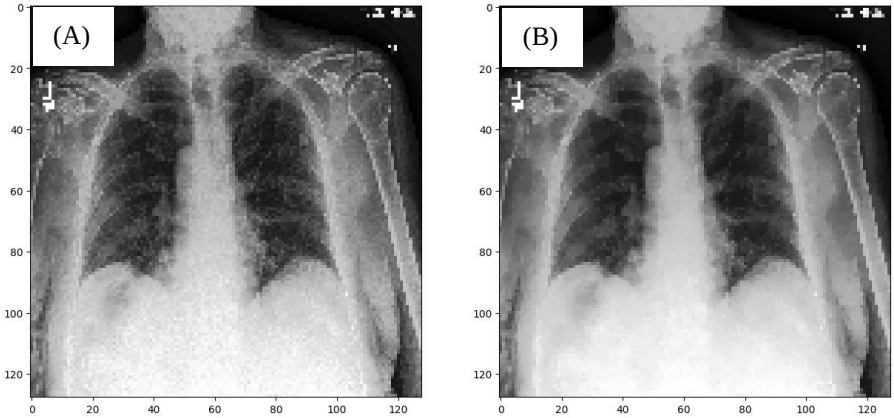

Figure 8 - Image before (A) and after (B) noise removal.

The pre-processing step shown in Figure 9, diaphragm removal, successfully eliminated the diaphragm region, as can be seen in after figure, and removing this region of the image reduces the potential distraction when analyzing the dataset images.

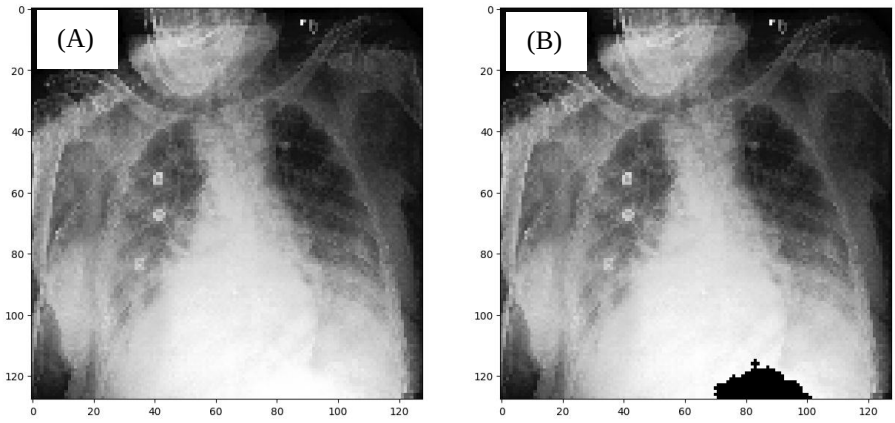

Figure 9 - Image before (A) and after (B) diaphragm removal.

The last pre-processing step was the contrast enhancement, and this step improved the visual distinction between the various regions of the image, highlighting the most important parts of the image, which is essential for an accurate diagnosis in the next stages.

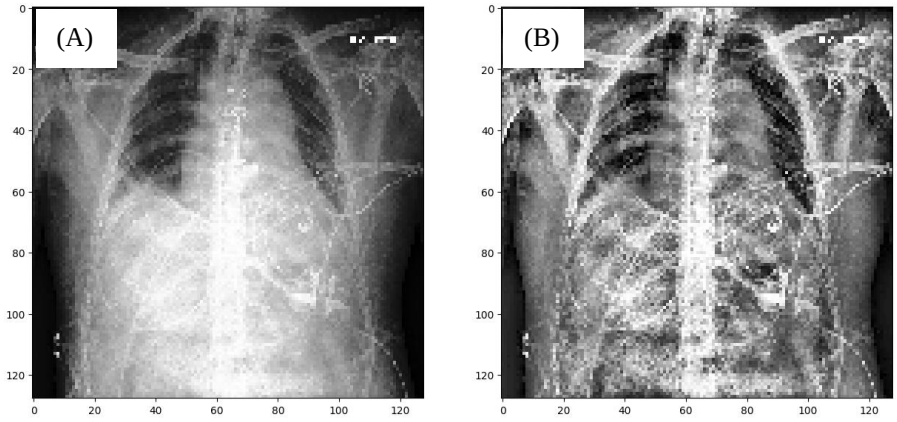

Figure 10 - Image before (A) and after (B) contrast enhancement.

Overall, all of the pre-processing steps were effective to clear the CheXpert dataset images for the upcoming model's analysis.

4.2. VGG19 Model

4.2.1. VGG19: Group 1

Group 1 consists of three conditions: Cardiomegaly, Pleural Effusion, and Edema. These conditions are critical when analyzing chest radiographies due to their implications for cardiovascular and respiratory health problems. This section's goal is to evaluate the use of the VGG19 model together with data augmentation techniques. By expanding the number of images inside the dataset, the goal was to improve the AUC values of Group 1 conditions.

Table 6 presents the AUC values for both baseline and augmented VGG19 models, related to the three conditions of this group and it can be observed that for all conditions there was an improvement in the AUC values after applying data augmentation.

Table 6 - AUC Values for Group 1 conditions using VGG10 model

	VGG19 Baseline		VGG19 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Cardiomegaly	0.84	0.84	0.97	0.97	0.13	0.13
Pleural Effusion	0.82	0.82	0.95	0.95	0.13	0.13
Edema	0.76	0.76	0.95	0.95	0.19	0.19

The AUC values for Cardiomegaly were already one of the highest without the data augmentation, but still, it can be seen that there was a solid improvement of 0.13, increasing from 0.84 to 0.97. For the second condition of this group, Pleural Effusion, the improvement was the same as the one of Cardiomegaly with also an increase of 0.13 from 0.82 to 0.95. The last condition, Edema, showed the biggest increase with 0.18, increasing from 0.77 to 0.95. Figure 11 shows the ROC curves for these three conditions, with the graphics 11a, 11c and 11e corresponding to the non-augmented ones and the graphics 11b, 11d and 11f corresponding to the augmented ones. By analyzing the Figure 11. it can also be seen in the increase in AUC with the curves on the right side having a higher area under the curve.

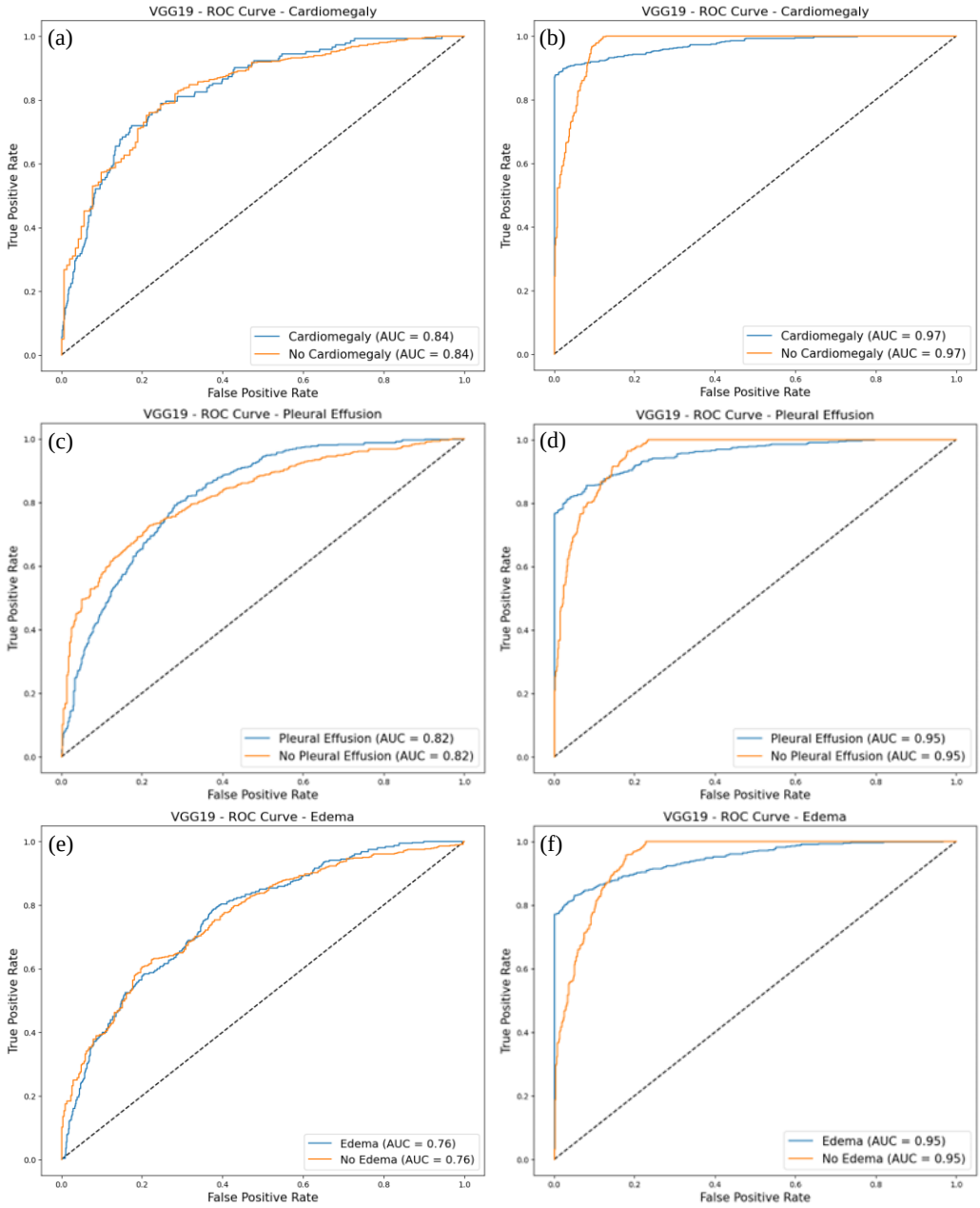

Figure 11 - ROC curves for Group 1 before (a, c, e) and after (b, d, f) data augmentation using VGG10 model

4.2.2. VGG19: Group 2

Group 2 only contains two conditions: Pneumothorax and Fracture. These conditions are two of the most common ones and demand accurate detection and characterization in chest imaging. The goal follows the same process as the ones before, using the VGG19 model with and without data augmentation to evaluate the improvement in the conditions of Group 2 of the CheXpert dataset. Table 7 presents the AUC values for both baseline and augmented VGG19 models, related to the 2 conditions of this group and it can be observed that for all conditions there was also an improvement in the AUC values after applying data augmentation. Compared to Group 1 conditions, Group 2 showed a higher improvement with all the conditions having an increase of over 0.20.

Table 7 - AUC Values for Group 2 conditions using VGG10 model

	VGG19 Baseline		VGG19 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Pneumothorax	0.76	0.76	0.96	0.96	0.20	0.20
Fracture	0.73	0.72	0.98	0.98	0.25	0.26

The AUC values for the Pneumothorax were lower than the values of Group 1 but it showed a greater improvement with an increase of 0.20 from 0.76 to 0.96. The AUC values for Fracture were lowest so far, without the data augmentation, but still, it showed an improvement of 0.25, increasing from 0.73 to 0.98. Figure 12 shows the ROC curves for these two conditions, with the graphics 12a and 12c corresponding to the non-augmented ones and the graphics 12b and 12d corresponding to the augmented ones. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve.

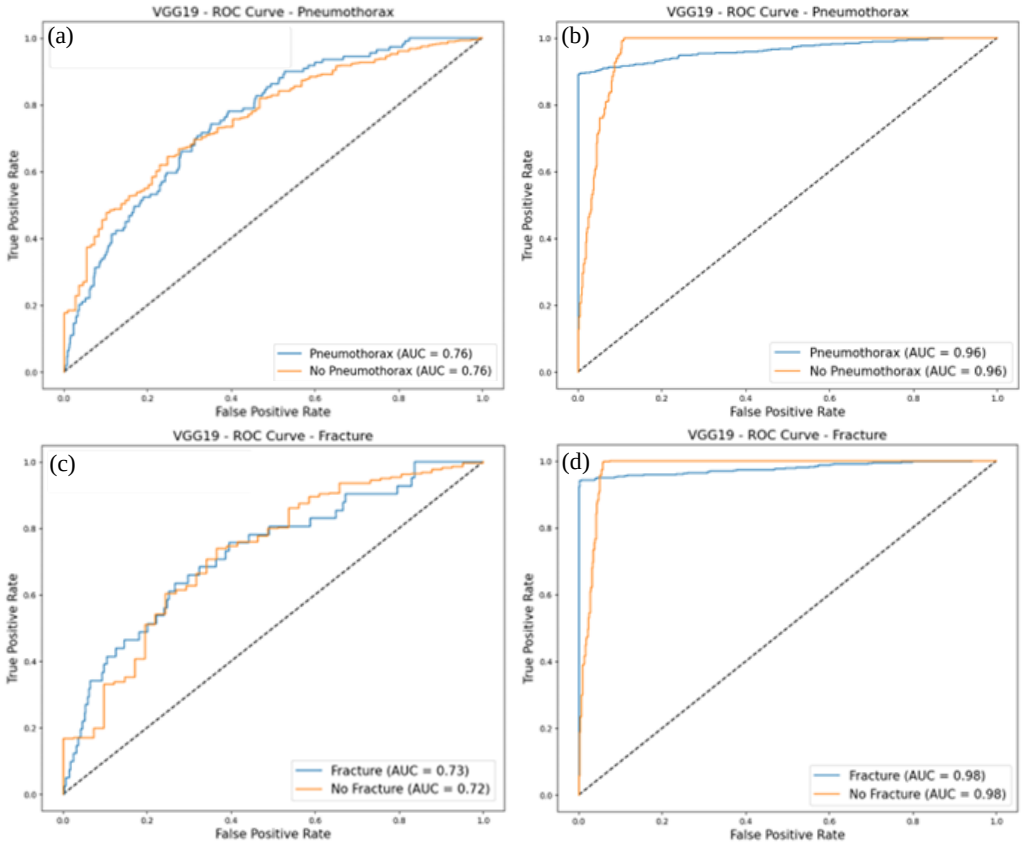

Figure 12 - ROC curves for Group 2 before (a, c) and after (b, d) data augmentation using VGG10 model

4.2.3. VGG19: Group 3

Group 3 contains four conditions: Lung Opacity, Enlarged Cardiomediastinum, Pleural Other, and Consolidation and this group consists of common pulmonary pathologies that present several diagnostic challenges. The goal is the same as before, using the VGG19 model with and without data augmentation to evaluate the improvement in the conditions of Group 3 of the CheXpert dataset.

Table 8 presents the AUC values for both baseline and augmented VGG19 models, related to the 4 conditions of this group and it can be observed that for all conditions there was also an improvement in the AUC values after applying data augmentation. Compared to the previous groups, Group 3 showed the conditions with the highest improvements so far, with all conditions having an increase of over 0.3.

Table 7 - AUC Values for Group 3 conditions using VGG10 model

	VGG19 Baseline		VGG19 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Lung Opacity	0.61	0.61	0.91	0.91	0.30	0.30
Enlarge Cardiomediastinum	0.67	0.67	0.97	0.97	0.30	0.30
Pleural Other	0.68	0.68	0.99	0.99	0.31	0.31
Consolidation	0.63	0.64	0.97	0.97	0.34	0.33

The conditions in Group 3, when compared to the groups before, show the worst baseline values of the 3 groups, but as happened in Group 2, as the value decreases before the data augmentation, the improvement increases, in this case, to values over 0.30. The AUC values for Lung Opacity were lower than 0.7, without the data augmentation, but still, it showed a great improvement, with an increase of 0.30, from 0.61 to 0.91. For the Enlarged Cardiomediastinum condition, the improvement was the same as before with an increase of 0.30 from 0.67 to 0.97.

The last 2 conditions of this group, Pleural Other and Consolidation, showed even lower baseline values than the other 2 conditions but also showed the greatest improvement of all. Pleural Other achieved an improvement of 0.31 from 0.68 to 0.99, while Consolidation had an improvement of 0.34 from 0.63 to 0.97. Figure 13 shows the ROC curves for these four conditions, with the graphics 13a, 13c, 13e and 13g corresponding to the non-augmented ones and the graphics 13b, 13d, 13f and 13h corresponding to the augmented ones. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve. In this group of graphics, it is possible to notice the increase in AUC better because the values before augmentation were lower than those of Group 1 and Group 2.

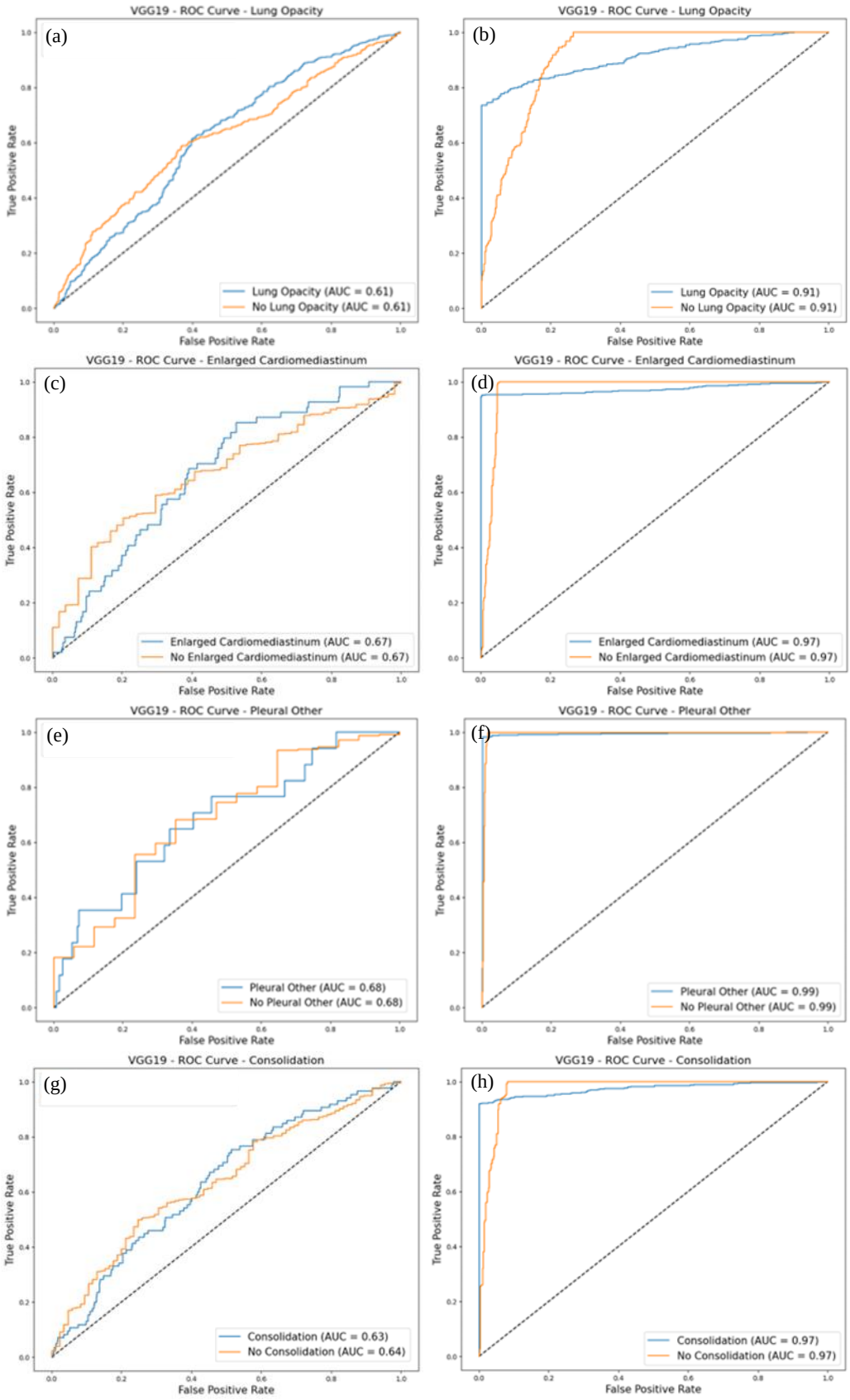

Figure 13 - ROC curves for Group 3 before (a, c, e, g) and after (b, d, f, h) data augmentation using VGG10 model

4.2.4. VGG19: Group 4

Group 4 is the last, and it contains three conditions: Lung Lesion, Atelectasis, and Pneumonia. This section's goal is also to evaluate the use of the VGG19 model with and without data augmentation techniques, to check if the AUC can be improved in Group 4 of the CheXpert dataset conditions. Table 9 presents the AUC values for both baseline and augmented VGG19 models, related to the 3 conditions of this group and it can be observed that for all conditions there was also an improvement in the AUC values after applying data augmentation. Compared to the previous groups, and as it is the last group, Group 4 showed the conditions with the highest improvements of all conditions on the CheXpert dataset with improvements of over 0.35.

Table 9 - AUC Values for Group 4 conditions using VGG10 model

	VGG19 Baseline		VGG19 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Lung Lesion	0.61	0.61	0.98	0.98	0.37	0.37
Atelectasis	0.53	0.55	0.91	0.91	0.38	0.36
Pneumonia	0.51	0.55	0.99	0.99	0.48	0.44

The conditions on Group 4, showed the worst baseline values of all conditions on the CheXpert dataset, but also were the conditions with the highest improvement. The AUC values for Lung Lesions were the highest of this group, without the data augmentation, and showed an improvement of 0.37, increasing from 0.61 to 0.98. For the Atelectasis condition, the improvement was similar to the one before with an increase of 0.38 from 0.53 to 0.91. The last condition of this group, Pneumonia, showed the biggest increase with 0.48, increasing from 0.51 to an almost perfect AUC of 0.99.

Figure 14 shows the ROC curves for these conditions, with the graphics 14a, 14c and 14e corresponding to the non-augmented ones and the graphics 14b, 14d and 14f corresponding to the augmented ones. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve. In this group of graphics, it is possible to notice the increase in AUC better because the values before augmentation were the lowest of all groups.

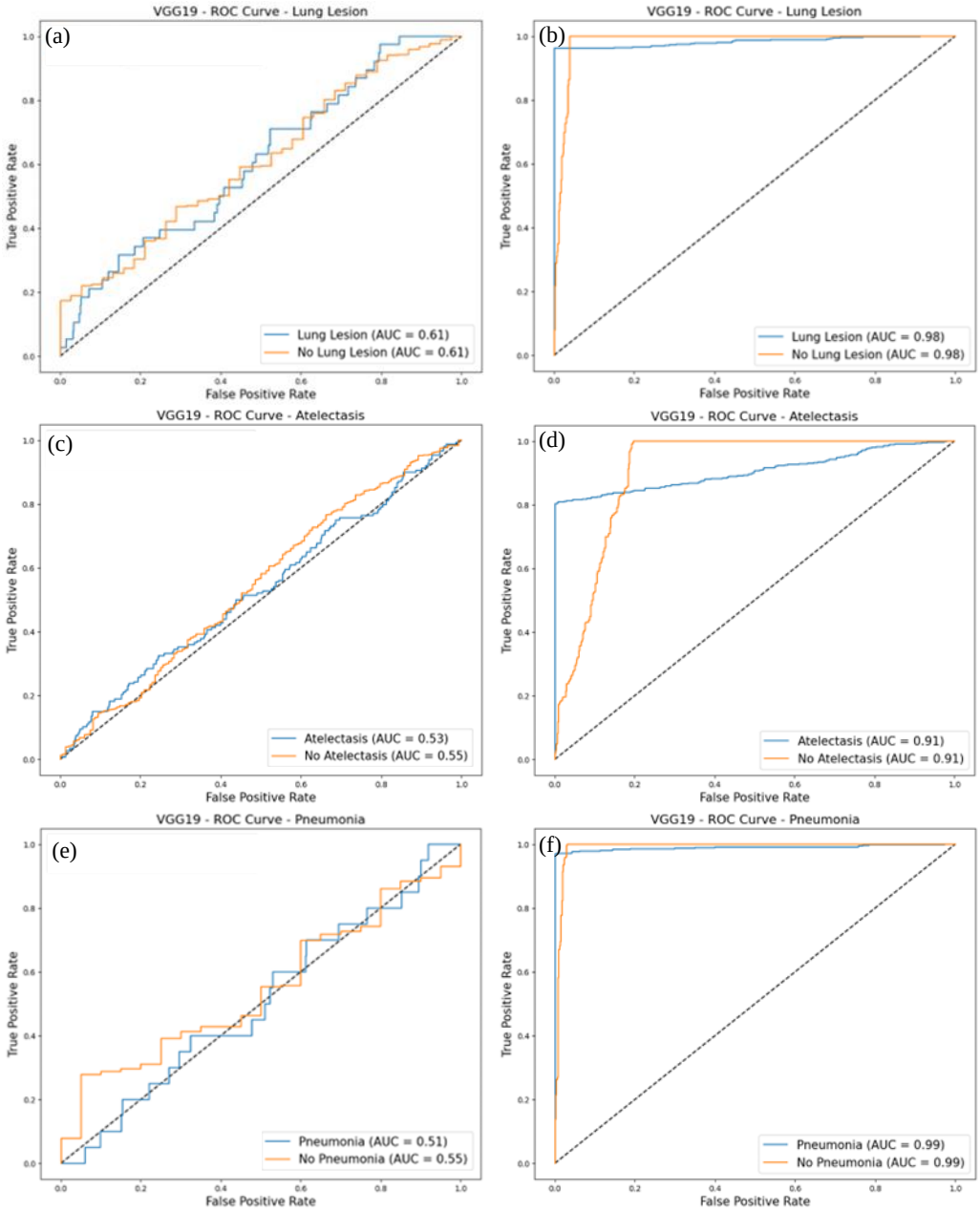

Figure 14 - ROC curves for Group 4 before (a, c, e) and after (b, d, f) data augmentation using VGG10 model

By analyzing all condition values, it can be seen that some conditions showed way better improvements than others, conditions like Cardiomegaly and Pleural Effusion, which already had a high baseline AUC value, 0.84 and 0.82 respectively, only achieved an improvement of 0.13 each. On the contrary, conditions like Atelectasis and Pneumonia which had the lowest baseline AUC value of all, achieved the highest improvement of 0.38 and 0.48. By comparing all values, it is possible to indicate that there is an inverse relationship between baseline value and improvement, as with higher baseline AUC values, there is less room for improvement, leading to a lower improvement, and with lower baseline values there is more room for improvement which leads to a higher improvement.

Comparing all graphics before data augmentation with the graphics after data augmentation it is possible to confirm that the data augmentation did a great job improving the AUC values of the conditions with some ROC curves achieving almost a perfect curve like in the graphic of Pleural Other and Pneumonia that achieved an AUC value of 0.99.

4.3. Dense121 Model

4.3.1. Dense121: Group 1

The groups for Dense121 will be divided similarly, with the same groups as the ones for the VGG19 model but with different conditions for each group. Group 1 consists of 4 conditions: Pleural Effusion, Cardiomegaly, Pneumothorax, and Edema. These conditions are common pulmonary conditions and this section's goal is to evaluate the use of the Dense121 model together with data augmentation techniques, to evaluate if the AUC can be improved in the Group 1 of the CheXpert dataset conditions.

Table 10 presents the AUC values for both baseline and augmented Dense121 models, related to the 4 conditions of this group and it can be observed that for all conditions there was an improvement in the AUC values after applying data augmentation.

Table 8 - AUC Values for Group 1 conditions using Dense121 model

	Dense121 Baseline		Dense121 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Pleural Effusion	0.84	0.84	0.95	0.95	0.11	0.11
Cardiomegaly	0.85	0.85	0.98	0.98	0.13	0.13
Pneumothorax	0.82	0.82	0.97	0.98	0.15	0.16
Edema	0.80	0.80	0.96	0.96	0.16	0.16

The AUC values for this whole group were already the highest baseline values, but still, as there was room for improvement, the data augmentation processes were applied, and the results managed to reach a better value than they had in the baseline. The values for Pleural Effusion were the second highest without the data augmentation, and showed an improvement of 0.11, increasing from 0.84 to 0.95. For the Cardiomegaly condition, which has the highest baseline value, the improvement is similar to the one before with an increase of 0.13 from 0.85 to 0.98. Pneumothorax values had an increase of 0.15, increasing from 0.82 to 0.97. The last condition of this group was Edema, which had the highest increase compared to the ones before, with 0.16 increasing from 0.80 to 0.96.

Figure 15 shows the ROC curves for these four conditions, with the graphics 15a, 15c, 15e and 15g corresponding to the non-augmented ones and the graphics 15b, 15d, 15f, and 15h corresponding to the augmented ones. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve.

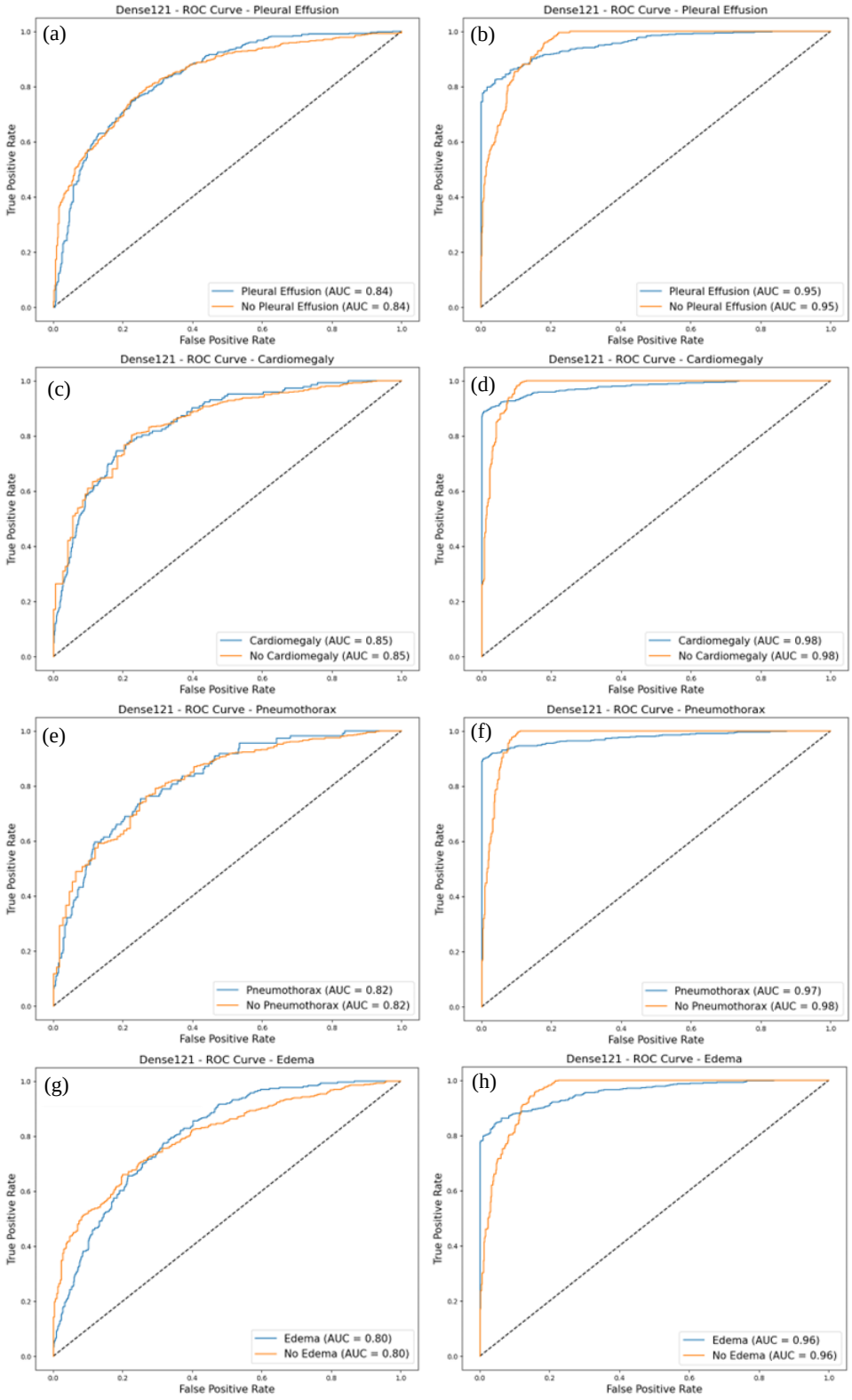

Figure 15 - ROC curves for Group 1 before (a, c, e, g) and after (b, d, f, h) data augmentation using Dense121 model

4.3.2. Dense121: Group 2

Group 2 consists of three conditions: Fracture, Pleural Other and Lung Lesion. This section's goal is also to evaluate the use of the Dense121 model with and without data augmentation techniques, to evaluate if the AUC can be improved in Group 2 of the CheXpert dataset conditions.

Table 11 presents the AUC values for both baseline and augmented Dense121 models, related to the 3 conditions of this group and it can be observed that for all conditions there is a higher improvement, when compared to the conditions in Group 1 after applying data augmentation. Compared to Group 1 conditions, Group 2 showed a higher improvement with all the conditions having an increase higher than 0.20.

Table 9 - AUC Values for Group 2 conditions using Dense121 model

	Dense121 Baseline		Dense121 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Fracture	0.76	0.77	0.98	0.98	0.22	0.21
Pleural Other	0.78	0.78	1.00	0.99	0.22	0.21
Lung Lesion	0.75	0.76	0.99	0.99	0.24	0.23

The AUC values for Group 2 follow the same pattern as the values for the VGG19 model, as the baseline value decreases, the improvement values increase. The lowest improvement of this group was for both Fracture and Pleural Other, with Fracture, having an increase of 0.22 from 0.76 to 0.98 and Pleural Other, achieving an increase of 0.22 from 0.78 to 1.00 The last condition of this group, Lung Lesion, showed the biggest increase of this group with 0.24, increasing from 0.75 to 0.99. Figure 16 shows the ROC curves for these 3 conditions, with the graphics 16a, 16c and 16e corresponding to the ones without the data augmentation and the graphics 16b, 16d and 16f corresponding to the graphics with the data augmentation applied. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve. Compared to the graphs for Group 1, the graphs after the data augmentation are similar, but as the graphs before the data augmentation are worse for Group 2, it is possible to see the improvement better.

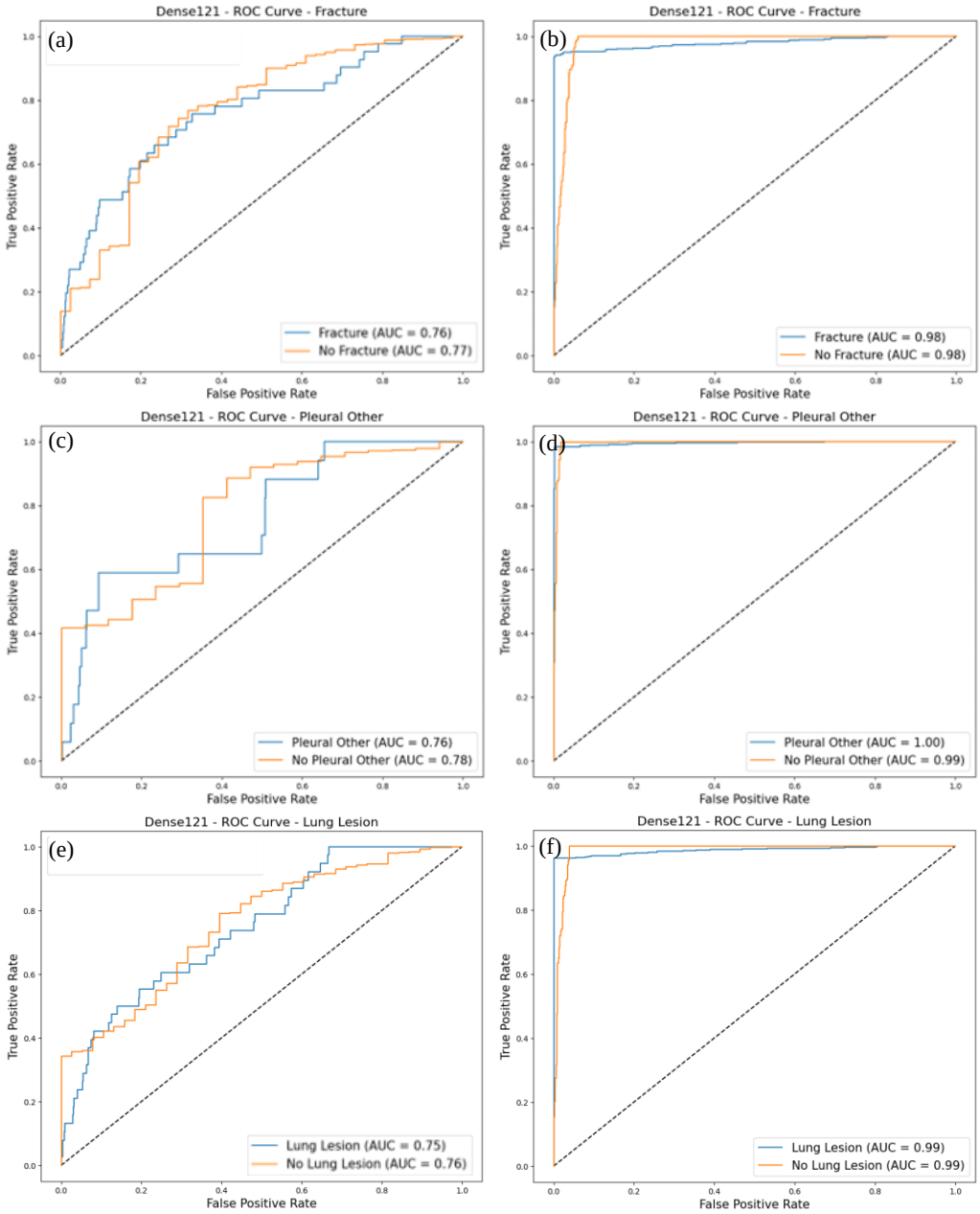

Figure 16 - ROC curves for Group 2 before (a, c, e) and after (b, d, f) data augmentation using Dense121 model

4.3.3. Dense121: Group 3

Group 3 only contains 2 conditions: Lung Opacity and Consolidation. The goal follows the same process as the ones before, using the Dense121 model with and without data augmentation to evaluate the improvement in the conditions of Group 4 of the CheXpert dataset.

Table 13 presents the AUC values for both baseline and augmented Dense121 models, related to the 2 conditions of this group and it can be observed that for all conditions there was a big improvement in the AUC values after applying data augmentation. Compared to the previous groups, and as it is the

last group, Group 4 showed the conditions with the highest improvements of all conditions on the CheXpert dataset with improvements almost reaching 0.3.

Table 10 - AUC Values for Group 3 conditions using Dense121 model

	Dense121 Baseline		Dense121 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Lung Opacity	0.65	0.65	0.92	0.91	0.27	0.26
Consolidation	0.68	0.67	0.97	0.97	0.29	0.30

The conditions in this group all had a lower baseline value compared to the conditions in the groups before, but showed the highest improvement of all, as the highest improvement before group 3 was 0.24 and the lowest improvement in this group was 0.27. The Lung Opacity condition was the condition with the highest improvement in this group with a 0.27 increase in value, from 0.65 to 0.92. Consolidation had a higher improvement than Fracture, with an increase of 0.29, from 0.68 to 0.97. Figure 17 shows the ROC curves for these 2 conditions, with graphics 17a and 17c corresponding to the ones without the augmentation and graphics 17b and 17d corresponding to the ones with data augmentation. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve. In this group of graphics, it is possible to notice the increase in AUC better because the values before augmentation were the lowest of all groups.

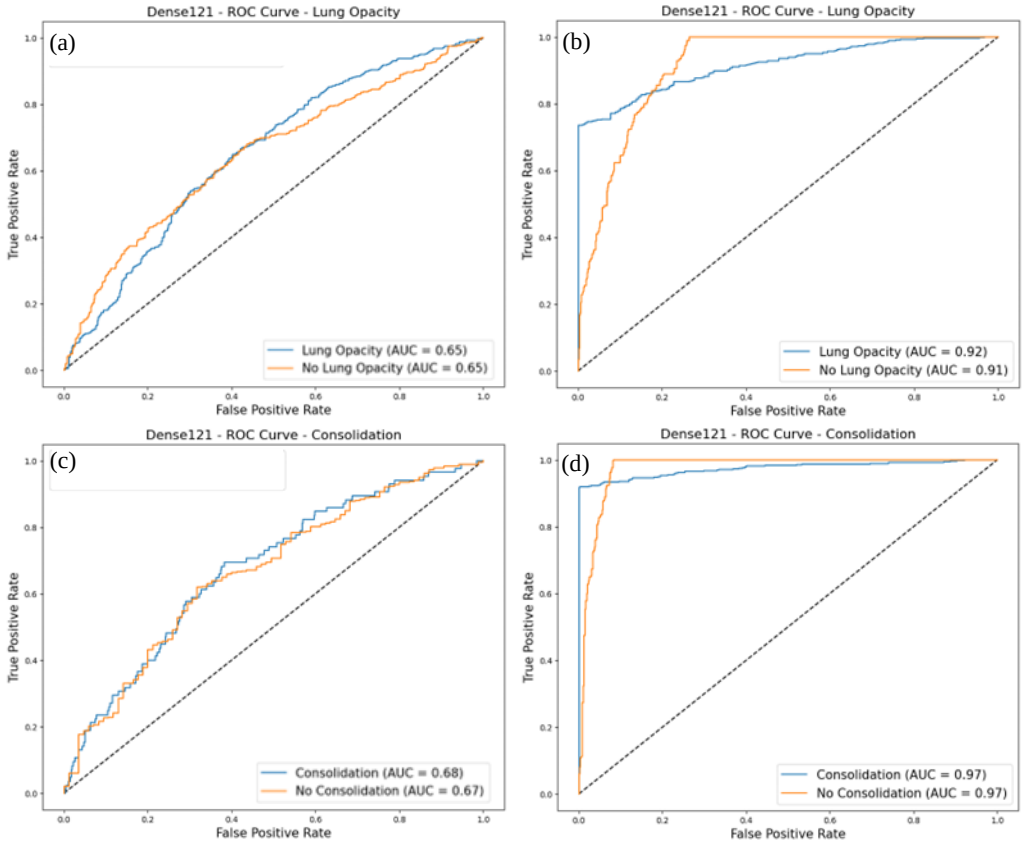

Figure 17 - ROC curves for Group 3 before (a, c) and after (b, d) data augmentation using Dense121 model

4.3.4. Dense121: Group 4

Group 4 contains three conditions: Atelectasis, Pneumonia and Enlarge Cardiomediastinum. These conditions have a highly complex diagnosis, and the goal of this section is to also evaluate the use of the Dense121 model with and without data augmentation to study the improvement of the conditions of Group 3 of the CheXpert dataset.

Table 12 presents the AUC values for both baseline and augmented Dense121 models, related to the 3 conditions in this group and it can be observed that for all conditions there was a higher improvement in the AUC values after applying data augmentation. Compared to the previous groups, Group 4 showed the conditions with the highest improvements, with all having an increase of over 0.30.

Table 11 - AUC Values for Group 4 conditions using Dense121 model

	Dense121 Baseline		Dense121 Augmented		Improvement	
	Condition	No Condition	Condition	No Condition	Condition	No Condition
Atelectasis	0.63	0.63	0.93	0.94	0.30	0.31
Pneumonia	0.69	0.69	0.99	0.99	0.30	0.30
Enlarge Cardiomediastinum	0.67	0.66	0.98	0.98	0.31	0.32

The conditions in Group 4, despite not having the lowest baseline values of all conditions using the Dense121 model, have the highest improvement. The AUC values for Atelectasis were lowest, without the data augmentation, but still, it showed the second-highest improvement so far, with an increase of 0.30, from 0.63 to 0.93. For the Pneumonia condition, the improvement was the same as the one before with an increase of only 0.30 from 0.69 to 0.99. The last condition of this group, Enlarge Cardiomediastinum, has the highest of all conditions using the Dense121 model, achieving an increase of 0.31 from 0.67 to 0.98.

Figure 18 shows the ROC curves for these three conditions, with the graphics 18a, 18c and 18e corresponding to the ones without the augmentation and the graphics 18b, 18d and 18f corresponding to the ones with the data augmentation. By analyzing the figure, it can also be seen the increase in AUC with the curves on the right side having a higher area under the curve. Analyzing each one, is possible to see that Pneumonia achieves an almost perfect AUC value.

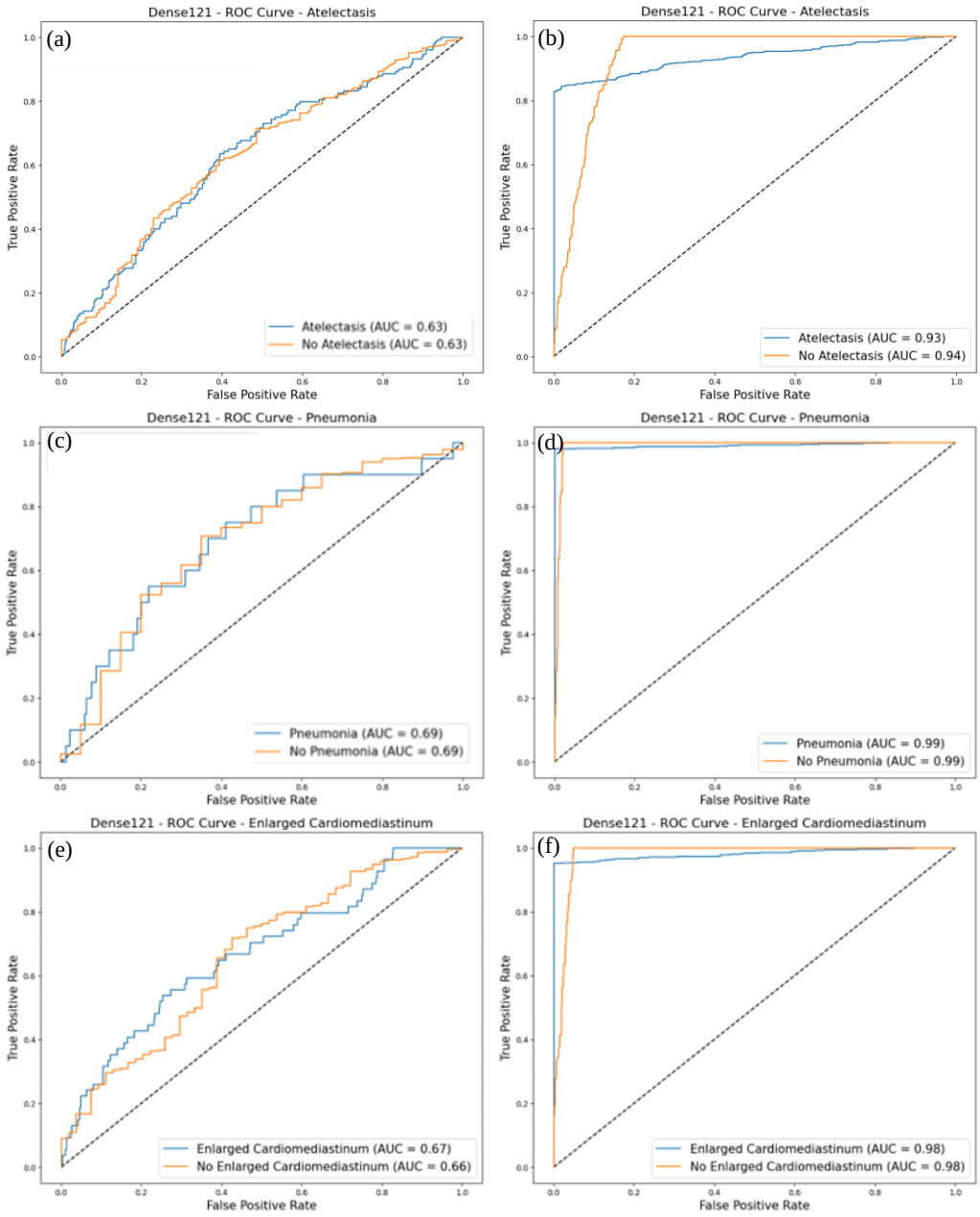

Figure 18 - ROC curves for Group 4 before (Left) and after (Right) data augmentation using Dense121 model

The values for all conditions using the Dense121 model, show similar behavior as the ones using the VGG19 model, with some conditions having a much higher improvement than others, with conditions like Pleural Effusion and Cardiomegaly having high baseline values, but lower improvement. Conditions like Consolidation and Fracture had a lower baseline value but showed the highest improvement. By comparing all values, it is possible to indicate that there is an inverse relationship between baseline value and improvement, as with higher baseline AUC values, there is less room for improvement, leading to a lower improvement, and with lower baseline values there is more room for

improvement which leads to a higher improvement. This happens with some exceptions, as the Enlarge Cardiomediastinum condition had the highest improvement but didn't have the lowest baseline value.

Comparing all graphics before data augmentation with the graphics after data augmentation it is possible to confirm that the data augmentation did a great job improving the AUC values of the conditions with some ROC curves achieving almost a perfect curve like in the graphic of Pleural Other and Pneumonia that achieves an AUC value of 1.00 and 0.99 respectively.

Comparing both models and as can be seen in Table 14, the model Dense121 started with a higher average baseline value of 0.74 when compared to the VGG19 model of only 0.68. which indicates that using only baseline values, the model Dense121 is better suited for the job. Analyzing the augmented values, both models achieved a similar augmented value of 0.96 for the VGG19 and 0.97 for the Dense121, this demonstrates that both models are capable of achieving a very high average performance for chest X-Ray images analysis. Comparing the improvement, it is clear that the VGG19 model had a higher improvement, achieving a 0.28 increase compared to the 0.23 increase of the Dense121 model. These values indicate that the VGG19 model was more responsive to the data augmentation techniques, which resulted on higher values after augmentation. Overall, both models have an inverse relationship between baseline and augmented values, having bigger improvements for conditions with lower initial AUC values.

In summary, when working with lower baseline values, VGG19 model proves to be better. This is because the model showed higher improvement when paired with data augmentation techniques. On the other hand, DenseNet121 showed a stronger initial performance, making it a good choice before any data augmentation is applied. Overall, data augmentation proved to be an essential technique to improve the performance of both models as it improved their classification capabilities, making the models reach an almost perfect average AUC.

Table 12 – Average Baseline, Augmented and Improvement values for VGG19 and Dense121 models

	VGG19			Dense121		
	Baseline	Augmented	Improvement	Baseline	Augmented	Improvement
Average Value	0.68	0.96	0.28	0.74	0.97	0.23

4.4. Comparison with Previous Work

The CheXpert dataset has been used to support several studies and this subsection goal is to present a comparative analysis of this work's findings with other recently reported studies mentioned earlier in this work.

The studies used to compare with the one developed in this work are, the one by (Rahimiaghdam & Alemdar, 2024) that uses the dataset to train CNN architectures, the study by (Zhang et al., 2022) that

developed a two-stage deep learning model combining region proposal, feature pyramid, and multi-instance learning, the study by (Pham et al., 2020) that used different CNNs, along with a label smoothing technique and finally the study by (Allaouzi & Ben Ahmed, 2019) that compared models created by the authors to the DenseNet-121 model in terms of AUC.

The results of this work, due to data augmentation, reached very interesting values and, compared to the study developed by (Rahimiaghdam & Alemdar, 2024), the same DenseNet121 model was used, and this model achieved a performance of 0.97 is higher than the one reached on their study of 0.88, which demonstrated the effectiveness of the used data augmentation techniques. Despite not being used in the indicated study, the VGG19 model used in this work, which reached a 0.96 AUC value, outperformed all models used in their study. When comparing this work to the one developed by (Zhang et al., 2022), it is possible to see that they reached a mean AUC value of 0.912, which is a very high value, but the augmented models used in this study achieved an even better value of 0.97 and 0.96. The work by (Pham et al., 2020)achieved a mean AUC of 0.94 using multiple CNNs together with label smoothing techniques, but the VGG19 and DenseNet121 models paired with data augmentation techniques also outperformed the 0.94 AUC value by reaching 0.96 and 0.97 respectively. The last study, the one performed by (Allaouzi & Ben Ahmed, 2019) used different models, and they reported AUC values of 0.812, 0.808, and 0.785 for their models, which are significantly lower than the 0.96 and 0.97 mean AUC that this work reached.

Overall, and as can be seen in Table 15, it is possible to see that the deep learning models used in this work, paired together with data augmentation techniques, achieved better results than the ones mentioned above. This indicated that the data augmentation techniques used in this work were crucial to achieve an almost perfect performance and that the approach used in this work has the potential to improve images classification for future research in chest X-ray interpretation.

Table 13 - Comparison of AUC Values Across Different Studies and Models

Study	Model	AUC Value Achieved
Our Work	DenseNet121	0.97
	VGG19	0.96
Rahimiaghdam & Alemdar (2024)	DenseNet121	0.88
Zhang et al. (2022)	Various	0.912 (mean)
Pham et al. (2020)	Multiple CNNs	0.94
Allaouzi & Ben Ahmed (2019)	Various	0.812
		0.808
		0.785

5. Conclusion and Future Work

The development of this thesis demonstrates, in accordance with the current state of the art, that machine (deep) learning models trained with benchmarking and curated datasets are able to automatically detect and classify pathological lesions (diseases) in chest X-ray images, e.g., the VGG19 and Dense121 models. In this sense, we developed a new framework that appropriately combines computer vision, machine (deep) learning and data augmentation techniques. First, it was improved image quality / features with simple / traditional digital imaging filter (i.e., noise reduction, contrast enhancing, etc.) and afterwards we implemented a data augmentation process (images rotations) to increment the dataset (CheXpert) with new synthetic images to feed and fine-tuning deep learning previously developed models, which were experimentally tested, demonstrating in all conditions a high performance (accuracy) when compared with previously developed models using the same baseline dataset (CheXpert).

One of the limitations that this study has is the generalizability, as this work was only trained and evaluated on the CheXpert dataset and, despite its size and diversity, doesn't include all possible pathologies and problems that can be found in clinical practice, therefore the models and data augmentation techniques may not work on different datasets. Another limitation is the fact that the models used are complex models that require a lot of time and resources to train and test and in institutions with limited computational capabilities the models the computer will have a hard time running these models. The final limitation is the fact that the models did not have clinical validation, and despite having high-performance metrics, they were not validated in a clinical environment.

It is required future work, on some aspects mentioned in the limitations, like the clinical validation part as this work with a clinical validation can be used safely in a clinical environment, the study can also be adapted to other datasets to evaluate if the data augmentation techniques that were used also work on different datasets, and if not, different ones can be added to try to adapt to different conditions.

References

Ahmad, H. K., Milne, M. R., Buchlak, Q. D., Ektas, N., Sanderson, G., Chamtie, H., Karunasena, S., Chiang, J., Holt, X., Tang, C. H. M., Seah, J. C. Y., Bottrell, G., Esmaili, N., Brotchie, P., & Jones, C. (2023). Machine Learning Augmented Interpretation of Chest X-rays: A Systematic Review. *Diagnostics*, *13*(4), 743. https://doi.org/10.3390/diagnostics13040743

Ait Nasser, A., & Akhloufi, M. A. (2023). A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography. *Diagnostics*, *13*(1), 159. https://doi.org/10.3390/diagnostics13010159

Allaouzi, I., & Ben Ahmed, M. (2019). A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases. *IEEE Access*, *7*, 64279–64288. https://doi.org/10.1109/ACCESS.2019.2916849

Ayalew, A. M., Bezabh, Y. A., Abuhayi, B. M., & Ayalew, A. Y. (2024). Atelectasis detection in chest X-ray images using convolutional neural networks and transfer learning with anisotropic diffusion filter. *Informatics in Medicine Unlocked*, *45*, 101448. https://doi.org/10.1016/j.imu.2024.101448

Bansal, M., Kumar, M., Sachdeva, M., & Mittal, A. (2023). Transfer learning for image classification using VGG19: Caltech-101 image data set. *Journal of Ambient Intelligence and Humanized Computing*, *14*(4), 3609–3620. https://doi.org/10.1007/s12652-021-03488-z

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica*, *83*, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

Brady, A. P., & Neri, E. (2020). Artificial Intelligence in Radiology—Ethical Considerations. *Diagnostics*, *10*(4), 231. https://doi.org/10.3390/diagnostics10040231

Broder, J. (2011). Imaging the Chest. In *Diagnostic Imaging for the Emergency Physician* (pp. 185–296). Elsevier. https://doi.org/10.1016/B978-1-4160-6113-7.10005-5

Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2019). *PadChest: A large chest x-ray image dataset with multi-label annotated reports*. https://doi.org/10.1016/j.media.2020.101797

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., & Zhou, C. (2020). *Deep Learning in Medical Image Analysis* (pp. 3–21). https://doi.org/10.1007/978-3-030-33128-3_1

Cillóniz, C., Torres, A., & Niederman, M. S. (2021). Management of pneumonia in critically ill patients. *BMJ*, e065871. https://doi.org/10.1136/bmj-2021-065871

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., & Bertrand, H. (2022). TorchXRayVision: A library of chest X-ray datasets and models. *Medical Imaging with Deep Learning*. https://openreview.net/forum?id=_5iri84DJmE

Colangelo, M. (2022, June 1). *AI In Healthcare Milestones Q3 2022*. https://www.linkedin.com/pulse/ai-healthcare-highlights-milestones-q3-2022-margaretta-colangelo

Dhar, T., Dey, N., Borra, S., & Sherratt, R. S. (2023). Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society*, *4*(1), 68–75. https://doi.org/10.1109/TTS.2023.3234203

Dreyer, R. G., Van der Merwe, C. M., Nicolaou, M. A., & Richards, G. A. (2023). Assessing and comparing chest radiograph interpretation in the Department of Internal Medicine at the University of the Witwatersrand medical school, according to seniority. *African Journal of Thoracic and Critical Care Medicine*, 12–17. https://doi.org/10.7196/AJTCCM.2023.v29i1.265

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *RadioGraphics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Farhud, D. D., & Zokaei, S. (2021). Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iranian Journal of Public Health*. https://doi.org/10.18502/ijph.v50i11.7600

Ferreiro, L., Toubes, M. E., San José, M. E., Suárez-Antelo, J., Golpe, A., & Valdés, L. (2020). Advances in pleural effusion diagnostics. *Expert Review of Respiratory Medicine*, *14*(1), 51–66. https://doi.org/10.1080/17476348.2020.1684266

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Borondy Kitts, A., Birch, J., Shields, W. F., van den Hoven van Genderen, R., Kotter, E., Wawira Gichoya, J., Cook, T. S., Morgan, M. B., Tang, A., Safdar, N. M., & Kohli, M. (2019). Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *Radiology*, *293*(2), 436–440. https://doi.org/10.1148/radiol.2019191586

Getty, N., Brettin, T., Jin, D., Stevens, R., & Xia, F. (2021). Deep medical image analysis with representation learning and neuromorphic computing. *Interface Focus*, *11*(1). https://doi.org/10.1098/rsfs.2019.0122

Giełczyk, A., Marciniak, A., Tarczewska, M., & Lutowski, Z. (2022). Pre-processing methods in chest X-ray image classification. *PLOS ONE*, *17*(4), e0265949. https://doi.org/10.1371/journal.pone.0265949

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Huang, S.-C., Shen, L., Lungren, M. P., & Yeung, S. (2021). GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3922–3931. https://doi.org/10.1109/ICCV48922.2021.00391

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*.

Jin, M. X., Gilotra, K., Young, A., & Gould, E. (2022). Call to Action: Creating Resources for Radiology Technologists to Capture Higher Quality Portable Chest X-rays. *Cureus*. https://doi.org/10.7759/cureus.29197

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019a). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317. https://doi.org/10.1038/s41597-019-0322-0

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019b). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317. https://doi.org/10.1038/s41597-019-0322-0

Jones, C. M., Buchlak, Q. D., Oakden-Rayner, L., Milne, M., Seah, J., Esmaili, N., & Hachey, B. (2021). Chest radiographs and machine learning – Past, present and future. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 538–544. https://doi.org/10.1111/1754-9485.13274

Juluru, K., Shih, H.-H., Keshava Murthy, K. N., Elnajjar, P., El-Rowmeim, A., Roth, C., Genereaux, B., Fox, J., Siegel, E., & Rubin, D. L. (2021). Integrating Al Algorithms into the Clinical Workflow. *Radiology: Artificial Intelligence*, *3*(6). https://doi.org/10.1148/ryai.2021210013

Kim, D., Chung, J., Choi, J., Succi, M. D., Conklin, J., Longo, M. G. F., Ackman, J. B., Little, B. P., Petranovic, M., Kalra, M. K., Lev, M. H., & Do, S. (2022). Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. *Nature Communications*, *13*(1), 1867. https://doi.org/10.1038/s41467-022-29437-8

Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H., & Kim, N. (2019). Deep Learning in Medical Imaging. *Neurospine*, *16*(4), 657–668. https://doi.org/10.14245/ns.1938396.198

Kotter, E., & Ranschaert, E. (2021). Challenges and solutions for introducing artificial intelligence (AI) in daily clinical workflow. *European Radiology*, *31*(1), 5–7. https://doi.org/10.1007/s00330-020-07148-2

Kufel, J., Bielówka, M., Rojek, M., Mitręga, A., Lewandowski, P., Cebula, M., Krawczyk, D., Bielówka, M., Kondoł, D., Bargieł-Łączek, K., Paszkiewicz, I., Czogalik, Ł., Kaczyńska, D., Wocław, A.,

Gruszczyńska, K., & Nawrat, Z. (2023). Multi-Label Classification of Chest X-ray Abnormalities Using Transfer Learning Techniques. *Journal of Personalized Medicine*, *13*(10), 1426. https://doi.org/10.3390/jpm13101426

Kumari, S., & Singh, P. (2023). *Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives*.

Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., Eswaran, K., Cameron Chen, P.-H., Liu, Y., Kalidindi, S. R., Ding, A., Corrado, G. S., Tse, D., & Shetty, S. (2020a). Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology*, *294*(2), 421–431. https://doi.org/10.1148/radiol.2019191293

Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., Eswaran, K., Cameron Chen, P.-H., Liu, Y., Kalidindi, S. R., Ding, A., Corrado, G. S., Tse, D., & Shetty, S. (2020b). Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology*, *294*(2), 421–431. https://doi.org/10.1148/radiol.2019191293

Mazzone, P. J., & Lam, L. (2022). Evaluating the Patient With a Pulmonary Nodule. *JAMA*, *327*(3), 264. https://doi.org/10.1001/jama.2021.24287

Meedeniya, D., Kumarasinghe, H., Kolonne, S., Fernando, C., Díez, I. D. la T., & Marques, G. (2022a). Chest X-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, *126*, 109319. https://doi.org/10.1016/j.asoc.2022.109319

Meedeniya, D., Kumarasinghe, H., Kolonne, S., Fernando, C., Díez, I. D. la T., & Marques, G. (2022b). Chest X-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, *126*, 109319. https://doi.org/10.1016/j.asoc.2022.109319

Milecki, L., Kalogeiton, V., Bodard, S., Anglicheau, D., Correas, J.-M., Timsit, M.-O., & Vakalopoulou, M. (2023). *MEDIMP: 3D Medical Images with clinical Prompts from limited tabular data for renal transplantation*.

Moon, S. M., Choi, H., Kim, S. H., Kang, H. K., Park, D. W., Jung, J. H., Han, K., Shin, D. W., & Lee, H. (2023). Increased Lung Cancer Risk and Associated Risk Factors in Tuberculosis Survivors: A Korean Population-Based Study. *Clinical Infectious Diseases*, *77*(9), 1329–1339. https://doi.org/10.1093/cid/ciad373

Mould, R. F. (1995). The early history of X-ray diagnosis with emphasis on the contributions of physics 1895-1915. *Physics in Medicine and Biology*, *40*(11), 1741–1787. https://doi.org/10.1088/0031-9155/40/11/001

Mustafa, Z., & Nsour, H. (2023). Using Computer Vision Techniques to Automatically Detect Abnormalities in Chest X-rays. *Diagnostics*, *13*(18), 2979. https://doi.org/10.3390/diagnostics13182979

Naik, N., Hameed, B. M. Z., Shetty, D. K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B. P., Chlosta, P., & Somani, B. K. (2022). Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Frontiers in Surgery*, *9*. https://doi.org/10.3389/fsurg.2022.862322

Najjar, R. (2023). Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics*, *13*(17), 2760. https://doi.org/10.3390/diagnostics13172760

Nooreldeen, R., & Bach, H. (2021). Current and Future Development in Lung Cancer Diagnosis. *International Journal of Molecular Sciences*, *22*(16), 8661. https://doi.org/10.3390/ijms22168661

Oren, O., Gersh, B. J., & Bhatt, D. L. (2020). Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*, *2*(9), e486–e488. https://doi.org/10.1016/S2589-7500(20)30160-6

Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., & Nguyen, H. Q. (2020). *Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels*.

Quý, N. (2023, May). *Chest-Xray14-code*. https://www.kaggle.com/code/nguynquy/chest-xray14-code/notebook

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

Rubin, G. D., Ryerson, C. J., Haramati, L. B., Sverzellati, N., Kanne, J. P., Raoof, S., Schluger, N. W., Volpi, A., Yim, J.-J., Martin, I. B. K., Anderson, D. J., Kong, C., Altes, T., Bush, A., Desai, S. R., Goldin, onathan, Goo, J. M., Humbert, M., Inoue, Y., … Leung, A. N. (2020). The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. *Radiology*, *296*(1), 172–180. https://doi.org/10.1148/radiol.2020201365

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Subramaniam, K., Palanisamy, N., Sinnaswamy, R. A., Muthusamy, S., Mishra, O. P., Loganathan, A. K., Ramamoorthi, P., Gnanakkan, C. A. R. C., Thangavel, G., & Sundararajan, S. C. M. (2023). A comprehensive review of analyzing the chest X-ray images to detect COVID-19 infections using deep learning techniques. *Soft Computing*, *27*(19), 14219–14240. https://doi.org/10.1007/s00500-023-08561-7

Sun, J., Shi, W., Giuste, F. O., Vaghani, Y. S., Tang, L., & Wang, M. D. (2023). Improving explainable AI with patch perturbation-based evaluation pipeline: a COVID-19 X-ray image analysis case study. *Scientific Reports*, *13*(1), 19488. https://doi.org/10.1038/s41598-023-46493-2

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., & Rajpurkar, P. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, *6*(12), 1399–1406. https://doi.org/10.1038/s41551-022-00936-9

Tiwari, P., Viswanath, S., Lee, G., & Madabhushi, A. (2011). Multi-modal data fusion schemes for integrated classification of imaging and non-imaging biomedical data. *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 165–168. https://doi.org/10.1109/ISBI.2011.5872379

Wiener, R. S., Gould, M. K., Arenberg, D. A., Au, D. H., Fennig, K., Lamb, C. R., Mazzone, P. J., Midthun, D. E., Napoli, M., Ost, D. E., Powell, C. A., Rivera, M. P., Slatore, C. G., Tanner, N. T., Vachani, A., Wisnivesky, J. P., & Yoon, S. H. (2015). An Official American Thoracic Society/American College of Chest Physicians Policy Statement: Implementation of Low-Dose Computed Tomography Lung Cancer Screening Programs in Clinical Practice. *American Journal of Respiratory and Critical Care Medicine*, *192*(7), 881–891. https://doi.org/10.1164/rccm.201508-1671ST

World Health Organization. (2021). *Global Tuberculosis Report 2021*.

Wu, J. T., Wong, K. C. L., Gur, Y., Ansari, N., Karargyris, A., Sharma, A., Morris, M., Saboury, B., Ahmad, H., Boyko, O., Syed, A., Jadhav, A., Wang, H., Pillai, A., Kashyap, S., Moradi, M., & Syeda-Mahmood, T. (2020). Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, *3*(10), e2022779. https://doi.org/10.1001/jamanetworkopen.2020.22779

Yamasaki, Y., Kamitani, T., Sagiyama, K., Hino, T., Kisanuki, M., Tabata, K., Isoda, T., Kitamura, Y., Abe, K., Hosokawa, K., Toyomura, D., Moriyama, S., Kawakubo, M., Yabuuchi, H., & Ishigami, K. (2024). Dynamic chest radiography for pulmonary vascular diseases: clinical applications and correlation with other imaging modalities. *Japanese Journal of Radiology*, *42*(2), 126–144. https://doi.org/10.1007/s11604-023-01483-2

Yan, R., Qu, L., Wei, Q., Huang, S.-C., Shen, L., Rubin, D. L., Xing, L., & Zhou, Y. (2023). Label-Efficient Self-Supervised Federated Learning for Tackling Data Heterogeneity in Medical Imaging. *IEEE Transactions on Medical Imaging*, *42*(7), 1932–1943. https://doi.org/10.1109/tmi.2022.3233574

Zeng, C., Lagier, D., Lee, J.-W., & Vidal Melo, M. F. (2022). Perioperative Pulmonary Atelectasis: Part I. Biology and Mechanisms. *Anesthesiology*, *136*(1), 181–205. https://doi.org/10.1097/ALN.0000000000003943

Zhang, S., Tang, T., Peng, X., Zhang, Y., Yang, W., Li, W., Xin, X., Zhang, J., Wang, W., & Zhang, B. (2022). Automatic Localization and Identification of Thoracic Diseases from Chest X-rays with Deep Learning. *Current Medical Imaging Formerly Current Medical Imaging Reviews*, *18*(13), 1416–1425. https://doi.org/10.2174/1573405618666220518110113

Zheng, R., Zhong, Y., Yan, S., Sun, H., Shen, H., & Huang, K. (2023). MsVRL: Self-Supervised Multiscale Visual Representation Learning via Cross-Level Consistency for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, *42*(1), 91–102. https://doi.org/10.1109/TMI.2022.3204551