Deep Learning for Clinical Practice: Enhancing Chest X-ray Diagnostics

Ecem Soganciogl u

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Deep Learning for Clinical Practice: Enhancing Chest X-ray Diagnostics

Ecem Sogancioglu

This publication has been made possible by the Dutch Technology Foundation STW, which formed the NWO Domain Applied and Engineering Sciences and partly funded by the Ministry of Economic Affairs (Perspectief programme P15-26 'DLMedIA: Deep Learning for Medical Image Analysis').

Author: Ecem Sogancioglu
Title: Deep Learning for Clinical Practice: Enhancing Chest X-ray Diagnostics

Radboud Dissertations Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Cover : Proefschrift AIO | Guntra Laivacuma
Printing : DPN Rikken/Pumbo
ISBN : 9789493296770
DOI : 10.54195/9789493296770

Free download at: www.boekenbestellen.nl/radboud-university-press/dissertations

RADBOUD
UNIVERSITY
PRESS

Deep Learning for Clinical Practice: Enhancing Chest X-ray Diagnostics

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 6 november 2024
om 12.30 uur precies

door

Ecem Sogancioglu

geboren op 15 Maart 1990
te Izmir (Turkije)

Promotoren:	**Prof. dr. B. van Ginneken**
	Prof. dr. ir. C.I. Sánchez Gutiérrez (Universiteit van Amsterdam)
Copromotor:	**Dr. K. Murphy**
Manuscriptcommissie:	**Dr. F. Ciompi**
	Dr. Y. Güçlütürk
	Prof. dr. P.A. de Jong (Universiteit Utrecht)

Deep Learning for Clinical Practice: Enhancing Chest X-ray Diagnostics

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Wednesday, November 6, 2024
at 12:30 pm

by

Ecem Sogancioglu

born on Maart 15, 1990
in Izmir (Türkiye)

PhD supervisors:	**Prof. dr. B. van Ginneken**
	Prof. dr. ir. C.I. Sánchez Gutiérrez (University of Amsterdam)
PhD co-supervisors:	**Dr. K. Murphy**
Manuscript Committee:	Dr. F. Ciompi
	Dr. Y. Güçlütürk
	Prof. dr. P.A. de Jong (Utrecht University)

Dedicated to my sister and me as children, who endured hardships with courage, strength, and play.

Zorlukları cesaret, güç ve oyunla karşılayan Ecem ve Gizem'in çocukluğuna adanmıştır.

TABLE OF CONTENTS

Introduction

1

1.1 Chest X-ray

Chest X-rays (CXR), after Wilhelm Röntgen discovered x-ray in the late 19th-century [1], have become a cornerstone in diagnostic imaging. This technology fundamentally transformed healthcare, offering a non-invasive window into the body's internal structures. Its impact was profound, shifting medical paradigms from invasive exploratory procedures to precise visual diagnostics. The first medical X-ray, illustrated in Figure 1.1 [2, 3], is not just a historical artifact but a milestone, underscoring the evolution of medical technology and its pivotal role in modern diagnostics.

Chest radiographs, in particular, are employed to visualize the thoracic cavity, which includes the heart, lungs, airways, blood vessels, and the bones of the spine and chest. They are critical in diagnosing and monitoring an array of chest abnormalities and diseases including lung diseases, cardiac conditions, bone abnormalities, foreign objects and accumulations, cancer detection, and preoperative and postoperative assessments.

In addition to their clinical applications, the relatively low cost, accessibility, low radiation dose and simplicity of chest X-rays have made them a cornerstone of medical diagnosis worldwide. This is evident from the 3.6 billion X-ray exams that were performed annually worldwide according to the 2008 United Nations Scientific Committee on Effects of Atomic Radiation (UNSCEAR) report [4]. In 2020, almost 17 million X-ray studies were performed in the UK, comprising 48% of all medical imaging exams [5].

Figure 1.1: The inaugural X-ray image by Röntgen (1895), showing Anna Bertha Ludwig's hand with her wedding ring, symbolizing the advent of medical imaging [2, 3].

1.1.1 Types of Chest X-ray in Healthcare

Chest X-rays are conducted using a machine that emits X-ray radiation. The patient is positioned so that the X-ray beams pass through the chest to create an image on a specialized digital plate or film lo-

cated on the opposite side. As the X-rays pass through the body, they are absorbed by different tissues to varying degrees; bones, for example, absorb more radiation and thus appear white on the image, while air in the lungs absorbs little and appears black. The resulting images offer a visualization of the structures within the chest.

There are primarily three types of standard chest X-ray images: Posteroanterior (PA), Anteroposterior (AP), and Lateral. An example of different types of CXR can be seen in Figure 1.2, and the types of CXRs are explained below in detail.

Posteroanterior (PA): In PA images, the X-ray beams travel from the back (posterior) of the patient to the front (anterior), and the image is captured from the front. The patient stands facing the image plate with their chest pressed against it and arms raised. This is the standard positioning as it offers a clear and accurate representation of the chest structures, with minimal distortion of the heart and lungs.

Anteroposterior (AP): In contrast, in AP images, the X-ray beams travel from the front (anterior) to the back (posterior) of the patient. These images are often taken when a patient is unable to stand or is bedridden, such as in intensive care units or emergency settings. In this case, the image plate is placed behind the patient's back, and the X-ray machine is positioned in front of the chest. AP images may have some magnification of the heart and mediastinum compared to PA images.

Lateral: Lateral images are taken from the side, typically together with PA view. The patient stands with arms raised and the left side pressed against the image plate. This view helps in providing additional information about the structures in the chest from a different angle, particularly helpful in localizing abnormalities seen in PA view. It can reveal abnormalities not visible in the PA or AP view, especially those located in the posterior or anterior chest wall, or those obscured by the heart and diaphragm.

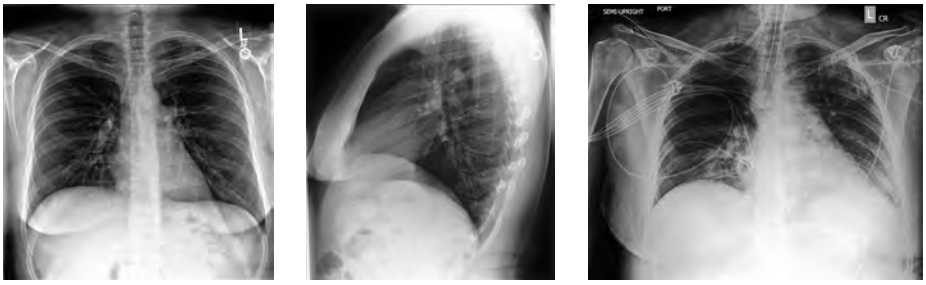


Figure 1.2: Left: posterior-anterior (PA) view frontal chest radiograph. Middle: lateral chest radiograph. Right: Anterior-posterior (AP) view chest radiograph. All three CXRs are taken from the CheXpert dataset [6], patient 184.

In this thesis, posteroanterior (PA) images are primarily utilized in each chapter. However, Chapter 4 also incorporates lateral images alongside PA images.

1.1.2 Typical workflow of CXR Interpretation in Healthcare

In the realm of chest X-ray (CXR) interpretation, the workflow of a radiologist is characterized by a blend of complex and diverse tasks. These tasks, while varied, often involve a methodical and systematic approach. This workflow is significantly influenced by the specific symptoms and clinical

context presented by each patient. Typically, a radiologist commences their analysis by methodically assessing CXRs for a range of common abnormalities. These may include lung pathologies such as pneumonia, pleural effusions, lung masses or nodules, and chronic conditions like COPD. The process entails a detailed and systematic review of various anatomical areas, which is a consistent step in the examination of each X-ray.
In addition to lung assessment, radiologists also evaluate the heart size to detect potential cardiac issues and inspect the bones for any fractures or lesions. In this thesis, we propose an automated method to assess heart size, as detailed in Chapter 3. A critical and frequent aspect of radiologists' work is comparing current images with previous scans, an essential practice for tracking the progression of diseases or the effectiveness of treatments. Depending on the patient's specific clinical scenario, such as in ICU settings, the radiologist's focus may shift. In these cases, they might concentrate on routinely checking the placement of medical devices like tubes and lines, or in instances of trauma, prioritize identifying acute injuries. Throughout their workflow, radiologists typically document their findings in a standardized format, often employing templated language to consistently convey observations and conclusions.

1.1.3 Challenges with Chest X-rays

Given their accessibility, relatively low cost, and non-invasive nature, CXRs have become indispensable in clinical practice worldwide. However, despite the utility of CXRs, several challenges are associated with their interpretation. CXRs are notoriously challenging to analyze (e.g overlapping anatomy, and variations in image quality), and the detection of specific pathologies requires a meticulous analysis from trained radiologists. Additionally, the subtle nature of certain abnormalities can make them easy to overlook, even for experienced radiologists [7, 8]. These challenges are further compounded by the sheer volume of CXRs being taken, which can lead to substantial workloads for radiologists and, in turn, delayed reporting and increased likelihood of errors due to fatigue. According to the American College of Radiology, a significant proportion of radiologists report feelings of burnout, in part due to the heavy workload associated with interpreting medical images, including CXRs [9].
Furthermore, the global disparity in the availability of trained radiologists is a significant issue, especially in developing countries. This shortage often leads to situations where chest X-rays (CXRs) are interpreted by healthcare workers who may not have specialized training in radiology. This can result in potential inaccuracies in diagnoses. In some cases, particularly in remote or under-resourced areas, there might even be instances where there is no one available to analyze these images at all.
In light of these challenges, automated algorithms for the analysis of CXRs have emerged as a promising avenue to augment diagnostic accuracy and efficiency. The anticipated benefits of automated Chest X-ray (CXR) analysis encompass enhanced detection of subtle abnormalities (e.g., nodules), triage of urgent cases, automated reporting of tedious daily tasks, and delivering analytical support in scenarios with limited or no radiologist availability.

1.2 Automated Chest X-ray analysis

Automated CXR analysis systems hold the potential to address numerous challenges inherent in conventional CXR interpretation to enhance diagnostic capabilities. These systems can potentially offer a range of benefits, including but not limited to:

1. **Increase Efficiency:** By rapidly analyzing images and generating preliminary reports, automated systems can greatly reduce the time radiologists need to spend on each case.
2. **Enhance Accuracy:** Automated algorithms can be trained to recognize subtle features that might be overlooked by the human eye, potentially reducing diagnostic errors.
3. **Prioritize Urgent Cases:** Automated algorithms can alleviate the workload on radiologists by acting as a first-level filter, identifying normal cases and flagging potential abnormalities for expert review as triage. This can reduce the turnaround time for reports, enabling timely clinical intervention, which is often critical for patient outcomes.
4. **Alleviate Workloads:** By handling the more routine and straightforward cases, automated systems can alleviate the workload on radiologists, allowing them to focus their expertise on more complex cases or those requiring urgent attention. This not only optimizes the allocation of human resources but also ensures that patients receive timely and accurate assessments.
5. **Facilitate Remote Diagnostics:** In areas where radiologists are not readily available, automated systems can provide critical diagnostic support.
6. **Identify New Health Markers:** Advanced algorithms in automated systems can potentially identify new health markers, thereby facilitating earlier detection of pathologies and enhancing our understanding of patient health.

1.3 Deep Learning

In the past years, deep learning has emerged as the go-to method for image analysis, and it has significantly transformed the realm of medical imaging including automated chest X-ray analysis [10, 11]. Deep learning, which differs from traditional machine learning, can learn complex patterns directly from data without needing specific feature engineering. This is achieved through its use of layered neural networks. These networks can extract and learn features at multiple levels of abstraction, making them particularly adept at handling the intricate details present in medical images.

The foundational concepts of neural networks date back to the 1940s and 1950s [12, 13]. Despite their early inception, the surge in deep learning (DL) popularity did not occur until the late 2000s [14]. This resurgence was largely fueled by the convergence of two critical factors: the availability of extensive datasets and the advent of robust computational hardware, particularly graphics processing units (GPUs). These GPUs facilitated large-scale training of deep neural networks, a task that was previously unattainable.

Additionally, the evolution of cloud computing technologies has played a pivotal role. It has democratized access to powerful computing resources, allowing researchers and practitioners to engage in DL without the prohibitive costs and complexities associated with establishing a personal computing infrastructure. The rise of DL has also been propelled by the emergence of open-source frameworks, like TensorFlow and PyTorch. These platforms have simplified the development and experimentation processes for DL models, contributing significantly to its widespread adoption.

In recent years, DL has been achieving unprecedented, state-of-the-art results across various domains. Notably, in the field of computer vision, these advancements are attributed to the progression in convolutional neural networks (CNNs) and the more recent integration of transformer models, a technology already prevalent in areas like natural language processing. This thesis focuses exclusively on the advancements and applications of Convolutional Neural Networks (CNNs) for CXR applications.

1.3.1 CNN

CNNs have become a dominant method for processing data with a grid-like structure, particularly images. This makes them markedly different from fully connected neural networks, especially in the context of image processing.
In fully connected networks, each input node is connected to every node in the following layer, which becomes computationally heavy and less efficient for high-dimensional data like images. CNNs, however, solve this problem by employing weight sharing with an approach called convolution. They use small, learnable filters that move across the input image to extract key features such as edges and textures. The same filter is applied everywhere in the image and this significantly reduces the number of weights, enhancing efficiency and reducing overfitting risks.
Further, the convolutional process allows them to detect features irrespective of where they appear in the input field, making them highly effective for tasks where the object of interest might vary in size or be located in different parts of the image. Additionally, the introduction of pooling layers in CNNs helps in reducing data dimensionality by downsampling the output from the convolutional layers, further decreasing computational demands.
Historically, CNNs gained prominence in the 1990s [15, 16] but truly came into the spotlight with the success of AlexNet in the 2012 ImageNet competition [17, 18], a landmark moment in the field of deep learning. This model showcased the power of CNNs in handling complex image recognition tasks with unprecedented accuracy. Since then, several other models like VGGNet, ResNet, and DenseNet [19–21] have continued to advance the field, each introducing novel concepts that have improved performance in various image processing tasks. Further details on overview of deep learning methods can be found in Section 2.2.

1.4 Deep Learning for CXR Analysis

Deep learning, notable for its extensive data requirements, has profoundly impacted chest X-ray research, an area that has attracted significant research efforts. In this thesis, we conducted a thorough review of over 290 research papers employing deep learning in CXR analysis. This review, presented in Chapter 2, identifies the trends, the gaps and suggests potential future directions in the field.
The significant increase in the number of publications happened especially following the release of large, labeled datasets. These datasets' labels have been predominantly generated through automated analysis of radiology reports. However, the emergence of public datasets and the consequent boom in CXR research has inadvertently led to a deviation in research focus. Much of the effort has been concentrated on identifying a broad spectrum of abnormalities, often more than ten, within a single CXR as a classification task, where the model's output is confined to the probability of an abnormality. This trend, while beneficial for benchmarking, reproducibility, and research acceleration, overlooks several crucial aspects. Firstly, there is a strong need to address data quality and inherent biases in the **annotation/labels** of these datasets. Often, dataset labels are derived using natural language processing techniques, which may not be sufficiently accurate, particularly for evaluating and comparing model performances. To accurately assess model performance, 'gold-standard' annotations or labels are necessary. Such high-quality labels could be obtained through expert radiological analysis of chest X-rays (CXR), ideally involving multiple readers, or through associated CT scans, laboratory test results, or other relevant measurements. Moreover, the majority of public datasets originate from **single institutions**, leading to a research focus on models trained and tested on data from a single source. While valuable, this approach does not address the critical need for models to perform consistently

across different clinical environments, a key requirement for real-world application.
Additionally, the research emphasis on detecting multiple abnormalities within a single CXR image has broadened **the gap between academic research and the practical clinical needs** for focused detection of significant abnormalities. There is a need for research to realign with clinical relevance and utility. This necessitates a rigorous evaluation of the potential benefits provided by the AI systems developed, as detailed in Section 1.2.
To transition from current research to more clinically relevant CXR analysis, a comprehensive approach is crucial. This entails a deep understanding of how AI systems will be utilized in clinical settings and their interaction with radiologists, which in turn will shape the development of these systems. It involves addressing radiologists' specific needs, building trust through transparent and explainable AI functionalities, ensuring that AI systems enhance efficiency without disrupting existing workflows, and considering their adaptability and scalability across diverse clinical environments. Key focuses include the quality of annotations, the relevance of the tasks AI is designed to perform, how the interaction of this AI system with radiologist will be and the generalizability of systems to various patient demographics and healthcare infrastructures.
This thesis focuses on addressing several of these considerations aiming to move towards the development of **clinically relevant AI systems for CXR analysis**, integrating insights from our exhaustive literature review to inform the creation of systems that align closely with the real-world clinical needs.

1.5 Outline

In this thesis, we identify and address the challenges and gaps within the realm of deep learning for CXR analysis, as outlined in Chapter 2. Drawing from these insights, our research efforts have been geared towards solutions that contribute to the advancement of **clinically relevant CXR analysis systems**, detailed in Chapters 3, 4, and 5. Additionally, by recognizing the distinct characteristics of the problems in this domain, we have contributed to the field through fostering open-source and collaborative research by organizing a research challenge (Chapter 5).
Chapter 2 conducts a systematic review of **296** research papers published in the domain of CXR analysis employing deep learning. This literature review provides a comprehensive introduction to CXRs and deep learning. We explain the deep learning methods frequently employed in this field, provide an exhaustive list of all the publicly available datasets, and identify the issues and **challenges** associated with these datasets. Further, we identify **gaps** and challenges in the field, and propose potential **future directions** towards the direction of building **clinically relevant** CXR systems.
Chapter 3 delves into the detection of cardiomegaly on frontal chest radiographs utilizing two distinct deep learning strategies - anatomical segmentation and image-level classification. Although the image-level classification method was the more prevalent approach in earlier literature for this task, our findings indicate that the segmentation-based technique we propose surpasses image-level classification in terms of both enhanced accuracy and superior interpretability. Moreover, the approach a based on segmentation, trained on a moderately sized set of chest radiographs, exhibits performance akin to that of an independent radiologist. In contrast to the classification-based solutions used previously, our proposed method establishes a more **clinically relevant** system by yielding **quantitative measurements** for cardiomegaly and producing a more **explainable** solution with enhanced **performance**.
Chapter 4 delves into the potential of utilizing deep learning for estimating a critical quantitative biomarker - total lung volume - directly from chest X-rays. This study marks, to the best of our

knowledge, one of the first instances of demonstrating that state-of-the-art deep learning solutions have the capability to accurately predict total lung volume from standard chest radiographs. The model developed is **openly accessible** and can be employed to determine total lung volume from routinely captured chest X-rays. This deep learning system can serve as a valuable instrument for tracking trends over time in patients who undergo regular chest X-ray examinations.

Chapter 5 examines cutting-edge nodule detection and generation methodologies by organizing an **open-source** and **collaborative** research challenge. We established a public challenge, **NODE21**, with the objective of pinpointing state-of-the-art techniques in nodule detection and generation on chest X-rays. Furthermore, our work systematically assess the utility of nodule generation methodologies for the task of nodule detection. To achieve this, additional comprehensive experiments were conducted with the winning solutions from both tracks to evaluate the influence of image generation on detection methods. Our results demonstrate that employing generated images can improve the performance of detection methods, with this impact being especially pronounced when there is a scarcity of real nodule images available. The structure of this challenge was designed to accept submissions exclusively in the form of open-source solutions using Docker containers, which guarantees the **reproducibility** of all methods submitted.

Deep Learning for Chest X-ray Analysis: A Survey

2

Authors: Ecem Sogancioglu*, Erdi Çallı*, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy

Original title: Deep learning for chest X-ray analysis: A survey

Published in: Medical Image Analysis, 72(8):102125, 2021

** contributed equally*

2.1 Introduction

A cornerstone of radiological imaging for many decades, chest radiography (chest X-ray, CXR) remains the most commonly performed radiological exam in the world with industrialized countries reporting an average 238 erect-view chest X-ray images acquired per 1000 of population annually [22]. In 2006, it is estimated that 129 million CXR images were acquired in the United States alone [23]. The demand for, and availability of, CXR images may be attributed to their cost-effectiveness and low radiation dose, combined with a reasonable sensitivity to a wide variety of pathologies. The CXR is often the first imaging study acquired and remains central to screening, diagnosis, and management of a broad range of conditions [24].

Chest X-rays may be divided into three principal types, according to the position and orientation of the patient relative to the X-ray source and detector panel: posteroanterior, anteroposterior, lateral. The posteroanterior (PA) and anteroposterior (AP) views are both considered as frontal, with the X-ray source positioned to the rear or front of the patient respectively. The AP image is typically acquired from patients in the supine position, while the patient is usually standing erect for the PA image acquisition. The lateral image is usually acquired in combination with a PA image, and projects the X-ray from one side of the patient to the other, typically from right to left. Examples of these image types are depicted in Figure 2.1.

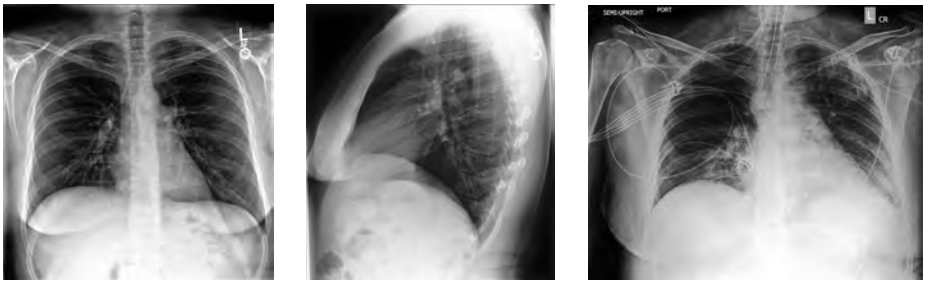


Figure 2.1: Left: posterior-anterior (PA) view frontal chest radiograph. Middle: lateral chest radiograph. Right: Anterior-posterior (AP) view chest radiograph. All three CXRs are taken from the CheXpert dataset [6], patient 184.

The interpretation of the chest radiograph can be challenging due to the superimposition of anatomical structures along the projection direction. This effect can make it very difficult to detect abnormalities in particular locations (for example, a nodule posterior to the heart in a frontal CXR), to detect small or subtle abnormalities, or to accurately distinguish between different pathological patterns. For these reasons, radiologists typically show high inter-observer variability in their analysis of CXR images [25–27].

The volume of CXR images acquired, the complexity of their interpretation, and their value in clinical practice have long motivated researchers to build automated algorithms for CXR analysis. Indeed, this has been an area of research interest since the 1960s when the first papers describing an automated abnormality detection system on CXR images were published [28–32]. The potential gains from automated CXR analysis include increased sensitivity for subtle findings, prioritization of time-sensitive cases, automation of tedious daily tasks, and provision of analysis in situations where radiologists are not available (e.g., the developing world).

In recent years, deep learning has become the technique of choice for image analysis tasks and made a tremendous impact in the field of medical imaging [10]. Deep learning is notoriously data-hungry and the CXR research community has benefited from the publication of numerous large labeled databases in recent years, predominantly enabled by the generation of labels through automatic parsing of radiology reports. This trend began in 2017 with the release of 112,000 images from the NIH clinical center [33]. In 2019 alone, more than 755,000 images were released in 3 labelled databases (CheXpert [6], MIMIC-CXR [34], PadChest [35]). In this work, we demonstrate the impact of these data releases on the number of deep learning publications in the field.
There have been previous reviews on the field of deep learning in medical image analysis [10, 36–38] and on deep learning or computer-aided diagnosis for CXR [39–41]. However, recent reviews of deep learning in chest radiography are far from exhaustive in terms of the literature and methodology surveyed, the description of the public datasets available, or the discussion of future potential and trends in the field. The literature review in this work includes 296 papers, published between 2015 and March 2021, and categorized by application. A comprehensive list of public datasets is also provided, including numbers and types of images and labels as well as some discussion and caveats regarding various aspects of these datasets. Trends and gaps in the field are described, important contributions discussed, and potential future research directions identified. We additionally discuss the commercial software available for chest radiograph analysis and consider how research efforts can best be translated to the clinic.

2.1.1 Literature Search

The initial selection of literature to be included in this review was obtained as follows: A selection of papers was created using a PubMed search for papers with the following query.

```
chest and ("x-ray" or xray or radiograph) and
("deep learning" or cnn or "convolutional" or
"neural network")
```

A systematic search of the titles of conference proceedings from SPIE, MICCAI, ISBI, MIDL and EMBC was also performed, searching paper titles for the same search terms listed above. In the case of multiple publications of the same paper, only the latest publication was included. Relevant peer-reviewed articles suggested by co-authors and colleagues were added. The last search was performed on March 3rd, 2021.
This search strategy resulted in 767 listed papers. Of these, 61 were removed as they were duplicates of others in the list. A further 260 were excluded as their subject matter did not relate to deep learning for CXR, they were commentary or evaluation papers (did not describe a deep-learning architecture, but rather just evaluated it) or they were not written in English. Publications that were not peer-reviewed were also excluded (8). Finally, during the review process 142 papers were excluded as the scientific content was considered unsound, as detailed further in Section 2.6, leaving 296 papers in the final literature review.
The remainder of this work is structured as follows: Section 2.2 provides a brief introduction to the concept of deep learning and the main network architectures encountered in the current literature. In Section 2.3, the public datasets available are described in detail, to provide context for the literature study. The review of the collected literature is provided in Section 2.4, categorized according to the major themes identified. Commercial systems available for chest radiograph analysis are described in Section 2.5. The paper concludes in Section 2.6, with a comprehensive discussion of the current

state of the art for deep learning in CXR as well as the potential for future directions in both research and commercial environments.

2.2 Overview of Deep Learning Methods

This section provides an introduction to deep learning for image analysis, and particularly the network architectures most frequently encountered in the literature reviewed in this work. Formal definitions and more in-depth mathematical explanations of fully-connected and convolutional neural-networks are provided in many other works, including a recent review of deep learning in medical image analysis [10]. In this work, we provide only a brief overview of these fundamental details and refer the interested reader to previous literature.
Deep learning is a branch of machine learning, which is a general term describing learning algorithms. The algorithm underpinning all deep learning methods is the neural network, in this case, constructed with many hidden layers ('deep'). These networks may be constructed in many ways with different types of layers included and the overall construction of a network is referred to as its 'architecture'. Sections 2.2.3 to 2.2.6 describe commonly used architectures categorized by types of application in the CXR literature.

2.2.1 Convolutional Neural Networks

In the 1980s, networks using convolutional layers were first introduced for image analysis [15], and the idea was formalized over the following years [16]. These convolutional layers now form the basis for all deep learning image analysis tasks, almost without exception. Convolutional layers use neurons that connect only to a small 'receptive field' from the previous layer. These neurons are applied to different regions of the previous layer, operating as a sliding window over all regions, and effectively detecting the same local pattern in each location. In this way, spatial information is preserved and the learned weights are shared.

2.2.2 Transfer Learning

Transfer learning investigates how to transfer knowledge extracted from one domain (source domain) to another (target) domain. One of the most commonly used transfer learning approaches in CXR analysis is the use of pre-training.
With the pre-training approach, the network architecture is first trained on a large dataset for a different task, and the trained weights are then used as an initialization for the subsequent task for fine-tuning [42]. Depending on data availability from the target domain, all layers can be re-trained, or only the final (fully connected) layer can be re-trained. This approach allows neural networks to be trained for new tasks using relatively smaller datasets since useful low-level features are learned from the source domain data. It has been shown that pre-training on the ImageNet dataset (for classification of natural images) [43] is beneficial for chest radiography analysis and this type of transfer learning is prominently used in the research surveyed in this work. ImageNet pre-trained versions of many architectures are publicly available as part of popular deep learning frameworks. The pre-trained architectures may also be used as feature extractors, in combination with more traditional methods, such as support vector machines or random forests. Domain adaptation is another subfield of transfer learning and is discussed thoroughly in Section 2.2.7.

2.2.3 Image-level Prediction Networks

In this work we use the term 'image-level prediction' to refer to tasks where prediction of a category label (classification) or continuous value (regression) is implemented by analysis of an entire CXR image. These methods are distinct from those which make predictions regarding small patches or segmented regions of an image. Classification and regression tasks are grouped together in this work since they typically use the same types of architecture, differing only in the final output layer. One of the early successful deep convolutional architectures for image-level prediction was AlexNet [17], which consists of 5 convolutional layers followed by 3 fully connected layers. AlexNet became extremely influential in the literature when it beat all other competitors in the ILSVRC (ImageNet) challenge [18] by a large margin in 2012. Since then many deep convolutional neural network architectures have been proposed. The VGG family of models [19] use 8 to 19 convolutional layers followed by 3 fully-connected layers. The Inception architecture was first introduced in 2015 [44] using multiple convolutional filter sizes within layered blocks known as Inception modules. In 2016, the ResNet family of models [20] began to gain popularity and improve upon previous benchmarks. These models define residual blocks consisting of multiple convolution operations, with skip connections which typically improve model performance. After the success of ResNet, skip connections were widely adopted in many architectures. DenseNet models [21], introduced in 2017, also use skip connections between blocks, but connect all layers to each other within blocks. A later version of the Inception architecture also added skip connections (Inception-Resnet) [45]. The Xception network architecture [46] builds upon the Inception architecture but separates the convolutions performed in the 2D image space from those performed across channels. This was demonstrated to improve performance compared to Inception V3.

The majority of works surveyed in this review use one or more of the model architectures discussed here with varying numbers of hidden layers.

2.2.4 Segmentation Networks

Segmentation is a task where pixels are assigned a category label, and can also be considered as a pixel classification. In natural image analysis, this task is often referred to as 'semantic segmentation' and frequently requires every pixel in the image to have a specified category. In the medical imaging domain these labels typically correspond to anatomical features (e.g., heart, lungs, ribs), abnormalities (e.g., tumor, opacity) or foreign objects (e.g., tubes, catheters). It is typical in the medical imaging literature to segment just one object of interest, essentially assigning the category 'other' to all remaining pixels.

Early approaches to segmentation using deep learning used standard convolutional architectures designed for classification tasks [47]. These were employed to classify each pixel in a patch using a sliding window approach. The main drawback to this approach is that neighboring patches have huge overlap in pixels, resulting in inefficiency caused by repeating the same convolutions many times. It additionally treats each pixel separately which results in the method being computationally expensive and only applicable to small images or patches from an image.

To address these drawbacks, fully convolutional networks (FCNs) were proposed, replacing fully connected layers with convolutional layers [48]. This results in a network which can take larger images as input and produces a likelihood map output instead of an output for a single pixel. In 2015, a fully convolutional architecture known as the U-Net was proposed [49] and this work has become the most cited paper in the history of medical image analysis. The U-Net consists of several convolutional lay-

ers in a contracting (downsampling) path, followed by further convolutional layers in an expanding (upsampling) path which restores the result to the input resolution. It additionally uses skip connections (feature forwarding) between the same levels on the contracting and expanding paths to recover fine details that were lost during the pooling operation. The majority of image segmentation works in this review employ a variant of the FCN or the U-Net.

2.2.5 Localization Networks

This survey uses the term localization to refer to identification of a specific region within the image, typically indicated by a bounding box, or by a point location. As with the segmentation task, localization, in the medical domain, can be used to identify anatomical regions, abnormalities, or foreign object structures. There are relatively few papers in the CXR literature reviewed here that deal specifically with a localization method, however, since it is an important task in medical imaging, and may be easier to achieve than a precise segmentation, we categorize these works together.
In 2014, the RCNN (Region Convolutional Neural Network) was introduced [50], identifying regions of interest in the image and using a CNN architecture to extract features of these regions. A support vector machine (SVM) was used to classify the regions based on the extracted features. This method involves several stages and is relatively slow. It was later superseded by fast-RCNN [51] and subsequently by faster-RCNN [52] which streamlined the processing pipeline, removing the need for initial region identification or SVM classification, and improving both speed and performance. In 2017, a further extension was added to faster-RCNN to additionally enable a precise segmentation of the item identified within the bounding box. This method is referred to as Mask R-CNN [53]. While this is technically a segmentation network, we mention it here as part of the RCNN family. Another architecture which has been popular in object localization is YOLO (You Only Look Once), first introduced in 2016 [54] as a single-stage object detection method, and improved in subsequent versions in 2017 and 2018 [55, 56]. The original YOLO architecture, using a single CNN and an image-grid to specify outputs was significantly faster than its contemporaries but not quite as accurate. The improved versions leveraged both classification and detection training data and introduced a number of training improvements to achieve state of the art performance while remaining faster than its competitors. A final localization network that features in medical imaging literature is RetinaNet [57]. Like YOLO, this is a single stage detector, which introduces the concept of a focal loss function, forcing the network to concentrate on more difficult examples during training. Most of the localization works included in this review use one of the architectures described above.

2.2.6 Image Generation Networks

One of the tasks deep learning has been commonly used for is the generation of new, realistic images, based on information learned from a training set. There are numerous reasons to generate images in the medical domain, including generation of more easily interpretable images (by increasing resolution, or removal of projected structures impeding analysis), generation of new images for training (data augmentation), or conversion of images to emulate appearances from a different domain (domain adaptation). Various generative schemes have also been used to improve the performance of tasks such as abnormality detection and segmentation.
Image generation was first popularized with the introduction of the generative adversarial network (GAN) in 2014 [58]. The GAN consists of two network architectures, an image generator, and a discriminator which attempts to differentiate generated images from real ones. These two networks are

trained in an adversarial scheme, where the generator attempts to fool the discriminator by learning to generate the most realistic images possible while the discriminator reacts by progressively learning an improved differentiation between real and generated images.

The training process for GANs can be unstable with no guarantee of convergence, and numerous researchers have investigated stabilization and improvements of the basic method [59–62]. GANs have also been adapted to conditional data generation [63, 64] by incorporating class labels, image-to-image translation (conditioned on an image in this case) [65], and unpaired image-to-image translation (CycleGAN [66]).

GANs have received a lot of attention in the medical imaging community and several papers were published for medical image analysis applications in recent years [67]. Many of the image generation works identified in this review employed GAN based architectures.

2.2.7 Domain Adaptation Networks

In this work we use the term 'Domain Adaptation', which is a subfield of transfer learning, to cover methods attempting to solve the issue that architectures trained on data from a single 'domain' typically perform poorly when tested on data from other domains. The term 'domain' is weakly defined; In medical imaging it may suggest data from a specific hardware (scanner), set of acquisition parameters, reconstruction method or hospital. It could, less frequently, also refer to characteristics of the population included, for example the gender, ethnicity, age or even strain of some pathology included in the dataset.

Domain adaptation methods consider a network trained for an image analysis task on data from one domain (the source domain), and how to perform this analysis accurately on a different domain (the target domain). These methods can be categorized as supervised, unsupervised, and semi-supervised depending on the availability of labels from the target domain and they have been investigated for a variety of CXR applications from organ segmentation to multi-label abnormality classification. There is no specific architecture that is typical for domain adaptation, but rather architectures are combined in various ways to achieve the goal of learning to analyze images from unseen domains. The approaches to this problem can be broadly divided into three classes (following the categorization of [68]); discrepancy-based, reconstruction-based and adversarial-based.

Discrepancy-based approaches aim to induce alignment between the source and target domain in some feature space by fine-tuning the image analysis network and optimizing a measurement of discrepancy between the two domains. Reconstruction-based approaches, on the other hand, use an auxiliary encoder-decoder reconstruction network that aims to learn domain invariant representation through a shared encoder. Adversarial-based approaches are based on the concept of adversarial training from GANs, and use a discriminator network which tries to distinguish between samples from the source and target domains, to encourage the use of domain-invariant features. This category of approaches is the most commonly used in CXR analysis for domain adaptation, and consists of generative and non-generative models. Generative models transform source images to resemble target images by operating directly on pixel space whereas non-generative models use the labels on the source domain and leverage adversarial training to obtain domain invariant representations.

2.3 Datasets

Deep learning relies on large amounts of annotated data. The digitization of radiological workflows enables medical institutions to collate and categorize large sets of digital images. In addition, advances in natural language processing (NLP) algorithms mean that radiological reports can now be automatically analyzed to extract labels of interest for each image. These factors have enabled the construction and release of multiple large labelled CXR datasets in recent years. Other labelling strategies have included the attachment of the entire radiology report and/or labels generated in other ways, such as radiological review of the image, radiological review of the report, or laboratory test results. Some datasets include segmentations of specified structures or localization information.
In this section we detail each public dataset that is encountered in the literature included in this review as well as any others available to the best of our knowledge. Details are provided in Table 2.1. Each dataset is given an acronym which is used in the literature review tables (Tables 2.2 to 2.7) to indicate that the dataset was used in the specified work.

1. ChestX-ray14 (C) is a dataset consisting of $112,120$ CXRs from $30,805$ patients [33]. The CXRs are collected at the (US) National Institute of Health. The images are distributed as 8-bit grayscale images scaled to 1024×1024 pixels. The dataset was automatically labeled from radiology reports, indicating the existence of 14 types of abnormality.

2. CheXpert (X) is a dataset consisting of $224,316$ CXRs from $65,240$ patients [6]. The CXRs are collected at Stanford Hospital between October 2002 and July 2017. The images are distributed as 8-bit grayscale images with original resolution. The dataset was automatically labeled from radiology reports using a rule-based labeler, indicating the presence, absence, uncertainty, and no-mention of 12 abnormalities, no findings, and the existence of support devices.

3. MIMIC-CXR (M) is a dataset consisting of $371,920$ CXRs from and $64,588$ patients [34]. The CXRs are collected from patients admitted to the emergency department of Beth Israel Deaconess Medical Center between 2011 and 2016. In version 1 (V1) the images are distributed as 8-bit grayscale images in full resolution. The dataset was automatically labeled from radiology reports using the same rule-based labeler system (described above) as CheXpert. A second version (V2) of MIMIC-CXR was later released including the anonymized radiology reports and DICOM files.

4. PadChest (P) is a dataset consisting of $160,868$ CXRs from $109,931$ studies and $67,000$ patients [35]. The CXRs are collected at San Juan Hospital (Spain) from 2009 to 2017. The images are stored as 16-bit grayscale images with full resolution. $27,593$ of the reports were manually labeled by physicians. Using these labels, an RNN was trained and used to label the rest of the dataset from the reports. The reports were used to extract 174 findings, 19 diagnoses, and 104 anatomic locations. The labels conform to a hierarchical taxonomy based on the standard Unified Medical Language System (UMLS) [82].

5. PLCO (PL) is a screening trial for prostate, lung, colorectal and ovarian (PLCO) cancer [69]. The lung arm of this study has $185,421$ CXRs from $56,071$ patients. The NIH distributes a standard set of $25,000$ patients and $88,847$ frontal CXRs. This dataset contains 22 disease labels with 4 abnormality levels and the locations of the abnormalities.

6. Open-i (O) is a dataset consisting of $7,910$ CXRs from $3,955$ studies and $3,955$ patients [70]. The CXRs are collected from the Indiana Network for Patient Care [83]. The images are distributed

Table 2.1: CXR datasets available for research. Values above $10,000$ are rounded and shortened using K, indicating thousand (such as 10K for $10,000$).
Labeling Methods: RP=Report Parsing, RIR=Radiologist Interpretation of Reports, RI=Radiologist Interpretation of Chest X-Rays, RCI=Radiologist Cohort agreement on Chest X-Rays, LT=Laboratory Tests.
Annotation Types: BB=Bounding Box, CL=Classification, CLoc=Classification with Location label, R=Report, SE=Segmentation.
Gold Standard Data: This refers to the number of images labeled by methods other than Report Parsing

	Patients (P) Studies (S) Images (I)	View Positions	Annotation Types	 Labels	 Studies	Image Format	Labeling method	Gold Standard Data
ChestX-ray14(C) [33]	P: 31K I: 112K	PA: 67K AP: 45K	CL BB	14 8	112K 983	PNG	RP RI	 984
CheXpert (X) [6]	P: 65K S: 188K I: 224K	PA: 29K AP: 162K LL: 32K	CL RP	14	224K	JPEG	RCI	235
MIMIC-CXR (M) [34]	P: 65K S: 224K I: 372K	PA+AP: 250K LL: 122K	CL (V1) R (V2)	14	372K 372K	JPEG (V1) DICOM (V2)	RP	
PadChest (P) [35]	P: 67K S: 110K I: 160K	PA: 96K AP: 20K LL: 51K	CL R	193	110K 110K	DICOM	RIR RP	27593
PLCO (PL) [69]	P: 25K I: 89K	PA: 89K	CL CLoc	22 17	89K 89K	TIFF	RI	All
Open-i (O) [70]	P: 3,955 I: 7,910	PA: 3,955 LL: 3,955	R		3,955	DICOM	RI	All
Ped-pneumonia (PP) [71]	I: 5,856		CL	2	5,856	JPEG	RI	All
JSRT+SCR (J) [72]	I: 247	PA: 247	SE	3	247	DICOM	RI	All
RSNA-Pneumonia (RP) [73]	I: 30K	PA:16K AP:14K	BB CL	1	30K	DICOM	RI	All
Shenzhen (S) [74]	I: 340	PA: 340	CL	2	340 340	DICOM	RI	All
Montgomery (MO) [74]	I: 138	PA: 138	CL SE	2	138 138	PNG	RI	All

continued on the next page

continued from the previous page

	Patients (P) Studies (S) Images (I)	View Positions	Annotation			Image Format	Labeling method	Gold Standard Data
			Types	Labels	Studies			
BIMCV (B) [75]	P: 9,129 S: 18,430 I: 25,554	PA: 8,748 AP: 10,469 LL: 6,337	CL	1	25,554	PNG	LT	All
COVIDDSL (CD) [76]	P: 1,725 S: 4,943	PA AP (most) LL	CL	1	4,943	DICOM	LT	All
SIIM-ACR (SI) [77]	I: 16K P:16K	PA: 11K AP: 4,799	SE	1	16K	DICOM	RI	All
CXR14-Rad-Labels (CR) [78]	P: 1,709 I: 4,374	AP: 3,244 PA: 1,132	CL	4	4,374	PNG	RCI	All
COVID-CXR (CC) [79]	I:866 P:449	PA:344 AP:438 LL:84	CL BB SE			PNG JPEG	Various	
NLST (N) [80]	I: 5493	No public information available Number of images is reported by [81]						
Object-CXR (OB)	I:10K	No longer at original download location						
Belarus (BL)	I: 300	No longer at original download location						

end of table

as anonymized DICOMs. The radiological findings obtained by radiologist interpretation are available in MeSH format[1].

7. Ped-Pneumonia (PP) is a dataset consisting of 5,856 pediatric CXRs [71]. The CXRs are collected from Guangzhou Women and Children's Medical Center, Guangzhou, China. The images are distributed in 8-bit grayscale images scaled in various resolutions. The labels include bacterial and viral pneumonia as well as normal.

8. JSRT dataset (J) consists of 247 images with a resolution of 2048×2048, 0.175mm pixel-size and 12-bit depth [72]. It includes nodule locations (on 154 images) and diagnosis (malignant or benign). The reference standard for heart and lung segmentations of these images are provided by the SCR dataset [84] and we group these datasets together in this work.

9. RSNA-Pneumonia (RP) is a dataset consisting of $30,000$ CXRs with pneumonia annotations [73]. These images are acquired from ChestX-ray14 and are 8-bit grayscale with 1024×1024 resolution. Annotations are added by radiologists using bounding boxes around lung opacities and 3 classes indicating normal, lung opacity, not normal.

[1] `https://www.nlm.nih.gov/mesh/meshhome.html`

10. Shenzhen (S) is a dataset consisting of 662 CXRs [74]. The CXRs are collected at Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China in September 2012. The images, including some pediatric images, are distributed as 8-bit grayscale with full resolution and are annotated for signs of tuberculosis.
11. Montgomery (MO) is a dataset consisting of 138 CXRs [74]. The CXRs are collected by the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. The images are distributed as anonymized DICOMs, annotated for signs of tuberculosis and additionally include lung segmentation masks.
12. BIMCV (B) is a COVID-19 dataset released by the Valencian Region Medical ImageBank (BIMCV) in 2020 [75]. It includes CXR images as well as CT scans and laboratory test results. The dataset includes 3,293 CXRs from 1,305 COVID-19 positive subjects. CXR images are 16-bit PNG format with original resolution.
13. COVIDDSL (CD) is a COVID-19 dataset released by the HM Hospitales group in Spain [76]. It includes CXR images for 1,725 patients as well as detailed results from laboratory testing, vital signs etc. All subjects are stated to be confirmed COVID-19 positive.
14. COVIDGR (CG) is a dataset consisting of 852 PA CXR images where half of them are labeled as COVID-19 positive based on corresponding RT-PCR results obtained within at most 24 hours [85]. This dataset was collected from Hospital Universitario Clínico San Cecilio, Granada, Spain, and the level of severity of positive cases is provided.
15. SIIM-ACR (SI) This dataset was released for a Kaggle challenge on pneumothorax detection and segmentation [77]. Researchers have determined that at least some (possibly all) of the images are from the ChestX-ray14 dataset although the challenge organizers have not confirmed the data sources. They are supplied in 1024×1024 resolution as DICOM files. Pixel segmentations of the pneumothorax in positive cases are provided.
16. CXR14-Rad-Labels (CR) supplies additional annotations for a subset of ChestX-ray14 data [78]. It consists of 4 labels for 4,374 studies and 1,709 patients. These labels are collected by the adjudicated agreement of 3 radiologists. These radiologists were selected from a cohort of 11 radiologists for the validation split (2,412 studies from 835 patients), and 13 radiologists for the test split (1,962 studies from 860 patients). The individual labels from each radiologist as well as the agreement labels were provided.
17. COVID-CXR (CC) is a dataset consisting of 930 CXRs at the time of writing (the dataset remains in continuous development) [79]. The CXRs are collected from a large variety of locations using different methods including screenshots from papers researching COVID-19. Available labels vary accordingly, depending on what information is available from the source where the image was obtained. Images do not have a standard resolution and are published as 8-bit PNG or JPEG files.
18. NLST (N) is a dataset of publicly available CXRs collected during the NLST screening trial [80]. This trial aimed to compare the use of low-dose computed tomography (CT) with CXRs for lung cancer screening in smokers. The study had 26,732 participants in the CXR arm and a part of this data is available upon request.
19. Object-CXR (OB) is a dataset of 10,000 CXR images from hospitals in China with foreign objects annotated on the images. The download location[2] is no longer available at the time of writing.

[2] `https://jfhealthcare.github.io/object-CXR/`

Further detail is not provided since it cannot be verified from the image source.

20. Belarus (BL) This dataset is included since it is used in a number of reviewed papers however the download location (http://tuberculosis.by) is no longer available at the time of writing. The dataset consisted of approximately 300 frontal chest X-rays with confirmed TB. Further detail is not provided since it can no longer be verified from the image source.

The rapid increase in the number of publicly available CXR images in recent years has positively impacted the number of deep learning studies published in the field. Figure 2.2 illustrates the cumulative number of publicly available CXR images and the number of publications on deep learning with CXR per year.

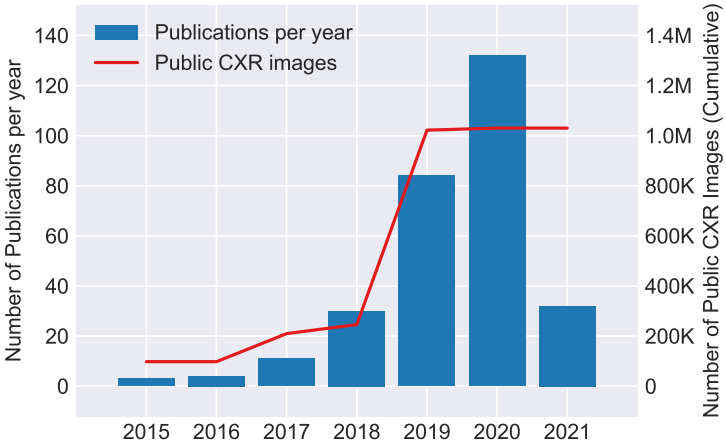


Figure 2.2: Number of publications that were reviewed in this work, by year, compared with the number of publicly available CXR images. Data for 2021 is until March 3rd of that year.

2.3.1 Public Dataset Caution

Publication of medical image data is extremely important for the research community in terms of advancing the state of the art in deep learning applications. However, there are a number of caveats that should be considered and understood when using the public datasets described in this work. Firstly, many datasets make use of Natural Language Processing (NLP) to create labels for each image. Although this is a fast and inexpensive method of labeling, it is well known that there are inaccuracies in labels acquired this way [6, 86, 87]. There are a number of causes for such inaccuracies. Firstly, some visible abnormalities may not be mentioned in the radiology report, depending on the context in which it was acquired [88]. Further, the NLP algorithm can be erroneous in itself, interpreting negative statements as positive, failing to identify acronyms, etc. Finally, many findings on CXR are subtle or doubtful, leading to disagreements even among expert observers [88]. Acknowledging some of these issues, [6] includes labels for uncertainty or no-mention in the labels on the CheXpert dataset. One particular cause for concern with NLP labels is the issue of systematic or structured mislabeling,

where an abnormality is consistently labeled incorrectly in the same way. An example of this occurs in the ChestX-ray14 dataset where subcutaneous emphysema is frequently identified as (pulmonary) 'emphysema' [86, 89].
It has been demonstrated that deep neural networks can tolerate reasonable levels of label inaccuracy in the training set without a significant effect on model performance [89, 90]. Although such labels can be used for training, for an accurate evaluation and comparison of models it is desirable that the test dataset is accurately labelled. In the literature reviewed in this work, many authors rely on labels from NLP algorithms in their test data, while others use radiologist annotations, laboratory tests and/or CT verification for improved test set labelling. We refer to data that uses these improved labelling techniques as gold standard data (Table 2.1).
The labels defined in the public datasets should also be considered carefully and understood by the researchers using them. Many labels have substantial dependencies between them. For example, some datasets supply labels for both 'consolidation' and 'pneumonia'. Consolidation (blocked airspace) is an indicator of a patient with pneumonia, suggesting there will be significant overlap between these labels. A further point for consideration is that, in practice, not all labels can be predicted by a CXR image alone. Pneumonia is rarely diagnosed by imaging alone, requiring other clinical signs or symptoms to suggest that this is the cause for a visible consolidation.
Many public datasets release images with a lower quality than is used for radiological reading in the clinic. This may be a cause for decreased performance in deep learning systems, particularly for more subtle abnormalities. The reduction in quality is usually related to a decrease in image size or bit-depth prior to release. This is typically carried out to decrease the overall download size of a dataset. However, in some cases, CXR data has been collected by acquiring screenshots from online literature, which results in an unquantifiable degradation of the data. In the clinical workflow, DICOM files are the industry standard for storing CXRs, typically using 12 bits per pixel and with image dimensions of approximately 2 to 4 thousand pixels in each of the X and Y directions. In the event that the data is post-processed before release it would be desirable that a precise description of all steps is provided to enable researchers to reproduce them for dataset combination.

2.4 Deep Learning for Chest Radiography

In this section we survey the literature on deep learning for chest radiography, dividing it into sections according to the type of task that is addressed (Image-level Prediction, Segmentation, Image Generation, Domain Adaptation, Localization, Other). For each of these sections a table detailing the literature on that task is provided. Some works which have equal main focus on two tasks may appear in both tables. For Segmentation and Localization, only studies that quantitatively evaluate their results are included in those categories. Figure 2.3 shows the number of studies for each of the tasks.

Image-level Prediction

Image-level prediction refers to the task of predicting a label (classification) or a continuous value (regression) by analyzing an entire image. Classification labels may relate to pathology (e.g. pneumonia, emphysema), information such as the subject gender, or orientation of the image. Regression values might, for example, indicate a severity score for a particular pathology, or other information such as the age of the subject.

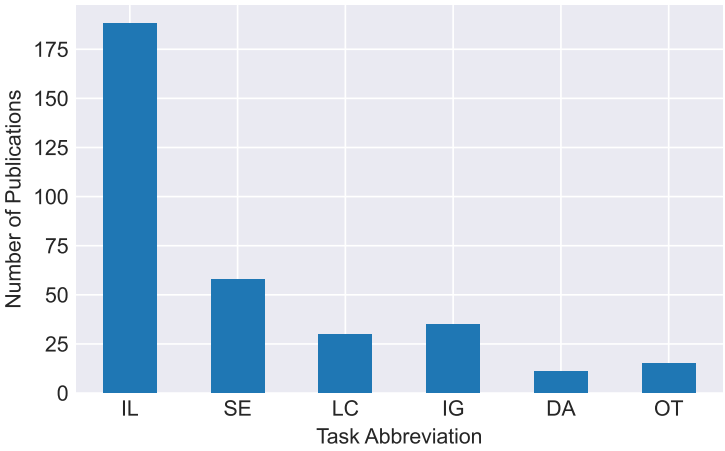


Figure 2.3: Number of publications reviewed for each task. 296 studies are included, each study may perform at most two tasks.
Tasks: IL=Image-level Predictions, SE=Segmentation, LC=Localization, IG=Image Generation, DA=Domain Adaptation, OT=Other.

We classified 188 studies, fully detailed in Table 2.2 as image-level predictions. Most of these studies make use of off-the shelf deep learning models to predict a pathology, metadata information or a set of labels provided with a dataset. The number of studies for each label are provided in Figure 2.4.

The most commonly studied image-level prediction task is predicting the labels of the ChestX-ray14 dataset (31 studies). For example, [146] compares the performance of various approaches to classify the 14 disease labels provided by the ChestX-ray14 dataset. [189] compares the performance of an ensemble of deep learning models to board-certified and resident radiologists, showing that their models achieve a performance comparable to expert observers in most of the 14 labels provided by ChestX-ray14. Following this, pneumonia is second most studied subject (26 studies). Of the 26 studies that worked with pneumonia, 12 studied pediatric chest X-rays and 11 of those used the Ped-Pneumonia dataset for training and evaluation [141, 157, 179, 230–234, 272–274]. Classification to Normal/Abnormal (or Triage) is another commonly studied topic (20 studies). Here, studies aim to distinguish normal CXRs or prioritize urgent/critical cases with the goal of reducing the radiologist workload or improving the reporting time. For example, [140] develops a triaging pipeline based on the urgency of exams. Similarly, [162] compares the performance of various deep learning models applied to several public chest X-ray datasets for distinguishing abnormal cases. Pneumothorax is another commonly studied condition (18 studies). For example, [238] aims to detect potentially critical patients and proposes that such models can be used to alert clinicians. Another common topic is tuberculosis detection (18 studies). The first studies that use deep learning to detect this infectious disease are [252, 253]. Performance of a deep learning model and how the assistance of this model improves the radiologist performance is studied by [257]. This study in particular evaluates the use of extra clinical information such as age, white blood cell count, patient temperature and oxygen saturation to assist the deep learning model. Diagnosis or evaluation of COVID-19 from CXR is another topic that has attracted a lot of interest from researchers (17 studies). For example, [150]

Table 2.2: Image-Level Prediction Studies (Section 2.4).
Tasks: AA=Adversarial Attack, DA=Domain Adaptation, IC=Interval Change, IG=Image Generation, IR=Image Retrieval, LC=Localization, OT=Other, PR=Preprocessing, RP=Report Parsing, SE=Segmentation, WS=Weak Supervision. **Bold font** in tasks implies that this additional task is central to the work and the study also appears in another table in this paper.
Labels: C=ChestX-Ray14, CM=Cardiomegaly, CV=COVID, E=Edema, GA=Gender/Age, L=Lung, LC=Lung Cancer, LO=Lesion or Opacity, M=MIMIC-CXR, MN=Many, ND=Nodule, OR=Orientation, P=PadChest, PE=Effusion, PL=PLCO, PM=Pneumonia, PT=Pneumothorax, Q=Image Quality, T=Triage/Abnormal, TB=Tuberculosis, TU=Catheter or Tube, X=CheXpert, Z=Other.
Datasets: BL=Belarus, C=ChestX-ray14, CC=COVID-CXR, CG=COVIDGR, J=JSRT+SCR, M=MIMIC-CXR, MO=Montgomery, O=Open-i, P=PadChest, PL=PLCO, PP=Ped-pneumonia, PR=Private, RP=RSNA-Pneumonia, S=Shenzen, SI=SIIM-ACR, SM=Simulated CXR from CT, X=CheXpert.

Citation	Method	Other Tasks	Labels	Datasets
[91]	Combines lung cropped CXR model and a CXR model to improve model performance	**SE**,LC,PR	C,L	C,J
[92]	Comparison of image-level prediction and segmentation models for cardiomegaly	**SE**	CM	C
[93]	A network with DenseNet and U-Net for classification of cardiomegaly	**SE**	CM	C
[94]	U-Net based model for heart and lung segmentation for cardiothoracic ratio	**SE**	CM	PR
[95]	Combines lung cropped CXR model and a CXR model using the segmentation quality	**SE**	E,LO,PE, PT,Z	M
[96]	Pneumonia detection is improved by use of lung segmentation	**SE**	PM	J,MO,PP, PR
[97]	U-Net based model to segment pneumonia	**SE**	PM	RP
[98]	Multi-scale DenseNet based model for pneumothorax segmentation	**SE**	PT	PR
[99]	DenseNet based U-Net for segmentation of the left and right humerus of the infant	**SE**	Z	PR
[100]	Uses a database of the intermediate ResNet-50 features to find similar studies	**OT**,IR	TB	MO,S
[101]	Uses activation and gradient based attention for localization and classification	**LC**	C,X	C

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[102]	Detects and localizes COVID-19 using various networks and ensembling	**LC**	CV	C,CC,PP, RP,X
[103]	GoogleNet trained with CXR patches, correlates with COVID-19 severity score	**LC**	CV,PM	C,PR
[104]	Proposes a segmentation and classification model compares with radiologist cohort	**LC**	LO,ND,PE, PT	PR
[105]	Trains a semisupervised network on a large CXR dataset with CT-confirmed nodule cases	**LC**	ND	PR
[106]	Defines a loss that minimizes the saliency map errors to improve model performance	**LC**	ND	PR
[107]	A weakly supervised localization with variational model, leverages attention maps	**LC**	PM	C
[108]	Attention guided CNN for pneumonia detection with bounding boxes	**LC**	PM	RP
[109]	A CNN for identification of abnormal CXRs and localization of abnormalities	**LC**	T	PR
[110]	Introduces a visualization method to identify regions of interest from classification	**LC**	TB	PR
[111]	Weakly supervised framework jointly trained with localization and classification	**LC**	TB	PR
[112]	Combines classification loss and autoencoder reconstruction loss	**IG**,SE	T	J,MO,O, S
[113]	Wasserstein GAN to permute diseased radiographs to appear healthy	**IG**,LC	Z	PR
[114]	Novel GAN model trained with healthy and abnormal CXR to predict difference map	**IG**	PE	SM,X
[115]	GANs with U-Net autoencoder and CNN discriminator and encoder for one-class learning	**IG**	T	C
[116]	Autoencoder uses uncertainty for reconstruction error in one-class learning setting	**IG**	T	PP,RP
[117]	Continual learning methods to classify data from new domains	**DA**	C,M	C,M
[118]	CycleGAN model to adapt adult to pediatric CXR for pneumonia classification	**DA**	PM	PP,RP
[119]	Trains a Variational Autoencoder, uses encoded features to train models	WS	X	X

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[120]	Predicts labels for unlabeled data using latent space similarity for semisupervision	WS	X	X
[121]	Y-Net to normalize image geometry for preprocessing	SE,PR	OR	C,M,X
[122]	COVID-19 opacity localization and severity detection on CXRs	SE,LC	CV	PR
[123]	ResNet-18 backbone for Covid-19 classification with limited data availability	SE,LC	CV,PM	CC,J,MO
[85]	Proposes a new dataset COVIDGR and a novel method using transformations with GANs	SE,IG	CV	CG
[124]	DenseNet for cardiomegaly detection given lung cropped CXR	SE	CM	O,P
[125]	Multiple models and combinations of CXR datasets used for COVID-19 detection	SE	CV	C,CC,PR, RP
[126]	ResNet-101 trained for COVID-19, heatmaps are generated for lung-segmented regions	SE	CV,PM	PR
[127]	Multiple architectures considered for two-stage classification of pediatric pneumonia	SE	PM	PP
[128]	Compares visualization methods for pneumonia localization	SE	PM	PP
[129]	Classifies patches and uses the positive area size to classify the image	SE	PT	PR
[130]	Feature extraction from CNN models and ensembling methods	SE	TB	MO,PR,S
[131]	Detection of central venous catheters using segmentation shape analysis	SE	TU	C
[132]	Detection of air-trapping in pediatric CXRs using Stacked Autoencoders	SE	Z	PR
[133]	Pneumoconiosis detection using Inception-v3 and evaluation against two radiologists	SE	Z	PR
[6]	Introduces CheXpert dataset and model performance on radiologist labeled test set	RP,LC	X	X
[134]	Curates data for interval change detection, proposes method comparing local features	RP,IC	Z	PR
[135]	Parses reports to define a topic model and predicts those using CXRs	RP	CM,PE,Z	O

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[136]	Trains model using image and reports to improve image only performance	RP	E	M
[137]	Extracts ambiguity of labels from reports, proposes model that uses this information	RP	E,PT,Z	M
[138]	Creates and parses reports for ChestX-ray14 AP data to obtain 73 labels for training	RP	MN	C
[139]	Obtains findings by tagging common report sentences to train models	RP	MN	PR
[140]	An ensemble of two CNNs to predict priority level for CXR queue management	RP	T	PR
[43]	Evaluates bone suppression and lung segmentation, detection of 8 abnormalities	PR,SE	CM,PE,PT,Z	O
[141]	Classification of pediatric pneumonia types using adaped VGG-16 architecture	PR,SE	PM	PP
[142]	Evaluates various image preprocessing algorithms on the performance of DenseNet-121	PR	T	C,MO,PR,S
[143]	Detects 8 findings and analyzes how these can improve workflow prioritization	OT	CM,PE,PT,Z	C,O
[144]	Proposes a model for weakly supervised classification and localization	LC,WS	C	C
[145]	Proposes a recurrent attention mechanism to improve model performance	LC	C	C
[146]	Evaluates the use of various model configurations for classification	LC	C	C
[147]	Attention mining and knowledge preservation for classification with localization	LC	C	C
[148]	Attention based model compared with well-known architectures	LC	C	C
[149]	Minimizes the encoding differences of a CXR from multiple models	LC	C	C,X
[150]	DenseNet used to predict COVID-19 severity as scored by radiologists	LC	CV	CC
[151]	Uses a ResNet-50 backed segmentation model to detect healthy, pneumonia, COVID-19	LC	CV,PM	CC,RP
[152]	Uses multi instance learning for classification with localization	LC	E,PM,PT	M,PR,RP
[153]	Lung cancer and nodule prediction using ResNet-34	LC	LC,ND	PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[154]	GradCam based attention mining loss, compared with labels extracted from reports	LC	LO	M
[78]	Trains Xception using >750k CXRs, compares results with radiologist labels	LC	LO,ND,PT, Z	C,PR
[155]	ResNet and VGG used to distinguish AP from PA images	LC	OR	PR,RP
[156]	DenseNet-121 trained on public data evaluated using CT-based labels	LC	PE,PM	C,PL
[157]	Evaluates the performance of various models trained on pediatric CXRs on adult CXRs	LC	PM	PP
[158]	Evaluates ensembling methods and visualization on pediatric CXRs	LC	PM,PT,Z	PR
[159]	Compares a ResNet-152 against radiologists and shows the statistical significance	LC	PT	C
[160]	Compares GradCAM with radiologist segmentations for evaluation of VGG-19	LC	PT	C,PR
[161]	Apical regions and patches from them extracted to detect pneumothorax	LC	PT	PR
[162]	Detection of abnormality, various networks compared with radiologist labeling	LC	T	C,O,PP,RP
[163]	Evaluates pre-training on ImageNet and CheXpert on various models/settings	LC	T	RP
[164]	Proposes a GAN-based model trained only with healthy images for anomaly detection	LC	T	RP
[165]	Proposes a new model for faster classification of TB	LC	TB	BL,MO,S
[166]	Evalutes the use of a ResNet based model on a large gold standard dataset	LC	TB	PR
[167]	Evaluates multiple models for detection of feeding tube malpositioning	LC	TU	PR
[168]	Graph CNN solution with ensembling which models disease dependencies	LC	X	X
[169]	Curates a dataset of heart failure cases and evaluates VGG-16 on it	LC	Z	C
[170]	CNN for identifiying the presence of subphrenic free air from CXR	LC	Z	PR
[171]	Evaluates several models to predict hypertension and artery systolic pressure	LC	Z	PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[172]	Uses ResNet-18 to measure the Brassfield Score, predicts Cystic Fibrosis based on it	LC	Z	PR
[173]	Simulates CXRs from CT scans and predicts emphysema scores	LC	Z	PR
[174]	Inception network to predict pulmonary to systemic flow ratio from pediatric CXR	LC	Z	PR
[175]	Predicts COVID-19 severity by comparing CXRs to previous ones	IC,LC	CV	PR
[176]	Addresses domain and label discrepancies in multi-dataset training	DA	C,TB,X	C,PR,X
[177]	Method to increase robustness of CNN classifiers to adversarial samples	AA,IG	PM	RP
[178]	Uses the features extracted from the training dataset to detect adversarial CXRs	AA	C	C
[179]	Self-supervision and adversarial training improves on transfer learning	AA	PM	PP
[180]	Claims 0.99 AUC for predicting TB, uses complex feature engineering and ensembling			
[181]	ResNet model trained with frontal and lateral images to predict COPD with PFT results			PR
[182]	One-class identification of viral pneumonia cases compared with binary classification			PR
[183]	A distributed learning method that overcomes problems of multi-institutional settings		C	C
[184]	Geometric deep learning including metadata with graph structure. Application to CXR		C	C
[185]	Proposes a new weighting scheme to impove abnormality classification		C	C
[186]	ResNet-34 used with various training settings for multi-label classification		C	C
[187]	Investigates effect of data augmentations on classification with Inception-Resnet-v2		C	C
[188]	Proposes a variational/generative architecture, demonstrates performance on CXRs		C	C
[189]	Evaluates the performance of an ensemble against many radiologists		C	C
[190]	Novel method for multi-label classification, application to CXR		C	C

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[191]	Defines a few-shot learning method by extracting features from autoencoders		C	C
[192]	Mean teacher inspired a probablistic graphical model with a novel loss		C	C
[193]	Examines the effect of denoising on pathology classification using DenseNet-121		C	C
[194]	Proposes integrating three attention mechanisms that work at different levels		C	C
[195]	Step-wise trained CNN and saliency-based autoeencoder for few shot learning		C	C,O
[196]	Uses CT and CXR reports with CXR images during training to diagnose unseen diseases		C	C,PR
[35]	Proposes a new dataset PadChest with multilabel labels and radiology reports		C	P
[197]	Lesion detection network used to improve image-level classification		C	PR
[198]	Method to produce confidence measure alongside probability, uses DenseNet-121		C,PL	C,PL
[199]	Uses self-supervised learning for pretraining, compares with ImageNet pretraining		C,PT	C,SI
[200]	Proposes a new CXR pre-training method, compares with pre-training on ImageNet		C,X	C,RP,X
[201]	Proposes a graph convolutional network framework which models disease dependencies		C,X	C,X
[202]	Compares several models for the detection of cardiomegaly		CM	C
[203]	Tests four off-the-shelf networks for prediction of cardiomegaly		CM	PR
[204]	Inception v3 trained to detect 4 abnormalities and compared with expert observers		CM,E,LO,Z	PR
[205]	GoogLeNet to classify normal and 5 abnormalities on a large proprietary dataset		CM,E,PE,PT,Z	PR
[206]	Compares the performance of deep learning with traditional feature extraction methods		CM,PE	PR
[207]	ImageNet pre-training and feature extraction methods for pathology detection		CM,PE,Z	PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[208]	An ensemble of DenseNet-121 networks used for COVID-19 classification		CV	C,PR
[209]	Investigates the value of soft tissue CXR for training DenseNet-121 for COVID-19		CV	C,PR,RP
[210]	Labels and predicts COVID-19 severity stage using CNN		CV	CC
[211]	Uses a model trained on COVID-19 cases to evaluate the effect of an imaging parameter		CV	PR
[212]	COVID-19 detection based on RT-PCR labels, evaluates an ensamble against radiologists		CV	PR
[213]	Ensemble of ResNet models for COVID-19 detection		CV	PR
[214]	Compares the performance of a DenseNet-121 ensemble to radiologists		CV,PM	PR
[215]	Various models and use of semi-supervised labels for edema severity estimation		E	M
[216]	Age prediction on PA or AP images using DenseNet-169		GA	C
[217]	Gender prediction using features from deep-learning models in traditional classifiers		GA	J,MO,O, PR,S
[218]	Age prediction on AP images using DenseNet-121 and ResNet-50		GA	X
[219]	Combines the CXR with age/sex/smoking history to predict the lung cancer risk		LC	PL
[220]	Densenet-121 pre-trained with public data used to identify 6 classes		LC,T,TB,Z	PR
[221]	Evaluates deep learning on pictures of CXRs captured with mobile phones		M,X	M,X
[222]	Ensemble of VGGNet and ResNet to detect various findings from AP CXRs		MN	C,M
[223]	Investigates the domain and label shift across publicly available CXR datasets		MN	C,M,O, P,RP,X
[224]	Explores the use of the lateral view CXR for classification of 64 different labels		MN,P	P
[225]	Classification of CXRs as Frontal or Lateral using GoogLeNet architecture		OR	PR
[226]	Assesses the effect of imprinted labels on AP/PA classification		OR	PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[227]	Distinguishes the CXR orientation, bone CXRs and soft tissue CXRs from dual energy		OR,Z	PR
[228]	Compares PA and Lateral images for pathology detection with DenseNet		P	P
[229]	Introduces a loss term that uses the label hierarchy to improve model performance		PL	PL
[230]	Trains VGG-16 on Ped-pneumonia dataset		PM	PP
[231]	Methods to mitigate imbalanced class sizes. Applied to CXR using ResNet-18		PM	PP
[232]	Evaluation of MobileNet to detect pneumonia on pediatric CXRs		PM	PP
[233]	Compares multiple architectures for pneumonia detection		PM	PP
[234]	Evaluates various capsule network architectures for pediatric pneumonia detection		PM	PP
[235]	Uses ResNet-50 to classify paediatric pneumonia		PM	PR
[236]	Compares traditional and generative data augmentation techniques on CXRs		PM	RP
[237]	Addresses catastrophic forgetting, application to pneumothorax detection using VGG-13		PT	C
[238]	Construction of large dataset, multiple architectures and hyperparameters optimized		PT	PR
[239]	Model pre-trained with public data and fine-tuned for pneumothorax detection		PT	PR
[240]	DenseNet-121 used to detect CXRs with acquisition-based defects		Q	C
[241]	GoogleNet combined with rule-based approach to determine the image quality		Q	PR
[242]	Detects abnormal CXRs using several models. Evaluates on independent private data		T	C
[243]	Defines a model on top of features extracted from Inception-ResNet-v2 for triaging		T	C
[244]	Collects features from pretrained models and adds a CNN on top for triaging		T	C,M
[245]	Studies the effect of various label noise levels on classification with DenseNet-121		T	C,PR,X

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[246]	Various models for detection of abnormal CXRs, effect of different training set sizes		T	PR
[247]	Defines 10 abnormalities to define a triaging model and uses CT based test labels		T	PR
[248]	Ensembe of DenseNet and EffecientNet for identification of normal CXR		T	PR
[249]	Examines the use of data augmentation in small data setting		T	PR
[250]	Evaluation of extra supervision in the form of localized region of interest		T	PR
[251]	Evalutates various models and ensembling methods for the triage task		T	RP
[252]	Evaluates deep learning approaches for tuberculosis detection		TB	BL,MO,PR,S
[253]	Evaluates the use of transfer learning for tuberculosis detection		TB	MO,PR,S
[254]	Extracts feaures using off-the-shelf models and trains a model using those		TB	MO,PR,S
[255]	Combines hand-crafted features and CNN for tuberculosis diagnosis		TB	MO,S
[256]	Evaluates a Bayesian-based CNN for detection of TB		TB	MO,S
[257]	Evaluates assisting clinicians with an AI based system to improve diagnosis of TB		TB	PR
[258]	Various architectures, inclusion of patient demographics in model considered		TB	PR
[259]	Addresses preservation of learned data, application to TB detection using ResNet-21		TB	PR
[260]	Pre-training using CXR pathology and metadata labels, application to TB detection		TB	S
[261]	Compares various models using various pretraining and ensembling strategies		TB	S
[262]	Evaluates models on detecting the position of feeding tube in abdominal and CXRs		TU	PR
[263]	Comparison of seven architectures and ensembling for detection of nine pathologies		X	X
[264]	A method to incorporate label dependencies and uncertainty data during classification		X	X

continued on the next page

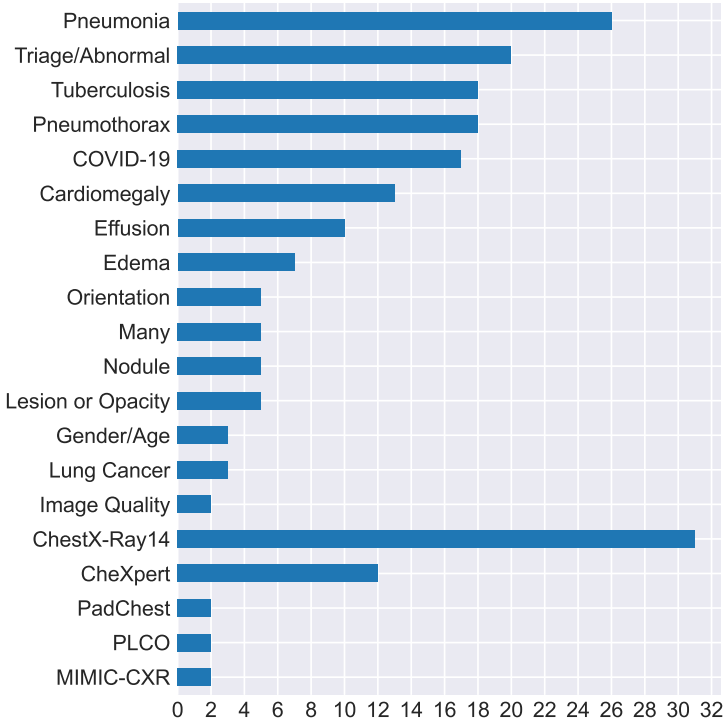


Figure 2.4: Number of studies for the Image-level Prediction labels. The studies that specifically work on a dataset and its labels are grouped together at the bottom. 188 papers are included, each may study more than one label.

predicts the disease severity, similarly [175] predicts the disease progression by comparing an exam with the previous exams of the patient, and [125] detects COVID-19 using a very limited amount of data. Other than these most common tasks, there are many studies using deep learning to make Image-level Predictions from CXRs. Other commonly utilized labels are illustrated in Figure 2.4 and listed in Table 2.2.

A large proportion of the studies use pre-trained standard architectures that can easily be found in deep learning libraries such as Tensorflow or Pytorch. These architectures are commonly Resnet [20], DenseNet [21], Inception [44], VGG [19], or AlexNet [17]. The choice of model depth (such as ResNet-18, ResNet-50, DenseNet-121, DenseNet-161) also varies between studies as there is no standard in this design choice. Most of those studies do not introduce methodological novelty but report or compare the performances of multiple architectures on a given task. For example, [266] compares various Resnet and Densenet models using both pretrained and randomly initialized weights on the performance of detecting the existence of foreign objects. Similarly many other studies compare the performance of different architectures with various depths on a given task, for example [162, 163, 167, 168]. Just like the model depth and architecture, there are many factors that affect the performance of a deep learning model. The effect of various data augmentation and input pre-processing methods

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[265]	Proposes self-training and student-teacher model for sample effeciency		X	X
[89]	Analyses the effect of label noise in training and test datasets		Z	C
[266]	Labels 6 different foreign object types and detects using various architectures		Z	M
[81]	Evaluates the use of CXRs to predict long term mortality using Inception-v4		Z	PL
[267]	Low-res segmentation is used to crop high-res lung areas and predict pneumoconiosis		Z	PR
[268]	Pneumoconiosis prediction with DenseNet-121 and SVMs applied to extracted features		Z	PR
[269]	Detection of coronary artery calcification using various CNN architectures		Z	PR
[270]	ResNet-50 for detection of the presence of elevated pulmonary arterial wedge pressure		Z	PR
[271]	A network is designed to identify subjects with elevated pulmonary artery pressure		Z	PR,RP

end of table

are evaluated by [142, 187]. The effect of increasing or decreasing the image size are evaluated in [146, 186]. Various pre-training schemes are evaluated by [146, 260]. More sophisticated pre-processing steps to improve model performance include bone suppression [43, 275] and lung cropping [91].
Some studies bring methodological novelty by making use of methods that are known to work well to improve model performance elsewhere. For example, it is known that an ensemble of many models improves performance compared to a single model [276]. Some studies that make use of this method are [189, 214, 251, 261]. Attention mining (or object-region mining, attention-based) models are also found in the literature [277]. Those models aim to improve performance and add localization capabilities to an image-level prediction model. Some studies making use of attention mining models are [145, 147]. Multiple-instance learning (multi-instance learning or MIL) [278] is another method that is used to add localization capabilities to image-level prediction models. MIL breaks the input image into smaller parts (instances), makes individual predictions relating to those instances and combines this information to make a prediction for the whole image. Some studies that make use of MIL are [152, 161]. Other topics within the literature include model uncertainty [198, 256], quality of the CXR [95, 121, 121, 240, 241] and defence against adversarial attack [177–179].
The different properties of datasets are also utilized to improve model capabilities or performance. Many of the public datasets make use of labels that are not mutually exclusive. This has resulted in a number of papers addressing the dependencies among abnormality labels [158, 168, 264]. Since many of the labels are common between datasets from different institutes there has been investigation of the issues related to domain and/or label shift in images from different sources [176, 223]. The effect of

dataset sizes is evaluated by [246]. Semi-supervised learning methods combine a small set of labeled and a large set of unlabeled data to train a model [119, 120, 192, 215].
Most of the studies working on image-level prediction tasks deal with frontal CXR images. The importance of lateral chest X-rays and models that can deal with multiple views are evaluated in [224, 228, 279].

2.4.1 Segmentation

Segmentation is one of the most commonly studied subjects in CXR analysis (58 papers) and includes literature focused on the identification of anatomy, foreign objects or abnormalities. The segmentation literature reviewed for this work is detailed fully in Table 2.3. Anatomical segmentation of the heart, lungs, clavicles or ribs, on chest radiographs, is a core part of many computer aided detection (CAD) pipelines. It is typically used as an initial step of such pipelines to define the region of interest for subsequent image analysis tasks to improve performance and efficiency [43, 91, 128, 132, 133, 258]. Further, the segmentation itself can be useful to quantify clinical parameters based on shape or area measurements. For example, cardiothoracic ratio, a clinically used measurement to assess heart enlargement (cardiomegaly), can be directly calculated from heart and lung segmentations [92, 94]. Organ segmentation has, for these reasons, become one of the most commonly studied subjects among CXR segmentation tasks as seen in Figure 2.5.
Another application found in the CXR literature is foreign object segmentation, i.e. catheter, tubes, lines, for which high performance levels have been reported using deep learning [294, 295, 327]. Interestingly, only a small number of works addressed segmentation of abnormalities. [97] focused on segmentation of pneumonia, and [323] developed a method to segment pneumothorax. Both of these works used recently published challenge datasets (hosted by Kaggle), namely RSNA-Pneumonia and SIIM-ACR. In general, the determination of abnormal locations on CXR is dominated by methods which addressed this as a localization task (i.e. via bounding-box type annotations) rather than exact delineation of abnormalities through segmentation. This is likely to be attributable to the difficulty of precise annotation on a projection image and to the high annotation cost for precise segmentations.
A small number of works tackled the segmentation task using a patch-based CNN, which is trained to classify the center of pixel in the patch as foreground or background by means of sliding-window approach [313, 315]. However, this approach is generally considered inefficient for segmentation and most works use fully convolutional networks (FCN) [48], which can take larger, arbitrary sized, images as input and produce a similar sized, per-pixel prediction, likelihood map in a single forward pass. In particular, the U-Net architecture [49], a type of FCN, dominates the field with 50% of segmentation works in literature (29/58) employing it or some similar variant. Successful applications were built with this architecture to segment organs [299, 319, 320], pneumonia [97] and foreign objects [294, 295]. For example, [299] compared three U-Net variant architectures for multi-class segmentation of the heart, clavicles and lungs on the JSRT dataset. Using regularization to prevent over-fitting and weighted cross entropy loss to balance the dataset, they outperformed the human observer at heart and lung segmentation. This result was in line with other works [292, 298, 301] employing FCN-type architectures which also achieved very high performance levels on this dataset.
One commonly encountered challenge is that many algorithms produce noisy segmentation maps. In order to tackle this, several works employed post-processing techniques. [327] used a probabilistic Hough line transform algorithm to remove false positives and produce a smoother segmentation of peripherally inserted central catheters (PICC). [324] used a heuristic approach to average cross-fold predictions with an optimized binarization threshold and a dilation technique for pneumothorax seg-

Table 2.3: Segmentation Studies (Section 2.4.1).
Tasks: DA=Domain Adaptation, IG=Image Generation, IL=Image-level Predictions, LC=Localization, PR=Preprocessing, WS=Weak Supervision. **Bold font** in tasks implies that this additional task is central to the work and the study also appears in another table in this paper.
Labels: C=ChestX-Ray14, CL=Clavicle, CM=Cardiomegaly, CV=COVID, E=Edema, H=Heart, L=Lung, LO=Lesion or Opacity, PE=Effusion, PM=Pneumonia, PT=Pneumothorax, R=Rib, TU=Catheter or Tube, Z=Other.
Datasets: BL=Belarus, C=ChestX-ray14, J=JSRT+SCR, M=MIMIC-CXR, MO=Montgomery, O=Open-i, PP=Ped-pneumonia, PR=Private, RP=RSNA-Pneumonia, S=Shenzen, SI=SIIM-ACR, SM=Simulated CXR from CT.

Citation	Method	Other Tasks	Labels	Datasets
[280]	A model based on U-Net and Faster R-CNN to detect PICC catether and its tip	**LC**,PR	TU	PR
[281]	Tailored Mask R-CNN for simultaneous detection and segmentation	**LC**	L	PR
[282]	Uses Mask R-CNN iteratively to segment and detect ribs.	**LC**	R	PR
[91]	Combines lung cropped CXR model and a CXR model to improve model performance	**IL**,LC,PR	C,L	C,J
[92]	Comparison of image-level prediction and segmentation models for cardiomegaly	**IL**	CM	C
[93]	A network with DenseNet and U-Net for classification of cardiomegaly	**IL**	CM	C
[94]	U-Net based model for heart and lung segmentation for cardiothoracic ratio	**IL**	CM	PR
[95]	Combines lung cropped CXR model and a CXR model using the segmentation quality	**IL**	E,LO,PE, PT,Z	M
[96]	Pneumonia detection is improved by use of lung segmentation	**IL**	PM	J,MO,PP, PR
[97]	U-Net based model to segment pneumonia	**IL**	PM	RP
[98]	Multi-scale DenseNet based model for pneumothorax segmentation	**IL**	PT	PR
[99]	DenseNet based U-Net for segmentation of the left and right humerus of the infant	**IL**	Z	PR
[283]	Attention-based network and CXR synthesis process for data augmentation	**IG**,IG	L	J,MO,PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[284]	Conditional GANs for multi-class segmentation of heart,clavicles and lungs	**IG**	CL,H,L	J
[285]	Processing method to produce scatter-corrected CXRs and segments masses with U-Net	**IG**	LO	SM
[286]	MUNIT based DA model for lung segmentation	**DA**	CL,H,L	J
[287]	Adversarial training of lung and heart segmentation for DA	**DA**	CM	J,PR
[288]	CycleGAN guided by a segmentation module to convert CXR to CT projection images	**DA**	H,L,Z	PR
[289]	CycleGAN based DA model with semantic aware loss for lung segmentation	**DA**	L	MO
[290]	Conditional GANs based DA for bone segmentation	**DA**	R	SM
[291]	FCN based novel model incorporating weak landmarks and bounding boxes annotations	WS	CL,H,L	J
[292]	U-Net segmentation model integrating unlabeled data through consistency loss	WS	CL,H,L	J
[293]	Attention masks derived from classification model to guide the segmentation model	IL	PT	PR
[294]	U-Net based network for classification and segmentation with simulated data	IL	Z	C,J
[295]	U-Net based model for segmentation and a classification for existance of lines	IL	Z	PR
[296]	U-Net for bone suppression given lung-segmented CXR image with patches			PR
[297]	Proposes teacher-student based learning with noisy segmentations		CL,H,L	J
[298]	Various FCN based models explored for simultaneous pixel and contour segmentation		CL,H,L	J
[299]	Investigates various FCN type architecture including U-Net for organ segmentation		CL,H,L	J
[300]	Capsule networks adapted for multi-class organ segmentation		CL,H,L	J
[301]	U-Net based architecture with residual connections for organ segmentation		CL,H,L	J,MO,S

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[302]	U-Net based architecture based on dense connections		CL,R	PR
[303]	CNN trained with CT projection images for quantification of airspace disease		CV	PR
[304]	Denoising autoencoder as post-processing to improve segmentations		H,L	J
[305]	Evaluates U-Net performance with various loss functions, and data augmentation		H,L	PR
[306]	Stacked denoising autoencoder model for space and shape parameter estimation		L	BL,J,PR
[307]	Investigates the effect of fine-tuning different layers for U-Net based model		L	J
[308]	Proposes a human-in-the-loop one shot anatomy segmentor		L	J
[309]	U-Net with conditional random field post processing for lung segmentation		L	J
[310]	Investigates U-Net with different optimizer and dropout		L	J
[311]	U-Net with dense connections for reducing network parameters for lung segmentation		L	J,MO
[312]	U-Net with self attention for lung segmentation		L	J,MO,S
[313]	Multi-scale and patch-based CNN to segment lungs		L	J,PR
[314]	U-Net based model for lung segmentation trained with CXR patches		L	MO
[315]	Two stage patch based CNN for refined lung field segmentation		L	MO
[316]	Encoder-decoder architecture with ConvLSTM and ResNet for segmentation		L	MO
[317]	Encoder-decoder based CNN with novel edge guidance module for lung segmentation		L	MO
[318]	Proposes a convolutional LSTM model for ultrasound, uses CXR as a secondary modality		L	MO
[319]	U-Net based segmentation model for dynamic CXRs		L	PR
[320]	U-Net for whole lung region segmentation including where heart overlaps		L	PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[321]	Cascaded U-net with sample selection with imperfect segmentations		L	S
[322]	ResNet-50 based architecture with segmentation and classification branches		PT	SI
[323]	Investigates U-Net based models with various backbone encoders for pneumothorax		PT	SI
[324]	Ensemble of three LinkNet based networks and with multi-step postprocessing		PT	SI
[325]	Cascaded network with Faster R-CNN and U-Net for aortic knuckle		Z	J
[326]	Multi-scale U-Net based model with recurrent module for foreign objects		Z	O
[327]	Two FCN to segment peripherally inserted central catheter line and its tip		Z	PR
[328]	Two Mask R-CNN to segment the spine and vertebral bodies and calculate the Cobb angle		Z	PR

end of table

mentation. Some authors proposed to learn post-processing by training an independent network, inputting segmentation predictions for refinement, rather than using conventional methods. For example, [304] used denoising autoencoders, trained to produce anatomically plausible segmentations from the initial predictions. Similarly, [315] used a FCN to refine segmentation predictions. The final segmentation was achieved by combining the initial and reconstructed segmentation results.

A number of researchers used a multi-stage training strategy, where network predictions are refined in several steps during training [282, 315, 321, 325]. For example, [325] employed faster-RCNN to produce coarse segmentation results, which were then used to crop the images to a region of interest, which was provided to a U-Net trained to predict the final segmentation result. Similarly, [315] employed two networks, where the second network received the predictions of the first to refine the segmentation results. [282] trained separate networks for segmentation of each rib in chest radiographs based on Mask R-CNN. The predicted segmentation results from the rib above was fed to each network as an additional input.

Although most of the works in the literature harnessed FCN architectures, a few authors employed recurrent neural networks (RNN) for segmentation tasks [316, 318, 326] and report good performance. [316] proposed a novel architecture where the decoding component was long short term memory (LSTM) architecture to obtain multi-scale feature integration. The proposed approach achieved a Dice score of 0.97 for lung segmentation on Montgomery dataset. Similarly, [67] developed a scale RNN, a network based on encoder and decoder architecture with recurrent modules, for segmentation of catheter and tubes on pediatric chest X-rays.

The high cost of obtaining segmentation annotations motivates the development of segmentation

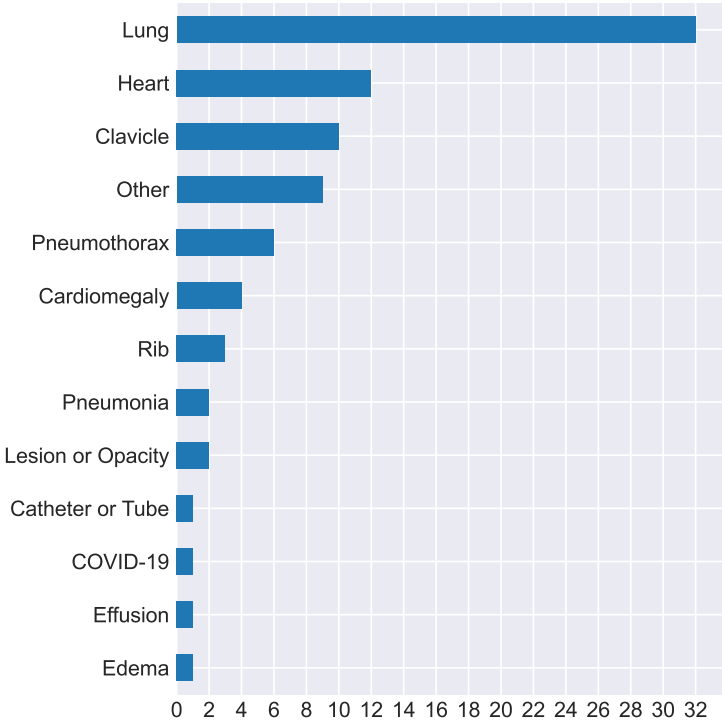


Figure 2.5: Number of studies for the Segmentation labels. 58 papers are included, each may study more than one label.

systems which incorporate weak-labels or simulated datasets with the aim of reducing annotation costs [293, 294, 308, 326]. Several works addressed this using weakly supervised learning approaches [293, 308]. [308] proposed a graph convolutional network based architecture which required only one labeled image and leveraged large amounts of unlabeled data (one-shot learning) through a newly introduced three contour-based loss function. [293] proposed a pneumothorax segmentation framework which incorporated both images with pixel level annotations and weak image-level annotations. The authors trained an image classification network, ResNet-101, with weakly labeled data to derive attention maps. These attention maps were then used to train a segmentation model, Tiramisu, together with pixel level annotations.

2.4.2 Localization

Localization refers to the identification of a region of interest using a bounding box or point coordinates rather than a more specific pixel segmentation. In this section we discuss only the CXR localization literature which provides a quantitative evaluation of this task. It should be noted that there are many other works which train networks for an image-level prediction task and provide some examples of heatmaps (e.g., saliency map or GradCAM) to suggest which region of the image determines the label. While this may be considered as a form of localization, these heatmaps are rarely quanti-

tatively evaluated and such works are not included here. Table 2.4 details all the reviewed studies where localization was a primary focus of the work.
The majority of CXR analysis papers performing localization focus on identifying abnormalities rather than objects (e.g., catheter) or anatomy (e.g., ribs). Localization of nodules, tuberculosis and pneumonia are commonly studied applications in the literature, as illustrated in Figure 2.6.

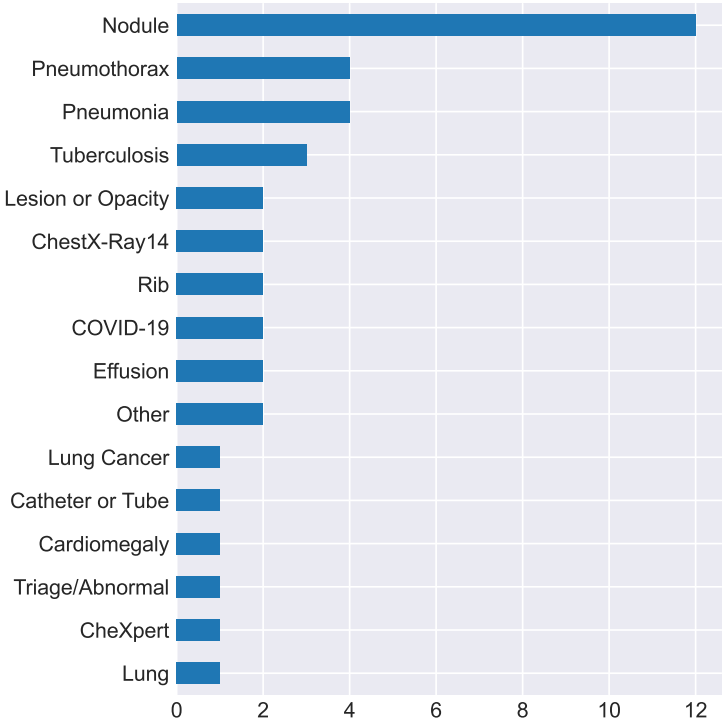


Figure 2.6: Number of studies for the Localization labels. 30 papers are included, each may study more than one label.

In recent years, a variety of specific architectures, i.e. YOLO, Mask R-CNN, Faster R-CNN, have been designed in computer vision research aiming at developing more accurate and faster algorithms for localization tasks [345]. Such state of the art architectures have been rapidly adapted for CXR analysis and shown to achieve high-level performance. For example, [341] demonstrated that the (original) YOLO architecture was successful at identifying the location of pneumothorax on chest radiographs. The model was evaluated on an external dataset with CXRs from 1,319 patients which were obtained after percutaneous transthoracic needle biopsy (PTNB) for pulmonary lesions; it achieved an AUC of 0.898 and 0.905 on 3-h and 1-day follow-up chest radiographs, respectively. Similarly, other studies [330, 334, 336, 338] harnessed architectures like RetinaNet, Mask R-CNN and RCNN for localization of nodules and masses. [334] trained RetinaNet and Mask R-CNN for detection of nodule and mass and investigated the optimal input size. The authors showed that, using a square image with 896 pixels as the edge length, RetinaNet and Mask R-CNN achieved FROC of 0.906 and 0.869, respectively.

Table 2.4: Localization Studies (Section 2.4.2).
Tasks: IC=Interval Change, IL=Image-level Predictions, PR=Preprocessing, RP=Report Parsing, SE=Segmentation, WS=Weak Supervision. **Bold font** in tasks implies that this additional task is central to the work and the study also appears in another table in this paper.
Labels: C=ChestX-Ray14, CM=Cardiomegaly, CV=COVID, L=Lung, LC=Lung Cancer, LO=Lesion or Opacity, ND=Nodule, PE=Effusion, PM=Pneumonia, PT=Pneumothorax, R=Rib, T=Triage/Abnormal, TB=Tuberculosis, TU=Catheter or Tube, X=CheXpert, Z=Other.
Datasets: C=ChestX-ray14, CC=COVID-CXR, J=JSRT+SCR, M=MIMIC-CXR, O=Open-i, PP=Ped-pneumonia, PR=Private, RP=RSNA-Pneumonia, S=Shenzen, X=CheXpert.

Citation	Method	Other Tasks	Labels	Datasets
[280]	A model based on U-Net and Faster R-CNN to detect PICC catether and its tip	**SE**,PR	TU	PR
[281]	Tailored Mask R-CNN for simultaneous detection and segmentation	**SE**	L	PR
[282]	Uses Mask R-CNN iteratively to segment and detect ribs.	**SE**	R	PR
[101]	Uses activation and gradient based attention for localization and classification	**IL**	C,X	C
[102]	Detects and localizes COVID-19 using various networks and ensembling	**IL**	CV	C,CC,PP, RP,X
[103]	GoogleNet trained with CXR patches, correlates with COVID-19 severity score	**IL**	CV,PM	C,PR
[104]	Proposes a segmentation and classification model compares with radiologist cohort	**IL**	LO,ND,PE, PT	PR
[105]	Trains a semisupervised network on a large CXR dataset with CT-confirmed nodule cases	**IL**	ND	PR
[106]	Defines a loss that minimizes the saliency map errors to improve model performance	**IL**	ND	PR
[107]	A weakly supervised localization with variational model, leverages attention maps	**IL**	PM	C
[108]	Attention guided CNN for pneumonia detection with bounding boxes	**IL**	PM	RP
[109]	A CNN for identification of abnormal CXRs and localization of abnormalities	**IL**	T	PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[110]	Introduces a visualization method to identify regions of interest from classification	**IL**	TB	PR
[111]	Weakly supervised framework jointly trained with localization and classification	**IL**	TB	PR
[329]	Extract nodule candidates using traditional methods and trains GoogleNet	SE,PR	ND	J
[330]	RetinaNet for detecting nodules incorporating lung segmentation	SE	ND	J,PR
[331]	Combines reports and CXRs for weakly supervised localization and classification	RP,WS	PM,PT	C,M
[332]	Proposes a model using LSTM and CNN, combining reports and images as inputs	IL,RP	CM,ND	C,O
[333]	Adversarially trained weakly supervised localization framework for interpretability	IL	C	C
[334]	Evaluates the effect of image size for nodule detection with Mask R-CNN and RetinaNet	IL	ND	PR
[335]	Evaluates the reproducibility of YOLO for disease localization in follow up exams	IC	LO,ND,PE, PT,Z	PR
[336]	Evaluates the reproducability of various detection architectures in follow up exams.	IC	ND	PR
[337]	ResNet model using CT and surgery based annotations for lung cancer prediction		LC	PR
[338]	R-CNN for localization of lung nodules		ND	J
[339]	Fuses AlexNet and hand-crafted features to improve random forest performance		ND	J
[340]	Patch-based nodule detectin, combines features from different resolutions		ND	J,PR
[341]	Evaluates the detection of pneumothorax before, 3h and 1d after biopsy		PT	PR
[342]	Proposes a U-Net based model for localizing and labeling individual ribs		R	O
[343]	AlexNet for localizing tuberculosis with patch-based approach		TB	S
[344]	Localizes anatomical features for image quality check		Z	PR

end of table

A number of papers adapted classification architectures (e.g., ResNet, DenseNet) to directly regress landmark locations for CXR localization tasks [109, 337]. One common way of tackling this is to adapt the networks to produce heatmap predictions and draw boxes around the areas that created the highest signals. For example, [109] tailored a DenseNet-based classifier to produce heatmap predictions for each of four types of CXR abnormalities. The network was trained with pixel-wise cross entropy between the predictions and annotations. Similarly, [337] adapted ResNet-50 and ResNet-101 architectures for localization of nodules and masses on CXR. Other studies [340, 343] tackled this problem using patch-based approaches, commonly referred as multiple instance learning, creating patches from chest X-rays and evaluating these for the presence of abnormalities.

One challenge in building robust deep learning localization systems is to collect large annotated datasets. Collecting such annotations is time-consuming and costly which has motivated researchers to build systems incorporating weaker labels during training. This research area is referred to as weakly supervised learning, and has been investigated by numerous works [105–107, 109, 111] for localization of a variety of abnormalities in CXR. Most of the works [105, 106, 109, 111] leveraged weak image-level labels by adapting a CNN architecture to create two branches for localization (heatmap predictions) and classification. A hybrid loss function was used, combining localization and classification losses, which enabled training of the networks using images without localization annotations.

2.4.3 Image Generation

There are 35 studies identified in this work whose main focus is Image Generation, as detailed in Table 2.5. Image generation techniques have been harnessed for a wide variety of purposes including data augmentation [346], visualization [113, 347], abnormality detection through reconstruction [114, 115], domain adaptation [288] or image enhancement techniques [348].

The generative adversarial network (GAN) [58, 67] has became the method of choice for image generation in CXR and over 50% of the works reviewed here used GAN-based models.

A number of works focused on CXR generation to augment training datasets [346, 353, 365] by using unconditional GANs which synthesize images from random noise. For example, [346] trained a DCGAN model, similar to [365], independently for each class, to generate chest radiographs with five different abnormalities. The authors demonstrated that this augmentation process improved the abnormality classification performance of DCNN classifiers (ResNet, GoogleNet, AlexNet) by balancing the dataset classes. Another work [353] proposed a novel GAN architecture to improve the quality of generated CXR by forcing the generator to learn different image representations. The authors proposed SkrGAN, where a sketch prior constraint is introduced by decomposing the generator into two modules for generating a sketched structural representation and the CXR image, respectively.

Abnormality detection is another task which has been addressed through a combination of image generation and one-class learning methods [115, 116]. The underlying idea of these methods is that a generative model trained to reconstruct healthy images will have a high reconstruction error if abnormal images are input at test time, allowing them to be identified. [115] harnessed GANs and employed a U-Net type autoencoder to reconstruct images (as the generator), and a CNN-based discriminator and encoder. The discriminator received both reconstructed images and real images to provide supervisory signal for realistic reconstruction through adversarial training. Similarly, [116] proposed an autoencoder for abnormality detection which was trained only with healthy images. In this case the autoencoder was tailored to not only reconstruct healthy images but also produce uncertainty predictions. By leveraging uncertainty, the authors proposed a normalized reconstruction error to distinguish abnormal CXR images from normal ones.

Table 2.5: Image Generation Studies (Section 2.4.3).
Tasks: DA=Domain Adaptation, IC=Interval Change, IG=Image Generation, IL=Image-level Predictions, LC=Localization, PR=Preprocessing, RE=Registration, SE=Segmentation, SR=Super Resolution. **Bold font** in tasks implies that this additional task is central to the work and the study also appears in another table in this paper.
Labels: BS=Bone Suppression, C=ChestX-Ray14, CL=Clavicle, CM=Cardiomegaly, CV=COVID, E=Edema, H=Heart, L=Lung, LO=Lesion or Opacity, PE=Effusion, PT=Pneumothorax, T=Triage/Abnormal, TB=Tuberculosis, Z=Other.
Datasets: C=ChestX-ray14, CC=COVID-CXR, J=JSRT+SCR, MO=Montgomery, O=Open-i, PL=PLCO, PP=Ped-pneumonia, PR=Private, RP=RSNA-Pneumonia, S=Shenzen, SM=Simulated CXR from CT, X=CheXpert.

Citation	Method	Other Tasks	Labels	Datasets
[283]	Attention-based network and CXR synthesis process for data augmentation	**SE**,IG	L	J,MO,PR
[284]	Conditional GANs for multi-class segmentation of heart,clavicles and lungs	**SE**	CL,H,L	J
[285]	Processing method to produce scatter-corrected CXRs and segments masses with U-Net	**SE**	LO	SM
[112]	Combines classification loss and autoencoder reconstruction loss	**IL**,SE	T	J,MO,O, S
[113]	Wasserstein GAN to permute diseased radiographs to appear healthy	**IL**,LC	Z	PR
[114]	Novel GAN model trained with healthy and abnormal CXR to predict difference map	**IL**	PE	SM,X
[115]	GANs with U-Net autoencoder and CNN discriminator and encoder for one-class learning	**IL**	T	C
[116]	Autoencoder uses uncertainty for reconstruction error in one-class learning setting	**IL**	T	PP,RP
[349]	Conditional GAN based DA for image registration using segmentation guidance	**DA**,RE,SE	L	C
[350]	Adversarial based method adapting new domains for abnormality classification	**DA**,IL	CM	PL
[351]	Proposes a patch-based CNN super resolution method	SR	Z	J

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[352]	Generates high resolution CXRs using multi-scale, patch based GANs	SR	Z	O
[353]	Novel GAN model with sketch guidance module for high resolution CXR generation	SR	Z	PP
[354]	AutoEncoder for bone suppression and segmentation with statistical similarity losses	SE,PR	BS	J
[355]	Uses neural architecture search to find a discriminator network for GANs	SE	H,L	J,PR
[356]	Proposes an iterative gradient based input preprocessing for improved performance	SE	L	S
[357]	Learns transformations to register two CXRs, uses the difference for interval change	RE,IC	Z	PR
[358]	Generates bone and soft tissue (dual energy) images from CXRs	PR	BS	PR
[359]	Proposes an CNN with multi-resolution decomposition for bone suppression images	PR	BS	PR
[360]	U-Net for bone generation with CT projection images, used for CXR enhancement	PR	BS	SM
[348]	U-Net based network to generate dual energy CXR	PR	Z	PR
[361]	GAN integrates edges of ribs and clavicles to guide DES-like images generation	PR	Z	PR
[362]	Generates diseased CXRs, evaluates their realness with radiologists and trains models	LC	C	C
[363]	Novel CycleGAN model to decompose CXR images incorporating CT projection images	IL	C	C,PR,SM
[346]	Uses DCGAN model to generate CXR with abnormalities for data augmentation	IL	CM,E,PE, PT	PR
[364]	U-Net based architecture to decompose CXR structures, application to TB detection	IL	TB	PR
[365]	Two DCGAN trained with normal and abnormal images for data augmentation	IL	Z	PL
[347]	Novel conditional GAN using lung function test results to visualize COPD progression	IL	Z	PR
[366]	Conditional GAN and two variational autoencoders designed for CXR generation			PR
[367]	Novel reconstruction algorithm for CXR enhancement			PR

continued on the next page

continued from the previous page

Citation	Method	Other Tasks	Labels	Datasets
[275]	Bone shadow suppression using conditional GANS with dilated U-Net variant		BS	J
[368]	Generates CXRs from CT to train CNN for bone suppression		BS	PR
[369]	Generates COVID-19 CXR images to improve network training and performance		CV	CC,RP
[279]	2D-to-3D encoder-decoder network for generating 3D spine models from CXR studies		Z	PR
[370]	Generates normal from abnormal CXRs, uses the deformations as disease evidence		Z	PR

end of table

The most widely studied subject in the image generation literature is image enhancement. Several researchers investigated bone suppression [275, 354, 359–361, 368] and lung enhancement [360, 363] techniques to improve image interpretability. A number of works [275, 361] employed GANs to generate bone-suppressed images. For example, [361] employed GANs and leveraged additional input to the generator to guide the dual-energy subtraction (DES) soft-tissue image generation process. In this study, bones, edges and clavicles were first segmented by a CNN model, and the resulting edge maps were fed to the generator with the original CXR image as prior knowledge. For building a deep learning model for bone suppressed CXR generation, the paired dual energy (DE) imaging is needed, which is not always available in abundance. Several other studies [360, 363] addressed this by leveraging digitally reconstructed radiographs for enhancing the lungs and bones in CXR. For instance, [363] trained an autoencoder for generating CXR with bone suppression and lung enhancement, and the knowledge obtained from DRR images were integrated through the encoder.

2.4.4 Domain Adaptation

Most of the papers surveyed in this work train and test their method on data from the same domain. This finding is inline with the previously reported studies [336, 371] and highlights an important concern: most of the performance levels reported in the literature might not generalize well to data from other domains [372]. Several studies [223, 372, 373] demonstrated that there was a significant drop in performance when deep learning systems were tested on datasets outside their training domain for a variety of CXR applications. For example, [373] investigated the performance of a DenseNet model for abnormality classification on CXR images using 10 diverse datasets varied by their location and patient distributions. The authors empirically demonstrated that there was a substantial drop in performance when a model was trained on a single dataset and tested on the other domains. [372] observed a similar finding for pneumonia detection on chest radiographs.

Domain adaptation (DA) methods investigate how to improve the performance of a model on a dataset from a different domain than the training set. In CXR analysis, DA methods have been investi-

gated in three main settings; adaptation of CXR images acquired from different hardware, adaptation of pediatric to adult CXR and adaptation of digitally reconstructed radiographs (generated by average intensity projections from CT) to real CXR images. All domain adaptation studies, and studies on generalization reviewed in this work are detailed in Table 2.6.

Table 2.6: Domain Adaptation Studies (Section 2.4.4).
Tasks: IG=Image Generation, IL=Image-level Predictions, RE=Registration, SE=Segmentation. **Bold font** in tasks implies that this additional task is central to the work and the study also appears in another table in this paper.
Labels: C=ChestX-Ray14, CL=Clavicle, CM=Cardiomegaly, H=Heart, L=Lung, M=MIMIC-CXR, PM=Pneumonia, R=Rib, TB=Tuberculosis, Z=Other.
Datasets: C=ChestX-ray14, J=JSRT+SCR, M=MIMIC-CXR, MO=Montgomery, O=Open-i, PL=PLCO, PP=Ped-pneumonia, PR=Private, RP=RSNA-Pneumonia, S=Shenzen, SM=Simulated CXR from CT.

Citation	Method	Other Tasks	Labels	Datasets
[286]	MUNIT based DA model for lung segmentation	**SE**	CL,H,L	J
[287]	Adversarial training of lung and heart segmentation for DA	**SE**	CM	J,PR
[288]	CycleGAN guided by a segmentation module to convert CXR to CT projection images	**SE**	H,L,Z	PR
[289]	CycleGAN based DA model with semantic aware loss for lung segmentation	**SE**	L	MO
[290]	Conditional GANs based DA for bone segmentation	**SE**	R	SM
[117]	Continual learning methods to classify data from new domains	**IL**	C,M	C,M
[118]	CycleGAN model to adapt adult to pediatric CXR for pneumonia classification	**IL**	PM	PP,RP
[349]	Conditional GAN based DA for image registration using segmentation guidance	**IG**,RE,SE	L	C
[350]	Adversarial based method adapting new domains for abnormality classification	**IG**,IL	CM	PL
[372]	Assessment of generalization to data from different institutes	IL	PM	C,O
[374]	Demonstrates the effect of training and test on data from different domains	IL	TB	S

Most of the research on DA for CXR analysis harnessed adversarial-based DA methods, which either use generative models (e.g., CycleGANs) or non-generative models to adapt to new domains using a variety of different approaches. For example, [287] investigated an unsupervised domain adaptation based on adversarial training for lung and heart segmentation. In this approach, a discriminator network, ResNet, learned to discriminate between segmentation predictions (heart and

lung) from the target domain and reference standard segmentations from the source domain. This approach forced the FCN-based segmentation network to learn domain invariant features and produce realistic segmentation maps. A number of works [288, 289, 375] addressed unsupervised DA using CycleGAN-based models to transform source images to resemble those from the target domain. For example, [288] used a CycleGAN-based architecture to adapt CXR images to digitally reconstructed radiographs (DRR) (generated from CT scans), for anatomy segmentation in CXR. A CycleGAN-based model was employed to convert the CXR image appearance and a U-Net variant architecture to simultaneously segment organs of interest. Similarly, CycleGAN-based models were adapted to transfer DRR images to resemble CXR images for bone segmentation [290] and to transform adult CXR to pediatric CXR for pneumonia classification [115].

Unlike most of the studies which utilized DA methods in unsupervised setting, a few studies considered supervised and semi-supervised approaches to adapt to the target domain. [286] employed a MUNIT-based architecture [376] to map target images to resemble source images, subsequently feeding the transformed images to the segmentation model. The authors investigated both unsupervised and semi-supervised approaches in this work, where some labels from the target domain were available. Another work by [117] studied several recently proposed continual learning approaches, namely joint training, elastic weight consolidation and learning without forgetting, to improve the performance on a target domain and to mitigate effectively catastrophic forgetting for the source domain. The authors evaluated these methods for 2 publicly available datasets, ChestX-ray14 and MIMIC-CXR, for a multi-class abnormality classification task and demonstrated that joint training achieved the best performance.

2.4.5 Other Applications

In this section we review articles with a primary application that does not fit into any of the categories detailed in Sections 2.4 to 2.4.4 (14 studies). These works are detailed fully in Table 2.7.

Image retrieval is a task investigated by a number of authors [100, 385–390]. The aim of image retrieval tools is to search an image archive to find cases similar to a particular index image. Such algorithms are envisaged as a tool for radiologists in their daily workflow. [387] proposed a ranked feature extraction and hashing model, while [390] proposed to use saliency maps as a similarity measure.

Another task that did not belong to previously defined categories is out-of-distribution detection. Studies working on this [382–384] aim to verify whether a test sample belongs to the distribution of the training dataset as model performance is otherwise expected to be sub-optimal. [384] propose using the training dataset statistics on different layers of a deep learning model and applying Mahalanobis distance to see the distance of a sample from the training dataset. [383] approach the problem differently and train an unsupervised autoencoder. Later they use the feature encodings extracted from CXRs to define a database of known encodings and compare new samples to this database.

Report generation is another task which has attracted interest in deep learning for CXR [377–380]. These studies aim to partially automate the radiology workflow by evaluating the chest X-ray and producing a text radiology report. For example, [377] first determines the findings to be reported and then makes use of a large dataset of existing reports to find a similar case. This case report is then customized to produce the final output.

One other task of interest is image registration [381]. This task aims to find the geometric transformation to convert a CXR so that it anatomically aligns with another CXR image or a statistically defined shape. The clinical goal of this task is typically to illustrate interval change between two images. De-

Table 2.7: Other Studies (Section 2.4.5).
Tasks: IL=Image-level Predictions, IR=Image Retrieval, OD=Out-of-Distribution, RE=Registration, RG=Report Generation, RP=Report Parsing. **Bold font** in tasks implies that this additional task is central to the work and the study also appears in another table in this paper.
Labels: C=ChestX-Ray14, H=Heart, L=Lung, Q=Image Quality, T=Triage/Abnormal, TB=Tuberculosis, X=CheXpert, Z=Other.
Datasets: C=ChestX-ray14, J=JSRT+SCR, M=MIMIC-CXR, MO=Montgomery, O=Open-i, PR=Private, S=Shenzen, X=CheXpert.

Citation	Method	Tasks	Labels	Datasets
[100]	Uses a database of the intermediate ResNet-50 features to find similar studies	**IL**,IR	TB	MO,S
[377]	Generate reports by classifying CXRs, and finding and modifying similar reports	RG,RP	Z	C,M
[378]	Extracts features from Chest X-rays and uses another network to write reports.	RG,IL	C	C,O
[379]	Generates radiology reports by training on classification labels and report text	RG,IL	Z	O,X
[380]	A novel recurrent generation network with attention mechanism	RG	Z	O
[381]	Anatomical priors to improve deep learning based image registration	RE	H,L	J,MO,S
[382]	Proposes a method to reject out-of-distribution images during test time	OD,IL	Z	C
[383]	Proposes to detect anomalies based on a dataset of autoencoder features	OD	Q,T	C
[384]	Mahalanobis distance on network layers to detect out-of-distribution samples	OD	Z	C
[385]	Compares the extracted feature and classification similarities for ranking	IR		PR
[386]	Uses extracted features to cluster similarly labeled CXRs across datasets	IR	C,X	C,X
[387]	Proposes a learnable hash to retrieve CXRs with similar pathologies	IR	Z	C
[388]	Residual network to retrieve images with similar abnormalities	IR	Z	O
[389]	Combines features extracted from CXRs and metadata for image retrieval	IR	Z	PR
[390]	Proposes to use the saliency maps as a similarity measure for image retrieval	IR	Z	X

tecting new findings, tracking the course of a disease, or evaluating the efficacy of a treatment are among the many uses of image registration [391]. To that end, [381] aims to create an anatomically plausible registration by using the heart and lung segmentations to guide the registration process.

2.5 Commercial Products

Computer-aided analysis of CXR images has been researched for many years, and in fact CXR was one of the first modalities for which a commercial product for automatic analyis became available in 2008. In spite of this promising start, and of the advances in the field achieved by deep learning, translation to clinical practice, even as an assistant to the reader, is relatively slow. There are a variety of legal and ethical considerations which may partly account for this [392, 393], however there is growing acceptance that artificial intelligence (AI) products have a place in the radiological workflow and attempts are underway to understand and address the issues to be overcome [394]. In this section we examine the currently available commercial products for CXR analysis.

An up to date list of commercial products for medical image analysis [395, 396] was searched for products applicable to chest X-ray. One product was excluded as it is not specifically a CXR diagnostic tool, but a texture analysis product for many modalities. The 21 remaining products are listed in Table 2.8. A number of these products have already been evaluated in peer-reviewed publications, as shown in Table 2.8 and it is beyond the scope of this work to make an assessment of their performance. All of the listed products are CE marked (Europe) and/or FDA cleared (United States) and are thus available for clinical use [395, 396].

The commercial products include applications for a wide range of abnormalities, with 6 of them reporting results for more than 5 (and up to 30) different labels. The most commonly addressed task is pneumothorax identification (8 products), followed by pleural effusion (7), nodules (6) and tuberculosis (4). In contrast with the literature, which is dominated by image-level prediction algorithms, 17 of 21 products in Table 2.8 claim to provide localization of one or more abnormalities which they are designed to detect, usually visualized with heatmaps or contouring of abnormalities. Two further products are designed for generation of bone suppression images, one for interval change visualization and one for identification and reporting of healthy images. Products contribute differently to the workflow of the radiologist. Five products focus on detecting acute cases to prioritize the worklist and speed up time to diagnosis. Draft reports are produced by five other products, for either the normal (healthy) cases only or for all cases. The production of draft reports, like workflow prioritization, is aimed at optimizing the speed and efficiency of the radiologist.

2.6 Discussion

In this work we have detailed datasets, literature and commercial products relevant to deep learning in CXR analysis. For researchers entering the field this study categorizes the existing data and literature for their ease of reference. In this section we further discuss how future research should be directed for higher quality and better clinical relevance.

It is clear that CXR deep learning research has thrived on the release of multiple large, public, labeled datasets in recent years, with 210 of 296 publications reviewed here using one or more public datasets in their research. The number of publications in the field has grown consistently as more public data becomes available, as demonstrated in Figure 2.2. However, although these datasets are extremely valuable, there are multiple caveats to be considered in relation to their use, as described in

Table 2.8: Commercial Products for CXR analysis. (Section 2.5)
Labels: T=Triage/Abnormal, PM=Pneumonia, CV=COVID, TB=Tuberculosis, LO=Lesion or Opacity, CM=Cardiomegaly, ND=Nodule, PE=Effusion, PT=Pneumothorax, TU=Catheter or Tube, LC=Lung Cancer, BS=Bone Suppression, E=Edema, Z=Other
Output: LOC=Localization, PRI=Prioritization, REP=Report, SCOR=Scoring

Company	Product	Literature (4 most recent)	Labels (Total number)	Output
Siemens Healthineers	AI-Rad Companion Chest X-Ray	[397]	LO PE PT Z (5)	LOC, SCOR, REP
Samsung Healthcare	Auto Lung Nodule Detection	[398]	ND (1)	LOC
Thirona	CAD4COVID-XRay	[399]	CV (1)	LOC, SCOR
Thirona	CAD4TB	[400–403]	TB (1)	LOC, SCOR
Oxipit	ChestEye CAD		T (1)	REP (healthy)
Arterys	Chest | MSK AI		LO, ND, PE, PT (4)	LOC, SCOR, PRI
Quibim	Chest X-Ray Classifier	[404]	PM CM ND PE PT E Z (16)	LOC, SCOR, REP
GE	Critical Care Suite		PT (1)	LOC, SCOR
InferVision	InferRead DR Chest		TB PE PT LC Z (9)	LOC, SCOR
JLK	JLD-O2K		LC Z (16)	LOC, SCOR
Lunit	Lunit INSIGHT CXR	[109, 166, 402, 405]	TB CM ND PE PT Z (11)	LOC, SCOR, PRI, REP
qure.ai	qXR	[402, 406–408]	T CV TB Z (30)	LOC, SCOR, PRI, REP
Digitec	TIRESYA		BS (1)	Bone Suppressed Image
VUNO	VUNO Med-Chest X-Ray	[409]	LO ND PE PT Z (5)	LOC, SCOR
Riverain Technologies	ClearRead Xray - Bone Suppress	[410–413]	BS(1)	Bone Suppressed Image
Riverain Technologies	ClearRead Xray - Compare		LC(1)	Subtraction Image
Riverain Technologies	ClearRead Xray - Confirm		TU(1)	LOC
Riverain Technologies	ClearRead Xray - Detect	[411, 414, 415]	ND LC (2)	LOC
behold.ai	Red Dot		T PT (2)	LOC
Zebra Medical Vision	Triage Pleural Effusion		PE	LOC, PRI
Zebra Medical Vision	Triage Pneumothorax		PT	LOC, PRI

Section 2.3. In particular, the caution required in the use of NLP-extracted labels is often overlooked by researchers, especially for the evaluation and comparison of models. For accurate assessment of model performance, the use of 'gold-standard' test data labels is recommended. These labels can be acquired through expert radiological interpretation of CXRs (preferably with multiple readers) or via associated CT scans, laboratory test results, or other appropriate measurements.

Other important factors to be considered when using public data include the image quality (if it has been reduced prior to release, is this a limiting factor for the application?) and the potential overlap between labels. Although a few publications address label dependencies, this is most often overlooked, frequently resulting in the loss of valuable diagnostic information.

While the increased interest in CXR analysis following the release of public datasets is a positive development in the field, a secondary consequence of this readily available labeled data is the appearance of many publications from researchers with limited experience or understanding of deep learning or CXR analysis. The literature reviewed during the preparation for this paper was very variable in quality. A substantial number of the papers included offer limited novel contributions although they are technically sound. Many of these studies report experiments predicting the labels on public datasets using off-the-shelf architectures and without regard to the label inaccuracies and overlap, or the clinical utility of such generic image-level algorithms. A large number of works were excluded for reasons of poor scientific quality (142). In 112 of these the construction of the dataset gave cause for concern, the most common example being that the training dataset was constructed such that images with certain labels came from different data sources, meaning that the images could be easily differentiated by factors other than the label of interest. In particular, a large number of papers (61) combined adult COVID-19 subjects with pediatric (healthy and other-pneumonia) subjects in an attempt to classify COVID-19. Other reasons for exclusion included the presentation of results optimized on a validation set (without a held-out test set), or the inclusion of the same images multiple times in the dataset prior to splitting train and test sets. This latter issue has been exacerbated by the publication of several COVID-19 related datasets which combine data from multiple public sources in one location, and are then themselves combined by authors building deep-learning systems. Such concerns about dataset construction for COVID-19 studies have been discussed in several other works [125, 416–419].

Although a broad range of off-the-shelf architectures are employed in the literature surveyed for this review, there is little evidence to suggest that one architecture outperforms another for any specific task. Many papers evaluate multiple different architectures for their task but differences between the various architecture results are typically small, proper hyperparameter optimization is not usually performed and statistical significance or data-selection influence are rarely considered. Many such evaluations use inaccurate NLP-extracted labels for evaluation which serves to muddy the waters even further.

While it is not possible to suggest an optimal architecture for a specific task, it is observed that ensembles of networks typically perform better than individual models [276]. At the time of writing, most of the top-10 submissions from the public challenges (CheXpert [6], SIIM-ACR [77], and RSNA-Pneumonia [73]) consist of network ensembles. There is also promise in the development of self-adapting frameworks such as the nnU-Net [420] which has achieved an excellent performance in many medical image segmentation challenges. This framework adapts specifically to the task at hand by selecting the optimal choice for a number of steps such as preprocessing, hyperparameter optimization, architecture etc., and it is likely that a similar optimization framework would perform well for classification or localization tasks, including those for CXR images.

In spite of the pervasiveness of CXR in clinics worldwide, translation of AI systems for clinical use has

been relatively slow. Apart from legal and ethical considerations regarding the use of AI in medical decision making [392, 393], a discussion which is outside the scope of this work, there are still a number of technical hurdles where progress can be made towards the goal of clinical translation. Firstly, the generalizability of AI algorithms is an important issue which needs further work. A large majority of papers in this review draw training, validation and test samples from the same dataset. However, it is well known that such models tend to have a weaker performance on datasets from external domains. If access to reliable data from multiple domains remains problematic then domain adaptation or active learning methods could be considered to address the generalization issue. An alternative method to utilize data from multiple hospitals without breaching regulatory and privacy codes is federated learning, whereby an algorithm can be trained using data from multiple remote locations [421]. Further research is required to determine how this type of system will work in clinical practice.

A final issue for deep learning researchers to consider is frequently referred to as 'explainable AI'. Systems which produce classification labels without any indication of reasoning raise concerns of trustworthiness for radiologists. It is also significantly faster for experts to accept or reject the findings of an AI system if there is some indication of how the finding was reached (e.g., identification of nodule location with a bounding box, identification of cardiac and thoracic diameters for cardiomegaly detection). Every commercial product for detection of abnormality in CXR provides a localization feature to indicate the abnormal location, however the literature is heavily focused on image-level predictions with relatively few publications where localization is evaluated. Many studies provide an unvalidated visualization of the area of interest [128, 145, 155, 165, 171, 263, 272], using methods like grad-cam [422] or saliency maps [423] which output heatmaps indicating which regions are important in the network result. Although these heatmaps may be useful for conditions that are indicated by localized patterns or signs, the lack of comprehensive evaluation of their accuracy is problematic. Furthermore, many conditions may be difficult to explain with a heatmap, for example emphysema, which is identified by irregular radiolucency throughout the entire lung (among other features). One possible way to achieve clinically useful systems in such cases is to label an image (e.g. positive or negative) for a series of known radiological features relating to the condition being identified, or to use other (e.g. segmentation) information in the classification [92].

Beyond the resolution of technical issues, researchers aiming to produce clinically useful systems need to consider the workflow and requirements of the end-user, the radiologist or clinician, more carefully. At present, in the industrialized world, it is expected that an AI system will act, at least initially, as an assistant to (not a replacement for) a radiologist. As a 2D image, the CXR is already relatively quickly interpreted by a radiologist, and so the challenge for AI researchers is to produce systems that will save the radiologist time, prioritize urgent cases or improve the sensitivity/specificity of their findings. Image-level classification for a long list of (somewhat arbitrarily defined) labels is unlikely to be clinically useful. Reviewing such a list of labels and associated probabilities for every CXR would require substantial time and effort, without a proportional improvement in diagnostic accuracy. A simple system with bounding boxes indicating abnormal regions is likely to be more helpful in directing the attention of the radiologist and has the potential to increase sensitivity to subtle findings or in difficult regions with many projected structures. Similarly, a system to quickly identify normal cases has the potential to speed up the workflow as identified by multiple vendors and in the literature [143, 246, 248].

To further understand how AI could assist with CXR interpretation, we first must consider the current typical workflow of the radiologist, which notably involves a number of additional inputs beyond the CXR image, that are rarely considered in the research literature. In most scenarios (excluding bed-

side/AP imaging) both a frontal and lateral CXR are acquired as part of standard imaging protocol, to reduce the interpretation difficulties associated with projected anatomy. Very few studies included in this review made use of the lateral image, although there are indications that it can improve classification accuracy [224]. Furthermore, the reviewing radiologist has access to the clinical question being asked, the patient history and symptoms and in many cases other supporting data from blood tests or other investigations. All of this information assists the radiologist to not only identify the visible abnormalities on CXR (e.g., consolidation), but to infer likely causes of these abnormalities (e.g., pneumonia). Incorporation of data from multiple sources along with the CXR image information will almost certainly improve sensitivity and specificity and avoid an algorithm erroneously suggesting labels which are not compatible with data from external sources. Another extremely important and time-consuming element in the radiological review of CXR is comparison with previous images from the same patient, to assess changes over time. Interval change is a topic studied by very few authors and addressed by only a single commercial vendor (by provision of a subtraction image). Innovative AI systems for the visualization and quantification of interval change with one or more previous images could substantially improve the efficiency of the radiologist. Finally, the radiologist is required to produce a report as a result of the CXR review, which is another time-consuming process addressed by very few researchers and just a handful of commercial vendors. A system which can convert radiological findings to a preliminary report has the potential to save time and cost for the care provider.
In many areas of the world, medical facilities that do perform CXR imaging do not have access to radiological expertise. This presents a further opportunity for AI to play a role in diagnostic pathways, as an assistant to the clinician who is not trained in the interpretation of CXR. Researchers and commercial vendors have already identified the need for AI systems to detect signs of tuberculosis (TB), a condition which is endemic in many parts of the world, and frequently in low-resource settings where radiologists are not available. While such regions of the world could potentially benefit from AI systems to detect other conditions, it is important to identify in advance what conditions could be feasibly both detected and treated in these areas where resources are severely limited.
The findings of this work suggest that while the deep learning community has benefited from large numbers of publicly available CXR images, the direction of the research has been largely determined by the available data and labels, rather than the needs of the clinician or radiologist. Future work, in data provision and labelling, and in deep learning, should have a more direct focus on the clinical needs for AI in CXR interpretation. More accurate comparison and benchmarking of algorithms would be enabled by additional public challenges using appropriately annotated data for clinically relevant tasks.

Acknowledgements

This work was supported by the Dutch Technology Foundation STW, which formed the NWO Domain Applied and Engineering Sciences and partly funded by the Ministry of Economic Affairs (Perspectief programme P15-26 'DLMedIA: Deep Learning for Medical Image Analysis'.
We would like to acknowledge and thank Gabrielle Ras and Gizem Sogancioglu who have supported us with their wise counsel and sympathetic ears.

Cardiomegaly Detection on Chest Radiographs: Segmentation versus Classification

3

Authors: Ecem Sogancioglu, Keelin Murphy, Erdi Çallı, , Ernst Th. Scholten, Steven Schalekamp, and Bram van Ginneken

Original title: Cardiomegaly Detection on Chest Radiographs: Segmentation versus Classification

Published in: IEEE Access, vol. 8, pp. 94631-94642, 2020

3.1 Introduction

Recent literature on the automatic interpretation of chest X-ray (CXR) images has been dominated by methods which learn to predict labels indicating the presence or absence of a specific abnormality in the CXR [6, 146, 189]. Such labels are frequently referred to as 'image-level' labels since they refer to the image as a whole and provide no more specific information, for example, regarding the location or severity of the abnormality. The popularity of this method of analysis is likely related to the recent release of numerous large public datasets, each of which provides multiple image-level labels for a variety of abnormalities [6, 34, 35, 339]. However, image-level labels may not be the optimal way to learn to recognise specific abnormalities. Since these labels provide no information on the shape or location of the abnormality, it is likely that a very large number of labelled samples will be needed to train a supervised-learning system. Furthermore, the trained system provides no insight or intuition into how it infers labels. Such a 'black-box' system is more difficult to trust and less likely to find acceptance in a clinical setting.

In this work, we investigate how a more intuitive and interpretable segmentation-based method to detect abnormality compares with the state of the art in deep-learning using image-level labels. The abnormality investigated in this case is cardiomegaly, one of the most frequently mentioned findings in radiology reports for chest radiography exams. Cardiomegaly refers to an enlargement of the heart and can be used as a marker for heart disease [424, 425]. Due to its wide availability, high cost-effectiveness, and low radiation dose, chest X-rays are often the first imaging study acquired and can be utilized as a fast screening tool for cardiomegaly. In order to detect this condition, radiologists examine the cardiac silhouette and calculate the cardiothoracic ratio (CTR), a commonly used radiographic index measured as the ratio of maximum horizontal cardiac diameter to the maximum horizontal thoracic diameter [426] (Figure 3.1). A CTR greater than 0.5 is the generally accepted threshold considered to indicate an enlarged cardiac silhouette, referred to as cardiomegaly.

A vast number of studies have addressed the cardiomegaly detection task along with other abnormalities in a multi-label classification scenario [146, 189, 427–429], predicting all available labels from the datasets used. Many of these works use the ChestX-ray14 dataset [339] which was released by the National Institutes of Health in 2017 with 112,120 CXRs, each labelled with binary labels for 14 different abnormalities. The labels are automatically extracted from the text analysis of radiology reports. These studies employed widely used state-of-the-art classification architectures, and applied slightly different augmentation and preprocessing techniques to tackle the classification problem. In particular, Baltruschat et al. [146] investigated the performance of different network architectures, namely ResNet-38, ResNet-50, and ResNet-101, for classification of 14 abnormalities on the ChestX-ray14 dataset [339]. They achieved a similar level of performance as other recently published studies [427, 428], but all these studies were limited due to their evaluation on the noisy held-out evaluation set where the labels were extracted from radiology reports using natural language processing [86]. In order to address this, Rajpurkar et al. [189] annotated a held-out evaluation set from ChestX-ray14 with the majority vote of 3 radiologists (not publicly available), and employed a 121-layer DenseNet architecture. The images were resized to 512 x 512 and normalized with the mean and standard deviation of images in the ImageNet training set before being fed into the network. They reported state-of-the-art results where the proposed algorithm achieved radiologist-level performance on 11 abnormalities in their held-out evaluation set, however, performed significantly worse than the radiologists for 3 abnormalities, one of which was cardiomegaly.

Some earlier works attempted to detect cardiomegaly through segmentation-based solutions via measuring CTR. Ginneken et al. [430] investigated the performance of three supervised segmentation

methods for anatomical segmentations, namely active shape models, pixel classification, and active appearance models. They showed that both active shape models and active appearance models reached a mean absolute error of 0.012 for cardiothoracic ratio measurement on their 247 held-out set. Candemir et al. [431] proposed a graph-cut lung field segmentation method which was then adapted to localize the heart region using heart models in order to measure the CTR. They reported 0.77 sensitivity and 0.76 specificity for the detection of cardiomegaly on 500 held-out evaluation images. Similarly, Dallal et al. [432] proposed a method that employed the same lung segmentation method proposed by Candemir et al. [431] and using the Harris operator to detect the heart boundaries from the resulting lung field segmentation in order to measure the CTR. They reported a root mean squared error of 0.06 on their 103 held-out images. Recent work by Li et al. [94] used a deep learning system for heart and lung field segmentation and showed improved performance for detection of cardiomegaly achieving a sensitivity of 0.97 and specificity of 0.92 on their 500 held-out set.

This study is the first to directly compare segmentation-based and classification-based solutions for cardiomegaly detection. We implement state-of-the-art deep learning methods for heart and lung segmentation, through which we calculate CTR directly, and also for image-level classification of cardiomegaly. Hyperparameter optimization is applied in all cases to ensure the best possible solution is obtained. We investigate the performance differences between the segmentation-based and classification-based systems for cardiomegaly detection, and the effect of varying the training-set size in each case.

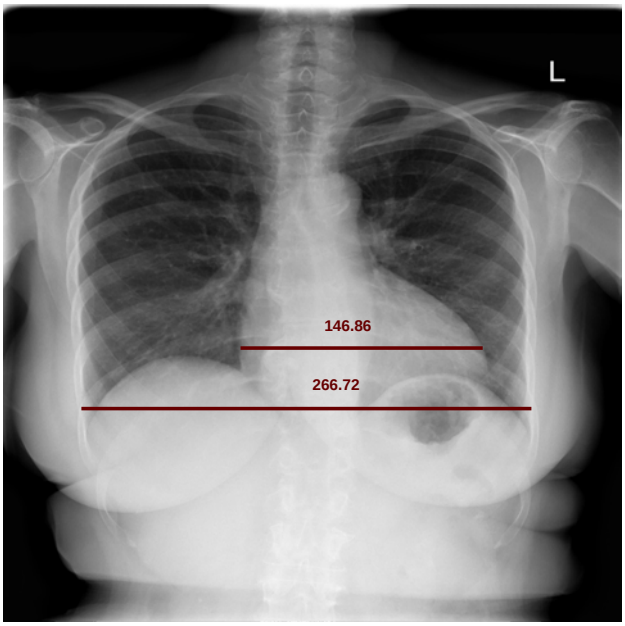


Figure 3.1: Measurement of the cardiothoracic ratio in chest radiographs. Maximum horizontal thoracic diameter = 266.72 (in mm), maximum horizontal cardiac diameter = 146.86 (in mm), CTR = 0.55 (146.86/266.72). CTR $>$ 0.5 and therefore this is a case of cardiomegaly.

3.2 Data

The data used in this study was retrospectively obtained from the publicly available ChestX-ray14 dataset [339]. It is composed of 112,120 frontal view chest radiographs from 30,805 patients stored as 8-bit grayscale images with dimensions of 1024x1024. The dataset was automatically labeled from text reports, indicating the presence or absence of 14 different thoracic abnormalities including cardiomegaly.

Heart enlargement, i.e. cardiomegaly, cannot reliably be assessed on AP view chest radiographs since the distance between the X-ray source and the patient is non-standardized on AP view, which causes a variable magnification of the heart. Hence, we selected only posteroanterior (PA) studies. This resulted in 67,310 PA images of 28,868 patients, 44% male, 41% abnormal.

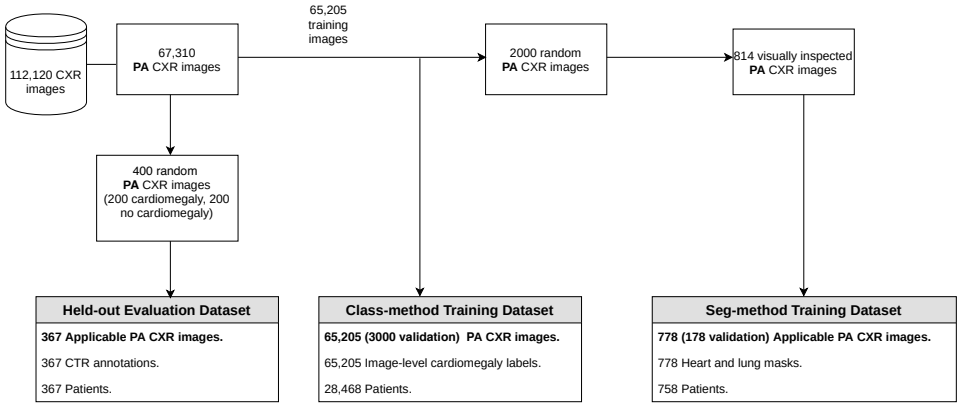


Figure 3.2: Flowchart of the data selection procedure. CXR = chest x-ray, PA = posteroanterior, CTR = cardiothoracic ratio, class-method = image-level cardiomegaly classification, seg-method = heart and lung segmentation. Images are from the publicly available ChestX-ray14 dataset.

3.2.1 Held-out Evaluation Set

For the final model evaluation, we created a class-balanced set of 400 images (Figure 3.2). Using the labels provided we randomly sampled 200 cases with cardiomegaly (200/1563) and 200 without cardiomegaly (200/65,747).

A chest radiologist with over 30 years of experience and another chest radiologist with over 5 years of experience independently annotated the maximal horizontal cardiac and thoracic diameters on all evaluation cases. Cases where radiologists could not reliably locate the heart borders were excluded from the study, leaving 367 cases. The annotations of the more experienced radiologist are used as the reference standard throughout this work, while the other radiologist is used as a second reader, for comparison with our automated methods.

3.2.2 Training & Validation Set

Classification-based Method

After the selection of only posteroanterior (PA) studies as seen in Figure 3.2, there was a total of 65,205 chest radiographs from 28,468 patients (excluding the patients in held-out evaluation set). This set was used as our training&validation set (3000 for validation), using the publicly available image-level cardiomegaly labels for training the classification-based method.

Segmentation-based Method

To develop deep neural networks to segment the heart and lungs we first set out to obtain manual segmentations of heart and lung boundaries. In order to select challenging cases for annotation of heart and lung boundaries, we developed a standard U-net [49] architecture which segments the heart and lung area, trained on a separate publicly available dataset, namely JSRT [72]. The JSRT dataset consists of 247 images from scanned films with a resolution of 2048x2048 and 12-bit depth. The reference standard for the heart and lung boundaries of those images are provided with the SCR dataset [430]. Our deep learning system was trained on a randomly selected 200 cases (200/247) and the remaining 47 cases were used as the validation set. The images were scaled to a dimension of 256 x 256, and the network was trained with Adam optimizer with a learning rate of 10^{-5}.
Further, a set of 2000 radiographs was randomly selected from the 65,205 remaining images in the ChestX-ray14 dataset (Figure 3.2). The JSRT-trained system was tested on those cases and visual inspection was used to select 814 cases most of which the algorithm performed sub-optimally. Those 814 cases were presented to a medical student and a computer scientist (with experience analyzing chest radiographs) who were instructed to annotate the heart and lung areas. An experienced radiologist was consulted for difficult cases and cases where the heart boundaries could not be inferred were excluded. This resulted in 778 radiographs (178 for validation) with lung and heart area annotations to be used as the segmentation training & validation set.

3.3 Methods

Two approaches for cardiomegaly detection are described in this section: firstly a classification approach based on image level labels (class-method) and secondly the segmentation-based approach (seg-method). For each approach hyperparameter optimization was run for 200 experiments. The final hyperparameters chosen were those that yielded the highest performance on the validation set.

3.3.1 Classification-based Method

To classify cardiomegaly using image-level labels we implemented three state-of-the-art classification architectures, ResNet18, ResNet50 [20], and DenseNet121[21], which have achieved excellent performance in several computer vision and medical image analysis tasks. Particularly, they were previously shown to achieve high-performance levels on the ChestX-ray14 dataset with multi-label classification settings [146, 189]. Training and architecture related hyperparameters of the class-method were systematically optimized to ensure optimum performance.
All the network architectures were pretrained on ImageNet, and a fully connected layer (2 output units with SoftMax activations) was added after the global average pooling layer. The networks

were trained with 65,205 frontal standard chest radiographs (3000 for validation) from ChestX-ray14 dataset, as in Figure 3.2, using categorical cross-entropy loss. Since there is a class imbalance problem in such a scenario (1156 images with cardiomegaly among 65k), we employed an over-sampling technique [433] by sampling the positive cardiomegaly cases until the dataset was balanced.
All images underwent per sample mean-standard deviation normalization. Data augmentation was applied to the training samples by means of inception-like preprocessing [44, 89]. This consists of applying a random rotation up to 7 degrees, random resizing with a scale in the range [0.7, 1], and random cropping a 4:3 or 3:4 part of the chest X-ray.

Class-method Hyperparameter Optimization

Several aspects of the hyperparameters were optimized for the class-method for 200 experiments.
Due to the very long training time of the class-method (which can take from 2 hours to 23 hours for one experiment depending on the network architecture and other hyperparameters), the hyperopt library [434] was used for 50 experiments. In every experiment during the optimization using hyperopt, the model being optimized is trained from scratch with the candidate hyperparameters for a maximum number of epochs predefined for each model. The selection of the candidate hyperparameters are based on Bayesian optimization, i.e., the hyperparameters were selected based on a trade-off between the results of the previous iterations, the regions of unexplored hyperparameter space, and their underlying distribution.
Further, we also optimized the hyperparameters through grid search, which can be run in parallel unlike hyperopt, for an additional 150 experiments.
The hyperparameters range and the values selected after the optimization can be seen in Table 3.1. We used three commonly used architectures, DenseNet121, ResNet50, and ResNet18, as a hyperparameter value in order to optimize the network architecture for our problem settings. Due to memory constraints, we made sure that the batch size was set to 8 when the network architecture was DenseNet121 or ResNet50 with an input resolution of 512 otherwise to 16.
Based on the hyperparameter optimization results, after every 100 iterations, the validation loss was calculated on the whole validation set. If the validation loss did not decrease compared to the previous step, the learning rate was reduced by multiplying it with 0.2. The model which showed the least validation error was selected as our final model.
After the hyperparameter optimization, the best model found for the class-method was ResNet50 trained with the largest input resolution of 512. During the experiments, we observed that all the deep learning models were powerful and achieved a high level of performance and that the most crucial hyperparameters on performance were learning-related, i.e. learning rate.
The hyperparameter optimization procedure took around 23 days with hyperopt on a PC equipped with TitanX GPU, and 6 days for grid search optimization (run in parallel) for 150 experiments using several GPU, TitanX, GTX1080, GTX1080ti, GTXTitanx, and TitanV. The code was implemented in Tensorflow [435].

3.3.2 Segmentation-based Method

The segmentation-based approach (seg-method) is designed to address the cardiomegaly detection task on chest radiographs, through segmentation of the heart and lungs and subsequent calculation of the cardiothoracic ratio (CTR). As illustrated in Figure 3.3, two different models were developed for heart and lung field segmentation respectively. After segmentation, the maximum horizontal cardiac

	Hyperparameter	Range	Best class-model
Learning	Optimizer	[Adam, SGD, Adagrad, RMSprop]	RMSprop
	Learning rate	{0.00001, 0.1}	0.012
	Initializer	[Orthogonal, Glorot, He, LeCun]	He
	LR reduced factor	[0.2, 0.9]	0.2
Architecture	Model	[ResNet18, ResNet50, DenseNet121]	ResNet50
	Input resolution	[64, 128, 256, 512]	512

Table 3.1: Optimized hyperparameters for the class-method. The naming convention follows [436]. LR=learning rate. LR reduced factor indicates the factor to multiply learning rate with in case of no improvement is seen on the validation set performance during training.

and thoracic diameters were calculated and used to calculate CTR and hence the presence or absence of cardiomegaly based on the clinically used CTR threshold of 0.5.

For the development of heart and lung segmentation models, a U-net-like fully convolutional network architecture [49] was implemented and its training, regularization, and architecture-related hyperparameters were systematically optimized for the best model selection.

The U-net architecture [49] is a state-of-the-art segmentation network, which has achieved promising results on a variety of medical image segmentation tasks [432, 437]. It consists of contracting and expanding paths, where the contracting path is composed of convolution operations decreasing the spatial resolution and the expanding path consists of transposed convolutions increasing the resolution. Further, the details that were lost through downsampling operations are recovered through skip connections which pass feature maps from the contracting to the expanding path.

During training, each model was trained by optimizing the binary cross-entropy loss between the predicted masks and the reference standard (heart or lung masks), which is formulated as follows:

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)$$

Where N denotes the number of images, y_i represents the reference standard for the sample i, $\hat{y}_i$ represents the model prediction for the sample i.

All images underwent per sample mean-standard deviation normalization. Data augmentation with random rotation, vertical and horizontal shift, zooming, and brightness was applied to improve system robustness. The model was trained for a maximum of 300 epochs, terminating if there was no improvement in the validation set performance for 20 successive epochs. We selected the epoch with the best performance on the validation set.

Seg-method Hyperparameter Optimization

Similar to the class-method, the hyperparameters of the seg-method were optimized using the hyperopt library [434] for 200 experiments.

The heart and lung segmentation models were optimized separately. A specific set of learning, architecture and regularization-related parameters [438–441] were selected for the hyperparameter search as listed in Table 3.2 (with the naming convention as in [436]). The learning rate was the only continuous hyperparameter and was sampled from a log uniform distribution. The other hyperparameters

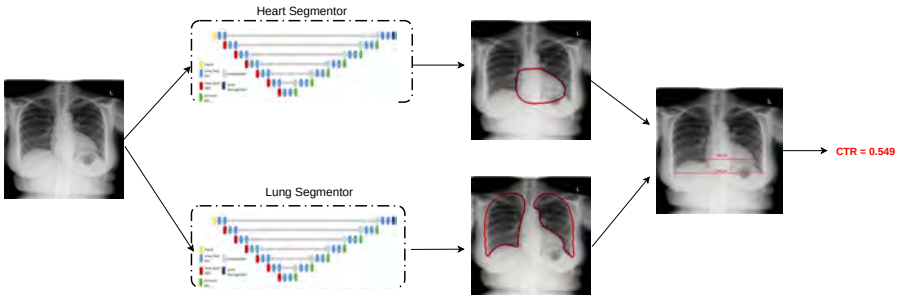


Figure 3.3: Illustration of the architecture pipeline for the seg-method. CTR = cardiothoracic ratio. Two different models are trained, for heart and lung field segmentation, respectively. CTR is derived from those predicted segmentation maps by determining the maximum horizontal thoracic and cardiac diameter and computing the ratio.

were sampled from a discrete uniform distribution between the defined choices.

As a regularization hyperparameter, the selection of dropout (with a probability of 0.5) [441] before each convolution in the expanding path was introduced as a binary hyperparameter. We used batch normalization [442] after every convolution layer as it improved performance by enabling more efficient learning.

Due to the limitations of computational memory, some restrictions on the combinations of hyperparameter settings were required. While a large batch size helps to stabilize the training, the depth of the network and the number of convolution operations per layer increase the capacity of the network, and the receptive field and the higher resolution images allow the network to see more details within the image. However, not all these conditions can be satisfied at the same time due to memory constraints. Therefore, the selection of these hyperparameters was conditioned on each other: when the input resolution was 512, the batch size was chosen as 4, and when the depth of the network was larger than 4, the number of convolution operations per depth was limited to 2 and the number of initial feature maps limited to 32.

The best models found after the hyperparameter optimization for both heart and lung segmentation were U-net architecture with the highest depth 6 as in Table 3.2. During the experiments, we observed that a larger input resolution yielded better performance.

The hyperparameter optimization procedure took around 13 days for each of the lung and heart segmentation models on a PC equipped with TitanX GPU and with the code implemented in Keras [436] with Tensorflow backend [435].

3.4 Experiments

The seg-method and class-method performance were investigated for cardiomegaly classification. Further, since the seg-method additionally produces a clinically relevant measure, CTR, the performance of this system was also evaluated in terms of CTR accuracy.

	Hyperparameter	Range	Heart model	Lung model
Regularization	Dropout	[True, False]	False	False
Learning	Batch size	[4,8,16]	4	4
	Optimizer	[Adam, SGD, Adagrad, RMSprop]	Adam	RMSprop
	Activation function	[ReLU, SELU, ELU]	SELU	ELU
	Learning rate	{0.00001, 0.01}	0.00018	0.00076
	Initializer	[Orthogonal, Glorot, He, LeCun]	LeCun	he_normal
Architecture	Convolutions per depth	[1, 2, 3]	2	1
	Depth of the network	[2, 3, 4, 5, 6]	6	6
	Input resolution	[64, 128, 256, 512]	512	512
	Initial feature maps	[32, 64]	32	32

Table 3.2: Hyperparameter optimization for the seg-method. Regularization, learning and architecture related hyperparameters are optimized and ranges are demonstrated. The naming convention follows [436].

3.4.1 Cardiomegaly Classification

We evaluate the performance of the two methods and of the second reader by calculating the area under the receiver operating characteristic curve (AUC). To construct ROC curves the reference standard CTR values were thresholded at 0.5 in order to obtain binary cardiomegaly labels. The sensitivity and specificity of each system and the reader performance is then computed at all possible operating points by applying various thresholds on the CTR output score (second-reader and seg-method) or SoftMax prediction for cardiomegaly (class-method) in order to produce an ROC curve.

It is important to note that the class-method was trained on a considerably larger dataset compared to the seg-method. This was done considering the different levels of annotation efforts between the two methods in order to have a fair comparison, and to investigate the performance of the class-method in its full potential. To validate our experimental design, we have also included the performance of the class-method when being trained with the same small dataset as the seg-method in our ROC analysis.

The kappa statistic [443] between the reference standard and the second reader and the models are calculated. Further, the sensitivity, specificity, positive predictive value, and negative predictive value and their 95% confidence intervals [444, 445] are reported, based on a fixed threshold of 0.5.

3.4.2 Training Set Size Analysis

In order to investigate the effect of the number of training images on the cardiomegaly classification performance, we constructed learning curves. We train both the seg-method and the class-method networks with varying numbers of training images and determine the effect of this on the method performance. The seg-method was trained with 50, 100, 200, 300, 400, 500, 600 images using 178 images in the validation set for each experiment and the class-method was trained with 2.5k, 5k, 10k, 20k, 40k, 62k images each using 3000 images as the validation set. We analyzed the results with the AUC score.

3.4.3 CTR Analysis

Heart and Lung segmentation

Since the seg-method detects cardiomegaly through lung and heart segmentation, the segmentation performance of the final models, which were found through hyperparameter optimization, were evaluated on the full JSRT dataset (247 images). We used intersection over union (IOU), also known as Jaccard index, as a performance measure which is calculated as follows:

$$\text{IOU} = \frac{| X \cap Y |}{| X \cup Y |}$$

where X represents the output of the network, and Y is the reference standard segmentation output. IOU quantifies the overlap between X and Y as the ratio between the number of pixels that are common between X and Y (cardinality of the intersection set) and the total number of pixels present across both of them (cardinality of the union set).

CTR calculation

The performance of the seg-method was analyzed as a regression task in order to evaluate the performance in terms of CTR accuracy. Segmentation predictions can directly be used to calculate maximal horizontal cardiac and thoracic diameter, and used to calculate CTR as their respective ratio. The reference standard is created from the first radiologist CTR annotations to which the performance of the seg-method and the second reader can be compared.
The mean absolute error was used to evaluate the accuracy of CTR predictions with respect to the reference standard as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} | \epsilon_t |,$$

where N denotes the number of images, and ϵ_t represents the difference between the predicted CTR and the reference standard CTR.
Moreover, CTR performance was also evaluated with Pearson correlation coefficient to summarize the strength of the linear relationship between the reference standard and the CTR predictions. The differences in CTR measurements, and the cardiac (in mm) and thoracic diameters (in mm) between the reference standard and the seg-method and the second reader were analyzed.

3.5 Results

3.5.1 Cardiomegaly Classification

As shown in Figure 3.4, the class-method performed reasonably well, but with clearly much lower specificity at all sensitivity settings compared to the seg-method. The performance of the second reader and the seg-method are very similar to each other on this dataset with an AUC of 0.978 (95% confidence interval [CI]: 0.969, 0.988) and 0.977 (95% [CI]: 0.966, 0.988), respectively. In contrast, the class-method obtained an AUC of only 0.941 (95% confidence interval [CI]: 0.922, 0.959) when it was trained on a large dataset (62k). Further, the performance of the class-method decreased considerably achieving an AUC of 0.830 (95% confidence interval [CI]: 0.789, 0.863) when it was trained on the same small dataset as the seg-method (600).

The kappa statistic for cardiomegaly classification (at a threshold of 0.5) between the reference standard and the second reader was 0.856 while for the seg-method and class-method were 0.870 and 0.683, respectively. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) (at a fixed threshold of 0.5) on the held-out evaluation set are provided in Table 3.3. The seg-method and the second reader showed similar performance levels with the sensitivity of 0.97 and 0.91 and specificity of 0.90 and 0.95, respectively.

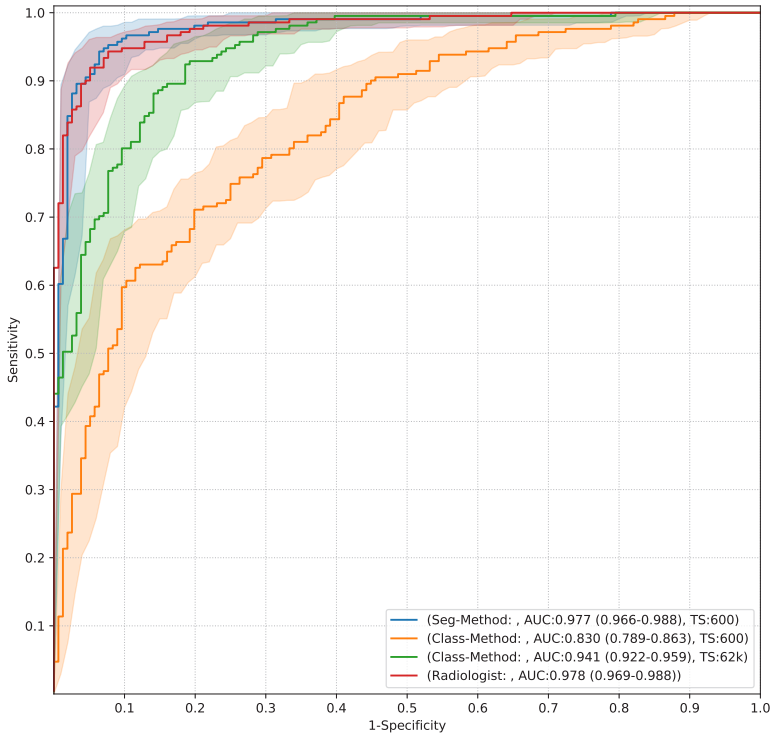


Figure 3.4: Receiver operating characteristic curves for detection of cardiomegaly in the held-out evaluation set (N=367). Reference = Radiologist 1, TS: Number of training samples, The second reader (Radiologist 2). Shaded areas represent the 95% confidence intervals. The reference standard CTR values were thresholded at 0.5 in order to obtain binary cardiomegaly labels.

3.5.2 Training Set Size Analysis

The impact of the number of training images on the classification performance is illustrated in Figure 3.5a and 3.5b for both seg-method and the class-method. Figure 3.5a illustrates that seg-method benefits from an increased number of training images until the number of training images reaches 500. It seems that increasing this number further does not bring any performance gain.
The effect of the number of training images for the performance of the class-method appears to be more crucial compared to the seg-method in Figure 3.5b. The performance continues to increase

	MAE	Sensitivity	Specificity	PPV	NPV
Seg-Method	**0.0135**	0.97 [0.93, 0.99]	0.90 [0.84,0.94]	0.93 [0.88,0.96]	0.95 [0.90,0.98]
Class-Method		0.81 [0.75,0.86]	0.89 [0.83,0.93]	0.91 [0.86,0.94]	0.77 [0.70, 0.83]
Second Reader	**0.0135**	0.91 [0.87,0.95]	0.95 [0.90,0.98]	0.96 [0.92,0.98]	0.89 [0.83,0.93]

Table 3.3: Comparison of the seg-method with the class-method and the second reader. PPV = positive predictive value, NPV = negative predictive value, MAE = mean absolute error. The number between brackets denote 95% confidence intervals. MAE is calculated against the reference standard for CTR. Since the class-method produces a binary output, MAE can not be calculated. All other measures relate to binary classification of cardiomegaly status.

substantially with the addition of more training data even after 40k training images.
Moreover, Figure 3.5a and 3.5b demonstrates that only 100 training images were sufficient for the seg-method to achieve a better performance than the class-method which was trained with 62k training images.

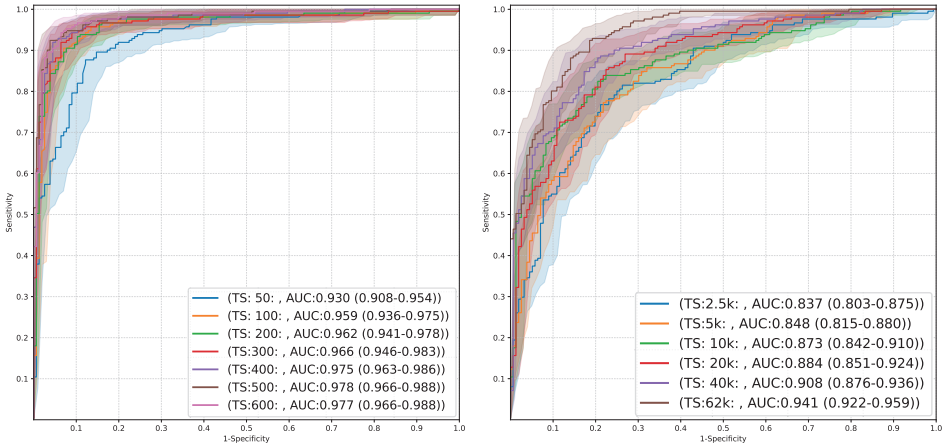
3.5.3 CTR Analysis

Heart and Lung segmentation

The seg-method achieved 0.87 and 0.95 intersection over union (IOU) on the full JSRT dataset (247 images) for heart and lung segmentation, respectively.

CTR calculation

The mean absolute error between both the seg-method and the second reader against the CTR reference standard was 0.0135 as seen in Table 3.3. The scatter plots of the reference standard CTR against the predicted CTR values of the model and the second reader are provided in Figure 3.6a and 3.6b, respectively. In line with our expectations, the misclassified cases for both the second reader and the seg-method are consistently those cases where the CTR is close to the threshold value of 0.5. Both the model and the second reader CTR predictions against the reference standard appear highly correlated, showing 0.960 and 0.965 Pearson correlation coefficient, respectively.
The histogram of the differences between the reference standard CTR values and the seg-method and the second reader are illustrated in Figure 3.6c and 3.6d, respectively. For both the seg-method and the second reader, the majority of the differences were less than 0.06. In particular, there were 7 cases out of 367 where the differences between both the seg-method and the second reader to the reference standard were higher than this value.
The range of differences between the reference standard maximal horizontal cardiac and thoracic diameters and the model and the second reader are shown in Figure 3.6e and 3.6f, respectively. The measurement differences for both the cardiac and thoracic diameters were in a similar range for the model and the second reader.



(a) Effect of training set size for the seg-method

(b) Effect of training set size for the class-method

Figure 3.5: The performance of the seg-method and the class-method for various training set sizes. TS = number of training samples. All curves are computed for the held-out evaluation set (N=367). Shaded areas represent the 95% confidence intervals.

Difficult Case Analysis

Example cases for the predictions of seg-method and class-method are shown in Figure 3.7. Misclassified cases where the reference standard is close to the CTR threshold of 0.5 are less interesting since these differences can be caused by inter-reader variability. Therefore we analyzed the misclassified cases where the reference standard was higher than 0.55 or lower than 0.45. There were no misclassified cases for both seg-method and class-method when the reference standard was lower than 0.45. However, class-method misclassified 8 cases where the reference standard was higher than 0.55 while the seg-method misclassified only one single case.

3.6 Discussion

In this work, it was demonstrated that a segmentation-based model trained on a modestly sized collection of chest radiographs (778 images) achieves an AUC of 0.977 for the detection of cardiomegaly, which is comparable to an independent second reader with an AUC of 0.978. The seg-method reached a high sensitivity and specificity on this task at 97% and 90%, respectively. In contrast, the class-method of image-level classification for cardiomegaly achieves a significantly lower performance with an AUC of 0.941 although it has been trained on 65,205 images. The performance achieved by the class-method is nonetheless representative of the state-of-the-art for classification-based solutions since several studies [6, 146, 189, 446–448] reported similar or lower cardiomegaly classification performance which were evaluated on a variety of datasets.
Experimental results demonstrated that the seg-method trained on only 100 annotated images can still outperform the class-method (Figure 3.5), trained on 65k images. This result highlights the difference

between the methods in several aspects. First, it reveals that integrating domain knowledge from segmentations in subsequent image analysis may greatly reduce the volume of annotated training data required to achieve high performance. It additionally suggests that much improved accuracy can be obtained on these tasks, even with very limited training data. Finally, the seg-method opens the black-box solution of the class-method by producing the heart and lung segmentation and the diameters making up the CTR measure, rather than producing a single classification output. This is likely to be useful in clinical settings where the use of black-box algorithms is typically viewed as a high-risk solution.

It is notable that the class-method continued to improve in performance as additional training data was added. We hypothesize that with enough training samples it would eventually obtain a similar performance to the seg-method and the second reader. Further, the performance of class-method might be improved if the training labels did not contain any noise, although deep-learning systems have been shown to be robust to training label noise in recent studies [89, 449]. However the method would remain, nonetheless, inexplicable to clinicians.

Compared to the previous studies using segmentation-based solutions for cardiomegaly classification [431, 432], our seg-method showed a substantially improved performance. Considering the fact that the heart and lung field segmentation performance is the key to the algorithm performance, it is clear that the improved performance of our seg-method relies heavily on our segmentation methodology. Unlike earlier studies, we employed a deep learning model, a state-of-the-art segmentation network [49], and systematically optimized its hyperparameters to segment the lung and heart field with optimal accuracy. This can be seen with the intersection over union (IOU) score reported for the heart and lung field segmentation in these studies. For instance, Candemir et al. [431] showed that they achieved IOU of 0.70 and 0.95 for the heart and lung segmentation, respectively whereas Dallal et al. [432] achieved IOU of 0.57 with their heart segmentation approach. However, our model achieved IOU of 0.87 and 0.95 on the JSRT dataset for heart and lung field segmentation, respectively, outperforming the results reported in [430].

This result suggests that there is a difference with a large margin in terms of heart segmentation performance between our proposed deep learning approach and the earlier studies. Recent work by Li et al. [94] which also used a deep learning segmentation model supports this result. In this work, the obtained CTR values are comparable with manual measurements although they required 5000 manually segmented scans for training, compared to just 778 in this work. Our work is the first to provide a direct comparison between segmentation-based and end-to-end image-level cardiomegaly classification demonstrating the advantages of the former, both in terms of clinical interpretation and performance. We also provide an online demo[1] where interested readers can test out our seg-method algorithm.

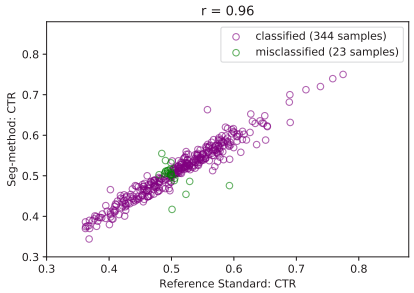
While it is clear that annotations of heart and lung boundaries are more time-consuming to obtain than image-level labels (which are often extracted using automatic methods from radiology reports), we believe that segmentation of anatomy is important not only for cardiomegaly detection, but also in the identification and quantification of many other abnormalities. Our manual segmentations took an average of 2 minutes per image (for both heart and lung boundaries) and we expect that our trained segmentation networks could now serve as guidance in many clinically interpretable abnormality detection systems. Future work will investigate the incorporation and importance of anatomical segmentation in other clinically relevant tasks.

This study has several limitations. First, all chest radiographs were retrieved from a single institution,

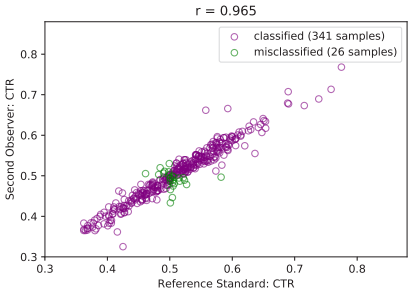
[1] https://grand-challenge.org/algorithms/cxr-cardiomegaly-detection/

which may affect the robustness of the system in evaluating images from other sources. Second, lateral view chest radiographs were not considered in our study although they might potentially be used, when in doubt, as complementary information to accept or reject cardiomegaly. Further, the cases for which the determination of CTR measurements was not possible (due to invisibility of anatomical boundaries) were manually excluded from our held-out evaluation set. In clinical practice, such images cannot be used for the determination of cardiomegaly. The automated rejection of such cases by the model would be a useful tool in clinical settings and might be a good future research direction.

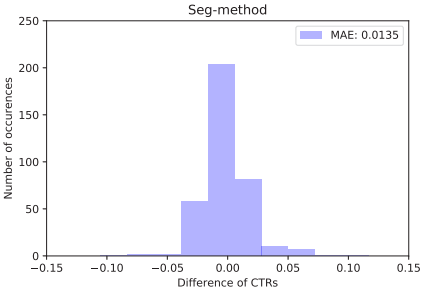
We conclude that we have implemented a segmentation-based cardiomegaly algorithm with performance comparable to a human reader, and with the advantages of improved accuracy and better interpretability compared to the image-level classification method. Future work will investigate extending the segmentation-based approach to other diagnostic tasks.



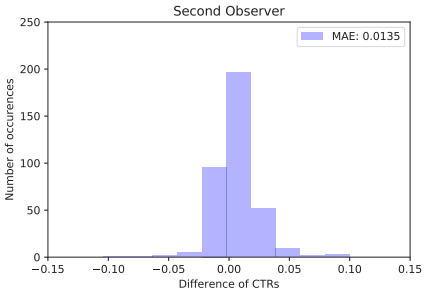
(a) CTR predictions of the seg-method against the reference standard.



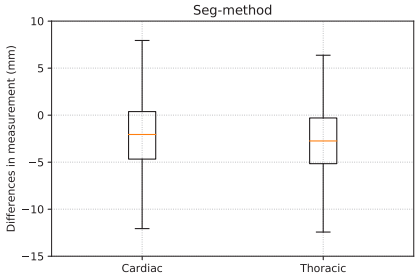
(b) CTR predictions of the second reader against the reference standard.



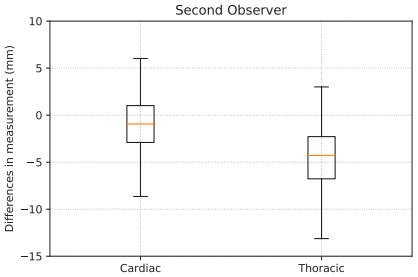
(c) The differences between the reference standard CTR and seg-method predictions.



(d) The differences of the second reader and the reference standard CTR.



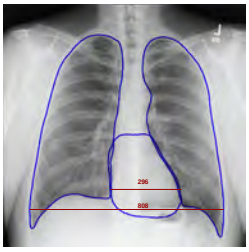
(e) The cardiac and thoracic diameter differences between the reference standard and the seg-method predictions.



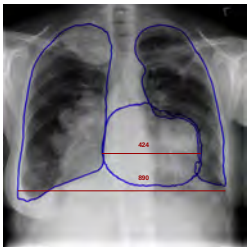
(f) The cardiac and thoracic diameter differences between the reference standard and the second reader.

Figure 3.6: MAE = mean absolute error, CTR = cardiothoracic ratio. (a) and (b): The scatter plots of the reference standard CTR values against the CTR values of the seg-method and the second reader respectively. Correctly classified and misclassified samples are visualized in purple and green, respectively. (c) and (d): The histogram of the CTR differences between the reference standard and the seg-method and second reader respectively. (e) and (f): The box plot of the differences between the maximal horizontal cardiac and thoracic diameters between the reference standard and the seg-method and the second reader in mm.

(a) CTR = 0.604 (RS = 0.605).

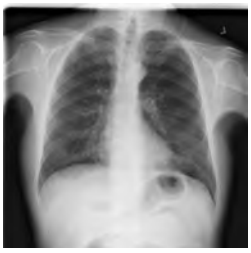


(b) CTR = 0.366 (RS = 0.362)



(c) CTR = 0.476 (RS = 0.592)

(d) RS CTR = 0.610



(e) RS CTR = 0.425

(f) RS CTR = 0.569

Figure 3.7: Example cases of the model predictions. CTR = cardiothoracic ratio, RS = reference standard. (a)-(c): Three example cases of the seg-method predictions. Model prediction CTR (reference standard CTR). Cases a and b are correctly classified and case c is misclassified. (d)-(i): Example cases of the class-method predictions. d and e are the correctly classified cases, whereas f is an example of misclassification.

Automated Estimation of Total Lung Volume using Chest Radiographs and Deep Learning

4

Authors: Ecem Sogancioglu, Keelin Murphy, Ernst T. Scholten, Luuk H. Boulogne, Mathias Prokop, Bram van Ginneken

Original title: Automated Estimation of Total Lung Volume using Chest Radiographs and Deep Learning

Published in: Medical Physics, 49:4466- 4477, 2022

4.1 Introduction

Chest radiography (CXR) remains the most commonly performed imaging technique and one of the most often repeated exams because of its low cost, rapid acquisition and low radiation exposure [24]. It was estimated that 129 million chest radiographs were performed in 2006 in the United States alone [23]. Chest radiographs play an important role in screening, monitoring, diagnosis, and management of thoracic diseases.

Wide availability of CXR has motivated researchers to build artificial intelligence (AI) systems that can automatically detect a variety of abnormalities [140, 189, 399] and extract quantitative clinical measurements from them [92, 94]. AI systems have potential use for routine quantification of numerous biomarkers related to lung diseases, cardiac health, or osteoporosis. Applying such systems, whenever a chest radiograph is acquired, would be a step towards routine quantitative radiology reporting.

This work focuses on an important quantitative biomarker, total lung volume, and investigates whether it can be measured automatically from plain chest radiographs using state-of-the-art deep learning approaches. Total lung volume (TLV) is used for assessing severity, progression and response to treatment in restrictive lung diseases [450, 451]. Specific temporal changes in TLV can be identified in patients with obstructive and restrictive lung diseases, such as emphysema, pulmonary fibrosis or asthma. Further, TLV has been shown to correlate with mortality and health status [452].

Currently, the gold standard for measurement of TLV is the pulmonary function test (PFT), using special techniques such as body plethysmography, helium, or nitrogen dilution techniques [451]. Several studies [453–455] demonstrated that TLV measured from CT strongly correlates to TLV obtained from PFTs. Alternatively, several studies investigated TLV estimation from CXR using predictive equations. In fact, this has been a research interest for a century, with the first paper appearing in 1918 ([456]) demonstrating the correlation of external measurements from CXR to the pulmonary function test (gas dilution technique). All such previous literature, investigating predictive equations, was either based on the use of planimetric techniques [457–460], or made assumption of a given a geometry [461–463], or required several manual linear measurements to estimate TLV from CXR. However, all these studies required manual measurements to estimate TLV and used small sample sizes, making it unclear whether the techniques could be generalized to other populations.

In this study, we investigate, to the best of our knowledge, for the first time, whether chest radiography can be used to automatically predict TLV in a fully automated fashion using large datasets and deep learning. We examine the role of TLV labels derived from thoracic CT imaging in training deep learning systems. In order to account for variations in inspiration and dataset complexity, experiments with simulated and real chest radiographs in three different datasets were designed in a step-wise fashion. For each experiment, we optimized various state-of-the-art deep learning regression approaches to predict TLV using only posterioranterior (PA) view, lateral view or both views. The purpose of our study was to determine the accuracy of fully automatic measurement of TLV from CXR using deep learning based models.

4.2 Materials and Methods

4.2.1 Data and Preprocessing

The data used in this study was obtained from two sources; the COPDGene study [464] and Radboud University Medical Center (RUMC). To facilitate our stepwise experimentation, demonstrating sources of error, we experimented with simulated CXR images (digitally reconstructed radiographs), which are obtained from average intensity projections (AIP) on thoracic CT, as well as with true CXR images. Reference total lung volume labels were obtained by two means; through segmentation of the lungs in CT and from pulmonary function tests (PFT). The datasets constructed are described in detail in the sections below and in Figure 4.1.

COPDGene-sim

Inspiration chest CT studies (1000) from unique patients were randomly selected from the COPDGene study, [464] which is publicly available on request for research purposes. The images in this study are acquired from patients with Chronic Obstructive Pulmonary Disorder (COPD), varying from mild to very severe. From the 1000 randomly selected CT studies, 800 (600 for training and 200 for validation) were used for training and validation, and 200 were retained as a held-out test set as illustrated at the top of Figure 4.1.
Lung segmentations were obtained by an automated algorithm and manually corrected by trained analysts with radiologist supervision [465]. Reference TLV was calculated for each CT scan by multiplying CT image spacing by the number of voxels segmented.
Simulated CXRs were generated from CT by creating AIP [173] from coronal and sagittal planes, resulting in frontal and lateral view simulated CXR. This dataset, which we refer to as COPDGene-sim, was used to demonstrate model performance in an ideal scenario where there is no inspiration difference between the label source (CT) and the (simulated) CXR image, CT segmentations are manually corrected, and the variety of pathologies is limited.

RUMC Datasets

This data was obtained from routine clinical care in Radboud University Medical Center, Nijmegen, the Netherlands (RUMC). This study was approved by the research ethics committee of the Radboud University Nijmegen Medical Centre. Dataset was collected and anonymized according to local guidelines and informed consent was obtained from all participants. All research was performed in accordance with relevant guidelines and regulations.
We retrospectively collected CXR studies and chest CT acquired between 2003 and 2019 resulting in 321k CXR studies and 120k CT studies. Patients with both CT and CXR (with PA and lateral view), performed a maximum of 15 days apart, were selected (4420 patients). The reference standard TLV measurements were obtained by a CT lung segmentation algorithm [465] and segmentation failure cases were visually identified and excluded (284 CT). This resulted in 7621 CXR studies and 5305 CT studies from 4275 patients (Figure 4.1). Multiple CXR studies from a single patient could be matched to a single CT reference standard.
A group of patients being assessed for lobectomy was used to provide subjects with both PFT and CXR data acquired within 15 days of each other. This resulted in 928 CXR studies from 485 patients. Reference TLV was determined using the helium delusion technique [466].

From this dataset, we created two sets for experimentation. The first is referred as RUMC-sim and used simulated CXR generated from CT as described in Section 4.2.1. The second is RUMC-real, consisting of real CXR with CT-derived and PFT-derived labels for TLV. To investigate the relationship between CT-derived and PFT-derived labels, we created a dataset, CT-evaluation, where both CT and PFT were acquired within 15 days of each other. We made sure that there was no patient overlap between training and held-out evaluation sets for all the datasets. These datasets are detailed below and illustrated in Figures 4.1 and 4.2.

RUMC-sim In this dataset, both frontal and lateral view chest radiographs were simulated from 5305 CT studies (4275 patients). Of these, 389 patients (590 CT studies) were randomly selected and used as a held-out evaluation set, whereas the remaining 3886 patients (3236 for training, 650 for validation) were used for training. This dataset, with CT-derived lung volume labels, was used to illustrate the model performance in a set of images with a large variety of abnormalities (compared to COPDGene-sim), e.g., pleural fluid, large masses, widespread interstitial abnormalities. The use of simulated CXR images removes any possibility of error related to inspiration effort, or patient position between the label source (CT) and the (simulated) CXR.

RUMC-real This dataset consists of patients with real CXR studies (PA and lateral) and with lung volume reference standard measurements from two sources, namely CT and PFT. For CT-based data, the same patient partitioning was used as in RUM-sim, but using the CXR with the study time closest to that of the corresponding CT study rather than a simulated CXR. This resulted in 7621 CXR studies with CT-derived labels, whereas PFT-derived labels were used for 928 CXR studies as seen in Figure 4.1. As a held-out evaluation set, 590 patients with 1008 CXRs with CT-derived labels, and 291 CXR from 150 patients with PFT-derived labels were randomly selected. We made sure there was no patient overlap between the PFT-based evaluation set and any training set (with CT-labels or PFT-labels).

CT evaluation dataset We identified patients with PFT results available that were also in the RUMC-sim dataset, and selected patients with PFT results obtained a maximum of 15 days apart from their CT study. This resulted in 137 CT studies from 130 patients. CT lung volume was calculated by means of an automated CT lung segmentation algorithm [465], and the results were visually inspected, identifying no obvious failed segmentations. This set was used to demonstrate the relationship between CT-derived and PFT-derived labels.
All CT scans used in the COPDGene-sim and RUMC-sim datasets were first resampled to 1mm isotropic spacing before generating simulated CXRs by average intensity projection. Similarly, real CXRs were resampled to have 1mm $\times$ 1mm spacing. Resampling of all CXR to the same spacing is crucial All real and simulated CXR images were padded with zeros to reach a fixed size of 512 x 512 pixels. Images underwent standard normalization to the range of -1 to 1.

4.2.2 Methods

We experiment with 5 different deep-learning architectures, 4 of which are widely used popular classification architectures (DenseNet121 [21], ResNet34, ResNet50 [20], VGGNet [467]), and one, referred as 6-layer CNN, was designed to represent a shallow architecture. The 6-layer CNN consisted of 6 CNN layers, each followed by RELU, batch normalization and a pooling layer. The first CNN layer had 32 feature maps, and the number of feature maps was doubled in each layer. The final CNN

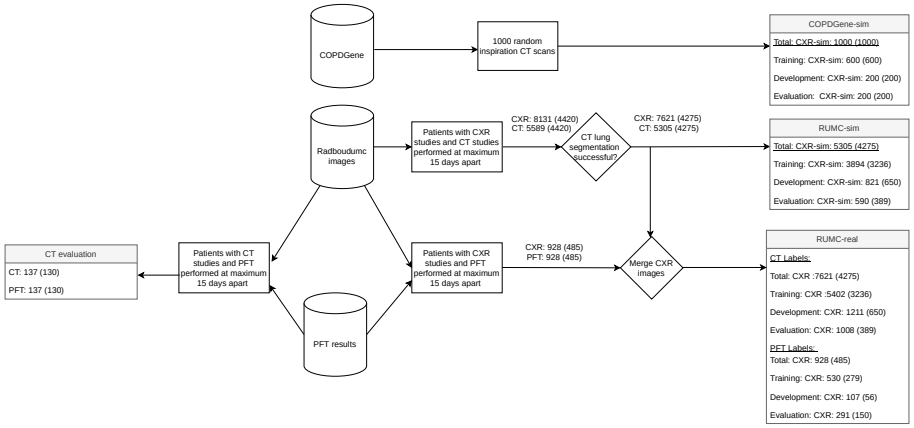
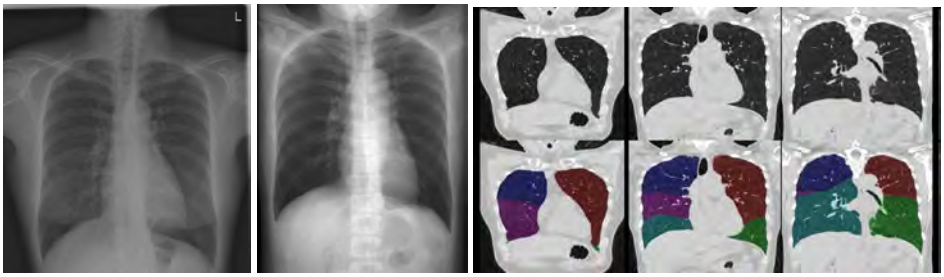


Figure 4.1: Flowchart that shows the criteria to select the data to be used in the experiments. Numbers of images are shown with numbers of patients in brackets. Abbreviations: CXR = chest radiographs, CXR-sim = simulated chest radiographs from CT, PFT = pulmonary function test.

		Experimental Datasets			
		COPDGene-sim	RUMC-sim	RUMC-real (CT-labels)	RUMC-real (PFT-labels)
	Label type	CT-derived	CT-derived	CT-derived	PFT-derived
Possible sources of label error	Patient position difference			**Y**	**Y**
	Inspiration effort difference			**Y**	**Y**
	CT segmentation inaccuracy		**Y**	**Y**	
	Diverse pathologies		**Y**	**Y**	**Y**

Table 4.1: Datasets characteristics in step-wise experiments. RUMC-real (PFT-labels) was used to finetune the models which were pretrained on the RUMC-real (CT-derived) dataset. Y indicates that the condition holds true.



(a) Real CXR **(b)** Fake CXR **(c)** CT scan

Figure 4.2: Real CXR (a), Simulated CXR (b) and coronal CT slices (c) from a patient in the RUMC-real dataset. Lobe segmentation results in CT are illustrated in the bottom row of (c). CT-derived TLV is calculated as the sum of the lobe volumes. CT-derived TLV for this subject was 3.8 liters, while PFT-derived TLV was 4.3 liters.

layer was followed by 3 fully connected layers which mapped the number of features to 512, 128 and 1, respectively.

The dual CNN architecture, which receives both frontal and lateral radiographs as input, consists of two branches with a backbone architecture that is either VGG-Net, ResNet34 or 6-layer CNN, and concatenates the features from these branches before the first fully connected layer. Due to memory limitations, Densenet121, and ResNet50 architectures were not investigated for the dual CNN model. These network architectures were trained with 3 possible inputs (PA CXR, lateral CXR or both, and methods of combining their outputs (see Figure 4.3). Each network outputs a regression value representing TLV in liters.

For each model trained, a hyperparameter optimization was carried out to ensure the best possible result for that architecture/input combination on the validation set. A variety of aspects for training a convolutional neural network were considered as hyperparameters: they were learning rate, optimizer, oversampling technique, and data augmentation as seen in Figure 4.4. Random hyperparameter optimization was employed given a predefined range for hyperparameters for each model (frontal, lateral, and dual CNN) separately.

Each model was trained by optimizing the mean squared error loss between the predicted TLV and the reference standard TLV. The model was trained for a maximum of 300 epochs, terminating if there was no improvement in the validation set performance for 50 successive epochs. We selected the epoch that yielded the least mean squared error in the validation set.

For each of our 3 datasets, the optimal combination of architecture and hyperparameters was identified for each of the 3 possible input types on the validation set. These models were then applied to the held-out evaluation set. In addition, the average of the 2 outputs from the networks using single (frontal or lateral) inputs is calculated and presented as Ensemble CNN output (Figure 4.3).

Our TLV prediction experiments were constructed in a step-wise fashion, to identify potential sources of error as the task becomes increasingly difficult. This is illustrated in Table 4.1. CT-derived volume labels are used in all experiments except the final one where the network is additionally fine-tuned with PFT-derived labels. We begin with the COPDGene-sim dataset, where errors related to patient position and inspiration effort as well as errors related to CT segmentation accuracy and diversity of underlying pathologies are eliminated. In RUMC-sim we introduce the potential for errors from

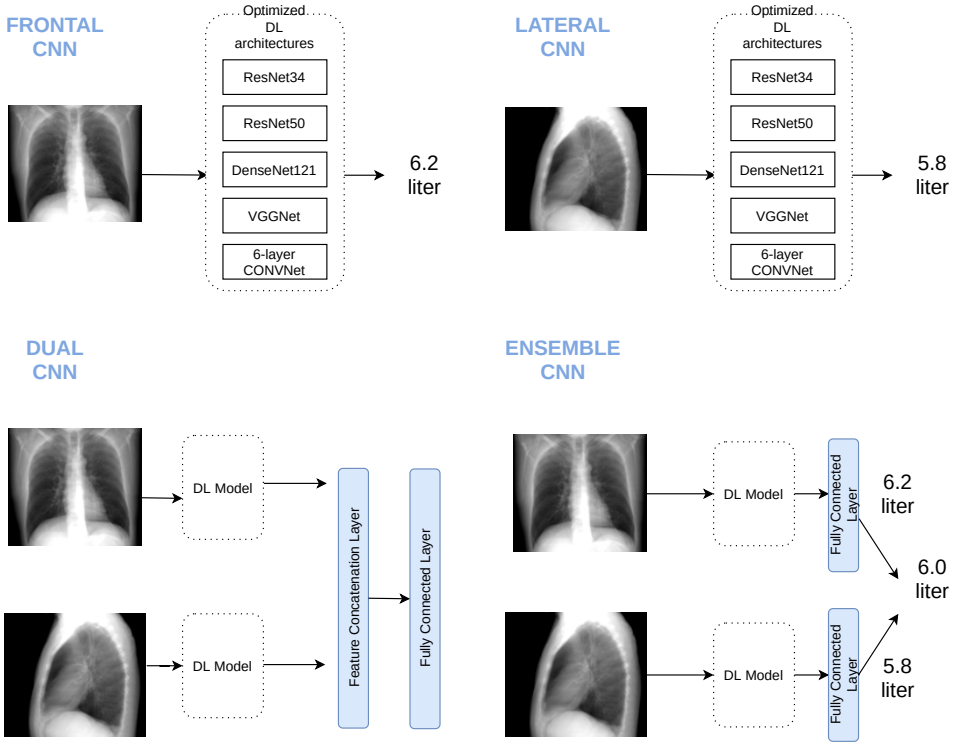


Figure 4.3: Illustration of architecture pipelines. Four different experimental designs were considered: frontal CNN, lateral CNN, dual CNN (combining frontal and lateral models by layer concatenation) and ensemble CNN (combining optimal frontal and lateral models by averaging their outputs).

minor CT segmentation inaccuracies, and from the diverse pathology within the dataset, which is likely to increase the variability in image appearance. Finally in RUMC-real, we first experiment with predicting CT-derived TLV from chest radiographs (RUMC-real (CT-labels)), and subsequently with PFT-derived TLV (RUMC-real (PFT labels)). In this last experiment, since there is only a small number of gold-standard PFT labels available (487 patients), the network trained with CT-labels is used as pre-trained model, and fine-tuned using CXR images with associated PFT-labels.
As an additional experiment, we investigate the relationship between PFT-derived TLV and CT-derived TLV, in a scenario where they are acquired at most 15 days apart from each other, using the CT-evaluation dataset.

4.2.3 Statistical Analysis

Mean absolute error (MAE), mean absolute percentage error (MAPE) and Pearson correlation coefficient were computed to demonstrate the relationship between predicted and reference TLV values. The 95% limits of agreement were estimated by means of a non-parametric method for Bland-Altman plot since the data distribution was not normal, as assessed with Shapiro-Wilk test [468] and quantile-

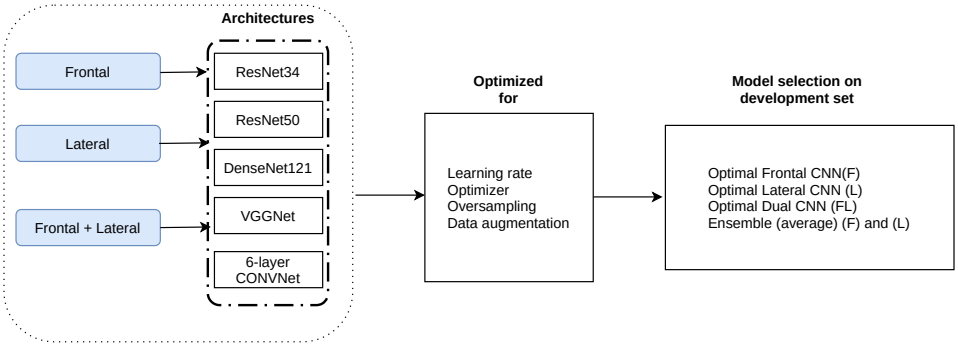


Figure 4.4: Illustration of our model selection process on validation set. Different network architectures were systematically optimized for three different inputs, namely frontal, lateral, and dual (frontal+lateral), separately. Each of them was optimized systematically for hyperparameters, and the model with the least mean absolute percentage error on the validation set was selected.

quantile plot [469].

4.3 Results

Model training for each model, namely frontal CNN, lateral CNN, and dual CNN, took between 8 to 14 hours on the RUMC-sim and RUMC-real datasets (depending on the network architecture), and 2 to 4 hours on the COPDGene-sim dataset using a variety of GPUs such as TitanX, GTX1080, GTX1080ti, GTXTitanx, and TitanV. The mean processing time per test image was 0.3 seconds.

Three trained models (frontal, lateral, dual) were selected for each dataset, based on optimization using the validation set, and applied to the held-out evaluation data. Additionally the outputs of the optimized frontal and lateral models were averaged and presented as "Ensemble" model. The selected architectures, and their performance on the held-out evaluation data are provided in Table 4.2.

In the COPDGene-sim dataset, where chest radiographs were simulated from CT and potential sources of label error were minimal, VGG-Net, 6-layer CNN, and Densenet121 architectures were selected. On the held-out evaluation set the model with the lowest error according to all 3 metrics was the dual CNN with 6-layer CNN architecture. This model achieved a mean absolute percentage error (MAPE) of 2.2% and mean absolute error (MAE) of only 112ml. The scatter plot of model predictions against the reference standard from CT volumes and Bland-Altman-like plot for analyzing differences between the reference standard and predicted TLV measurements are shown in Figure 4.5 (a) and (b), respectively. As shown in Figure 4.5 (b), 95% of differences between predicted and reference standard TLV were from -351ml to 261ml.

On the RUMC-sim dataset, which contains more abnormal images compared to COPDGene-sim, Densenet121, and ResNet architectures were selected from the validation set experiments. As in the COPDGene-sim experiments, the lateral CNN model performed better than the frontal CNN model and the best performance on the evaluation set was, once again, achieved by the dual CNN with MAPE of 2.9% and MAE of 112ml as seen in Table 2 and plotted in Figure 4.5 (c). Limits of agreement

of the differences between the predicted and reference standard TLV measurements was between -348ml to 235ml as shown in Figure 4.5 (d).

Finally, in the RUMC-real dataset, where real chest radiographs were used, dual CNN and ensemble CNN performed very similarly, and the best result obtained (with the least MAPE) with CT-derived labels was achieved by the ensemble CNN, as shown in Table 2. This model achieved 15.7% MAPE, and MAE of 597ml. The model predictions and references for the evaluation set of 1008 CXRs are plotted in Figure 4.6 (a); and the differences between predicted TLV and reference standard is analyzed in Figure 4.6 (b). As shown in Figure 4.6 (b), the model tended to underestimate TLV where reference standard was higher than 6 liters, and overestimate TLV where reference standard was lower than 4 liters.

For the final experiment using PFT-derived labels, the best models trained on the RUMC-real (CT-labels) data for frontal, lateral, dual CNN were used as pretrained models and further fine-tuned on 637 CXR images with PFT-derived labels. The results achieved on 291 CXR images with PFT-derived labels are shown in in Table 2 (RUMC-real (PFT-labels)). The best model on the held-out evaluation set was the dual CNN with ResNet34 architecture and achieved MAE of 408ml and MAPE of 8.1%. The model predictions and PFT-derived reference standard were highly correlated with Pearson correlation coefficient of 0.92 as illustrated in Figure 4.6 (c); 95% of differences between predicted and reference standard TLV measurements were from -1 liters to 938 ml (Figure 4.6 (d)).

Figure 4.7 (a) and (b) shows the results of the comparison between CT-derived TLV and PFT-derived TLV on the CT evaluation set of 137 subjects. These two measurements were well correlated with Pearson's r of 0.78, however, considerable variations were observed between the two measurements for some patients. TLV was consistently underestimated by CT-based measurements where median differences (bias) between CT-derived and PFT-derived was -560ml as shown in Figure 4.7 (b).

Evaluation Datasets (#images)	Model	Architecture	MAPE(%)	MAE(ml)	Pearson's r
COPDGene-sim (200)	Frontal CNN	DenseNet121	4.3	226	0.978
	Lateral CNN	VGG-Net	3.6	198	0.983
	Dual CNN	**6-layer CNN**	**2.2**	**112**	**0.995**
	Ensemble CNN	DenseNet121&VGG-net	2.6	139	**0.992**
RUMC-sim (590)	Frontal CNN	DenseNet121	5.5	220	0.978
	Lateral CNN	DenseNet121	5.0	200	0.984
	Dual CNN	**ResNet34**	**2.9**	**112**	**0.993**
	Ensemble CNN	DenseNet121& DenseNet121	3.8	154	0.989
RUMC-real (CT-labels) (1008)	Frontal CNN	VGG-Net	16.9	650	0.826
	Lateral CNN	DenseNet121	16.8	639	0.831
	Dual CNN	ResNet34	16.1	**592**	**0.855**
	Ensemble CNN	**VGG-Net & DenseNet121**	**15.7**	597	0.851
RUMC-real (PFT-labels) (291)	Frontal CNN	VGG-Net	10.3	509	0.870
	Lateral CNN	DenseNet121	9.2	472	0.875
	Dual CNN	**ResNet34**	**8.1**	**408**	**0.922**
	Ensemble CNN	VGG-Net & DenseNet121	8.5	420	0.907

Table 4.2: Results of the selected models on the held-out evaluation sets. Mean absolute error is calculated against the reference standard for TLV measurements. MAE = mean absolute error (in milliliters), MAPE = mean absolute percentage error, Pearson's r = Pearson correlation coefficient. Bold font indicates best performance per dataset and metric.

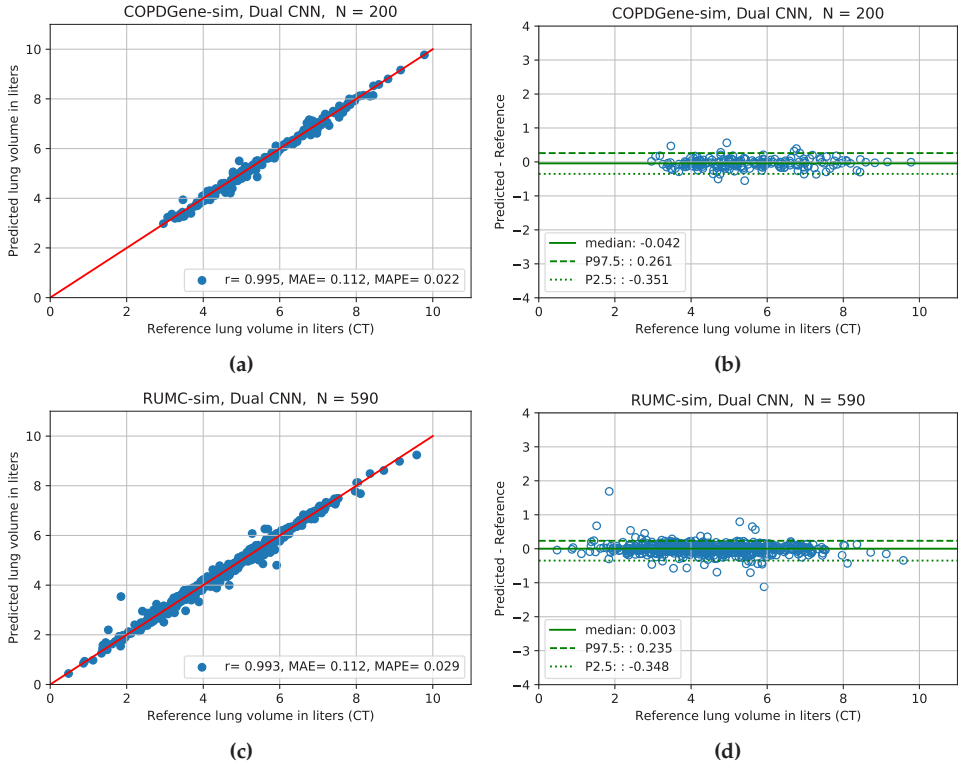


Figure 4.5: Results on simulated datasets in step-wise experiments. Left: The TLV predictions of the best model against the reference standard measurements on the held-out evaluation sets. (a) COPDGene, (c) RUMC-sim. Red line is line of identity (ideal agreement). Right: Bland-Altman-like plot to analyze the differences between predicted and reference standard TLV measurements. Non-parametric method was used to estimate 95% limits of agreement. Abbreviations: r = Pearson correlation coefficient, MAE = mean absolute error, MAPE = mean absolute percentage error, N = number of data, P2.5 = 2.5th percentile P97.5= 97.5th percentile.

4.4 Discussion

This study demonstrated that state-of-the-art deep learning solutions can measure TLV from PA and lateral CXRs, using primarily CT-derived labels and a small number of PFT-derived measures. To demonstrate the sources of error, the experiments were conducted in a step-wise fashion with increasing levels of complexity. Using simulated CXRs eliminated potential error related to the patient position or inspiration level between the CT and CXR image acquisition. Results on both simulated datasets show extremely low error (MAPE of 2.2% and 2.9%) and high correlation with the reference labels (r=0.99 and r=0.99). The slightly better performance on the COPDGene-sim dataset may be attributed to the fact that this dataset contains a limited range of pathologies and that the CT segmentations were manually corrected, meaning that even very small inaccuracies were eliminated.

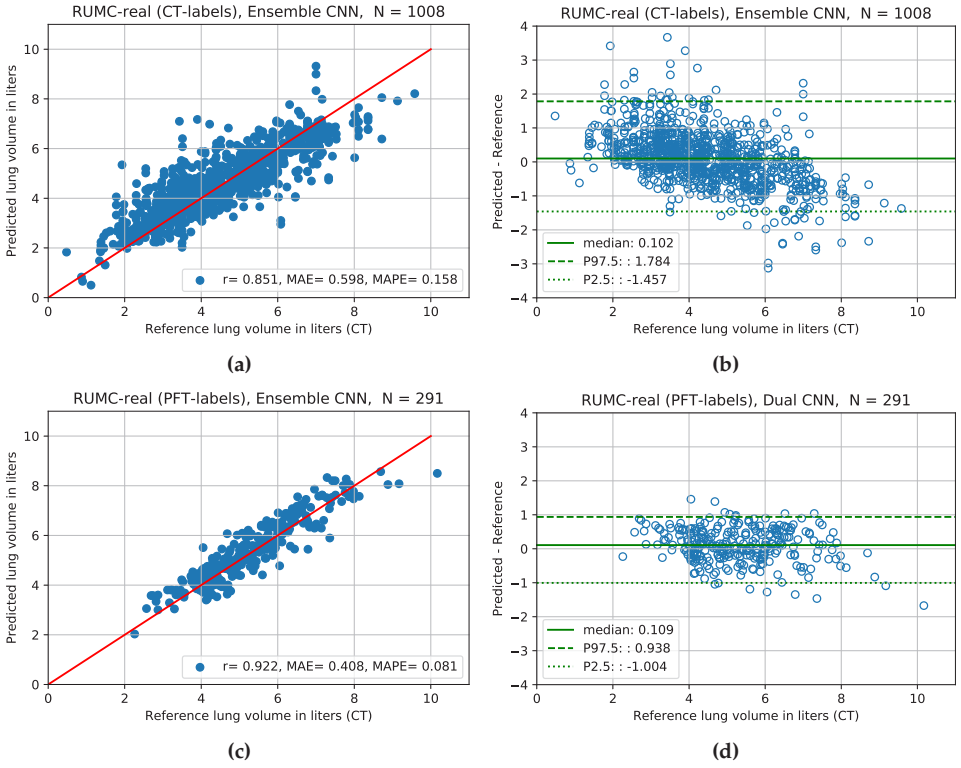
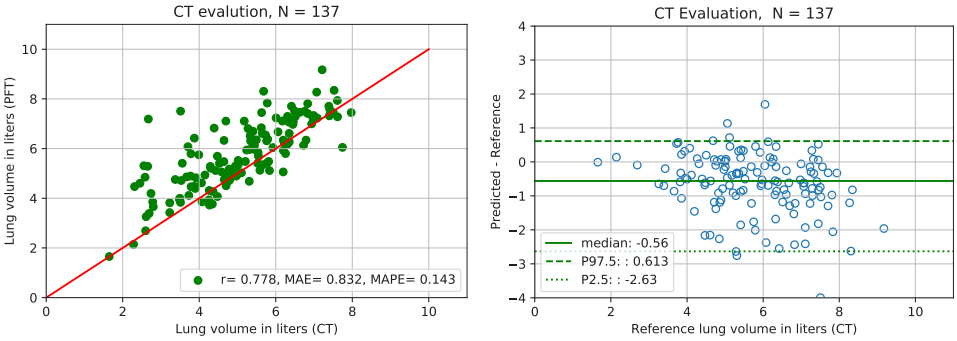


Figure 4.6: Results on real datasets in step-wise experiments. Left: The TLV predictions of the best model against the reference standard measurements on the held-out evaluation sets. (a) RUMC-real, (c) RUMC-real (PFT-labels). Red line is line of identity (ideal agreement). Right: Bland-Altman-like plot to analyze the differences between predicted and reference standard TLV measurements. Non-parametric method was used to estimate 95% limits of agreement. Abbreviations: r = Pearson correlation coefficient, MAE = mean absolute error, MAPE = mean absolute percentage error, N = number of data, P2.5 = 2.5th percentile P97.5= 97.5th percentile.

In the dataset of clinical CXR with CT-derived volumes (RUMC-real dataset) we see a substantial increase in the prediction error with MAPE of 15.7%, which we attribute largely to the difference in patient position and inspiration effort between the CT and the CXR image acquisition. It is likely that the degree of inspiration in the CXR and CT images is different, particularly given that there is known to be a high intra-individual deviation in TLV between routine CT scans ([470]). The indication from this experiment is that CT-derived labels are useful, but not optimal, to learn the TLV from CXR. As an additional check we investigated the relationship between CT-derived and PFT-derived volumes in 137 cases where both were available. This provides results in line with previous studies on CT-derived lung volumes [470, 471]: although CT-derived lung volume and TLV are well correlated (r=0.78), there are considerable differences in some patients.



(a) **(b)**

Figure 4.7: CT-derived TLV against PFT-derived TLV on CT evaluation dataset. Left: Comparison of CT-derived total lung volumes with Pulmonary Function Test total lung volumes on the CT evaluation set. Right: Bland-Altman-like plot to analyze differences between CT-derived and PFT-derived total lung volume. Abbreviations: r = Pearson correlation coefficient, MAE = mean absolute error, MAPE = mean absolute percentage error, N = number of data, PFT = pulmonary function test, P97.5 = 97.5th percentile, P2.5 = 2.5th percentile.

To overcome the issues with the CT-derived labels on the RUMC-real dataset we further fine-tuned the best networks from that experiment with PFT-derived labels. Evaluation on an independent dataset of 291 subjects that were not used for training showed that the error of the estimated TLV from CXR relative to the measured TLV from PFT is reduced considerably, achieving MAPE of 8.1% and Pearson's correlation coefficient of 0.92. This algorithm is publicly available at "https://grand-challenge.org/algorithms/cxr-total-lung-volume-measurement/".

In all experiments the model was optimized to use the best performing architecture and input. In the experiments using simulated CXR images, it is notable that the networks using lateral images as input perform better than the networks using frontal images. This may indicate that the lateral projection image contains more information related to CT-derived TLV. However we note also that in all experiments the combination of frontal and lateral images produced the optimal results, either by use of a dual-CNN or through an ensemble.

Previous literature has investigated predictive equations for measurement of TLV from chest radiographs using manual measurements. One study [472] investigated performance with simulated chest radiographs to predict CT-derived TLV. Their method, which required manual measurements, had an inferior performance (MAPE of 5.7%) on their dataset compared to our results obtained in the COPDGene-sim and RUMC-sim datasets (MAPE of 2.2% and 2.9%). For studies which investigated predictive equations to estimate PFT-derived TLV from real CXR [458, 461, 462], the coefficient of correlation between predictions and reference standard (body plethysmography or helium dilution technique) generally ranged from 0.80 to 0.93 (compared to our method with 0.92). Sample sizes in these papers ranged from 21 to 100 patients. However, it should be noted that many of these studies used spirometric control to regulate the level of inspiration during CXR acquisition. In fact, one study [473] has shown that without spirometric control the correlation of predicted TLV and PFT-derived reference standard was only 0.47, compared to 0.82 with spirometric control. In this work, however,

we experiment with routinely taken chest radiographs (with no spirometric control), and produce TLV predictions which are highly correlated (r=0.92) to PFT-derived results. Our work is the first to demonstrate automated measurement of TLV from chest radiographs and achieves a comparable or lower error range with a remarkably larger sample size compared to previous literature.
There are several limitations in this study. First, the algorithms were evaluated on an internal dataset from a single institution; validation of the models on an external dataset is an important next step to assess the algorithm robustness. Second, the datasets were constructed from routinely taken studies with the assumption that TLV would not change in 15 days, which might not hold true for extreme cases. This selection criterium also yielded an under-representation of healthy subjects but reflects a clinical population in which TLV measurements are of clinical interest. The PFT-derived reference standard measurements were obtained using the helium dilution technique which might underestimate TLV in patients with severe airway obstruction. Furthermore, inspiration levels were not controlled in a similar fashion to PFT in these routine chest radiographs, which could have introduced a source of error in our predictions, but this represents regular clinical practice. One possible solution to address this issue would be to develop an automated algorithm to assess the inspiration level on CXR, for example by rib counting [342]. Moreover, our held-out evaluation set was constructed with patients assessed for lobectomy since their PFT results were readily available; future research should address the evaluation of the algorithm on a population with other clinically relevant pathologies, including fibrosis.
In conclusion, we demonstrated that TLV can be automatically estimated from CXR using a deep-learning approach, with an accuracy that is superior or comparable to the previous literature using semi-automated methods. Further, we showed that the deep learning system can be trained primarily with CT-derived labels from automatically segmented chest CT images, and fine-tuned on gold-standard PFT-derived labels. This automated system could be routinely applied to clinical chest radiographs and serve as a tool for identifying temporal change in total lung volume in patients with restrictive and obstructive lung diseases.

Acknowledgements

We thank Weiyi Xie for his help in dataset collection, and Jesus Lago for useful discussion on statistical analysis. This work was supported by the Dutch Technology Foundation STW, which formed the NWO Domain Applied and Engineering Sciences and partly funded by the Ministry of Economic Affairs (Perspectief programme P15-26 'DLMedIA: Deep Learning for Medical Image Analysis').

Nodule detection and generation on chest X-rays: NODE21 Challenge

5

Authors: Ecem Sogancioglu, Bram van Ginneken, Finn Behrendt, Marcel Bengs, Alexander Schlaefer, Miron Radu, Di Xu, Ke Sheng, Fabien Scalzo, Eric Marcus, Samuele Papa, Jonas Teuwen, Ernst Th. Scholten, Steven Schalekamp, Nils Hendrix, Colin Jacobs, Ward Hendrix, Clara I Sánchez, Keelin Murphy

Original title: Nodule detection and generation on chest X-rays: NODE21 Challenge

Published in: IEEE Transactions on Medical Imaging, 10.1109/TMI.2024.3382042, 2024

5.1 Introduction

Lung nodules may be an early manifestation of lung cancer, the biggest cancer killer among both women and men [474]. Because early stage lung cancers are often asymptomatic, most lung cancers are diagnosed when the disease is already metastasized. However, the mortality rate varies significantly depending on the stage of the cancer when it was detected. While the 5-year survival rate of localized lung cancer is 59.0%, it is only 5.8% when the disease has metastasized [475]. This statistic highlights the crucial role of early detection of lung cancer in reducing mortality rates.

While chest CT scans are preferred over chest X-rays for lung cancer screening [476, 477], the inclusion criteria for CT screening programs are typically strict and a considerable number of patients who develop lung cancer in their lifetime might not be eligible for such screening programs. In contrast, Chest X-rays (CXR), being the most common imaging study acquired, play a crucial role for the detection of early lung nodules through routine clinical practice. Pulmonary nodules are frequently encountered as incidental findings in patients undergoing CXR as routine examination or for issues unrelated to lung cancer.

The detection of lung nodules on CXR is a challenging task because superimposition of anatomical structures may obscure lung lesions as seen in Figure 5.1. In fact, several studies [25, 478] show that radiologist sensitivity for detecting nodules can vary from 36% to 84% on various datasets. Other work [7, 8] shows that 19%–26% of lung cancers visible on chest radiographs were, in fact, missed at their first readings.

Considering its high clinical relevance and potential impact, nodule localization has been one of the most widely studied topics on automated CXR analysis for decades [36]. This trend has changed in the last few years, however, with the release of publicly available CXR datasets (Chest X-ray14, CheXpert, MIMIC-CXR) [6, 33, 34] of which many publications made use [11]. The annotations of these datasets were obtained using natural language processing (NLP) techniques on radiology reports, and image-level labels are generated with more than 10 different abnormalities including lung nodules.

The volume of publications inspired by these datasets demonstrates their value to the research community, particularly as large-scale training sets. For development of clinically applicable algorithms, however, evaluation must be extremely rigorous and evaluation dataset labels must be of a very high standard. Many of the works using these public datasets used the NLP-generated labels [146, 194, 195, 197, 446] for evaluation or, at best, radiologist assessment of CXR images [78, 104, 153, 189]. The pitfalls of NLP labeling have been well documented [86], failing largely because the radiology report is not always a complete description of the entire image, but often refers only to a specific clinical question. Similarly named conditions such as pulmonary emphysema and subcutaneous emphysema are also known to be confused by such labelling systems [11, 86]. Radiological reading of CXR, while substantially better, also has limitations as a gold-standard; many nodules have a very subtle appearance on CXR and radiologist sensitivity and agreement is low in the absence of a CT or pathology-based gold standard.

For clinically relevant results a reference standard based on CT or on proven lung cancer is optimal, while algorithms should ideally pinpoint nodule locations to provide explainability and improve efficiency if acting as a second reader. Several studies have described work including such datasets used for evaluation of either a commercial or research algorithm for nodule localization [105, 337, 398, 479]. However, since this data (or the evaluation platforms) remains inaccessible to the public, direct comparison with other nodule detection algorithms on the same dataset is not possible. The time, cost, and patient privacy issues associated with the collection of large datasets with strong reference

standards limits the number of available dataset for evaluation of algorithms for many researchers. The data-hungry nature of modern deep-learning technologies means that researchers also require large training datasets. The labeling requirements in training data are generally less stringent than those in evaluation sets, however, NLP labels are generally insufficient for extracting useful cases of solitary pulmonary nodules, and radiological reading is certainly required to obtain training data for localization algorithms (those which identify the location of the nodule(s)). Given the challenges and expenses of acquiring radiological labels and the scarcity of cases with subtle or concealed nodules, there is significant interest in exploring the insertion of simulated nodules into CXR images as a training tool. Such a dataset could be constructed to meet the needs of the user in terms of numbers of images, nodule sizes, locations, and conspicuity. However, this area, while holding strong promise, is still a newly explored domain, indicating an early stage in its development and application in the field of CXR [480, 481].

Motivated by these observations, we organized a public challenge, NODE21, which consists of two tracks: nodule detection and nodule generation. The aim of the NODE21 challenge is to improve the state of the art for the detection of solitary nodules on CXR. The nodule detection track assesses the performance of state-of-the-art nodule detection systems for CXR whereas the nodule generation track determines the utility of simulated nodule training data on the performance of nodule detection systems. Radiologist-annotated training data is made publicly available and private test sets have a CT-based reference standard. Algorithm evaluation is provided through the Grand-Challenge platform [482] and the challenge design ensures that the algorithms and code are publicly available and reproducible.

In this paper, we discuss the results of the detection and generation tracks of the NODE21 challenge. Additional extensive experiments are performed using various combinations of detection and generation algorithms to analyze the impact of the generated images on the detection performance and provide guidance on how best to incorporate simulated data as part of training data.

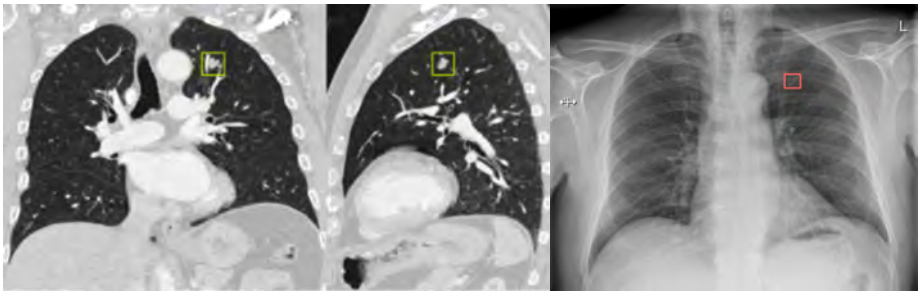


Figure 5.1: Left and Middle: Coronal and sagittal CT images showcasing a lung cancer lesion highlighted in a yellow box. Right: A posterior-anterior chest X-ray of the same patient, conducted a few days after the CT scans, with the identical lesion marked in a red box. The CT scans distinctly reveal the tumor, making it hard to overlook. Conversely, the Chest X-ray image, even though the nodule is not obscured by major organs like the heart, only faintly displays it due to the overlay of surrounding structures. This illustrates the inherent difficulty in detecting nodules through chest radiographs.

5.2 Data

There are three datasets associated with NODE21: the training set, the experimental test set, and the final test set. The training dataset is made public, while both test sets are private and only accessible to challenge organizers. Participants could evaluate their method against the experimental test set multiple times during the preparation of their algorithm to ensure correct working and to allow method or parameter tuning. Submission to the final test set was allowed only once, at the end of the challenge, to ensure that participants could not optimize their method for that set.
The aim of the NODE21 challenge is to improve the state of the art for the detection of solitary nodules on CXR, a key factor in early lung cancer diagnosis. With this in mind, data selection for the challenge deliberately excluded images with a predominantly abnormal pattern of consolidation or infiltrates, clusters of nodules or more than three visible nodules. This ensures the challenge focuses on detecting small and clinically relevant nodules, while excluding clear abnormal cases that could cause models to overfit to obvious signs instead of learning the finer details necessary for early detection. Data selection was strategic, focusing on nodules sized 6mm-30mm to maintain clinical relevance and excluding obvious abnormalities, as they often indicate advanced disease stages [483].
All datasets were pre-processed using the publicly available OpenCXR library [484] image standardization process. This process first removes homogeneous border regions and then applies energy-based normalization of image intensity values to standardize image appearance [485] using a lung segmentation [92] as an intermediate step. The images were then cropped to the region of the lung fields and resized to 1024 x 1024 pixels preserving aspect ratio and using padding on the shorter side. The training set is provided with both the original images and the pre-processed versions available, however, participants were advised that their algorithm would be tested on images that had been pre-processed in this way. Previous work [485] has demonstrated that such preprocessing makes image analysis systems more robust to variation in test data from different X-ray equipment, for example.
The NODE21 training dataset was made available on the Zenodo data sharing platform [486], with a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
The following subsections describe the three datasets in more detail.

5.2.1 Training dataset

This dataset consists of postero-anterior (PA) chest radiographs (both with and without nodules) with bounding boxes provided to identify the nodule locations. The original data selection was made from public datasets where we had explicit permission to redistribute or where the dataset license provided permits it. These public datasets are as follows:

- JSRT [72]
- PadChest [35]
- ChestX-ray14 [339]
- Open-I [487]

To select images likely to contain nodules, data from each of these sets was chosen to include PA images with a label indicating nodule and (where possible) other labels selected to exclude confounding abnormalities such as consolidation or infiltrates. For a detailed description of the filtering process please refer to the annotation page [488].

Since the JSRT labels were provided by radiological examination with CT as the reference standard and with nodule location information available we did not re-label or re-annotate any of these cases, however, five cases were excluded as the nodules were outside the desired size range. All other selected data consist of chest X-ray images (without corresponding CT scan) and was reviewed in a reader study on the Grand-Challenge platform [482] by a chest radiologist with over 30 years of experience (ETS). The radiologist was asked to identify visible nodules on a chest X-ray by drawing bounding boxes around them. Images, where no nodule could be seen, where the nodule was outside the desired size range, or where there were significant confounding abnormalities, were excluded. Figure 5.2 displays all lung nodule bounding boxes from the training data superimposed on a single normalized chest X-ray image. While a majority of the nodules are located in the central lung regions, the dataset also includes instances of nodules positioned behind the diaphragm and heart.

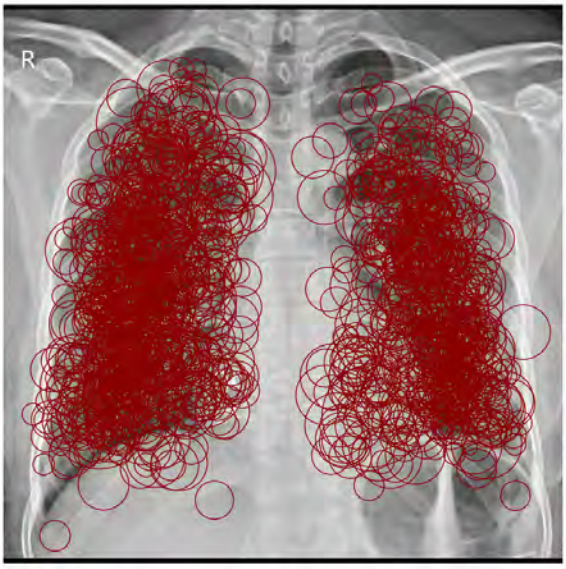


Figure 5.2: Normalized chest X-ray with lung nodule bounding boxes from training data, showcasing the approximate distribution of locations and size variations of nodules.

In addition to the nodule images, a selection of normal images was also included for training. These images were chosen from the same four public datasets using the label of non-nodule for JSRT and of 'normal' or 'no finding' for the other datasets (please refer to [488] for a detailed explanation of filtering). Since the PadChest and ChestX-ray14 datasets had very large numbers of images matching the applied filters (28,688 and 22,452 respectively) a random selection of 1500 normal images was chosen from each in the initial selection. All selected images were then reviewed in a 3-step process to reduce the number of FP as follows: 1) A member of the NODE21 team specializing in deep learning for chest X-ray image analysis with over four years of experience briefly reviewed each image and rejected any with obvious nodules or abnormalities. 2) The baseline nodule detection system (see section 5.3.1 for further details of this system) was run on the remaining images to identify suspicious regions, which were again reviewed by a team member. 3) Any case where the team member was

uncertain whether a nodule was present was presented to a radiologist (ETS) for review and rejected if an abnormality was present.
Following this selection and review process, 1134 images (containing 1476 nodules) and 3748 non-nodule (normal) images were obtained. The numbers per dataset pre- and post-review are provided in Table 5.1.
Participants received the original CXR images, complete with identifiers for tracing them back to their respective public datasets, alongside the OpenCXR standardized image versions. All related metadata is publicly available and readily accessible. The dataset features a balanced gender distribution, with 48% female and 52% male participants, and an average age of 58.6. Additionally, the dataset encompasses a variety of nodule sizes, ranging from 6mm to 30mm.

Table 5.1: Training data selection process. Nodule and non-nodule (normal) CXR images were selected from 4 public datasets: JSRT(J), PadChest(P), ChestX-ray14(C) and Open-I(O) and reviewed before inclusion in the challenge. Review steps are described in detail in the text. Figures indicate the number of images with the number of nodules in brackets.

Source	J	P	C	O	Total(Nodules)
Nodule Data					
Initial Selection	154	908	1586	82	2730
After radiology review	149	314	617	54	1134(1476)
Non-Nodule (Normal) Data					
Initial Selection	93	1500	1500	1164	4257
After 3-step review	93	1366	1187	1102	3748

For the generation track, we randomly sample 1000 images from the non-nodule images in the training dataset. Nodule bounding boxes were generated using our nodule location generator as shown in Figure 5.3. Up to 3 bounding boxes (1-3) were selected per image to be used to generate nodules inside. The images and the bounding boxes were provided to the participants at the test time where their submitted generation algorithm was expected to generate nodules inside the requested bounding boxes. Further, for the generation track, we provided a public set of NODE21 CT patches which participants were free to use as part of their generation algorithm. These are cropped 3D patches containing nodules from the public Luna16 CT dataset [489]. The patches were 50 x 50 x 50 mm, resampled to voxels of 1 x 1 x 1 mm. A total of 1186 nodule patches are provided together with associated nodule segmentations via the Zenodo data sharing platform [486].

5.2.2 Experimental test set

The images for the experimental test set were collected for this study is derived from standard clinical procedures at Radboud University Medical Center (RUMC) in Nijmegen, the Netherlands. For this

set a reference standard of CT was required to confirm the presence of a nodule, so patients who had undergone both frontal CXR and a CT scan within a maximum of 60 days of each other were identified. These were further filtered to those patients whose corresponding CT report contained the word 'nodule'. A nodule detection system [490] was used to identify nodules on CT scans as a means of assisting the annotating radiologist (ETS). The radiologist was then provided with the 200 CXR images alongside the CT slices with detected nodules identified and asked to find the corresponding nodules on CXR and annotate them. The radiologist could also access the full CT scan if needed.
The annotator was asked to annotate solitary nodules, solid or subsolid, located in a region of otherwise normal appearance. Images with a predominant pattern of abnormal tissue, or clusters of nodules were excluded, as such very abnormal cases are not of interest. The cases where a nodule was only visible on CT but could not be seen on CXR even by estimating the approximate location were excluded from the study to maintain a high level of specificity and relevance for CXR-based detection systems. This approach ensures that only confirmed nodules, which are consistent across both imaging modalities, are included, thus enhancing the reliability of the dataset.
To select normal CXR images for inclusion in the experimental test set the Radboudumc CXR reports were searched for the text 'Normal image of heart and lungs' or the text 'Normal cardiopulmonary image' since these phrases had been observed to be used frequently to report a completely normal CXR image. From the results, a random selection of 120 PA CXR images from unique patients was made and provided to a chest radiologist for review. Images with any suspicion of nodule or other confounding abnormality were rejected.
A total of 281 CXR images were selected in this way, 166 of which contained (248) nodules and the remaining 115 were normal. The details are provided in Table 5.2

5.2.3 Final test set

The final test set was originally collected for a previous study [413]. It consists of 300 CXR images from 4 different hospitals in the Netherlands. A positive case is defined by the presence of a solitary pulmonary nodule visible on the PA image and confirmed by CT acquired within 3 months of the CXR. A negative case is one without nodules or other substantial pathology, confirmed by CT within 6 months of the CXR acquisition. The nodule locations were provided through the original study data and used for the NODE21 challenge. In addition, the findings of a total of 12 independent readers (6 radiologists and 6 radiology residents) are available, indicating whether or not they believe a nodule is visible on the CXR. These additional findings are used in this paper to provide an independent comparison between computer algorithms and human experts, but not as the reference standard for the challenge evaluation process. The final dataset excluded two images (which could not be obtained) from the original set in the paper and consists of 111 nodule-positive images (each with one nodule) and 187 non-nodule images as summarized in Table 5.2.

5.2.4 Additional Experiments Data

For additional experiments performed in this paper, in order to experiment with larger datasets, we utilized VinDr-CXR dataset [491]. VinDr-CXR is a publicly available dataset that contains 18k posterior-anterior view chest X-rays with both localization and classification labels for thoracic diseases. The images were labeled by a group of 17 experienced radiologists for the presence of 22 critical findings and 6 diagnoses. From this dataset, we selected 10606 images which were labeled with 'No finding', (meaning the selected CXR images are expected to contain no abnormalities). Nod-

Table 5.2: Experimental and Final test set statistics. Figures indicate the number of images with numbers of nodules in brackets. Before and After indicates the number of images before and after the radiology review, respectively. Details of the selection and review process are provided in the text.

		Nodule Data	**Non-nodule Data**
Experimental Test Set	Before	200	120
	After	166 (248)	115
Final Test Set	After	111 (111)	187

ule bounding boxes were generated using our nodule location generator as shown in Figure 5.3. Up to 3 bounding boxes (1-3, 7-17mm) were selected per image to be used to generate nodules inside. This size selection is deliberately smaller than the training dataset criteria to represent more subtle nodule characteristics in the dataset.

5.3 Challenge Setup

The NODE21 challenge was hosted on the Grand-Challenge platform [482], which has hosted over 330 medical image analysis challenges since 2007. The challenge website is publicly accessible online at `https://node21.grand-challenge.org/`([483]), providing access to all information and functionality, including data, evaluation, and leaderboards. On Grand-Challenge, interested parties could register and find a general overview of the challenge including the deadlines, a description of the datasets, the evaluation metrics, and the preprocessing code. Through the website, the participants could submit their algorithms and access a forum to post questions or comments. The NODE21 challenge continues to accept unlimited submissions for the experimental test set and limited number of submissions on final test set, supporting continuous benchmarking efforts.

One aim of the NODE21 challenge was for competing algorithms to be fully reproducible and publicly available. To this end, only algorithm submissions in the form of a docker container were accepted. Docker containers encapsulate the software, allowing it to run uniformly and consistently on any system that supports Docker which plays a significant role in bridging the reproducibility gap in scientific research, reinforcing the credibility of the study, and fostering an environment conducive to collaborative scientific exploration and advancement. Once submitted, the container would automatically run on the private test set and generate results for the leaderboard. The submitted solutions were required to be linked to a public GitHub repository with a version tag and an Apache 2.0 or MIT license. The submitted algorithms are thus open-source and publicly available and can be tested out by interested users on the Grand-Challenge platform. The GitHub repository is available at https://github.com/node21challenge.

The NODE21 challenge was divided into two tracks, a detection track, where participants submitted algorithms for detecting nodules in CXR, and a generation track, where participants submitted algorithms for generating realistic nodules on normal CXR images. Interested parties could enter either or both tracks.

The challenge was open from October 19, 2021, when the training dataset was released, until January 25, 2022. Participants were allowed to submit their methods for evaluation on the experimental test

set starting from December 2nd, to test their model performance as well as to make sure their docker submission worked as expected. Repeated submissions for evaluation on the experimental test set were permitted. From January 10th to 25th, participants were able to submit their final best algorithm to the final test set where only a single submission per participant was allowed. Submissions (to either phase) were not permitted after January 25, 2022.

5.3.1 Baseline models

For each of the detection and generation tracks, a baseline model was provided with code available at https://github.com/node21challenge. This provided a benchmark performance for each track as well as template code for participants to demonstrate how to build working docker containers for submission to Grand-Challenge. The baseline methods are described in more detail in section 5.4.1

5.3.2 Detection track

The detection track participants were required to submit an algorithm that reads a chest X-ray as input, and returns a list of bounding boxes for identified pulmonary nodules, with a likelihood score associated with each one.

Evaluation Metrics

For each algorithm submitted to the detection track, the following metrics were calculated: Area under the receiver operating curve (AUC) and sensitivities at average false positive (FP) rates of 0.125, 0.25, and 0.50 nodules per image.

To calculate the AUC, an image score was assigned to each chest X-ray by choosing the maximum bounding-box probability among detected nodules in that image. If there was no nodule prediction for an image, the image score was set at 0. These image scores were thresholded to obtain the receiver operating curve and, hence, the AUC.

To obtain sensitivities at different FP rates, free-response operating curve (FROC) analysis was used. If more than one predicted bounding box overlapped a reference bounding box with intersection-over-union (IOU) > 0.2 then only the prediction bounding box with the maximum probability among them was retained. Any prediction bounding box was then considered as a true positive if it overlapped with a reference standard bounding box at IOU > 0.2, otherwise, it was considered a false positive. Using the numbers of true and false-positives, we then calculated the average sensitivity at 3 predefined false positive rates: $1/8$, $1/4$, and $1/2$ FPs per image. For cases where the FROC curve did not extend to the specified false positive rate the highest sensitivity value from the curve was used.

The final metric used to rank participants on the leaderboard was calculated as follows:

$$rank\_metric = (0.75 * AUC) + (0.25 * S) \tag{5.1}$$

where S is the sensitivity at 0.25 FP per image. This gives a heavier weighting to the algorithm's ability to identify images containing nodules (which is the most clinically important task) but also considers its ability to correctly pinpoint the nodule locations.

5.3.3 Generation track

The generation track participants were required to submit algorithms that take a frontal chest X-ray and one or more bounding box locations as inputs and return the same X-ray image with synthetically generated nodules inserted at the requested locations.
The locations of the nodules to be generated were pre-determined by the challenge organizers. In order to select plausible locations on the input chest X-rays, deep learning based lung and heart segmentation algorithms [92] were run. The resulting segmentation maps were used to select the region where nodules could potentially appear, including the entire lung segmentation and the heart segmentation to the lowest detected point of each lung (see Figure 5.3). In order to include the lung regions obscured by the heart and diaphragm, for each lung the most upper point of the heart segmentation, the lowest detected point of the lung, and the leftmost point of the right lung (rightmost for the left lung) were used, which creates rectangular squared like region at the bottom of the lung areas. Up to three square bounding box locations (random from 1 to 3 nodules per image) with random sizes (7-17mm) were selected from this region and their locations and sizes were provided with the image for the generation algorithms to be trained with. We made sure that the nodule boxes fit inside the boundaries.

Figure 5.3: Process used to identify locations where nodules should be generated. Step 1 applies heart and lung segmentation on a given CXR image and indicates the boundaries including heart and lungs to below the diaphragm. Step 2 receives the segmented CXR image and randomly places 1-3 square boxes (in the size range of 7-17mm) in the bounded regions.

Evaluation Metrics

Generation track algorithms were evaluated by training a detection system with the generated images, including synthetic nodules, and evaluating the resulting nodule detection system as described in section 5.3.2. This evaluation metric is based on the principle that a high-quality generation system should create images that can improve the performance of a detection system when included as training data.
For the evaluation of the generation algorithms, 1000 chest X-rays that are free of nodules were randomly selected from the NODE21 training dataset, and bounding box locations where nodules should be generated were pre-determined. This set of images and the locations were kept private and only visible to challenge organizers.
Once a generation algorithm was submitted, it was run on this dataset to output 1000 chest X-rays with generated nodules. The resulting generated images were used to train our baseline nodule

detection system. This trained nodule detection system was then evaluated on the appropriate test set (experimental or final, depending on which phase the participant submitted to). The same evaluation metrics as used in the detection track (see section 5.3.2 were calculated and detection performance was equated with generation performance for leaderboard ranking.

5.4 Challenge Submissions

In total, 302 participants from various countries joined the challenge before the submission deadline. There were over 230 submissions to the experimental test set from both tracks combined. In the final test phase, 10 teams from 7 countries (6 teams for the detection track and 4 teams for the generation track) submitted a solution.
For inclusion in this paper, the best ranking methods from each track are selected for analysis and further experimentation. From the detection track, we include the top three performing methods as well as the baseline. Most of the generation track methods had a poor performance compared with the baseline and only one additional method (the top ranking method) was selected, along with the baseline, based on its methodology and performance.

5.4.1 Baseline Methods

DB (Baseline Detection Algorithm) This model is the open-source baseline detection algorithm, which was provided by the challenge organizers before submissions were opened. It is based on a Faster R-CNN architecture [52] which uses ResNet50[20] as the backbone. The model was trained on the OpenCXR-preprocessed version of the NODE21 training dataset.
In order to tackle the data imbalance issue, images with nodules were oversampled until the number of negative images was reached. The model was trained for 30 epochs, and early stopping was used in case of no improvement in the validation set performance for 5 consecutive epochs.

GB (Baseline Generation Algorithm) This model is the open-source baseline generation algorithm that was provided by the challenge organizers before submissions were opened. The method requires 3D nodule templates segmented from CT scans. The algorithm is based on a simple cut and paste principle[480, 492], where nodules are generated from 3D nodule templates from CT scans and superimposed into a chest X-ray at the requested location. For each bounding box, a randomly selected nodule, which was cropped from a CT scan, was resampled so that it fit into the size of the given bounding box. As a next step, the resampled nodule was superimposed inside the bounding box, and the Poisson image blending technique [493] was applied to reduce local discrepancies around the corner regions.
This model used 3D nodule templates which were cropped from LUNA16 dataset and this dataset was also provided to the NODE21 participants together with the training dataset as described on Section 5.2.1.

5.4.2 Detection Track Top Submissions

In this section, we describe the top three detection solutions submitted. Methods D1, D2, D3 denote rank 1, rank 2 and rank 3 algorithms, respectively. Further details regarding the training strategies of these methods can be found online in the NODE21 challenge page [494].

D1 This model was placed as the rank 1 algorithm in the final leaderboard [495]. The submitted algorithm was an ensemble of 20 different models based on Faster RCNN [52], RetinaNet [57], YOLOv5 [55] and EfficientDet-D2 [496] architectures. Each model was trained using 5-fold cross-validation; Yolov5 were trained with a resolution of 640 x 640 and 1024 x 1024. The final ensemble used 5 models from each fold from Faster R-CNN, RetinaNet, Yolov5 (640 x 640 resolution), 4 models from Yolov5 (1024 x 1024 resolution), and 1 model from EfficientDet-D2 and the 20 model predictions were ensembled using weighted box fusion [497].
All the models were trained using the OpenCXR-preprocessed version of the NODE21 training dataset and no additional preprocessing steps were performed.
The Faster R-CNN and RetinaNet implementations utilized a pretrained ResNet-50 model as a backbone network. All the models except RetinaNet leveraged transfer learning, and pretrained the models on VinDr-CXR dataset [491]. All the model parameters were kept trainable during training.
In order to tackle data imbalance, the participants generated artificial nodules on 1000 randomly selected healthy images from the training dataset by using the GB.
For all the models except YOLOv5, various data augmentation schemes were applied such as cropping and padding, horizontal flipping, random rotation, blurring, and cutout augmentation. For the Yolov5 model, the original augmentation strategies were used, and test time data augmentation was applied.

D2 This model was ranked in 2nd place on the final leaderboard. The submitted algorithm is an ensemble of 33 YOLOv5 models, which were trained using 33 folds where each fold contains 85% of the nodules from the training dataset. The predictions of the 33 models were merged using a non-maximum suppression method.
Data balancing was tackled by undersampling the number of negative images in the training dataset. The model was trained from scratch using OpenCXR preprocessed version of the NODE21 training dataset. No further preprocessing steps were applied.

D3 This model was ranked 3rd on the final leaderboard. The algorithm is an ensemble of six models based on MaskRCNN and RetinaNet architectures with ResNet50 backbone. Three models per architecture were trained where each model was trained using different thresholding values for normalizing the dataset. The thresholding on the pixel intensities was performed based on predefined upper and lower quantile values, which was then followed by uniform normalization. Predictions of the six models were merged using non-maximum suppression.
All the models were trained using transfer learning. They were first pretrained with projected CT scans on the DeepLesion dataset [498], and then were further fine-tuned on the Luna16 dataset [489] to have better weight initialization. The resulting models were then trained on NODE21 training dataset where all the layers in the networks were kept trainable. Data imbalance was tackled by oversampling the positive cases for the training of Mask-RCNN model.

5.4.3 Generation Track Top Submissions

The generation track aims to assess whether the state-of-the-art generation algorithms can improve the performance of the detection systems. The algorithms should take a frontal chest radiograph and one or more bounding box locations as input and produce an image with generated nodules at the requested locations. Further details regarding these methods can be found online in the NODE21 challenge page [499].

Table 5.3: Network architecture and training details of NODE21 detection solutions. WBF= weighted box fusion, NMS= non max suppression

Method	Architecture	Pretrained	Ensemble Size	Ensemble Method	Resolution Resolution	Batch Size	Epochs	Data imbalance
D1	YOLO FRCNN RetNet	VinDr-CXR	20	WBF	1024 or 640	8 or 16	20-60	simulated data
D2	YOLO	ImageNet	33	NMS	1024	8	60	undersampling
D3	MaskRCNN RetNet	DeepLesion Luna16	6	NMS	1024	8	40	oversampling
DB	FRCNN	COCO	1	no	1024	4	30	oversampling

G1 This model was placed as rank 1 algorithm in the final leaderboard. The nodule generation task was tackled by generative inpainting, where a network learns to inpaint the mask region in a given patch.

This model uses a generative adversarial networks (GANs), which consists of generator and discriminator networks specialized for inpainting. It is based on a recently proposed CR-Fill architecture [500], where the generator network receives a masked patch along with the actual mask and gradually produces the inpainted region. The generator has two components, the coarse network and the refinement network, where the predictions produced by coarse network are refined in the next step by the refinement network. Several losses were calculated to train the network; L1 loss was calculated from both the coarse and refined inpainted patches, adversarial loss and structural similarity index measure (SSIM) were calculated from the refined image patch. It also used contextual reconstruction loss from the feature maps produced by the refinement network, which aims to select useful patches from the image to fill in the missing region.

In order to increase the number of nodule cases to train the network, CheXpert [6] and MIMIC datasets [34] were utilized. Since these two datasets do not have location annotations, the baseline detection network was run on them, and predictions with confidence higher than 0.7 were selected. This procedure resulted in 7000 nodule images from CheXpert, 6500 nodule images from MIMIC. Since NODE21 dataset contains higher quality nodule annotations, the images from NODE21 were oversampled ten times during training. All the images were preprocessed using the OpenCXR library and no additional preprocessing step was performed. Horizontal flipping was used as augmentation during training.

5.5 Experiments

In addition to presenting the challenge evaluation metrics for each detection and generation method, in this work, we also evaluate an ensemble model of the best solutions in each track. The four detection track algorithms (D1-D3 and DB) were ensembled using the weighted box fusion method [497]. For the generation track, we combine all the generated images from both methods (G1 and GB) and assess the impact of this simulated nodule data in training. The performance of these experiments was evaluated in the same way as in the challenge. For the ensembled detection method, AUC score, and sensitivity at various false positive rate (0.5, 0.25, 0.125) were computed. In the generation track experiment, the combination of generated images produced by G1 and GB methods were used to

train a baseline detection method. The resulting nodule detection system was then evaluated using the same detection evaluation metrics described above.

For this publication, additional experiments were performed to systematically assess the impact of the generated nodule images for building nodule detection systems. All models were trained from scratch without using any external data to make sure none of the images that were used in our test dataset were included in our training data. The experiments were designed to determine the impact of the dataset size, and the type of the detection and generation methods. In these additional experiments, we have utilized the large VinDr-CXR dataset [491] to generate nodules using G1 and GB methods aiding in assessing how dataset size affects performance. 10606 images which were labeled with 'No finding', were selected and up to 3 nodule bounding boxes (1-3) were generated using our nodule location generator(Figure 5.3, see Section 5.2.4 for details). Both G1 and GB generation methods were run on the images to synthesize nodules within the requested bounding boxes. These 10606 images are used for the experiments described in the remainder of this section to assess the impact of various factors when building a nodule detection system.

5.5.1 Impact of the generation methods

These experiments aim to compare the performance of two nodule generation methods, namely G1 and GB, to create data for training nodule detection systems. To evaluate the utility of simulated nodules generated by the G1 and GB methods, the baseline detection method, DB, was trained exclusively on images produced by G1 and GB. Using a fixed detection method architecture allows us to investigate only the impact of the generation methods. The training dataset for DB consisted solely of positive cases, each containing at least one nodule generated by G1 or GB.

In these experiments, the DB detection algorithm was trained solely with the generated CXR images (no real nodules) obtained using G1 or GB. For each generation model, we trained the detector firstly using all available images (10606 from VinDr-CXR), and further with various smaller dataset sizes set at 10%, 20%, 50% and 75% of the full simulated dataset. Finally, we used an ensemble approach and combined the images from both methods (G1+GB, resulting in 21212 images), and trained the detection model again using the full dataset and the specified subsets of 10%, 20%, 50%, and 75%.

The resulting detection models from each experiment were evaluated on the final test set, and AUC score and sensitivity at various false positive rate (0.5, 0.25, 0.125) were computed to measure the performance of the corresponding generation model or ensemble.

5.5.2 Impact of the real dataset size

In these experiments, we consider the importance of the availability of real CXR nodule images and investigate the added value of generated nodule images for boosting model performance.

The NODE21 training dataset (4882 images, 1134 with nodules) is used as the source of real (not generated) nodule images, and 20k images with generated nodules are used (10606 VinDr-CXR images with nodules generated by each of 2 methods). The nodule images in the real dataset were oversampled until they reached the same size as the non-nodule images to have a balanced dataset. For each experiment where we combine real and generated data, we make sure that data was balanced as well by oversampling the real dataset size until it reaches the same size as the generated data. We investigate the impact of these generated images on the detection model when the number of real NODE21 images is varied.

The baseline detection method, DB, was trained using 10%, 20%, 50%, 75%, and 100% of the real dataset respectively. Next, each of these training datasets was combined with all the available generated images and DB was re-trained with each of these combination datasets. To prevent data imbalance where the number of real images was too small compared to the number of generated images, we oversample the real dataset until it reaches the generated dataset size during training.

The resulting detection methods from each experiment were evaluated on the final test set using AUC, and sensitivity at various false positive rates (0.5, 0.25, 0.125).

5.5.3 Impact of combining detection and generation methods

In our final experiments, we analyze the impact of the generated data from different generators (G1 and GB) on each of the different detection methods described in this paper, namely D1, D2, D3, and DB. Each detection model was first trained with the real dataset (the NODE21 training dataset) to create a benchmark performance measure. Next, the detection method was trained by boosting the NODE21 training dataset with generated images, which were generated either by G1 or GB methods or a combination of both.

It is important to note that during the challenge the submitted algorithms were allowed to use external data sources and computational size or time was not limited for training. However, for these additional experiments, in order to compare the impact of the detection method (and not other factors such as external data or computational source), all three detection methods, D1, D2, and D3, were adapted to fit with the computational resource requirements. All models were trained from scratch without using any external data to make sure none of the images were used in our test dataset. The batch-size of the methods was decreased to be able to train each model with 12GB memory, and no other external data was used for training. For this reason, the benchmark performance measures are not identical to those achieved during the challenge. The specific modifications to each method were as follows: D1: The batch-size was reduced from 16 to 8 and the training set was limited to the provided NODE21 training data. D2: The batch-size was reduced from 16 to 8 D3: The method was trained with random weight initialization instead of pre-trained using external data-sources.

For all the experiments, we trained the corresponding detection model three times and the model with the best validation set performance was selected as the final model for evaluation. This was done in order to reduce the impact of the randomization process during training which arises from specific GPU computations.

The final model performance was evaluated on the final test set using AUC score, and sensitivity at various false positive rates (0.5, 0.25, 0.125).

5.5.4 Statistical Methods

To compare AUC scores achieved by different methods, DeLong's test is used [501] with statistical significance set at $p<0.05$.

Table 5.4: The performance of the nodule detection algorithms on the final test set of 298 images (111 with nodules). The ensemble model was obtained using weighted box fusion [497] on the four individual model predictions. AUC=Area under ROC curve, 'Outperforms' indicates the names of methods that have significantly lower AUC. S(0.5) indicates the algorithm sensitivity at an average of 0.5 false positives per image.

Training Dataset Source : NODE21 training set					
Training Dataset Size : 3748 images per generator					
Test Dataset Source : NODE21 Final Test Set					
Test Dataset Size : 187 (non-nodule) and 111 (nodule) images					
Method	AUC	S (0.5)	S (0.25)	S (0.125)	Outperforms
D1	0.868	0.800	0.750	0.603	DB
D2	0.862	0.771	0.723	0.600	DB
D3	0.833	0.761	0.704	0.590	
DB	0.816	0.714	0.635	0.504	
Ensemble	0.877	0.819	0.754	0.619	D3, DB

5.6 Results

5.6.1 Challenge Detection track

The performance of the detection algorithms on the final test data is provided in Table 5.4. As seen in the table, the D1 and D2 methods achieved a similar level of performance with an AUC of 0.868 and 0.862 and sensitivity of 0.800 and 0.771 at 0.5 FP rate, respectively. D3 method also achieved a high level of performance with an AUC score of 0.833 and sensitivity of 0.761 at 0.5 FP rate, however, only D1 and D2 algorithms showed significantly higher performance than DB ($p<0.05$).

The ensembled model which was created from the four detection track algorithms (D1, D2, D3 and DB) using weighted box fusion [497] achieved a significantly higher performance than D3 and DB methods ($p<0.05$) with an AUC of 0.877 and a sensitivity of 0.819 at 0.5 FP rate (Table 5.4). Further, the performance of the detection algorithms was compared to 12 experienced radiologists on the final evaluation set. Figure 5.6 illustrates the performance of each detection algorithm, the ensemble detection model, and the performance of the 12 observers. The final ensemble method showed a better performance than 3 radiologists and achieved a similar level of performance to 8 radiologists, underperforming only one radiologist.

Figure 5.4 shows example nodule cases from the final test set. As illustrated by the figure, while more obvious nodules are detected by all the detection methods, D1 and D2 methods perform better than D3 and DB methods for detecting small subtle nodules (nodules behind the rib and clavicles as in the example). Further, all the detection methods can miss very small or subtle nodules and nodules that appear in the region of vessels.

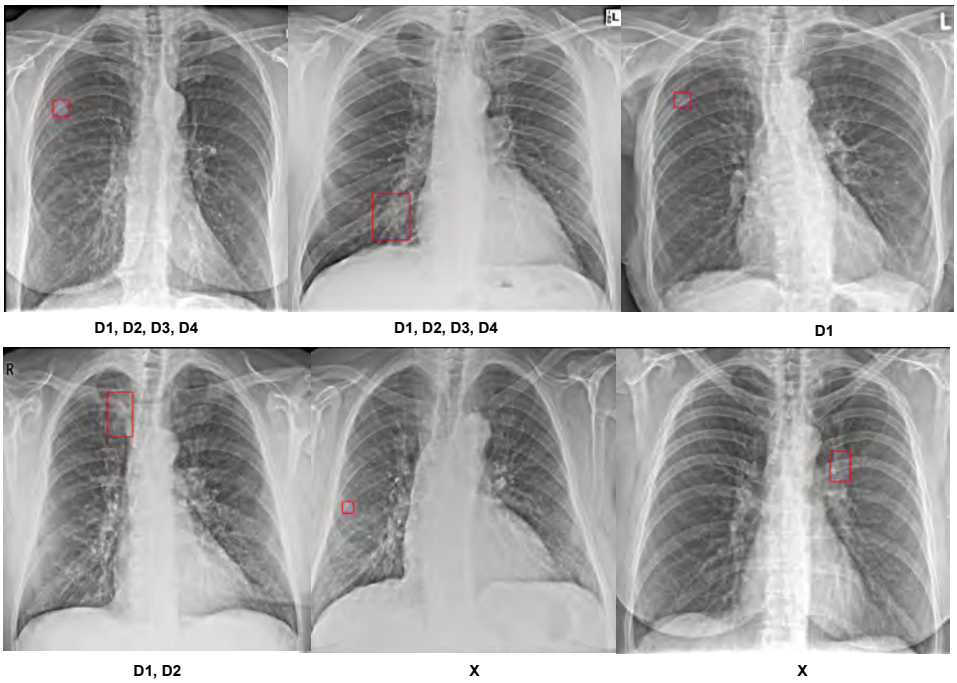


Figure 5.4: Example nodule cases in the final test set. Detection methods that detected the corresponding nodule in the image are displayed below each image. X denotes that the nodule was missed by all the detection methods.

5.6.2 Challenge Generation track

The generation algorithms were evaluated by running the baseline detection model with training data composed of CXRs with nodules generated by the submitted generation algorithm.

The baseline detection model trained with 1000 images generated by G1 and GB methods respectively achieved similar levels of performance with AUCs of 0.746 and 0.722 and a sensitivity of 0.524 and 0.505 at 0.5 FP rate, when evaluated on the final test data. These results are provided in Table 5.5. The impact of combining images from both generation methods, G1, and GB, was evaluated by creating a total of 2000 nodule images (1000 images from each generation method). The final detection method (Faster R-CNN model) trained with these combined images achieved a significantly higher performance ($p<0.05$) than those trained using data from individual generation methods, with an AUC of 0.783 on the final test data (Table 5.5.

In a qualitative evaluation, we visually inspected 200 randomly selected generated images from each method. Some examples of the generated images from both methods can be seen in Figure 5.5. While the G1 method consistently produced more visually realistic nodules compared to the GB method, the diversity of the nodule appearance (e.g. shape) was more limited. The GB method, on the other hand, tended to produce very bright nodules around the heart region and was more prone to producing extremely subtle nodules.

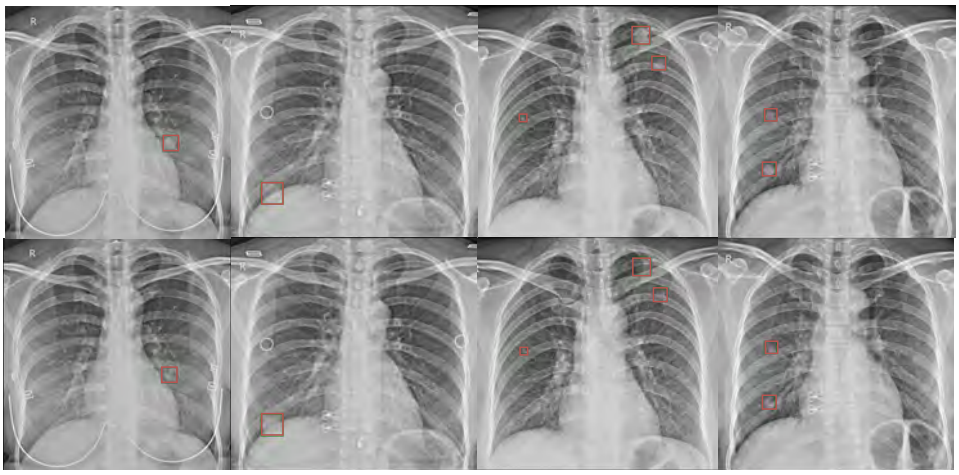


Figure 5.5: Examples of generated nodules using G1 and GB methods. The top row indicates nodules generated by G1, and the bottom row shows generated nodules using GB.

5.6.3 Experiment Results

Impact of the generation method

The first set of experimental results where the baseline detection system was trained using solely generated data can be seen in Table 5.6. As seen in the table, the baseline detection method, DB, achieved an AUC of 0.701 and 0.722 using data from G1 and GB methods, respectively, and using only 10% of the generated data (1060 images). Considering networks trained using data from just one generator (G1 or GB), the AUC does not increase significantly when the dataset size is increased gradually from 10% all the way to 100% and remains in similar ranges ($p>0.05$). Networks trained using G1 data alone also did not have a significantly different performance compared to those trained using GB data alone, regardless of dataset size.

It is, however, noteworthy that the performance consistently improved when G1 and GB generated datasets were combined for training, regardless of the dataset size. The model trained with both G1 and GB generated datasets surpassed the performance of models trained exclusively on data from either G1 or GB, given the same number of samples. The highest performance levels (0.778-0.798) were achieved when the model was trained with the combination of G1+GB images, and the model trained with just 10% of this data (n=2,121) has an AUC that is not statistically different from that of the model trained with any other percentage, or with all data from both models (n=21,212).

Impact of the real dataset size

In Table 5.7, the results of varying the size of the real dataset available for training the DB detection model are shown. When training with real data only the detection performance improves consistently as the size of the dataset is increased.

Adding generated images into the training data results in performance improvements when only part of the real dataset is available. All the experiments except when the full real dataset was used showed

Table 5.5: The performance of the generation algorithms on the final test set. Each generator, G1, and GB methods were run on 1000 (non-nodule) CXR images from the training data, and the resulting data was used to train the baseline Faster R-CNN model. G1+GB denotes the experiments where images generated from both methods are combined (results in 2000 generated images). AUC=Area under ROC curve. 'Outperforms' indicates the names of methods that have significantly lower AUC. S(0.5) indicates the algorithm sensitivity at an average of 0.5 false positives per image.

Training Dataset Source : NODE21 training set					
Training Dataset Size : 1000 (generated-nodule) images per generator					
Test Dataset Source : NODE21 Final Test Set					
Test Dataset Size : 187 (non-nodule) and 111 (nodule) images					
Nodule Generator	AUC	S (0.5)	S (0.25)	S (0.125)	Outperforms
G1	0.746	0.524	0.362	0.27	
GB	0.722	0.505	0.352	0.324	
G1+GB	0.783	0.591	0.51	0.463	G1, GB

significant improvement when generated images (G1+GB) were added ($p<0.05$).
The best performance was achieved when DB was trained using all NODE21 training data (real data) and all generated data which resulted in an AUC score of 0.844 and sensitivity of 0.762 at a 0.5 FP rate. The visual analysis comparing the outcomes of the two models revealed that augmenting the Faster R-CNN training with G1+GB generated images improved nodule detection in difficult regions such as near the clavicles or adjacent to blood vessels. Figure 5.7 presents few example cases where nodules were detected solely after the addition of generated images to the training data.

Impact of combining detection and generation methods

In this set of experiments, each of the detection algorithms (D1, D2, D3, and DB) was trained first using only real data (NODE21 training dataset), and then using a combination of real data with simulated nodule images generated by G1, G2 or both. The experiment results are displayed in Table 5.8.
The AUC values achieved by models with generated data added to the NODE21 training data are generally not significantly better than those achieved by the model using only NODE21 (real) training data. The only exception to this is for D3, where adding all generated images (G1+GB) to the training set yielded a significant performance improvement($p<0.05$), increasing the AUC score from 0.766 to 0.812.

5.7 Discussion

In this paper, we analyze the results of the two-track NODE21 challenge which was organized to collectively develop nodule detection and generation algorithms on chest X-rays. NODE21 was one

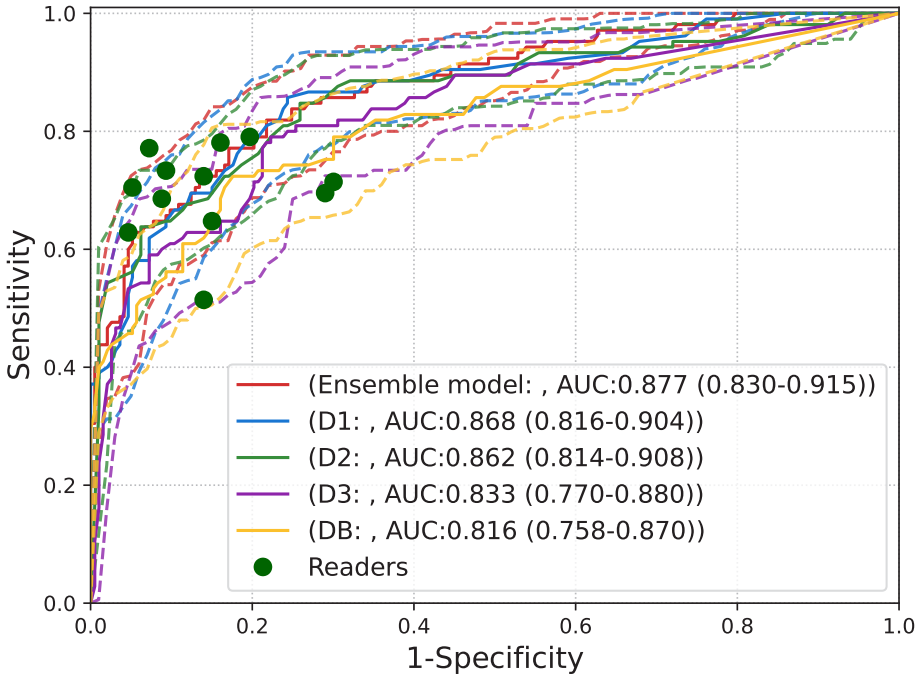


Figure 5.6: ROC curve for the D1, D2, D3, and DB methods, their respective ensemble performance (denoted as ensemble model), and the 12 observers on the final test set of 298 images (111 with nodules). The ensemble model was obtained using weighted box fusion [497] on the four individual model predictions. AUC=Area under ROC curve.

of the first challenges which included only fully reproducible and open algorithm solutions. The best-performing algorithms from both tracks were selected to be included in the paper, and additional experiments were performed to systematically analyze the impact of the generated images on the state-of-the-art detection model performance.

The detection track solutions achieve results comparable to the 12 radiological readers (comprising six radiologists and six residents). Based on the 95% confidence intervals of the ROC curves, the ensemble model (of D1, D2, D3, DB) is outperformed by only one reader and its performance surpasses three others. Only two readers exceed the performance of the top model, D1, while its performance surpasses that of three others, as indicated by the 95% confidence intervals. The commercial nodule detection solution tested in 2014 on the same dataset [413] achieved a sensitivity of 81% at 1.9FP per image, while D1 alone can obtain 80% sensitivity at 0.5FP per image. This demonstrates the advances in AI in the last decade, which allow researchers to quickly develop systems with radiologist-level performance. Comparison with recent work using different test sets is generally difficult since the criteria for data selection and annotation vary widely. In 2021, a commercial system (AI Rad Companion Chest X-Ray algorithm (Siemens Healthineers AG)) achieved an AUC of 0.82 on a dataset that included CT-visualised nodules considered challenging to detect on CXR, similar to our test dataset

Figure 5.7: Illustrative examples of nodules detected exclusively in the Faster R-CNN model trained with additional G1+GB generated images.

[479]. This suggests that the detection networks in this challenge (AUC=0.816-0.868) perform in the range of the current state-of-the-art technologies.

Interestingly, the top three solutions (D1-D3) all utilized an ensemble of numerous state-of-the-art models (ranging from 6 to 33 models), implying the advantages of this strategy in achieving superior performance. These findings are consistent with prior studies [11, 276], and also align with the majority of the top-10 public challenge submissions such as CheXpert [6], SIIM-ACR [77], and RSNA-Pneumonia [73], all of which have utilized ensemble methodologies. Pinpointing the reasons for performance differences between the detection solutions (D1-DB) is a complex task, as there could be various factors at play. These may range from the quantity of models used in an ensemble, the architecture of the models themselves, to the strategies deployed during training. While the number of models in an ensemble appears to suggest an improvement in performance (Table 5.3), this hypothesis requires additional validation through future research.

The generation track attracted a smaller number of participants and a scarcity of innovative methods. Most entries closely mirrored the provided baseline algorithm, with only one competitor (G1) employing a unique GAN method and demonstrating robust performance. This could suggest that the task of generating nodules was viewed as more challenging than that of nodule detection, a sentiment that aligns with previous research [11] where detection on CXR has been explored more comprehensively than the generation task.

Notably, our research demonstrates that combining nodules from different generators is considerably more beneficial than using a larger quantity from a single generator (as shown in Table 5.6). This underlines the importance of diverse generation techniques in real-world applications. The experiments, where the DB model was trained with a range of sizes of generated images (from G1, GB, or a combination of both), revealed that a model trained with only 2121 images from a combination of G1 and GB significantly outperforms one that is trained with a notably larger set of 10606 images from G1 alone. No model trained with data from a single generator could outperform a model trained on combined data, regardless of dataset size. One theory to explain this phenomenon is that when the data is generated using only a single method (G1 or GB), the detection method might learn features consistently produced by the generator, such as sharp nodule borders for example, which are not always present in real nodule data. We hypothesize that generating data with different methods can help the detection algorithm to focus less on these generator-specific features, and more on the important nodule characteristics. This theory is consistent with previous work on natural images [492]

Table 5.6: Impact of the generation methods: The baseline detection model, DB, is used; G1 and GB denote 10606 images with nodules generated by G1 and GB models, respectively. Each experiment was run three times, and the model with the best performance on the validation set was selected. AUC=Area under ROC curve. 'Outperforms' indicates which methods which have significantly lower AUC. (Methods that were never significantly worse than others are not listed in these additional columns). * indicates that the method with significantly worse performance was trained with a dataset that was larger or the same size. S(0.5) indicates the algorithm sensitivity at an average of 0.5 false positives per image.

Training Dataset Source : VinDr-CXR dataset																
Training Dataset Size : 10606 (generated-nodule) images per generator																
Test Dataset Source : NODE21 Final Test Set																
Test Dataset Size : 187 (non-nodule) and 111 (nodule) images																
Dataset	AUC	S (0.5)	S (0.25)	S (0.125)	Nb of images	Outperforms										
						10%		20%		50%		75%		100%		
						G1	GB	G1	GB	G1	GB	G1	GB	G1	GB	
10% G1	0.701	0.392	0.349	0.295	1060											
10% GB	0.722	0.504	0.400	0.307	1060											
10% G1 + GB	0.778	0.581	0.495	0.381	2121	✓	✓	✓*		✓*		✓*		✓*		
20% G1	0.708	0.438	0.371	0.324	2121											
20% GB	0.734	0.476	0.419	0.359	2121											
20% G1 + GB	0.797	0.590	0.472	0.381	4242	✓	✓	✓	✓	✓*		✓*		✓*		
50% G1	0.716	0.466	0.383	0.324	5303											
50% GB	0.743	0.505	0.429	0.343	5303											
50% G1 + GB	0.797	0.648	0.505	0.416	10606	✓	✓	✓	✓	✓		✓	✓	✓*		
75% G1	0.700	0.414	0.390	0.343	7954											
75% GB	0.743	0.571	0.457	0.354	7954											
75% G1 + GB	0.798	0.619	0.552	0.475	15909	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
G1	0.709	0.405	0.381	0.362	10606											
GB	0.745	0.545	0.466	0.352	10606											
G1+GB	0.782	0.638	0.524	0.438	21212	✓	✓	✓		✓		✓		✓		

which showed that combining generated images with different blending techniques performed better than using a single blending technique for an object localization task.

Another insight from our results is that increasing generated dataset size does not significantly increase the performance of the detection model (for either single generators or combined). Using only 10% of the available data produces a comparable result to using 100% for all datasets (G1, GB, G1+GB). This indicates that the generated nodules are useful but lack diversity, hence producing larger numbers of them does not aid performance. However, it is notable that DB detection systems trained only with synthesized nodule images can lead to high-performance levels (with AUC close to 0.8 for those trained with combined datasets) when evaluated on real nodule images. This is further emphasized in scenarios where the dataset size is limited (e.g., 2000 images), where a model trained with G1+GB datasets achieves a performance level (AUC of 0.783, Table 5.5) comparable to that of a model trained with a real dataset (AUC of 0.774, Table 5.7), highlighting the usefulness of the generation dataset.

In the next set of experiments (Table 5.7), the focus was placed on examining the enhancement provided by the inclusion of synthetically generated images when real data (chest X-rays with nodules)

Table 5.7: Impact of the real dataset size: The baseline detection model, DB, is used; G1 and GB denote 10606 images with nodules generated by G1 and GB models, respectively. Each experiment was run three times, and the model with the best performance on the validation set was selected. AUC=Area under ROC curve. S(0.5) indicates the algorithm sensitivity at an average of 0.5 false positives per image. * indicates that the AUC is significantly improved compared to the previous row (without generated training data).

Training Dataset Source : VinDr-CXR and NODE21 training dataset					
Training Dataset Size : 10606 (generated-nodule), 4882 NODE21 images					
Test Dataset Source : NODE21 Final Test Set					
Test Dataset Size : 187 (non-nodule) and 111 (nodule) images					
Dataset	AUC	S (0.5)	S (0.25)	S (0.125)	Nb of images
10% Real	0.742	0.409	0.314	0.276	488
10% Real+ G1 + GB	0.802*	0.668	0.584	0.499	21700
20% Real	0.774	0.571	0.438	0.350	976
20% Real+ G1 + GB	0.830*	0.704	0.629	0.556	22188
50% Real	0.789	0.676	0.584	0.504	2441
50% Real + G1 + GB	0.848*	0.795	0.635	0.523	23653
75% Real	0.797	0.685	0.590	0.512	3661
75% Real + G1 + GB	0.849*	0.779	0.693	0.584	24873
Real	0.816	0.714	0.635	0.504	4882
Real + G1 + GB	0.844	0.762	0.638	0.514	26094

Table 5.8: Impact of the generated images on the performance of detection methods. Real = 4882 NODE21 images, G1 and GB are 10606 images with simulated nodules generated by G1 and GB methods, respectively. AUC=Area under ROC curve. S(0.5) indicates the algorithm sensitivity at an average of 0.5 false positives per image.T denotes training time in days. *indicates the AUC values found to be significantly different (comparisons were only made within each detection method).

Training Dataset Source : VinDr-CXR and Node21 dataset					
Training Dataset Size : 10606 (generated-nodule), 4882 NODE21					
Test Dataset Source : NODE21 Final Test Set					
Test Dataset Size : 187 (non-nodule) and 111 (nodule) images					
Method	AUC	S (0.5)	S (0.25)	S (0.125)	T (days)
D1					
Real	0.845	0.781	0.707	0.594	2
Real+G1	0.838	0.759	0.674	0.543	5
Real+GB	0.844	0.792	0.713	0.584	5
Real+G1+GB	0.852	0.784	0.733	0.600	6
D2					
Real	0.846	0.782	0.704	0.619	1
Real+G1	0.849	0.784	0.712	0.639	2
Real+GB	0.851	0.779	0.711	0.623	2
Real+G1+GB	0.858	0.762	0.705	0.609	2
D3					
Real	0.766*	0.614	0.449	0.352	1
Real+G1	0.778	0.639	0.599	0.502	1
Real+GB	0.789	0.645	0.601	0.521	1
Real+G1+GB	0.812*	0.695	0.614	0.533	1
DB					
Real	0.816	0.714	0.635	0.504	4h
Real+G1	0.832	0.724	0.648	0.505	6h
Real+GB	0.827	0.719	0.644	0.500	6h
Real+G1+GB	0.844	0.762	0.638	0.514	6h

was accessible. For models using up to 75% of the real dataset, there was a statistically significant improvement in performance when the 21k synthetic nodule images were added to the dataset. When using the full real dataset for training the baseline model, the addition of the generated data did not result in a statistically improved model, even though the AUC value did rise from 0.816 to 0.844. Notably, however, this latter version of the DB model achieved comparable performance to that of D1 and D2 ($p > 0.05$) (in contrast to the original DB which had significantly worse performance compared to D1 and D2 (Table 5.4). This indicates that the addition of the generated data has elevated the performance of the simple Faster R-CNN model to be comparable to that of much more complex ensemble models.

The finding that generated data is most useful when real dataset size is limited is consistent with the additional findings in Table 5.8. While adding G1 and GB datasets into training yielded a slight increase in AUC for all the detection methods, only D3 showed a statistically significant improvement when retrained with all generated nodule data added to the real training dataset. This suggests that for the other three methods, the real data contains sufficient data diversity for the task.

In conclusion, the results from the NODE21 challenge demonstrate that the utilization of generated nodule data can improve the effectiveness of detection methods for identifying nodules in CXRs under certain scenarios. Given the data and methodologies applied in this study, the enhancement was most prominent when the size of the real dataset was restricted and when data generated from two different generation methods was combined. These findings suggest that employing various generation methods, or possibly even differing method parameters or blending techniques, can increase diversity and likely offer more benefits compared to merely using the same generator to produce a larger volume of images. Future efforts should concentrate on enhancing the diversity of the generated images to potentially achieve greater advances in performance.

General Discussion

6

In this thesis, we have focused on leveraging deep learning techniques for enhancing chest X-ray (CXR) analysis, encapsulating a series of studies that span from literature review to the development of clinically relevant diagnostic algorithms and the facilitation of a public research challenge.
This chapter outlines the main contributions of our work and proposes directions for future research towards clinical integration.

6.1 Towards clinically relevant AI systems for CXR

To advance AI systems for clinical use, our research highlights several essential factors. The next sections outline these critical steps, drawing on findings from our literature review in Chapter 2. We discuss how our thesis approaches these factors, aiming to contribute effectively to the field.

6.1.1 Assessing Benefits and Impacts of AI systems

An important conclusion from our comprehensive literature review shows a widening gap between academic research in automated CXR analysis with deep learning and practical application within the realm of clinical radiology, as mentioned in Section 1.4. To bridge this gap, it is essential to understand the intended use of AI systems within clinical settings and their interaction with radiologists. This thesis delves into several key areas where AI can significantly impact radiological practices, building on the areas identified in Section 1.2.

1. **Alleviating Workloads and Enhancing Efficiency:** A primary focus of our research, discussed in Chapter 3, centers on the development of a cardiomegaly detection algorithm using deep learning techniques. This algorithm automates the calculation of the cardiothoracic ratio, a clinically relevant measure often used in assessing heart enlargement. This application exemplifies how AI can automate routine, time-consuming tasks, and aims to let radiologists to concentrate on more complex diagnostic challenges.

2. **Prioritizing Urgent Cases and Enhancing Diagnostic Accuracy:** Chapter 5 delves into one of the most important diagnostic challenges in chest X-ray (CXR) analysis: the detection of lung nodules. Given the critical role early detection plays in patient outcomes and the inherent complexity of accurately identifying subtle lung nodules, this area of CXR analysis is paramount. Despite its significance, recent years have seen a noticeable decline in research efforts directed towards lung nodule detection, a trend that our review paper in Chapter 2 attributes to a broader shift in research focus. To address these challenges and move the field forward, Chapter 5 introduces a public research challenge focused on the nodule detection task where models identify the location of the nodule with bounding boxes. The challenge also provides a public CXR nodule dataset with annotations from radiologists to accelerate research in this area. This application demonstrates an example of an AI system which can be used to enhance diagnostic precision, through the early identification of subtle nodules, and potentially prioritize urgent cases to ensure they receive prompt attention.

3. **Uncovering New Health Indicators:** The exploration of AI applications in identifying previously unavailable health markers from chest X-rays (CXR) represents a relatively unexplored territory. The premise here is to leverage deep learning to uncover health indicators that are not immediately apparent to radiologists through traditional visual analysis of CXRs. An illustrative example is presented in Chapter 4, where we develop a deep learning model capable of

estimating total lung volume from CXR images. To our knowledge, this study marks the first successful demonstration of such a capability, highlighting the potential for AI applications to extend the scope of insights from CXR beyond the current clinical practice.

6.1.2 Explainable AI through Transparent Model Outputs

The development of deep learning models for chest X-ray (CXR) analysis is increasingly focused on not just achieving high accuracy but also on ensuring that the models' outputs are interpretable and actionable for clinicians. Explainability is important in the medical field, as it provides insights into the logic behind a model's predictions, enabling healthcare professionals to make informed decisions. This is particularly important when dealing with binary outcomes, such as the presence or absence of a specific condition, where understanding the underlying reasoning might significantly impact radiologist's decision making process.

A key challenge in the current landscape is the deep learning models' "black box" nature, which can erode trust among medical practitioners and hinder the adoption of AI tools in clinical settings. One common method to address this issue in CXR analysis has been the use of saliency maps, which aim to visually represent the areas of an image most influential to the model's decision. While these maps offer some insights into the model's focus areas, the effectiveness and interpretability of these visual maps have not been thoroughly evaluated.

Recognizing the limitations of simple binary outputs and the need for more granular insights, a shift towards models that provide detailed localization and segmentation of abnormalities is likely to be more beneficial in clinical settings. Chapter 3 illustrates this shift by introducing a segmentation-based model for cardiomegaly detection on CXR, diverging from the classification-based models that have been predominantly used in the literature. This model calculates and visually presents the cardiothoracic ratio, offering a quantifiable and visually interpretable output that directly supports clinical decision-making. The study highlights how refining the training objectives of deep learning models not only enhances performance but also yields results with significantly improved explainability.

Extending this approach, Chapter 5 delves into lung nodule detection, framing it as a bounding box detection challenge to emphasize explainability and clinical applicability. A public challenge was crafted to assess algorithms not merely on their ability to detect nodules within a CXR but also on the accuracy of their localization. To facilitate this, a public dataset annotated with bounding box representations of nodules was provided for algorithm training. This method of specifying nodule locations directly supports diagnostic processes and is likely to foster a more streamlined and effective workflow for radiologists. By reducing the time devoted to uncertain searches for nodules and allowing clinicians to easily disregard model outputs for obvious false positives, this approach has potential to significantly speed up the diagnostic process compared to classification based models with binary outputs. It empowers clinicians to make informed decisions about the model's findings, ensuring that they do not expend effort searching for nodules merely because the model indicates their presence.

These advancements mark a significant transition in the application of deep learning within medical imaging. We are moving from basic detection tasks to creating deep learning models that deliver detailed, understandable, and clinically useful insights. This shift is vital for building confidence in deep learning among medical professionals, ensuring these technologies can be seamlessly integrated into daily clinical workflows. Ultimately, this progress aims to enhance radiologists' workflow by providing precise, actionable information that supports timely and informed decision-making.

6.1.3 Quality of labels and annotations

The reliability and accuracy of labels and annotations play an important role in the development and assessment of deep learning models in medical imaging. Chapter 2 reveals that a substantial portion of the studies in the literature relies on publicly available labeled datasets for either training or evaluation. However, these datasets frequently depend on automated techniques to parse radiology reports for labels, introducing a layer of vagueness due to the automated labeling process. Moreover, radiology reports themselves may not comprehensively list all findings, given their contextual nature, leading to labels that might not fully capture the spectrum of possible diagnoses. This process, compounded by the limitations of natural language processing (NLP) technologies, often results in label inaccuracy or noise.

The utilization of these public datasets, while beneficial for training purposes, introduces significant challenges when employed for evaluating the performance of chest X-ray analysis systems. The reliance on such datasets tends to skew performance metrics, rendering comparisons between models and the selection of the most effective model unreliable. The hidden pitfalls of using labels extracted via NLP methods are frequently underestimated, especially when it comes to model evaluation and comparison.

In response to these challenges, Chapter 2 advocates for the use of evaluation datasets grounded in a well-defined reference standard. Achieving a more accurate model performance assessment necessitates high-quality labels, ideally obtained from thorough radiological examinations by multiple expert reviewers or supported by additional diagnostic evidence, including CT scans and laboratory tests. Building on this principle, Chapter 5 introduces the NODE21 CXR dataset, annotated by experienced radiologists with bounding box annotations to pinpoint nodule locations. To ensure the highest levels of accuracy, two distinct test datasets were employed: an experimental test set and a final test set, both of which were augmented with CT scan data. By making the training dataset publicly accessible and keeping the leaderboard open for experimental test set submissions, we aim to foster research in this area and facilitate robust model comparisons. This strategy not only enhances the quality of research but also supports the development of more reliable and effective diagnostic tools in medical imaging.

Additionally, each chapter of this thesis emphasizes the creation of high quality annotated evaluation datasets to ensure the reliability and accuracy of performance assessments. In Chapter 3, while utilizing the publicly accessible Chest X-ray14 dataset, we enhanced the evaluation dataset by having radiologists precisely annotate cardiomegaly, marking the cardiothoracic ratio directly on each chest X-ray image. A similar approach was adopted in Chapter 4, where the task was to train a deep learning system to estimate the total lung volume from chest X-ray images. Given the challenge of deriving total lung volume directly from a CXR, we compiled a dataset from patients who had undergone both a CXR and a pulmonary function test (PFT) within a short time frame. We then utilized the total lung volume measurements from the PFTs as the labels for our dataset, ensuring our model could be trained and evaluated against a concrete, clinically validated standard.

6.1.4 Benchmarking

Public datasets and challenges hold significant importance in the field of deep learning for medical imaging, providing a unified platform for the assessment and comparison of various methodologies and models. These resources enable researchers to benchmark their work against state-of-the-art techniques, highlighting opportunities for enhancement and guiding the evolution of innovative strate-

gies. They represent a critical mechanism for facilitating research advancement, fostering collaboration, and enabling reproducible, fair evaluations.
Leveraging this insight, Chapter 5 details our initiative to launch NODE21, a public research challenge focused on the detection and generation of lung nodules on CXR. To support this challenge, we compiled a training dataset from several public sources, all re-annotated by radiologists to ensure label accuracy. A unique aspect of this challenge was the requirement for participants to submit their solutions as Docker containers, promoting full reproducibility. This format allowed each submitted solution to be utilized not only for evaluating model performance but also for further training of models. To our knowledge, NODE21 represents one of the first examples of a research challenge that offers solutions which are accessible for both training and testing purposes, thereby fostering complete reproducibility and transparency. Furthermore, we have kept submissions for the experimental test data open, enabling ongoing comparison of new models against the established benchmarks.

6.1.5 Reproduciblity

Reproducibility is fundamental to scientific progress. It's vital for ensuring that other researchers can replicate results, which strengthens the trust in findings and facilitates further advancements.
In line with promoting reproducibility, we have made the algorithms developed in chapters 3, 4 and 5, in the thesis publicly available. This allows others in the field to apply our models to new datasets, test their effectiveness, and conduct comparisons with their own models. Such comparisons are crucial for identifying improvements and driving innovation in CXR analysis using deep learning.

6.1.6 Generalizability

In medical imaging, ensuring the generalizability of models represents a formidable challenge due to the inherent variability and complexity of the data. A model developed and trained using datasets from one medical institution may underperform when applied to data from another institution. This discrepancy can arise from differences in imaging techniques, equipment specifications, and patient demographics. This issue is made even worse by the fact that collection of large medical data to train these models is very difficult, limiting the potential for models to learn from a broad and representative sample of data.
In the spirit of this, Chapter 5 provides public training dataset from multiple resources. We have aimed to address model generalizability by compiling the training dataset from various public sources, while ensuring the test dataset comprises data from a hospital not involved in the training phase, thereby broadening the applicability of research findings to diverse clinical settings.

6.2 Future Work

The progress in deep learning has led to significant advancements across various research domains, with CXR analysis standing out as a field that has seen remarkable progress. Several recent applications demonstrate performance levels comparable to radiologists, marking a substantial leap forward. However, as we achieve these high performance benchmarks, it becomes apparent that in focusing solely on marginal gains, we may overlook broader aspects critical to translating these systems into clinical practice.

Moving forward, our focus should shift from simply making small improvements in performance to taking a broader look at the entire field of chest X-ray analysis. Our thorough review of existing research has highlighted how public datasets have already started to change where we focus our efforts. However, while these datasets have been invaluable, it's time to build on what we've learned from their limitations. We need to develop high-quality public datasets that not only address clinically important questions but also provide clear definitions of how these systems can be useful and ensure the labels used are accurate and reliable. The positive influence of well-designed public datasets and research challenges on advancing the field is undeniable—they push us forward and inspire new ideas and breakthroughs.

Another vital area for future work is the enhancement of label quality, possibly through the incorporation of data from CT scans or other diagnostic tests, thereby surpassing the limitations of traditional radiological assessments. Evaluating models against these refined datasets will not only improve accuracy but also push the boundaries of what is achievable with AI in medical imaging.

As we move towards the integration of AI systems into clinical settings, it is essential to consider more than just performance metrics. Understanding the specific information radiologists require, determining the desired outcomes of AI models, assessing the generalizability and fairness of model performance, and ensuring robustness are all critical factors. These considerations are fundamental to the successful deployment of AI systems in healthcare environments.

Lastly, there is a pressing need for studies that investigate the prospective impact of AI systems in clinical practice. Such research would provide invaluable insights into the practical benefits and challenges of implementing AI-assisted diagnostics, guiding future developments in a direction that maximizes utility and efficacy in real-world settings.

In conclusion, it is crucial that future developments not only strive for technical excellence but also for practical applicability in healthcare settings. By focusing on creating high-quality datasets, enhancing label accuracy, and considering the broader implications of AI integration into medical workflows, we can ensure that deep learning tools become valuable assets in improving patient care. The journey ahead is not just about achieving technological milestones but also about making real-world impacts that enhance the effectiveness and efficiency of clinical diagnostics.

Summary

7

In **Chapter 1**, we lay the groundwork by introducing the core concepts on the basics of Chest X-rays (CXR), detailing their varieties and applications in medical practice. Subsequently, we address the challenges in CXR interpretation in clinical settings, establishing a case for the necessity of automated systems in CXR analysis. This chapter also offers a brief overview of the deep learning frameworks employed throughout the thesis and highlights the current gaps and hurdles identified in existing research. It equips readers with the essential knowledge needed to understand the field of CXR analysis and introduces the solutions proposed in this thesis to address these challenges.

Chapter 2 presents an extensive review of the literature on CXR analysis from 296 peer reviewed studies, pinpointing critical gaps and suggesting directions for future research. It introduces deep learning and CXR fundamentals, categorizes key literature trends by their analytical tasks, and provides a detailed overview of public datasets and commercial CXR analysis products. Serving as a pivotal resource, this analysis is designed not only to orient new researchers to the field but also to offer insights for researchers from other disciplines seeking to understand the nuances of CXR analysis through deep learning. The chapter reflects on the collective achievements and challenges faced by the CXR deep learning research community, discussing common pitfalls and suggesting directions for future work. By doing so, it lays the groundwork for developing CXR analysis methodologies with clear clinical applicability, thereby setting the stage for the research discussed in subsequent chapters.

In **Chapter 3**, we critically examine the prevalent use of classification-based methodologies in the literature for detecting cardiomegaly in chest X-rays (CXR) and introduce an alternative strategy centered on anatomical segmentation. This chapter performs a comparative analysis between these two distinct deep learning tasks, namely anatomical segmentation and image-level classification, employing systematic hyperparameter optimization to optimize their performance. Our findings indicate that the segmentation-based technique not only surpasses image-level classification in performance but also enhances interpretability substantially. The proposed approach, when trained on a dataset of moderate size comprising chest radiographs, achieves a comparable performance to the radiologist. Importantly, the model generates a quantitative measurement that is clinically relevant, offering a pathway to consistent and reproducible assessments, as well as facilitating detailed reporting. The model developed is publicly available.

Chapter 4 delves into the utilization of deep learning methodologies for the measurement of a critical quantitative biomarker: total lung volume, using chest X-ray (CXR) images. To our knowledge, this is one of the first studies illustrating the ability of state-of-the-art deep learning techniques to accurately estimate total lung volume from conventional chest radiographs. The study demonstrates a potential application towards expanding the diagnostic capabilities of CXR beyond traditional visual assessments. The model developed is openly accessible and can be employed to determine total lung volume from routinely captured chest X-rays. This deep learning system has potential to serve as a valuable instrument for tracking trends over time in patients who undergo regular chest X-ray examinations.

Chapter 5 examines state-of-the-art nodule detection and generation methodologies through the orchestration of an open-source and collaborative research initiative, NODE21. We organized a public research challenge, **NODE21**, with the objective of benchmarking state-of-the-art techniques in nodule detection and generation task on chest X-rays. We additionally performed extensive experiments using the top performing solutions from each track to analyze the impact of synthetic nodule generation methodologies for the task of nodule detection on CXR. Our results demonstrate that employing generated images can improve the performance of detection methods, with this impact being especially pronounced when there is a scarcity of real nodule images available. Furthermore, the structure of this challenge was designed to accept submissions exclusively in the form of open-source solutions

using Docker containers, which guarantees the **reproducibility** of all methods submitted. The challenge also contributes to the field by providing a valuable public dataset, annotated by radiologists, to facilitate research and address this important clinical issue.

In **Chapter 6**, we reflect on the work presented in the previous chapters of this thesis. We examine the main contributions and applications of our research for CXR analysis. We further discuss potential future directions to move towards the development and integration of AI systems that can be effectively utilized in clinical settings.

Samenvatting

Dit proefschrift gaat over toepassingen van deep learning in de klinische praktijk en richt zich specifiek op het gebruik van deze moderne techniek om thoraxfoto's te analyseren.

In **Hoofdstuk 1** geven we een uitgebreide review van de onderwerpen die centraal staan in het proefschrift: thoraxfoto's (CXR, een afkorting die staat voor chest x-ray), de uitdagingen bij het interpreteren van deze beelden, en het belang dat automatische systemen kunnen spelen. We reviewen de literatuur op het gebied van deep learning en de onderliggende technieken voor het analyseren van thoraxfoto's. We geven ook een overzicht van publieke datasets die beschikbaar zijn voor onderzoekers. Het hoofdstuk stipt ook aan waar de grootste uitdagingen liggen in het onderzoek op dit terrein.

Hoofdstuk 2 bevat een uitgebreide literatuurstudie over CXR-analyse gebaseerd op 296 peer-reviewed studies, waarbij kritieke lacunes worden geïdentificeerd en richtingen voor toekomstig onderzoek worden gesuggereerd. Het introduceert deep learning en de basisprincipes van CXR, categoriseert de belangrijkste literatuurtrends op basis van hun analytische taken en biedt een gedetailleerd overzicht van openbare datasets en commerciële producten voor CXR-analyse. We reflecteren op de collectieve prestaties en uitdagingen waarmee de onderzoeksgemeenschap van CXR deep learning wordt geconfronteerd. Daarnaast stippen we veelvoorkomende valkuilen aan en suggereren we richtingen voor toekomstig onderzoek. Dit hoofdstuk legt de basis voor het ontwikkelen van methodologieën voor CXR-analyse met duidelijke klinische toepasbaarheid.

In **Hoofdstuk 3** onderzoeken we kritisch het heersende gebruik van op classificatie gebaseerde methodologieën in de literatuur voor het detecteren van cardiomegalie (een vergroot hart) in thoraxfoto's. We introduceren een alternatieve aanpak die het probleem benadert als een segmentatietaak. Dit hoofdstuk voert een vergelijkende analyse uit tussen deze twee verschillende aanpakken. Om de vergelijking goed te kunnen maken voeren we een systematische hyperparameter optimalisatie uit voor beide methoden. Onze resultaten geven aan dat de segmentatiegebaseerde techniek niet alleen de prestaties van classificatie op beeldniveau overtreft, maar ook de interpretatie aanzienlijk verbetert. De voorgestelde benadering is even nauwkeurig als een radioloog, terwijl deze is getraind op een relatief kleine dataset. Belangrijk is dat het model een klinisch relevante kwantitatieve meting genereert die opgenomen kan worden in het radiologierapport. Het ontwikkelde model is openbaar beschikbaar.

Hoofdstuk 4 gaat in op het gebruik van deep learning methodologieën voor het meten van een kritische kwantitatieve biomarker: het totale longvolume. We schatten dit volume met uitsluitend een thoraxfoto als input. Dit is een van de eerste studies die deze toepassing onderzoekt. De studie toont de potentie aan om voortaan bij elke thoraxfoto een schatting van het longvolume mee te leveren. Ook hier is het ontwikkelde model openlijk toegankelijk. Dit deep learning systeem zou een waardevol instrument kunnen zijn om patienten die regelmatig thoraxfoto's krijgen te volgen over de tijd.

Hoofdstuk 5 onderzoekt state-of-the-art methoden om nodules in thoraxfoto's automatisch te detecteren. Nodules kunnen wijzen op de aanwezigheid van longkanker. We hebben hiervoor een online competitie, NODE21, georganiseerd. De competitie bestond uit twee tracks: een track waarbij deelnemers methoden ontwikkelden om nodules te detecteren met de computer, en een tweede track om kunstmatig nodules toe te voegen in thoraxfoto's. De beelden met gesimuleerde nodules uit de tweede track konden vervolgens gebruikt worden in de eerste track om de detectiemethoden verder te verbeteren. Het hoofdstuk laat uitgebreide experimenten zien met de best presterende oplossingen van elke track. Onze resultaten tonen aan dat het gebruik van gegenereerde afbeeldingen de prestaties van detectiemethoden kan verbeteren, met name wanneer er een tekort is aan beschikbare echte nodule afbeeldingen. NODE21 was zo ontworpen dat inzendingen uitsluitend in de vorm van open-source oplossingen met Docker containers werden geaccepteerd, wat de reproduceerbaarheid van

alle ingediende methoden garandeert. Deze studie heeft ook een waardevolle openbare dataset opgeleverd, geannoteerd door radiologen, die toekomstig onderzoek kan faciliteren om dit belangrijke klinische probleem aan te pakken.
In **Hoofdstuk 6** sluiten we af met een discussie die de belangrijkste bijdragen van het proefschrift samenvat en ingaat op de belangrijkste punten voor het toepassen van AI systemen in de klinische praktijk. We kijken naar manieren om de voordelen en impact van AI te meten, hoe we AI meer uitlegbaar kunnen maken, hoe de kwaliteit van de data en annotaties waarmee AI wordt getraind kan worden verbeterd en hoe AI generaliseerbaar en reproduceerbaar gemaakt kan worden.

Publications

Papers in international journals

Sogancioglu E, van Ginneken B, Behrendt F, Bengs M, Schlaefer A, Radu M, Xu D, Sheng K, Scalzo F, Marcus E, Papa S, Teuwen J, Scholten ET, Schalekamp S, Hendrix, N, Jacobs C, Hendrix W, Sánchez CI, Murphy K. Nodule detection and generation on chest X-rays: NODE21 Challenge. *IEEE Transactions on Medical Imaging* 2024; doi: 10.1109/TMI.2024.3382042

Çallı E, **Sogancioglu E**, van Ginneken B, Murphy K. FRODO: An in-depth analysis of a system to reject outlier samples from a trained neural network. *IEEE Transactions on Medical Imaging*, 2023; 42(4): pp. 971–981.

Sogancioglu E, Murphy, K, Th.Scholten, E, Boulogne, LH, Prokop, M, van Ginneken, B. Automated estimation of total lung volume using chest radiographs and deep learning. *Medical Physics* 2022; 49: 4466–4477.

Noothout JMH, Lessmann N, Van Eede MC, van Harten LD, **Sogancioglu E**, Heslinga FG, Veta M, van Ginneken B, Išgum I. Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation, *Journal of Medical Imaging* 2022; 9(5):052407.

Sogancioglu E, Çallı E, van Ginneken B, van Leeuwen KG, Murphy K. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis* 2021; 72:102125.

Sogancioglu E, Murphy K, Çallı E, Scholten ET, Schalekamp S, Van Ginneken B. Cardiomegaly detection on chest radiographs: Segmentation versus classification. *IEEE Access* 2020; 8:94631–94642.

Papers in conference proceedings

Çallı E, Murphy K, **Sogancioglu E**, Van Ginneken B. FRODO: Free rejection of out-of-distribution samples: application to chest x-ray analysis. In *International Conference on Medical Imaging with Deep Learning*, pages 1–4, 2019.

Çallı E, **Sogancioglu E**, Scholten ET, Murphy K, van Ginneken B. Handling label noise through model confidence and uncertainty: application to chest radiograph classification. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, pages 289-296, 2019.

Bibliography

[1] Röntgen W. C. Eine neue art von strahlen (on a new kind of rays). *Physikalisch-medicinischen Gesellschaft of Würzburg*, 64:1–37, 1895.

[2] Röntgen W. C. "hand mit Ringen: Radiographic print", onview: Digital collections & exhibits, 1895.

[3] Zonneveld F. W. Spectacular rediscovery of the original prints of radiographs roentgen sent to lorentz in 1896. *Insights into Imaging*, 11(1):29, 2020.

[4] on the Effects of Atomic Radiation (UNSCEAR) U. N. S. C. 2008 report to the general assembly: Annex b exposures of the public and workers from various sources of radiation. *United Nations*, I, 2010.

[5] NHS England. Diagnostic imaging dataset statistical release 2020-2021, 2021.

[6] Irvin J., Rajpurkar P., Ko M., Yu Y., Ciurea-Ilcus S., Chute C., Marklund H., Haghgoo B., Ball R. L., Shpanskaya K. S., 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., and Ng A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[7] Li F., Arimura H., Suzuki K., Shiraishi J., Li Q., Abe H., Engelmann R., Sone S., MacMahon H., and Doi K. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. *Radiology*, 237(2):684–690, nov 2005.

[8] Austin J. H., Romney B. M., and Goldsmith L. S. Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect. *Radiology*, 182 (1):115–122, jan 1992.

[9] Harolds J. A., Parikh J. R., Bluth E. I., Dutton S. C., and Recht M. P. Burnout of radiologists: Frequency, risk factors, and remedies: A report of the acr commission on human resources. *Journal of the American College of Radiology*, 13(4):411–416, 2016.

[10] Litjens G., Kooi T., Bejnordi B. E., Setio A. A. A., Ciompi F., Ghafoorian M., van der Laak J. A., van Ginneken B., and Sánchez C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[11] Çallı E., Sogancioglu E., van Ginneken B., van Leeuwen K. G., and Murphy K. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021.

[12] Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review*, 65(6):386–408, 1958.

[13] Samuel A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959.

[14] LeCun Y., Bengio Y., and Hinton G. Deep learning. *Nature*, 521(7553):436–444, 2015.

[15] Fukushima K. and Miyake S. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In *Competition and Cooperation in Neural Nets*, pages 267–285, 1982.

[16] LeCun Y. and Bengio Y. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, 1998.

[17] Krizhevsky A., Sutskever I., and Hinton G. E. ImageNet classification with deep convolutional

neural networks. In *International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[18] Deng J., Dong W., Socher R., Li L., Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[19] Simonyan K. and Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

[20] He K., Zhang X., Ren S., and Sun J. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*, pages 630–645, 2016.

[21] Huang G., Liu Z., v. d. Maaten L., and Weinberger K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017.

[22] United Nations. United nations scientific committee on the effects of atomic radiation (UNSCEAR), 2008 report on sources and effects of ionizing radiation. `http://www.unscear.org/docs/publications/2008/UNSCEAR_2008_Annex-A-CORR.pdf`, 2008.

[23] Mettler F. A., Bhargavan M., Faulkner K., Gilley D. B., Gray J. E., Ibbott G. S., Lipoti J. A., Mahesh M., McCrohan J. L., Stabin M. G., Thomadsen B. R., and Yoshizumi T. T. Radiologic and nuclear medicine studies in the united states and worldwide: Frequency, radiation dose, and comparison with other radiation sources—1950–2007. *Radiology*, 253(2):520–531, 2009.

[24] Raoof S., Feigin D., Sung A., Raoof S., Irugulpati L., and Rosenow E. C. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

[25] Quekel L. G., Kessels A. G., Goei R., and van Engelshoven J. M. Detection of lung cancer on the chest radiograph: A study on observer performance. *European Journal of Radiology*, 39(2): 111–116, 2001.

[26] Balabanova Y., Coker R., Fedorin I., Zakharova S., Plavinskij S., Krukov N., Atun R., and Drobniewski F. Variability in interpretation of chest radiographs among russian clinicians and implications for screening programmes: Observational study. *The BMJ*, 331(7513):379–382, 2005.

[27] Young M. Interobserver variability in the interpretation of chest roentgenograms of patients with possible pneumonia. *Archives of Internal Medicine*, 154(23):2729, 1994.

[28] Lodwick G. S., Keats T. E., and Dorst J. P. The coding of roentgen images for computer analysis as applied to lung cancer. *Radiology*, 81(2):185–200, 1963.

[29] Becker H. C., Nettleton W. J., Meyers P. H., Sweeney J. W., and Nice C. M. Digital computer determination of a medical diagnostic index directly from chest x-ray images. *IEEE Transactions on Biomedical Engineering*, BME-11(3):67–72, 1964.

[30] Meyers P. H., Nice C. M., Becker H. C., Nettleton W. J., Sweeney J. W., and Meckstroth G. R. Automated computer analysis of radiographic images. *Radiology*, 83(6):1029–1034, 1964.

[31] Kruger R. P., Townes J. R., Hall D. L., Dwyer S. J., and Lodwick G. S. Automated radiographic diagnosis via feature extraction and classification of cardiac size and shape descriptors. *IEEE Transactions on Biomedical Engineering*, BME-19(3):174–186, 1972.

[32] Toriwaki J.-I., Suenaga Y., Negoro T., and Fukumura T. Pattern recognition of chest x-ray images. *Computer Graphics and Image Processing*, 2(3-4):252–271, 1973.

[33] Wang X., Peng Y., Lu L., Lu Z., Bagheri M., and Summers R. M. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017.

[34] Johnson A. E. W., Pollard T. J., Berkowitz S. J., Greenbaum N. R., Lungren M. P., ying Deng C., Mark R. G., and Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

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

[36] van Ginneken B. Fifty years of computer analysis in chest imaging: Rule-based, machine learning, deep learning. *Radiological Physics and Technology*, 10(1):23–32, 2017.

[37] Sahiner B., Pezeshk A., Hadjiiski L. M., Wang X., Drukker K., Cha K. H., Summers R. M., and Giger M. L. Deep learning in medical imaging and radiation therapy. *Medical Physics*, 46(1): e1–e36, 2018.

[38] Feng Y., Teh H. S., and Cai Y. Deep learning for chest radiology: A review. *Current Radiology Reports*, 7(8):24, 2019.

[39] Qin C., Yao D., Shi Y., and Song Z. Computer-aided detection in chest radiography based on artificial intelligence: A survey. *BioMedical Engineering OnLine*, 17(1):113, 2018.

[40] Kallianos K., Mongan J., Antani S., Henry T., Taylor A., Abuya J., and Kohli M. How far have we come? artificial intelligence for chest radiograph interpretation. *Clinical Radiology*, 74(5): 338–345, 2019.

[41] Anis S., Lai K. W., Chuah J. H., Shoaib M. A., Mohafez H., Hadizadeh M., Ding Y., and Ong Z. C. An overview of deep learning approaches in chest radiograph. *IEEE Access*, 8:182347–182354, 2020.

[42] Yosinski J., Clune J., Bengio Y., and Lipson H. How transferable are features in deep neural networks? In *International Conference on Neural Information Processing Systems*, volume 27, pages 1–9, 2014.

[43] Baltruschat I. M., Steinmeister L., Ittrich H., Adam G., Nickisch H., Saalbach A., von Berg J., Grass M., and Knopp T. When Does Bone Suppression And Lung Field Segmentation Improve Chest X-Ray Disease Classification? In *IEEE International Symposium on Biomedical Imaging*, pages 1362–1366, 2019.

[44] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., and Rabinovich A. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[45] Szegedy C., Ioffe S., Vanhoucke V., and Alemi A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, volume 31, page 4278–4284, 2017.

[46] Chollet F. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1800–1807, 2017.

[47] Chen L.-C., Papandreou G., Kokkinos I., Murphy K., and Yuille A. L. DeepLab: Semantic

image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[48] Shelhamer E., Long J., and Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.

[49] Ronneberger O., Fischer P., and Brox T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 9351, pages 234–241. Springer, 2015.

[50] Girshick R., Donahue J., Darrell T., and Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[51] Girshick R. Fast r-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[52] Ren S., He K., Girshick R., and Sun J. Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149, 2017.

[53] He K., Gkioxari G., Dollar P., and Girshick R. Mask r-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[54] Redmon J., Divvala S., Girshick R., and Farhadi A. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[55] Redmon J. and Farhadi A. YOLO9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.

[56] Redmon J. and Farhadi A. Yolov3: An incremental improvement. *arXiv*, 2018.

[57] Lin T.-Y., Goyal P., Girshick R., He K., and Dollar P. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[58] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., and Bengio Y. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, volume 63, pages 139–144, 2014.

[59] Salimans T., Goodfellow I., Zaremba W., Cheung V., Radford A., and Chen X. Improved techniques for training gans. In *International Conference on Neural Information Processing Systems*, volume 29, page 2234–2242, 2016.

[60] Heusel M., Ramsauer H., Unterthiner T., Nessler B., and Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *International Conference on Neural Information Processing Systems*, page 6629–6640, 2017.

[61] Karras T., Aila T., Laine S., and Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, pages 1–8, 2018.

[62] Arjovsky M., Chintala S., and Bottou L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

[63] Chen X., Duan Y., Houthooft R., Schulman J., Sutskever I., and Abbeel P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in International Conference on Neural Information Processing Systems*, volume 29, page 2180–2188,

2016.

[64] Odena A., Olah C., and Shlens J. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning*, volume 70, pages 2642–2651, 2017.

[65] Isola P., Zhu J.-Y., Zhou T., and Efros A. A. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[66] Zhu J.-Y., Park T., Isola P., and Efros A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[67] Yi X., Walia E., and Babyn P. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

[68] Wang M. and Deng W. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[69] Zhu C. S., Pinsky P. F., Kramer B. S., Prorok P. C., Purdue M. P., Berg C. D., and Gohagan J. K. The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial and Its Associated Research Resource. *Journal of the National Cancer Institute*, 105(22):1684–1693, 2013.

[70] Demner-Fushman D., Antani S., Simpson M., and Thoma G. R. Design and Development of a Multimodal Biomedical Information Retrieval System. *Journal of Computing Science and Engineering*, 6(2):168–177, 2012.

[71] Kermany D. Large dataset of labeled optical coherence tomography (oct) and chest x-ray images, 2018.

[72] Shiraishi J., Katsuragawa S., Ikezoe J., Matsumoto T., Kobayashi T., Komatsu K.-i., Matsui M., Fujita H., Kodera Y., and Doi K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[73] RSNA. RSNA Pneumonia Detection Challenge, 2018.

[74] Jaeger S., Candemir S., Antani S., Wáng Y.-X. J., Lu P.-X., and Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475–477, 2014.

[75] Vayá M. d. l. I., Saborit J. M., Montell J. A., Pertusa A., Bustos A., Cazorla M., Galant J., Barber X., Orozco-Beltrán D., García-García F., Caparrós M., González G., and Salinas J. M. BIMCV COVID-19+: A large annotated dataset of RX and CT images from COVID-19 patients. *arXiv*, 2020.

[76] HMHospitales. COVIDDSL, covid data save lives. `https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version`, 2020.

[77] ACR. SIIM-ACR Pneumothorax Segmentation, 2019.

[78] 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., and Shetty S. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology*, 294(2):421–

431, 2019.

[79] Cohen J. P., Morrison P., and Dao L. Covid-19 image data collection: Prospective predictions are the future. *arXiv*, 2020.

[80] National Lung Screening Trial Research Team N., Aberle D. R., Adams A. M., Berg C. D., Black W. C., Clapp J. D., Fagerstrom R. M., Gareen I. F., Gatsonis C., Marcus P. M., and Sicks J. D. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, 365(5):395–409, 2011.

[81] Lu M. T., Ivanov A., Mayrhofer T., Hosny A., Aerts H. J. W. L., and Hoffmann U. Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA Network Open*, 2(7):e197416, 2019.

[82] Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):267D–270, 2004.

[83] McDonald C. J., Overhage J. M., Barnes M., Schadow G., Blevins L., Dexter P. R., and Mamlin B. The Indiana Network For Patient Care: A Working Local Health Information Infrastructure. *Health Affairs*, 24(5):1214–1220, 2005.

[84] van Ginneken B., Stegmann M. B., and Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. *Medical Image Analysis*, 10:19–40, 2006.

[85] Tabik S., Gomez-Rios A., Martin-Rodriguez J. L., Sevillano-Garcia I., Rey-Area M., Charte D., Guirado E., Suarez J. L., Luengo J., Valero-Gonzalez M. A., Garcia-Villanova P., Olmedo-Sanchez E., and Herrera F. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. *IEEE Journal of Biomedical and Health Informatics*, 24 (12):3595–3605, 2020.

[86] Oakden-Rayner L. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology*, 27(1):106–112, 2020.

[87] Oakden-Rayner L. Half a million x-rays! first impressions of the stanford and MIT chest x-ray datasets. `https://lukeoakdenrayner.wordpress.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/`, 2019.

[88] Olatunji T., Yao L., Covington B., and Upton A. Caveats in generating medical imaging labels from radiology reports with natural language processing. In *International Conference on Medical Imaging with Deep Learning*, pages 1–4, 2019.

[89] Çallı E., Sogancioglu E., Scholten E. T., Murphy K., and Ginneken B. v. Handling label noise through model confidence and uncertainty: Application to chest radiograph classification. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, pages 289 – 296, 2019.

[90] Rolnick D., Veit A., Belongie S., and Shavit N. Deep Learning is Robust to Massive Label Noise, 2018.

[91] Liu H., Wang L., Nan Y., Jin F., Wang Q., and Pu J. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Computerized Medical Imaging and Graphics*, 75:66–73, 2019.

[92] Sogancioglu E., Murphy K., Çallı E., Scholten E. T., Schalekamp S., and Ginneken B. V. Car-

diomegaly detection on chest radiographs: Segmentation versus classification. *IEEE Access*, 8: 94631–94642, 2020.

[93] Que Q., Tang Z., Wang R., Zeng Z., Wang J., Chua M., Gee T. S., Yang X., and Veeravalli B. CardioXNet: Automated detection for cardiomegaly based on deep learning. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 612–615, 2018.

[94] Li Z., Hou Z., Chen C., Hao Z., An Y., Liang S., and Lu B. Automatic cardiothoracic ratio calculation with deep learning. *IEEE Access*, 7:37749–37756, 2019.

[95] Moradi M., Wong K. L., Karargyris A., and Syeda-Mahmood T. Quality controlled segmentation to aid disease detection. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11314, pages 362 – 368, 2020.

[96] E L., Zhao B., Guo Y., Zheng C., Zhang M., Lin J., Luo Y., Cai Y., Song X., and Liang H. Using deep-learning techniques for pulmonary-thoracic segmentations and improvement of pneumonia diagnosis in pediatric chest radiographs. *Pediatric Pulmonology*, 54(10):1617–1626, 2019.

[97] Hurt B., Yen A., Kligerman S., and Hsiao A. Augmenting Interpretation of Chest Radiographs With Deep Learning Probability Maps. *Journal of Thoracic Imaging*, 35(5):285–293, 2020.

[98] Wang Q., Liu Q., Luo G., Liu Z., Huang J., Zhou Y., Zhou Y., Xu W., and Cheng J.-Z. Automated segmentation and diagnosis of pneumothorax on chest X-rays with fully convolutional multi-scale ScSE-DenseNet: A retrospective study. *BMC Medical Informatics and Decision Making*, 20 (S14):317, 2020.

[99] Liu Y.-C., Lin Y.-C., Tsai P.-Y., Iwata O., Chuang C.-C., Huang Y.-H., Tsai Y.-S., and Sun Y.-N. Convolutional Neural Network-Based Humerus Segmentation and Application to Bone Mineral Density Estimation from Chest X-ray Images of Critical Infants. *Diagnostics*, 10(12): 1028, 2020.

[100] Owais M., Arsalan M., Mahmood T., Kim Y. H., and Park K. R. Comprehensive Computer-Aided Decision Support Framework to Diagnose Tuberculosis From Chest X-Ray Images: Data Mining Study. *Journal of Medical Informatics*, 8(12):e21790, 2020.

[101] Ouyang X., Karanam S., Wu Z., Chen T., Huo J., Zhou X. S., Wang Q., and Cheng J.-Z. Learning Hierarchical Attention for Weakly-supervised Chest X-Ray Abnormality Localization and Diagnosis. *IEEE Transactions on Medical Imaging*, 40(10):2698–2710, 2020.

[102] Rajaraman S., Sornapudi S., Alderson P. O., Folio L. R., and Antani S. K. Analyzing inter-reader variability affecting deep ensemble learning for COVID-19 detection in chest radiographs. *PLOS One*, 15(11):e0242301, 2020.

[103] Samala R. K., Hadjiiski L., Chan H.-P., Zhou C., Stojanovska J., Agarwal P., and Fung C. Severity assessment of COVID-19 using imaging descriptors: A deep-learning transfer learning approach from non-COVID-19 pneumonia. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11597, pages 426 – 430, 2021.

[104] Park S., Lee S. M., Lee K. H., Jung K.-H., Bae W., Choe J., and Seo J. B. Deep learning-based detection system for multiclass lesions on chest radiographs: Comparison with observer readings. *European Radiology*, 30(3):1359–1368, 2020.

[105] Nam J. G., Park S., Hwang E. J., Lee J. H., Jin K.-N., Lim K. Y., Vu T. H., Sohn J. H., Hwang S., Goo J. M., and Park C. M. Development and Validation of Deep Learning–based Automatic

Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology*, 290 (1):218–228, 2019.

[106] Pesce E., Joseph Withey S., Ypsilantis P.-P., Bakewell R., Goh V., and Montana G. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Medical Image Analysis*, 53:26–38, 2019.

[107] Taghanaki S. A., Havaei M., Berthier T., Dutil F., Di Jorio L., Hamarneh G., and Bengio Y. InfoMask: Masked Variational Latent Representation to Localize Chest Disease. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 739–747. Springer, 2019.

[108] Li B., Kang G., Cheng K., and Zhang N. Attention-Guided Convolutional Neural Network for Detecting Pneumonia on Chest X-Rays. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4851–4854, 2019.

[109] Hwang E. J., Park S., Jin K.-N., Kim J. I., Choi S. Y., Lee J. H., Goo J. M., Aum J., Yim J.-J., Cohen J. G., Ferretti G. R., and and C. M. P. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Network Open*, 2(3):e191095, 2019.

[110] Lenis D., Major D., Wimmer M., Berg A., Sluiter G., and Bühler K. Domain Aware Medical Image Classifier Interpretation by Counterfactual Impact Analysis. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 315–325. Springer, 2020.

[111] Hwang S. and Kim H.-E. Self-Transfer Learning for Weakly Supervised Lesion Localization. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 9901, pages 239–246. Springer, 2016.

[112] Wang C., Elazab A., Jia F., Wu J., and Hu Q. Automated chest screening based on a hybrid model of transfer learning and convolutional sparse denoising autoencoder. *BioMedical Engineering OnLine*, 17(1):63, 2018.

[113] Seah J. C. Y., Tang J. S. N., Kitchen A., Gaillard F., and Dixon A. F. Chest radiographs in congestive heart failure: Visualizing neural network learning. *Radiology*, 290(2):514–522, 2019.

[114] Wolleb J., Sandkühler R., and Cattin P. C. DeScarGAN: Disease-Specific Anomaly Detection with Weak Supervision. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12264, pages 14–24. Springer, 2020.

[115] Tang Y.-X., Tang Y.-B., Han M., Xiao J., and Summers R. M. Abnormal Chest X-Ray Identification With Generative Adversarial One-Class Classifier. In *IEEE International Symposium on Biomedical Imaging*, pages 1358–1361, 2019.

[116] Mao Y., Xue F.-F., Wang R., Zhang J., Zheng W.-S., and Liu H. Abnormality Detection in Chest X-Ray Images Using Uncertainty Prediction Autoencoders. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12266, pages 529–538, 2020.

[117] Lenga M., Schulz H., and Saalbach A. Continual Learning for Domain Adaptation in Chest X-ray Classification. In *International Conference on Medical Imaging with Deep Learning*, pages 121:413–423, 2020.

[118] Tang Y., Tang Y., Sandfort V., Xiao J., and Summers R. M. TUNA-Net: Task-Oriented UN-

supervised Adversarial Network for Disease Recognition in Cross-domain Chest X-rays. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 431–440. Springer, 2019.

[119] Gyawali P. K., Li Z., Ghimire S., and Wang L. Semi-supervised Learning by Disentangling and Self-ensembling over Stochastic Latent Space. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 766–774. Springer, 2019.

[120] Gyawali P. K., Ghimire S., Bajracharya P., Li Z., and Wang L. Semi-supervised Medical Image Classification with Global Latent Mixing. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 604–613. Springer, 2020.

[121] McManigle J. E., Bartz R. R., and Carin L. Y-Net for Chest X-Ray Preprocessing: Simultaneous Classification of Geometry and Segmentation of Annotations. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1266–1269, 2020.

[122] Blain M., T Kassin M., Varble N., Wang X., Xu Z., Xu D., Carrafiello G., Vespro V., Stellato E., Ierardi A. M., Di Meglio L., D Suh R., A Walker S., Xu S., H Sanford T., B Turkbey E., Harmon S., Turkbey B., and J Wood B. Determination of disease severity in COVID-19 patients using deep learning in chest X-ray images. *Diagnostic and Interventional Radiology*, 27(1):20–27, 2020.

[123] Oh Y., Park S., and Ye J. C. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Transactions on Medical Imaging*, 39(8):2688–2700, 2020.

[124] Ferreira Junior J. R., Cardenas D. A. C., Moreno R. A., Rebelo M. d. F. d. S., Krieger J. E., and Gutierrez M. A. A general fully automated deep-learning method to detect cardiomegaly in chest x-rays. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11597, pages 537 – 542, 2021.

[125] Tartaglione E., Barbano C. A., Berzovini C., Calandri M., and Grangetto M. Unveiling COVID-19 from CHEST X-Ray with Deep Learning: A Hurdles Race with Small Data. *International Journal of Environmental Research and Public Health*, 17(18):6933, 2020.

[126] Kusakunniran W., Karnjanapreechakorn S., Siriapisith T., Borwarnginn P., Sutassananon K., Tongdee T., and Saiviroonporn P. COVID-19 detection and heatmap generation in chest x-ray images. *Journal of Medical Imaging*, 8(S1):014001, 2021.

[127] Narayanan B. N., Davuluru V. S. P., and Hardie R. C. Two-stage deep learning architecture for pneumonia detection and its diagnosis in chest radiographs. In *SPIE Medical Imaging: Imaging Informatics for Healthcare, Research, and Applications*, volume 11318, pages 130 – 139, 2020.

[128] Rajaraman S., Thoma G., Antani S. ., and Candemir S. Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, pages 200 – 211, 2019.

[129] Blumenfeld A., Konen E., and Greenspan H. Pneumothorax detection in chest radiographs using convolutional neural networks. In *SPIE Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 15 – 20, 2018.

[130] Rajaraman S., Candemir S., Xue Z., Alderson P. O., Kohli M., Abuya J., Thoma G. R., and Antani S. A novel stacked generalization of models for improved TB detection in chest radiographs. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 718–721, 2018.

[131] Subramanian V., Wang H., Wu J. T., Wong K. C. L., Sharma A., and Syeda-Mahmood T. Automated Detection and Type Classification of Central Venous Catheters in Chest X-Rays. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 522–530. Springer, 2019.

[132] Mansoor A., Perez G., Nino G., and Linguraru M. G. Automatic tissue characterization of air trapping in chest radiographs using deep neural networks. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 97–100, 2016.

[133] Wang X., Yu J., Zhu Q., Li S., Zhao Z., Yang B., and Pu J. Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. *Occupational and Environmental Medicine*, 77(9):597–602, 2020.

[134] Oh D. Y., Kim J., and Lee K. J. Longitudinal Change Detection on Chest X-rays Using Geometric Correlation Maps. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 748–756. Springer, 2019.

[135] Daniels Z. A. and Metaxas D. N. Exploiting Visual and Report-Based Information for Chest X-RAY Analysis by Jointly Learning Visual Classifiers and Topic Models. In *IEEE International Symposium on Biomedical Imaging*, pages 1270–1274, 2019.

[136] Chauhan G., Liao R., Wells W., Andreas J., Wang X., Berkowitz S., Horng S., Szolovits P., and Golland P. Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12262, pages 529–539. Springer, 2020.

[137] Karargyris A., Wong K. C. L., Wu J. T., Moradi M., and Syeda-Mahmood T. Boosting the Rule-Out Accuracy of Deep Disease Detection Using Class Weight Modifiers. In *IEEE International Symposium on Biomedical Imaging*, pages 877–881, 2019.

[138] Syeda-Mahmood T., Ahmad H., Ansari N., Gur Y., Kashyap S., Karargyris A., Moradi M., Pillai A., Seshadhri K., Wang W., Wong K. C. L., and Wu J. Building a Benchmark Dataset and Classifiers for Sentence-Level Findings in AP Chest X-Rays. In *IEEE International Symposium on Biomedical Imaging*, pages 863–867, 2019.

[139] Laserson J., Lantsman C. D., Cohen-Sfady M., Tamir I., Goz E., Brestel C., Bar S., Atar M., and Elnekave E. TextRay: Mining Clinical Reports to Gain a Broad Understanding of Chest X-Rays. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11071, pages 553–561. Springer, 2018.

[140] Annarumma M., Withey S. J., Bakewell R. J., Pesce E., Goh V., and Montana G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1): 272–272, 2019.

[141] Ferreira J. R., Armando Cardona Cardenas D., Moreno R. A., de Fatima de Sa Rebelo M., Krieger J. E., and Antonio Gutierrez M. Multi-View Ensemble Convolutional Neural Network to Improve Classification of Pneumonia in Low Contrast Chest X-Ray Images. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1238–1241, 2020.

[142] Vidya M. S., Manikanda K. V., Anirudh G., Srinivasa R. K., and Vijayananda J. Local and global transformations to improve learning of medical images applied to chest radiographs. In *SPIE Medical Imaging: Image Processing*, volume 10949, pages 813 – 821, 2019.

[143] Baltruschat I., Steinmeister L., Nickisch H., Saalbach A., Grass M., Adam G., Knopp T., and Ittrich H. Smart chest X-ray worklist prioritization using artificial intelligence: A clinical workflow simulation. *European Radiology*, 31(6):3837–3845, 2020.

[144] Hermoza R., Maicas G., Nascimento J. C., and Carneiro G. Region Proposals for Saliency Map Refinement for Weakly-Supervised Disease Localisation and Classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12266, pages 539–549. Springer, 2020.

[145] Saednia K., Jalalifar A., Ebrahimi S., and Sadeghi-Naini A. An Attention-Guided Deep Neural Network for Annotating Abnormalities in Chest X-ray Images: Visualization of Network Decision Basis. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1258–1261, 2020.

[146] Baltruschat I. M., Nickisch H., Grass M., Knopp T., and Saalbach A. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Scientific Reports*, 9(1):6381, 2019.

[147] Cai J., Lu L., Harrison A. P., Shi X., Chen P., and Yang L. Iterative Attention Mining for Weakly Supervised Thoracic Disease Pattern Localization in Chest X-Rays. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11071, pages 589–598. Springer, 2018.

[148] Wang H., Jia H., Lu L., and Xia Y. Thorax-Net: An Attention Regularized Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography. *IEEE Journal of Biomedical and Health Informatics*, 24(2):475–485, 2020.

[149] Ma C., Wang H., and Hoi S. C. H. Multi-label Thoracic Disease Image Classification with Cross-Attention Networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 730–738. Springer, 2019.

[150] Cohen J. P., Dao L., Roth K., Morrison P., Bengio Y., Abbasi A. F., Shen B., Mahsa H. K., Ghassemi M., Li H., and Duong T. Predicting COVID-19 Pneumonia Severity on Chest X-ray With Deep Learning. *Cureus*, 12:e9448, 2020.

[151] Wang Z., Xiao Y., Li Y., Zhang J., Lu F., Hou M., and Liu X. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. *Pattern Recognition*, 110:107613, 2021.

[152] Schwab E., Goossen A., Deshpande H., and Saalbach A. Localization of Critical Findings in Chest X-Ray Without Local Annotations Using Multi-Instance Learning. In *IEEE International Symposium on Biomedical Imaging*, pages 1879–1882, 2020.

[153] Yoo H., Kim K. H., Singh R., Digumarthy S. R., and Kalra M. K. Validation of a Deep Learning Algorithm for the Detection of Malignant Pulmonary Nodules in Chest Radiographs. *JAMA Network Open*, 3(9):e2017135, 2020.

[154] Kashyap S., Karargyris A., Wu J., Gur Y., Sharma A., Wong K. C. L., Moradi M., and Syeda-Mahmood T. Looking in the Right Place for Anomalies: Explainable Ai Through Automatic Location Learning. In *IEEE International Symposium on Biomedical Imaging*, pages 1125–1129, 2020.

[155] Hosch R., Kroll L., Nensa F., and Koitka S. Differentiation Between Anteroposterior and

Posteroanterior Chest X-Ray View Position With Convolutional Neural Networks. *Röfo - Fortschritte Auf Dem Gebiet Der Röntgenstrahlen Und Der Bildgebenden Verfahren*, 193:168–176, 2020.

[156] Rueckel J., Kunz W. G., Hoppe B. F., Patzig M., Notohamiprodjo M., Meinel F. G., Cyran C. C., Ingrisch M., Ricke J., and Sabel B. O. Artificial Intelligence Algorithm Detecting Lung Infection in Supine Chest Radiographs of Critically Ill Patients With a Diagnostic Accuracy Similar to Board-Certified Radiologists. *Critical Care Medicine*, 48(7):e574–e583, 2020.

[157] Ureta J., Aran O., and Rivera J. P. Detecting pneumonia in chest radiographs using convolutional neural networks. In *International Conference on Machine Vision*, volume 11433, pages 541 – 548, 2020.

[158] Chen K.-C., Yu H.-R., Chen W.-S., Lin W.-C., Lee Y.-C., Chen H.-H., Jiang J.-H., Su T.-Y., Tsai C.-K., Tsai T.-A., Tsai C.-M., and Lu H. H.-S. Diagnosis of common pulmonary diseases in children by X-ray images and deep learning. *Scientific Reports*, 10(1):17374, 2020.

[159] Yi P. H., Kim T. K., Yu A. C., Bennett B., Eng J., and Lin C. T. Can AI outperform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. *Emergency Radiology*, 27(4):367–375, 2020.

[160] Crosby J., Chen S., Li F., MacMahon H., and Giger M. Network output visualization to uncover limitations of deep learning detection of pneumothorax. In *SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, volume 11316, pages 125 – 128, 2020.

[161] Crosby J., Rhines T., Li F., MacMahon H., and Giger M. Deep learning for pneumothorax detection and localization using networks fine-tuned with multiple institutional datasets. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11314, pages 70 – 74, 2020.

[162] Tang Y.-X., Tang Y.-B., Peng Y., Yan K., Bagheri M., Redd B. A., Brandon C. J., Lu Z., Han M., Xiao J., and Summers R. M. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine*, 3(1):70, 2020.

[163] Rajaraman S., Kim I., and Antani S. K. Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles. *PeerJ*, 8:e8693, 2020.

[164] Nakao T., Hanaoka S., Nomura Y., Murata M., Takenaga T., Miki S., Watadani T., Yoshikawa T., Hayashi N., and Abe O. Unsupervised Deep Anomaly Detection in Chest Radiographs. *Journal of Digital Imaging*, 34(2):418–427, 2021.

[165] Pasa F., Golkov V., Pfeiffer F., Cremers D., and Pfeiffer D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Scientific Reports*, 9(1): 6268, 2019.

[166] Hwang E. J., Park S., Jin K.-N., Kim J. I., Choi S. Y., Lee J. H., Goo J. M., Aum J., Yim J.-J., Park C. M., Deep Learning-Based Automatic Detection Algorithm Development and Evaluation Group, Kim D. H., Woo W., Choi C., Hwang I. P., Song Y. S., Lim L., Kim K., Wi J. Y., Oh S. S., and Kang M.-J. Development and Validation of a Deep Learning–based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. *Clinical Infectious Diseases*, 69(5):739–747, 2019.

[167] Singh V., Danda V., Gorniak R., Flanders A., and Lakhani P. Assessment of Critical Feeding

Tube Malpositions on Radiographs Using Deep Learning. *Journal of Digital Imaging*, 32(4):651–655, 2019.

[168] Chakravarty A., Sarkar T., Ghosh N., Sethuraman R., and Sheet D. Learning Decision Ensemble using a Graph Neural Network for Comorbidity Aware Chest Radiograph Screening. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1234–1237, 2020.

[169] Matsumoto T., Kodera S., Shinohara H., Ieki H., Yamaguchi T., Higashikuni Y., Kiyosue A., Ito K., Ando J., Takimoto E., Akazawa H., Morita H., and Komuro I. Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning. *International Heart Journal*, 61(4):781–786, 2020.

[170] Su C.-Y., Tsai T.-Y., Tseng C.-Y., Liu K.-H., and Lee C.-W. A Deep Learning Method for Alerting Emergency Physicians about the Presence of Subphrenic Free Air on Chest Radiographs. *Journal of Clinical Medicine*, 10(2):254, 2021.

[171] Zou X.-L., Ren Y., Feng D.-Y., He X.-Q., Guo Y.-F., Yang H.-L., Li X., Fang J., Li Q., Ye J.-J., Han L.-Q., and Zhang T.-T. A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. *PLOS One*, 15(7):e0236378, 2020.

[172] Zucker E. J., Barnes Z. A., Lungren M. P., Shpanskaya Y., Seekins J. M., Halabi S. S., and Larson D. B. Deep learning to automate Brasfield chest radiographic scoring for cystic fibrosis. *Journal of Cystic Fibrosis*, 19(1):131–138, 2020.

[173] Campo M. I., Pascau J., and Estepar R. S. J. Emphysema quantification on simulated X-rays through deep learning techniques. In *IEEE International Symposium on Biomedical Imaging*, pages 273–276, 2018.

[174] Toba S., Mitani Y., Yodoya N., Ohashi H., Sawada H., Hayakawa H., Hirayama M., Futsuki A., Yamamoto N., Ito H., Konuma T., Shimpo H., and Takao M. Prediction of Pulmonary to Systemic Flow Ratio in Patients With Congenital Heart Disease Using Deep Learning–Based Analysis of Chest Radiographs. *JAMA Cardiology*, 5(4):449, 2020.

[175] Li M. D., Arun N. T., Gidwani M., Chang K., Deng F., Little B. P., Mendoza D. P., Lang M., Lee S. I., O'Shea A., Parakh A., Singh P., and Kalpathy-Cramer J. Automated assessment of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence*, 2(4):e200079, 2020.

[176] Luo L., Yu L., Chen H., Liu Q., Wang X., Xu J., and Heng P.-A. Deep Mining External Imperfect Data for Chest X-Ray Disease Screening. *IEEE Transactions on Medical Imaging*, 39(11):3583–3594, 2020.

[177] Xue F.-F., Peng J., Wang R., Zhang Q., and Zheng W.-S. Improving Robustness of Medical Image Diagnosis with Denoising Convolutional Neural Networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 846–854. Springer, 2019.

[178] Li X. and Zhu D. Robust Detection of Adversarial Attacks on Medical Images. In *IEEE International Symposium on Biomedical Imaging*, pages 1154–1158, 2020.

[179] Anand D., Tank D., Tibrewal H., and Sethi A. Self-Supervision vs. Transfer Learning: Robust Biomedical Image Analysis Against Adversarial Attacks. In *IEEE International Symposium on Biomedical Imaging*, pages 1159–1163, 2020.

[180] Khatibi T., Shahsavari A., and Farahani A. Proposing a novel multi-instance learning model for tuberculosis recognition from chest X-ray images based on CNNs, complex networks and stacked ensemble. *Physical and Engineering Sciences in Medicine*, 44(1):291–311, 2021.

[181] Schroeder J. D., Bigolin Lanfredi R., Li T., Chan J., Vachet C., Paine R., Srikumar V., and Tasdizen T. Prediction of Obstructive Lung Disease from Chest Radiographs via Deep Learning Trained on Pulmonary Function Data. *International Journal of Chronic Obstructive Pulmonary Disease*, 15: 3455–3466, 2021.

[182] Zhang J., Xie Y., Pang G., Liao Z., Verjans J., Li W., Sun Z., He J., Li Y., Shen C., and Xia Y. Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection. *IEEE Transactions on Medical Imaging*, 40(3):879–890, 2021.

[183] Balachandar N., Chang K., Kalpathy-Cramer J., and Rubin D. L. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. *Journal of the American Medical Informatics Association*, 27(5):700–708, 2020.

[184] Burwinkel H., Kazi A., Vivar G., Albarqouni S., Zahnd G., Navab N., and Ahmadi S.-A. Adaptive Image-Feature Learning for Disease Classification Using Inductive Graph Networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 640–648. Springer, 2019.

[185] Nugroho B. A. An aggregate method for thorax diseases classification. *Scientific Reports*, 11(1): 3242, 2021.

[186] DSouza A. M., Abidin A. Z., and Wismüller A. Automated identification of thoracic pathology from chest radiographs with enhanced training pipeline. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, pages 855 – 861, 2019.

[187] Sirazitdinov I., Kholiavchenko M., Kuleev R., and Ibragimov B. Data Augmentation for Chest Pathologies Classification. In *IEEE International Symposium on Biomedical Imaging*, pages 1216–1219, 2019.

[188] Mao C., Yao L., Pan Y., Luo Y., and Zeng Z. Deep Generative Classifiers for Thoracic Disease Diagnosis with Chest X-ray Images. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1209–1214, 2018.

[189] Rajpurkar P., Irvin J., Ball R. L., Zhu K., Yang B., Mehta H., Duan T., Ding D., Bagul A., Langlotz C. P., Patel B. N., Yeom K. W., Shpanskaya K., Blankenberg F. G., Seekins J., Amrhein T. J., Mong D. A., Halabi S. S., Zucker E. J., Ng A. Y., and Lungren M. P. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, 2018.

[190] Kurmann T., Márquez-Neila P., Wolf S., and Sznitman R. Deep Multi-label Classification in Affine Subspaces. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11764, pages 165–173. Springer, 2019.

[191] Paul A., Tang Y.-X., and Summers R. M. Fast few-shot transfer learning for disease identification from chest x-ray images using autoencoder ensemble. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11314, pages 33 – 38, 2020.

[192] Unnikrishnan B., Nguyen C. M., Balaram S., Foo C. S., and Krishnaswamy P. Semi-supervised Classification of Diagnostic Radiographs with NoTeacher: A Teacher that is Not Mean. In

International Conference on Medical Image Computing and Computer Assisted Intervention, volume 12261, pages 624–634. Springer, 2020.

[193] Michael P. and Yoon H.-J. Survey of image denoising methods for medical image classification. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11314, pages 892 – 899, 2020.

[194] Wang H., Wang S., Qin Z., Zhang Y., Li R., and Xia Y. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*, 67:101846, 2021.

[195] Paul A., Tang Y.-X., Shen T. C., and Summers R. M. Discriminative ensemble learning for few-shot chest x-ray diagnosis. *Medical Image Analysis*, 68:101911, 2021.

[196] Paul A., Shen T. C., Lee S., Balachandar N., Peng Y., Lu Z., and Summers R. M. Generalized Zero-shot Chest X-ray Diagnosis through Trait-Guided Multi-view Semantic Embedding with Self-training. *IEEE Transactions on Medical Imaging*, 40(10):2642–2655, 2021.

[197] Li F., Shi J.-X., Yan L., Wang Y.-G., Zhang X.-D., Jiang M.-S., Wu Z.-Z., and Zhou K.-Q. Lesion-aware convolutional neural network for chest radiograph classification. *Clinical Radiology*, 76 (2):155.e1–155.e14, 2021.

[198] Ghesu F. C., Georgescu B., Gibson E., Guendel S., Kalra M. K., Singh R., Digumarthy S. R., Grbic S., and Comaniciu D. Quantifying and Leveraging Classification Uncertainty for Chest Radiograph Assessment. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 676–684. Springer, 2019.

[199] Haghighi F., Hosseinzadeh Taher M. R., Zhou Z., Gotway M. B., and Liang J. Learning Semantics-Enriched Representation via Self-discovery, Self-classification, and Self-restoration. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 137–147. Springer, 2020.

[200] Zhou H.-Y., Yu S., Bian C., Hu Y., Ma K., and Zheng Y. Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs by Comparing Image Representations. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 398–407. Springer, 2020.

[201] Chen B., Li J., Lu G., Yu H., and Zhang D. Label Co-Occurrence Learning With Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2292–2302, 2020.

[202] Zhou S., Zhang X., and Zhang R. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. *Studies in Health Technology and Informatics*, 264:482–486, 2019.

[203] Bougias H., Georgiadou E., Malamateniou C., and Stogiannos N. Identifying cardiomegaly in chest X-rays: A cross-sectional study of evaluation and comparison between different transfer learning methods. *Acta Radiologica*, 62:1601–1609, 2020.

[204] Brestel C., Shadmi R., Tamir I., Cohen-Sfaty M., and Elnekave E. RadBot-CXR: Classification of Four Clinical Finding Categories in Chest X-Ray Using Deep Learning. In *International Conference on Medical Imaging with Deep Learning*, pages 1–8, 2018.

[205] Cicero M., Bilbily A., Colak E., Dowdell T., Gray B., Perampaladas K., and Barfett J. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs:. *Investigative Radiology*, 52(5): 281–287, 2017.

[206] Bar Y., Diamant I., Wolf L., and Greenspan H. Deep learning with non-medical training used for chest pathology identification. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 9414, pages 215–221, 2015.

[207] Bar Y., Diamant I., Wolf L., Lieberman S., Konen E., and Greenspan H. Chest pathology detection using deep learning with non-medical training. In *IEEE International Symposium on Biomedical Imaging*, pages 294–297, 2015.

[208] Griner D., Zhang R., Tie X., Zhang C., Garrett J. W., Li K., and Chen G.-H. COVID-19 pneumonia diagnosis using chest x-ray radiograph and deep learning. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11597, page 1159706, 2021.

[209] Hu Q., Drukker K., and Giger M. L. Role of standard and soft tissue chest radiography images in COVID-19 diagnosis using deep learning. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11597, pages 1 – 7, 2021.

[210] Zhu J., Shen B., Abbasi A., Hoshmand-Kochi M., Li H., and Duong T. Q. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. *PLOS One*, 15(7):e0236621, 2020.

[211] Fricks R. B., Abadi E., Ria F., and Samei E. Classification of COVID-19 in chest radiographs: Assessing the impact of imaging parameters using clinical and simulated images. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11597, pages 53 – 64, 2021.

[212] Wehbe R. M., Sheng J., Dutta S., Chai S., Dravid A., Barutcu S., Wu Y., Cantrell D. R., Xiao N., Allen B. D., MacNealy G. A., Savas H., Agrawal R., Parekh N., and Katsaggelos A. K. DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large US Clinical Dataset. *Radiology*, 299(1):E167–E176, 2020.

[213] Castiglioni I., Ippolito D., Interlenghi M., Monti C. B., Salvatore C., Schiaffino S., Polidori A., Gandola D., Messa C., and Sardanelli F. Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: A first experience from Lombardy, Italy. *European Radiology Experimental*, 5(1):7, 2021.

[214] Zhang R., Tie X., Qi Z., Bevins N. B., Zhang C., Griner D., Song T. K., Nadig J. D., Schiebler M. L., Garrett J. W., Li K., Reeder S. B., and Chen G.-H. Diagnosis of Coronavirus Disease 2019 Pneumonia by Using Chest Radiography: Value of Artificial Intelligence. *Radiology*, 298 (2):E88–E97, 2021.

[215] Wang X., Schwab E., Rubin J., Klassen P., Liao R., Berkowitz S., Golland P., Horng S., and Dalal S. Pulmonary edema severity estimation in chest radiographs using deep learning. In *International Conference on Medical Imaging with Deep Learning*, pages 1,4, 2019.

[216] Karargyris A., Kashyap S., Wu J. T., Sharma A., Moradi M., and Syeda-Mahmood T. Age prediction using a large chest x-ray dataset. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, pages 468 – 476, 2019.

[217] Xue Z., Antani S., Long R., and Thoma G. R. Using deep learning for detecting gender in adult chest radiographs. In *SPIE Medical Imaging: Imaging Informatics for Healthcare, Research, and Applications*, page 10, 2018.

[218] Sabottke C. F., Breaux M. A., and Spieler B. M. Estimation of age in unidentified patients via chest radiography using convolutional neural network regression. *Emergency Radiology*, 27(5):

463–468, 2020.

[219] Lu M. T., Raghu V. K., Mayrhofer T., Aerts H. J., and Hoffmann U. Deep Learning Using Chest Radiographs to Identify High-Risk Smokers for Lung Cancer Screening Computed Tomography: Development and Validation of a Prediction Model. *Annals of Internal Medicine*, 173(9): 704–713, 2020.

[220] Thammarach P., Khaengthanyakan S., Vongsurakrai S., Phienphanich P., Pooprasert P., Yaemsuk A., Vanichvarodom P., Munpolsri N., Khwayotha S., Lertkowit M., Tungsagunwattana S., Vijitsanguan C., Lertrojanapunya S., Noisiri W., Chiawiriyabunya I., Aphikulvanich N., and Tantibundhit C. AI Chest 4 All. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1229–1233, 2020.

[221] Kuo P.-C., Tsai C. C., López D. M., Karargyris A., Pollard T. J., Johnson A. E. W., and Celi L. A. Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. *npj Digital Medicine*, 4(1):25, 2021.

[222] 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., and Syeda-Mahmood T. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, 3(10):e2022779, 2020.

[223] Cohen J. P., Hashir M., Brooks R., and Bertrand H. On the limits of cross-domain generalization in automated x-ray prediction. In *International Conference on Medical Imaging with Deep Learning*, pages 121:136–155, 2020.

[224] Hashir M., Bertrand H., and Cohen J. P. Quantifying the value of lateral views in deep learning for chest x-rays. In *International Conference on Medical Imaging with Deep Learning*, volume 121, pages 288–303, 2020.

[225] Rajkomar A., Lingam S., Taylor A. G., Blum M., and Mongan J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *Journal of Digital Imaging*, 30(1): 95–101, 2017.

[226] Crosby J., Rhines T., Duan C., Li F., MacMahon H., and Giger M. Impact of imprinted labels on deep learning classification of AP and PA thoracic radiographs. In *SPIE Medical Imaging: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, pages 92 – 96, 2019.

[227] Crosby J., Rhines T., Li F., MacMahon H., and Giger M. Deep convolutional neural networks in the classification of dual-energy thoracic radiographic views for efficient workflow: Analysis on over 6500 clinical radiographs. *Journal of Medical Imaging*, 7(01):1, 2020.

[228] Bertrand H., Hashir M., and Cohen J. P. Do lateral views help automated chest x-ray predictions? In *International Conference on Medical Imaging with Deep Learning*, page 1, 2019.

[229] Chen H., Miao S., Xu D., Hager G. D., and Harrison A. P. Deep Hierarchical Multi-label Classification of Chest X-ray Images. In *International Conference on Medical Imaging with Deep Learning*, volume 102, pages 109–120, 2019.

[230] Shah U., Abd-Alrazeq A., Alam T., Househ M., and Shah Z. An Efficient Method to Predict Pneumonia from Chest X-Rays Using Deep Learning Approach. *Studies in Health Technology and Informatics*, 272:457–460, 2020.

[231] Qu W., Balki I., Mendez M., Valen J., Levman J., and Tyrrell P. N. Assessing and mitigating the

effects of class imbalance in machine learning with application to X-ray imaging. *International Journal of Computer Assisted Radiology and Surgery*, 15(12):2041–2048, 2020.

[232] Yue Z., Ma L., and Zhang R. Comparison and Validation of Deep Learning Models for the Diagnosis of Pneumonia. *Computational Intelligence and Neuroscience*, 2020:1–8, 2020.

[233] Elshennawy N. M. and Ibrahim D. M. Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images. *Diagnostics*, 10(9):649, 2020.

[234] Mittal A., Kumar D., Mittal M., Saba T., Abunadi I., Rehman A., and Roy S. Detecting Pneumonia Using Convolutions and Dynamic Capsule Routing for Chest X-ray Images. *Sensors*, 20(4): 1068, 2020.

[235] Longjiang E., Baisong Zhao, Liu H., Zheng C., Song X., Cai Y., and Liang H. Image-based Deep Learning in Diagnosing the Etiology of Pneumonia on Pediatric Chest X-rays. *Pediatric Pulmonology*, 56(5):1036–1044, 2020.

[236] Ganesan P., Rajaraman S., Long R., Ghoraani B., and Antani S. Assessment of Data Augmentation Strategies Toward Performance Improvement of Abnormality Classification in Chest Radiographs. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 841–844, 2019.

[237] Ravishankar H., Venkataramani R., Anamandra S., Sudhakar P., and Annangi P. Feature Transformers: Privacy Preserving Lifelong Learners for Medical Imaging. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11767, pages 347–355. Springer, 2019.

[238] Taylor A. G., Mielke C., and Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. *PLOS Medicine*, 15(11):e1002697, 2018.

[239] Kitamura G. and Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. *Clinical Imaging*, 61:15–19, 2020.

[240] Kashyap S., Moradi M., Karargyris A., Wu J. T., Morris M., Saboury B., Siegel E., and Syeda-Mahmood T. Artificial intelligence for point of care radiograph quality assessment. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, page 109503K, 2019.

[241] Takaki T., Murakami S., Watanabe R., Aoki T., and Fujibuchi T. Calculating the target exposure index using a deep convolutional neural network and a rule base. *European Journal of Medical Physics*, 71:108–114, 2020.

[242] Pan I., Agarwal S., and Merck D. Generalizable Inter-Institutional Classification of Abnormal Chest Radiographs Using Efficient Convolutional Neural Networks. *Journal of Digital Imaging*, 32(5):888–896, 2019.

[243] Wong K. C. L., Moradi M., Wu J., and Syeda-Mahmood T. Identifying disease-free chest x-ray images with deep transfer learning. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, page 24, 2019.

[244] Wong K. C. L., Moradi M., Wu J., Pillai A., Sharma A., Gur Y., Ahmad H., Chowdary M. S., Chiranjeevi J., Reddy Polaka K. K., Wunnava V., Reddy D., and Syeda-Mahmood T. A Robust Network Architecture to Detect Normal Chest X-Ray Radiographs. In *IEEE International Symposium on Biomedical Imaging*, pages 1851–1855, 2020.

[245] Jang R., Kim N., Jang M., Lee K. H., Lee S. M., Lee K. H., Noh H. N., and Seo J. B. Assessment of the Robustness of Convolutional Neural Networks in Labeling Noise by Using Chest X-Ray Images From Multiple Centers. *Journal of Medical Informatics*, 8(8):e18089, 2020.

[246] Dunnmon J. A., Yi D., Langlotz C. P., Ré C., Rubin D. L., and Lungren M. P. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. *Radiology*, 290(2):537–544, 2019.

[247] Nam J. G., Kim M., Park J., Hwang E. J., Lee J. H., Hong J. H., Goo J. M., and Park C. M. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *European Respiratory Journal*, 57(5):2003061, 2020.

[248] Dyer T., Dillard L., Harrison M., Morgan T. N., Tappouni R., Malik Q., and Rasalingham S. Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm. *Clinical Radiology*, 76(6):473.e9–473.e15, 2021.

[249] Ogawa R., Kido T., Kido T., and Mochizuki T. Effect of augmented datasets on deep convolutional neural networks applied to chest radiographs. *Clinical Radiology*, 74(9):697–701, 2019.

[250] Ellis R., Ellestad E., Elicker B., Hope M. D., and Tosun D. Impact of hybrid supervision approaches on the performance of artificial intelligence for the classification of chest radiographs. *Computers in Biology and Medicine*, 120:103699, 2020.

[251] Rajaraman S., Sornapudi S., Kohli M., and Antani S. Assessment of an ensemble of machine learning models toward abnormality detection in chest radiographs. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3689–3692, 2019.

[252] Lakhani P. and Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2):574–582, 2017.

[253] Hwang S., Kim H.-E., Jeong J., and Kim H.-J. A novel approach for tuberculosis screening based on deep convolutional neural networks. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 9785, page 97852W, 2016.

[254] Sivaramakrishnan R., Antani S., Candemir S., Xue Z., Thoma G., Alderson P., Abuya J., and Kohli M. Comparing deep learning models for population screening using chest radiography. In *SPIE Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 322 – 332, 2018.

[255] Ayaz M., Shaukat F., and Raja G. Ensemble learning based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors. *Physical and Engineering Sciences in Medicine*, 44(1):183–194, 2021.

[256] Ul Abideen Z., Ghafoor M., Munir K., Saqib M., Ullah A., Zia T., Tariq S. A., Ahmed G., and Zahra A. Uncertainty Assisted Robust Tuberculosis Identification With Bayesian Convolutional Neural Networks. *IEEE Access*, 8:22812–22825, 2020.

[257] Rajpurkar P., O'Connell C., Schechter A., Asnani N., Li J., Kiani A., Ball R. L., Mendelson M., Maartens G., van Hoving D. J., Griesel R., Ng A. Y., Boyles T. H., and Lungren M. P. CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digital Medicine*, 3(1):115, 2020.

[258] Heo S.-J., Kim Y., Yun S., Lim S.-S., Kim J., Nam C.-M., Park E.-C., Jung I., and Yoon J.-H. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Ra-

diographs in Annual Workers' Health Examination Data. *International Journal of Environmental Research and Public Health*, 16(2):250, 2019.

[259] Kim H.-E., Kim S., and Lee J. Keep and Learn: Continual Learning by Constraining the Latent Space for Knowledge Preservation in Neural Networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11070, pages 520–528. Springer, 2018.

[260] Gozes O. and Greenspan H. Deep Feature Learning from a Hospital-Scale Chest X-ray Dataset with Application to TB Detection on a Small-Scale Dataset. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4076–4079, 2019.

[261] Rajaraman S. and Antani S. K. Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs. *IEEE Access*, 8:27318–27326, 2020.

[262] Lakhani P. Deep Convolutional Neural Networks for Endotracheal Tube Position and X-ray Image Classification: Challenges and Opportunities. *Journal of Digital Imaging*, 30(4):460–468, 2017.

[263] Mitra A., Chakravarty A., Ghosh N., Sarkar T., Sethuraman R., and Sheet D. A Systematic Search over Deep Convolutional Neural Network Architectures for Screening Chest Radiographs. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1225–1228, 2020.

[264] Pham H. H., Le T. T., Ngo D. T., Tran D. Q., and Nguyen H. Q. Interpreting chest x-rays via {cnn}s that exploit hierarchical disease dependencies and uncertainty labels. In *International Conference on Medical Imaging with Deep Learning*, pages 1–8, 2020.

[265] Rajan D., Thiagarajan J. J., Karargyris A., and Kashyap S. Self-training with improved regularization for sample-efficient chest x-ray classification. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11597, pages 418 – 425, 2021.

[266] Deshpande H., Harder T., Saalbach A., Sawarkar A., and Buelow T. Detection Of Foreign Objects In Chest Radiographs Using Deep Learning. In *IEEE International Symposium on Biomedical Imaging Workshops*, pages 1–4, 2020.

[267] Zhang L., Rong R., Li Q., Yang D. M., Yao B., Luo D., Zhang X., Zhu X., Luo J., Liu Y., Yang X., Ji X., Liu Z., Xie Y., Sha Y., Li Z., and Xiao G. A deep learning-based model for screening and staging pneumoconiosis. *Scientific Reports*, 11(1):2201, 2021.

[268] Devnath L., Luo S., Summons P., and Wang D. Automated detection of pneumoconiosis with multilevel deep features learned from chest X-Ray radiographs. *Computers in Biology and Medicine*, 129:104125, 2021.

[269] Liu X., Wang S., Deng Y., and Chen K. Coronary artery calcification (CAC) classification with deep convolutional neural networks. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10134, page 101340M, 2017.

[270] Hirata Y., Kusunose K., Tsuji T., Fujimori K., Kotoku J., and Sata M. Deep Learning for Detection of Elevated Pulmonary Artery Wedge Pressure using Standard Chest X-Ray. *Canadian Journal of Cardiology*, 37:1198–1206, 2021.

[271] Kusunose K., Hirata Y., Tsuji T., Kotoku J., and Sata M. Deep learning to predict elevated pulmonary artery pressure in patients with suspected pulmonary hypertension using standard

chest X ray. *Scientific Reports*, 10(1):19311, 2020.

[272] Rajaraman S., Candemir S., Kim I., Thoma G., and Antani S. Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. *Applied Sciences*, 8(10):1715, 2018.

[273] Liang G. and Zheng L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*, 187:104964, 2020.

[274] Behzadi-khormouji H., Rostami H., Salehi S., Derakhshande-Rishehri T., Masoumi M., Salemi S., Keshavarz A., Gholamrezanezhad A., Assadi M., and Batouli A. Deep learning, reusable and problem-based architectures for detection of consolidation on chest X-ray images. *Computer Methods and Programs in Biomedicine*, 185:105162, 2020.

[275] Zhou Z., Zhou L., and Shen K. Dilated conditional GAN for bone suppression in chest radiographs with enforced semantic features. *Medical Physics*, 47(12):6207–6215, 2020.

[276] Dietterich T. G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15, 2000.

[277] Wei Y., Feng J., Liang X., Cheng M.-M., Zhao Y., and Yan S. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. *arXiv*, 2018.

[278] Xu Y., Mo T., Feng Q., Zhong P., Lai M., and Chang E. I. Deep learning of feature representation with multiple instance learning for medical image analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1626–1630, 2014.

[279] Bayat A., Sekuboyina A., Paetzold J. C., Payer C., Stern D., Urschler M., Kirschke J. S., and Menze B. H. Inferring the 3D Standing Spine Posture from 2D Radiographs. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12266, pages 775–784. Springer, 2020.

[280] Yu D., Zhang K., Huang L., Zhao B., Zhang X., Guo X., Li M., Gu Z., Fu G., Hu M., Ping Y., Sheng Y., Liu Z., Hu X., and Zhao R. Detection of peripherally inserted central catheter (PICC) in chest X-ray images: A multi-task deep learning model. *Computer Methods and Programs in Biomedicine*, 197:105674, 2020.

[281] Zhang W., Li G., Wang F., E L., Yu Y., Lin L., and Liang H. Simultaneous Lung Field Detection and Segmentation for Pediatric Chest Radiographs. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 594–602. Springer, 2019.

[282] Wessel J., Heinrich M. P., von Berg J., Franz A., and Saalbach A. Sequential Rib Labeling and Segmentation in Chest X-Ray using Mask R-CNN. In *International Conference on Medical Imaging with Deep Learning*, 2019.

[283] Tang Y.-B., Tang Y.-X., Xiao J., and Summers R. M. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. In *International Conference on Medical Imaging with Deep Learning*, volume 102, pages 457–467. PMLR, 2019.

[284] Eslami M., Tabarestani S., Albarqouni S., Adeli E., Navab N., and Adjouadi M. Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography. *IEEE Transactions on Medical Imaging*, 39(7):2553–2565, 2020.

[285] Onodera S., Lee Y., and Tanaka Y. Evaluation of dose reduction potential in scatter-corrected bedside chest radiography using U-net. *Radiological Physics and Technology*, 13(4):336–347, 2020.

[286] Oliveira H. N., Ferreira E., and Santos J. A. D. Truly generalizable radiograph segmentation with conditional domain adaptation. *IEEE Access*, 8:84037–84062, 2020.

[287] Dong N., Kampffmeyer M., Liang X., Wang Z., Dai W., and Xing E. Unsupervised Domain Adaptation for Automatic Estimation of Cardiothoracic Ratio. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11071, pages 544–552. Springer, 2018.

[288] Zhang Y., Miao S., Mansi T., and Liao R. Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-ray Image Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11071, pages 599–607. Springer, 2018.

[289] Chen C., Dou Q., Chen H., and Heng P.-A. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International Workshop on Machine Learning in Medical Imaging*, volume 11046, pages 143–151. Springer, 2018.

[290] Oliveira H., Mota V., Machado A. M., and dos Santos J. A. From 3d to 2d: Transferring knowledge for rib segmentation in chest x-rays. *Pattern Recognition Letters*, 140:10–17, 2020.

[291] Shah M. P., Merchant S. N., and Awate S. P. MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11073, pages 379–387. Springer, 2018.

[292] Bortsova G., Dubost F., Hogeweg L., Katramados I., and de Bruijne M. Semi-supervised Medical Image Segmentation via Learning Consistency Under Transformations. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 810–818. Springer, 2019.

[293] Ouyang X., Xue Z., Zhan Y., Zhou X. S., Wang Q., Zhou Y., Wang Q., and Cheng J.-Z. Weakly Supervised Segmentation Framework with Uncertainty: A Study on Pneumothorax Segmentation in Chest X-ray. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 613–621. Springer, 2019.

[294] Frid-Adar M., Amer R., and Greenspan H. Endotracheal Tube Detection and Segmentation in Chest Radiographs Using Synthetic Data. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 784–792. Springer, 2019.

[295] Sullivan R. P., Holste G., Burkow J., and Alessio A. Deep learning methods for segmentation of lines in pediatric chest radiographs. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 11314, pages 577 – 583, 2020.

[296] Cardenas D. A. C., Jr J. R. F., Moreno R. A., Rebelo M. d. F. d. S., Krieger J. E., and Gutierrez M. A. Automated radiographic bone suppression with deep convolutional neural networks. In *SPIE Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11600, pages 317 – 323, 2021.

[297] Zhang M., Gao J., Lyu Z., Zhao W., Wang Q., Ding W., Wang S., Li Z., and Cui S. Characterizing Label Errors: Confident Learning for Noisy-Labeled Image Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages

721–730. Springer, 2020.

[298] Kholiavchenko M., Sirazitdinov I., Kubrak K., Badrutdinova R., Kuleev R., Yuan Y., Vrtovec T., and Ibragimov B. Contour-aware multi-label chest X-ray organ segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 15(3):425–436, 2020.

[299] Novikov A. A., Lenis D., Major D., Hladuvka J., Wimmer M., and Buhler K. Fully convolutional architectures for multiclass segmentation in chest radiographs. *IEEE Transactions on Medical Imaging*, 37(8):1865–1876, 2018.

[300] Bonheur S., Štern D., Payer C., Pienn M., Olschewski H., and Urschler M. Matwo-CapsNet: A Multi-label Semantic Segmentation Capsules Network. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11768, pages 664–672. Springer, 2019.

[301] Arsalan M., Owais M., Mahmood T., Choi J., and Park K. R. Artificial Intelligence-Based Diagnosis of Cardiac and Related Diseases. *Journal of Clinical Medicine*, 9(3):871, 2020.

[302] Wang W., Feng H., Bu Q., Cui L., Xie Y., Zhang A., Feng J., Zhu Z., and Chen Z. MDU-Net: A Convolutional Network for Clavicle and Rib Segmentation from a Chest Radiograph. *Journal of Healthcare Engineering*, 2020:1–9, 2020.

[303] Mortani Barbosa E. J., Gefter W. B., Ghesu F. C., Liu S., Mailhe B., Mansoor A., Grbic S., and Vogt S. Automated Detection and Quantification of COVID-19 Airspace Disease on Chest Radiographs: A Novel Approach Achieving Expert Radiologist-Level Performance Using a Deep Convolutional Neural Network Trained on Digital Reconstructed Radiographs From Computed Tomography–Derived Ground Truth. *Investigative Radiology*, 56(8):471–479, 2021.

[304] Larrazabal A. J., Martinez C., Glocker B., and Ferrante E. Post-DAE: Anatomically Plausible Segmentation via Post-Processing With Denoising Autoencoders. *IEEE Transactions on Medical Imaging*, 39(12):3813–3820, 2020.

[305] Holste G., Sullivan R. P., Bindschadler M., Nagy N., and Alessio A. Multi-class semantic segmentation of pediatric chest radiographs. In *SPIE Medical Imaging: Image Processing*, volume 11313, pages 323 – 330, 2020.

[306] Mansoor A., Cerrolaza J. J., Perez G., Biggs E., Okada K., Nino G., and Linguraru M. G. A Generic Approach to Lung Field Segmentation From Chest Radiographs Using Deep Space and Shape Learning. *IEEE Transactions on Biomedical Engineering*, 67(4):1206–1220, 2020.

[307] Amiri M., Brooks R., and Rivaz H. Fine-Tuning U-Net for Ultrasound Image Segmentation: Different Layers, Different Outcomes. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(12):2510–2518, 2020.

[308] Lu Y., Li W., Zheng K., Wang Y., Harrison A. P., Lin C., Wang S., Xiao J., Lu L., Kuo C.-F., and Miao S. Learning to Segment Anatomical Structures Accurately from One Exemplar. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 678–688. Springer, 2020.

[309] Li Y., Dong X., Shi W., Miao Y., Yang H., and Jiang Z. Lung fields segmentation in chest radiographs using Dense-U-Net and fully connected CRF. In *International Conference on Graphics and Image Processing*, volume 11720, page 1172011, 2021.

[310] Portela R. D. S., Pereira J. R. G., Costa M. G. F., and Filho C. F. F. C. Lung Region Segmentation

in Chest X-Ray Images using Deep Convolutional Neural Networks. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1246–1249, 2020.

[311] Yahyatabar M., Jouvet P., and Cheriet F. Dense-Unet: A light model for lung fields segmentation in Chest X-Ray images. In *International Conference of the IEEE Engineering in Medicine Biology Society*, pages 1242–1245, 2020.

[312] Kim M. and Lee B.-D. Automatic Lung Segmentation on Chest X-rays Using Self-Attention Deep Neural Network. *Sensors*, 21(2):369, 2021.

[313] Arbabshirani M. R., Dallal A. H., Agarwal C., Patel A., and Moore G. Accurate segmentation of lung fields on chest radiographs using deep convolutional networks. In *SPIE Medical Imaging: Image Processing*, volume 10133, pages 37–42, 2017.

[314] Rahman M. F., Tseng T.-L. B., Pokojovy M., Qian W., Totada B., and Xu H. An automatic approach to lung region segmentation in chest x-ray images using adapted U-Net architecture. In *SPIE Medical Imaging: Physics of Medical Imaging*, volume 11595, pages 894 – 901, 2021.

[315] Souza J. C., Bandeira Diniz J. O., Ferreira J. L., França da Silva G. L., Corrêa Silva A., and de Paiva A. C. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Computer Methods and Programs in Biomedicine*, 177:285–296, 2019.

[316] Milletari F., Rieke N., Baust M., Esposito M., and Navab N. CFCM: Segmentation via Coarse to Fine Context Memory. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11073, pages 667–674. Springer, 2018.

[317] Zhang Z., Fu H., Dai H., Shen J., Pang Y., and Shao L. ET-Net: A Generic Edge-aTtention Guidance Network for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11764, pages 442–450. Springer, 2019.

[318] Mathai T. S., Gorantla V., and Galeotti J. Segmentation of Vessels in Ultra High Frequency Ultrasound Sequences Using Contextual Memory. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11765, pages 173–181. Springer, 2019.

[319] Kitahara Y., Tanaka R., Roth H., Oda H., Mori K., Kasahara K., and Matsumoto I. Lung segmentation based on a deep learning approach for dynamic chest radiography. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10950, pages 909 – 914, 2019.

[320] Furutani K., Hirano Y., and Kido S. Segmentation of lung region from chest x-ray images using U-net. In *International Forum on Medical Imaging in Asia*, volume 11050, pages 165 – 169, 2019.

[321] Xue C., Deng Q., Li X., Dou Q., and Heng P.-A. Cascaded Robust Learning at Imperfect Labels for Chest X-ray Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12266, pages 579–588. Springer, 2020.

[322] Wang H., Gu H., Qin P., and Wang J. CheXLocNet: Automatic localization of pneumothorax in chest radiographs using deep convolutional neural networks. *PLOS One*, 15(11):e0242013, 2020.

[323] Tolkachev A., Sirazitdinov I., Kholiavchenko M., Mustafaev T., and Ibragimov B. Deep Learning for Diagnosis and Segmentation of Pneumothorax: The Results on The Kaggle Competition and Validation Against Radiologists. *IEEE Journal of Biomedical and Health Informatics*, 25 (5):1660–1672, 2020.

[324] Groza V. and Kuzin A. Pneumothorax Segmentation with Effective Conditioned Post-Processing in Chest X-Ray. In *IEEE International Symposium on Biomedical Imaging Workshops*, pages 1–4, 2020.

[325] Xue Z., Long R., Jaeger S., Folio L., George Thoma R., and Antani a. S. Extraction of Aortic Knuckle Contour in Chest Radiographs Using Deep Learning. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5890–5893, 2018.

[326] Yi X., Adams S., Babyn P., and Elnajmi A. Automatic Catheter and Tube Detection in Pediatric X-ray Images Using a Scale-Recurrent Network and Synthetic Data. *Journal of Digital Imaging*, 33(1):181–190, 2019.

[327] Lee H., Mansouri M., Tajmir S., Lev M. H., and Do S. A Deep-Learning System for Fully-Automated Peripherally Inserted Central Catheter (PICC) Tip Detection. *Journal of Digital Imaging*, 31(4):393–402, 2018.

[328] Pan Y., Chen Q., Chen T., Wang H., Zhu X., Fang Z., and Lu Y. Evaluation of a computer-aided method for measuring the Cobb angle on chest X-rays. *European Spine Journal*, 28(12):3035–3043, 2019.

[329] Chen S., Han Y., Lin J., Zhao X., and Kong P. Pulmonary nodule detection on chest radiographs using balanced convolutional neural network and classic candidate detection. *Artificial Intelligence in Medicine*, 107:101881, 2020.

[330] Schultheiss M., Schober S. A., Lodde M., Bodden J., Aichele J., Müller-Leisse C., Renger B., Pfeiffer F., and Pfeiffer D. A robust convolutional neural network for lung nodule detection in the presence of foreign bodies. *Scientific Reports*, 10(1):12987, 2020.

[331] Tam L. K., Wang X., Turkbey E., Lu K., Wen Y., and Xu D. Weakly Supervised One-Stage Vision and Language Disease Detection Using Large Scale Pneumonia and Pneumothorax Studies. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12264, pages 45–55. Springer, 2020.

[332] Moradi M., Madani A., Gur Y., Guo Y., and Syeda-Mahmood T. Bimodal Network Architectures for Automatic Generation of Image Annotation from Text. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11070, pages 449–456. Springer, 2018.

[333] Khakzar A., Albarqouni S., and Navab N. Learning Interpretable Features via Adversarially Robust Optimization. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 793–800. Springer, 2019.

[334] Kim Y.-G., Lee S. M., Lee K. H., Jang R., Seo J. B., and Kim N. Optimal matrix size of chest radiographs for computer-aided detection on lung nodule or mass with deep learning. *European Radiology*, 30(9):4943–4951, 2020.

[335] Cho Y., Kim Y.-G., Lee S. M., Seo J. B., and Kim N. Reproducibility of abnormality detection on chest radiographs using convolutional neural network in paired radiographs obtained within a short-term interval. *Scientific Reports*, 10(1):17417, 2020.

[336] Kim Y.-G., Cho Y., Wu C.-J., Park S., Jung K.-H., Seo J. B., Lee H. J., Hwang H. J., Lee S. M., and Kim N. Short-term Reproducibility of Pulmonary Nodule and Mass Detection in Chest Radiographs: Comparison among Radiologists and Four Different Computer-Aided Detections with

Convolutional Neural Net. *Scientific Reports*, 9(1):18738, 2019.

[337] Cha M. J., Chung M. J., Lee J. H., and Lee K. S. Performance of deep learning model in detecting operable lung cancer with chest radiographs. *Journal of Thoracic Imaging*, 34(2):86–91, 2019.

[338] Takemiya R., Kido S., Hirano Y., and Mabu S. Detection of pulmonary nodules on chest x-ray images using R-CNN. In *International Forum on Medical Imaging in Asia*, volume 11050, pages 147 – 152, 2019.

[339] Wang C., Elazab A., Wu J., and Hu Q. Lung nodule classification using deep feature fusion in chest radiography. *Computerized Medical Imaging and Graphics*, 57:10–18, 2017.

[340] Li X., Shen L., Xie X., Huang S., Xie Z., Hong X., and Yu J. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. *Artificial Intelligence in Medicine*, 103:101744, 2020.

[341] Park S., Lee S. M., Kim N., Choe J., Cho Y., Do K.-H., and Seo J. B. Application of deep learning–based computer-aided detection system: Detecting pneumothorax on chest radiograph after biopsy. *European Radiology*, 29(10):5341–5348, 2019.

[342] Mader A. O., von Berg J., Fabritz A., Lorenz C., and Meyer C. Localization and Labeling of Posterior Ribs in Chest Radiographs Using a CRF-regularized FCN with Local Refinement. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11071, pages 562–570. Springer, 2018.

[343] Xue Z., Jaeger S., Antani S., Long R., Karagyris A., Siegelman J., Folio L. R., and Thoma G. R. Localizing tuberculosis in chest radiographs with deep learning. In *SPIE Medical Imaging: Imaging Informatics for Healthcare, Research, and Applications*, page 28, 2018.

[344] von Berg J., Krönke S., Gooßen A., Bystrov D., Brück M., Harder T., Wieberneit N., and Young S. Robust chest x-ray quality assessment using convolutional neural networks and atlas regularization. In *SPIE Medical Imaging: Image Processing*, pages 391 – 398, 2020.

[345] Zhao Z.-Q., Zheng P., Xu S.-T., and Wu X. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019.

[346] Salehinejad H., Colak E., Dowdell T., Barfett J., and Valaee S. Synthesizing Chest X-Ray Pathology for Training Deep Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 38(5):1197–1206, 2019.

[347] Bigolin Lanfredi R., Schroeder J. D., Vachet C., and Tasdizen T. Adversarial Regression Training for Visualizing the Progression of Chronic Obstructive Pulmonary Disease with Chest X-Rays. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 685–693. Springer, 2019.

[348] Lee D., Kim H., Choi B., and Kim H.-J. Development of a deep neural network for generating synthetic dual-energy chest x-ray images with single x-ray exposure. *Physics in Medicine & Biology*, 64(11):115017, 2019.

[349] Mahapatra D. and Ge Z. Training Data Independent Image Registration with Gans Using Transfer Learning and Segmentation Information. In *IEEE International Symposium on Biomedical Imaging*, pages 709–713, 2019.

[350] Madani A., Moradi M., Karargyris A., and Syeda-Mahmood T. Semi-supervised learning with

generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In *IEEE International Symposium on Biomedical Imaging*, pages 1038–1042, 2018.

[351] Umehara K., Ota J., Ishimaru N., Ohno S., Okamoto K., Suzuki T., Shirai N., and Ishida T. Super-resolution convolutional neural network for the improvement of the image quality of magnified images in chest radiographs. In *SPIE Medical Imaging: Image Processing*, volume 10133, pages 488 – 494, 2017.

[352] Uzunova H., Ehrhardt J., Jacob F., Frydrychowicz A., and Handels H. Multi-scale GANs for Memory-efficient Generation of High Resolution Medical Images. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 112–120. Springer, 2019.

[353] Zhang T., Fu H., Zhao Y., Cheng J., Guo M., Gu Z., Yang B., Xiao Y., Gao S., and Liu J. SkrGAN: Sketching-Rendering Unconditional Generative Adversarial Networks for Medical Image Synthesis. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11767, pages 777–785. Springer, 2019.

[354] Lin C., Tang R., Lin D. D., Liu L., Lu J., Chen Y., Gao D., and Zhou J. Deep Feature Disentanglement Learning for Bone Suppression in Chest Radiographs. In *IEEE International Symposium on Biomedical Imaging*, pages 795–798, 2020.

[355] Dong N., Xu M., Liang X., Jiang Y., Dai W., and Xing E. Neural Architecture Search for Adversarial Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 828–836. Springer, 2019.

[356] Taghanaki S. A., Abhishek K., and Hamarneh G. Improved Inference via Deep Input Transfer. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 819–827. Springer, 2019.

[357] Fang Q., Yan J., Gu X., Zhao J., and Li Q. Unsupervised learning-based deformable registration of temporal chest radiographs to detect interval change. In *SPIE Medical Imaging: Image Processing*, volume 11313, pages 747 – 753, 2020.

[358] Yang W., Chen Y., Liu Y., Zhong L., Qin G., Lu Z., Feng Q., and Chen W. Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. *Medical Image Analysis*, 35:421–433, 2017.

[359] Zarshenas A., Liu J., Forti P., and Suzuki K. Separation of bones from soft tissue in chest radiographs: Anatomy-specific orientation-frequency-specific deep neural network convolution. *Medical Physics*, 46(5):2232–2242, 2019.

[360] Gozes O. and Greenspan H. Bone Structures Extraction and Enhancement in Chest Radiographs via CNN Trained on Synthetic Data. In *IEEE International Symposium on Biomedical Imaging*, pages 858–861, 2020.

[361] Liu Y., Liu M., Xi Y., Qin G., Shen D., and Yang W. Generating Dual-Energy Subtraction Soft-Tissue Images from Chest Radiographs via Bone Edge-Guided GAN. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12262, pages 678–687. Springer, 2020.

[362] Xing Y., Ge Z., Zeng R., Mahapatra D., Seah J., Law M., and Drummond T. Adversarial Pulmonary Pathology Translation for Pairwise Chest X-Ray Data Augmentation. In *International*

Conference on Medical Image Computing and Computer Assisted Intervention, volume 11769, pages 757–765. Springer, 2019.

[363] Li Z., Li H., Han H., Shi G., Wang J., and Zhou S. K. Encoding CT Anatomy Knowledge for Unpaired Chest X-ray Image Decomposition. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 275–283. Springer, 2019.

[364] Albarqouni S., Fotouhi J., and Navab N. X-Ray In-Depth Decomposition: Revealing the Latent Structures. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 10435, pages 444–452. Springer, 2017.

[365] Madani A., Moradi M., Karargyris A., and Syeda-Mahmood T. F. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *SPIE Medical Imaging: Image Processing*, volume 10574, page 57, 2018.

[366] Zarei M., Abadi E., Fricks R., Segars W. P., and Samei E. A probabilistic conditional adversarial neural network to reduce imaging variation in radiography. In *SPIE Medical Imaging: Physics of Medical Imaging*, volume 11595, page 115953Y, 2021.

[367] Gomi T., Hara H., Watanabe Y., and Mizukami S. Improved digital chest tomosynthesis image quality by use of a projection-based dual-energy virtual monochromatic convolutional neural network with super resolution. *PLOS One*, 15(12):e0244745, 2020.

[368] Matsubara N., Teramoto A., Saito K., and Fujita H. Bone suppression for chest X-ray image using a convolutional neural filter. *Physical and Engineering Sciences in Medicine*, 43(1):97–108, 2020.

[369] Zunair H. and Hamza A. B. Synthesis of COVID-19 chest X-rays using unpaired image-to-image translation. *Social Network Analysis and Mining*, 11(1):23, 2021.

[370] Bigolin Lanfredi R., Schroeder J. D., Vachet C., and Tasdizen T. Interpretation of Disease Evidence for Medical Images Using Adversarial Deformation Fields. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12262, pages 738–748. Springer, 2020.

[371] Prevedello L. M., Halabi S. S., Shih G., Wu C. C., Kohli M. D., Chokshi F. H., Erickson B. J., Kalpathy-Cramer J., Andriole K. P., and Flanders A. E. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, 1(1):e180031, 2019.

[372] Zech J. R., Badgeley M. A., Liu M., Costa A. B., Titano J. J., and Oermann E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018.

[373] Yao L., Prosky J., Covington B., and Lyman K. A strong baseline for domain adaptation and generalization in medical imaging. In *International Conference on Medical Imaging with Deep Learning*, pages 1–4, 2019.

[374] Sathitratanacheewin S., Sunanta P., and Pongpirul K. Deep learning for automated classification of tuberculosis-related chest X-Ray: Dataset distribution shift limits diagnostic performance generalizability. *Heliyon*, 6(8):e04614, 2020.

[375] Oliveira H. and dos Santos J. Deep transfer learning for segmentation of anatomical structures in chest radiographs. In *Conference on Graphics, Patterns and Images*, pages 204–211, 2018.

[376] Huang X., Liu M.-Y., Belongie S., and Kautz J. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, pages 179–196. Springer, 2018.

[377] Syeda-Mahmood T., Wong K. C. L., Gur Y., Wu J. T., Jadhav A., Kashyap S., Karargyris A., Pillai A., Sharma A., Syed A. B., Boyko O., and Moradi M. Chest X-Ray Report Generation Through Fine-Grained Label Learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12262, pages 561–571. Springer, 2020.

[378] Li X., Cao R., and Zhu D. Vispi: Automatic visual perception and interpretation of chest x-rays. In *International Conference on Medical Imaging with Deep Learning*, pages 1–8, 2020.

[379] Yuan J., Liao H., Luo R., and Luo J. Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11769, pages 721–729. Springer, 2019.

[380] Xue Y., Xu T., Rodney Long L., Xue Z., Antani S., Thoma G. R., and Huang X. Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11070, pages 457–466. Springer, 2018.

[381] Mansilla L., Milone D. H., and Ferrante E. Learning deformable registration of medical images with anatomical constraints. *Neural Networks*, 124:269–279, 2020.

[382] Márquez-Neila P. and Sznitman R. Image Data Validation for Medical Systems. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11767, pages 329–337. Springer, 2019.

[383] Bozorgtabar B., Mahapatra D., Vray G., and Thiran J.-P. SALAD: Self-supervised Aggregation Learning for Anomaly Detection on X-Rays. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 468–478. Springer, 2020.

[384] Çallı E., Murphy K., Sogancioglu E., and van Ginneken B. FRODO: Free rejection of out-of-distribution samples: Application to chest x-ray analysis. In *International Conference on Medical Imaging with Deep Learning*, pages 1–4, 2019.

[385] Anavi Y., Kogan I., Gelbart E., Geva O., and Greenspan H. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2015, pages 2940–2943, 2015.

[386] Haq N. F., Moradi M., and Wang Z. J. A deep community based approach for large scale content based X-ray image retrieval. *Medical Image Analysis*, 68:101847, 2021.

[387] Chen Z., Cai R., Lu J., Feng J., and Zhou J. Order-Sensitive Deep Hashing for Multimorbidity Medical Image Retrieval. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 11070, pages 620–628. Springer, 2018.

[388] Conjeti S., Roy A. G., Katouzian A., and Navab N. Hashing with Residual Networks for Image Retrieval. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 10435, pages 541–549. Springer, 2017.

[389] Anavi Y., Kogan I., Gelbart E., Geva O., and Greenspan H. Visualizing and enhancing a deep learning framework using patients age and gender for chest x-ray image retrieval. In *SPIE*

Medical Imaging: Computer-Aided Diagnosis, volume 9785, page 978510, 2016.

[390] Silva W., Poellinger A., Cardoso J. S., and Reyes M. Interpretability-Guided Content-Based Medical Image Retrieval. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, volume 12261, pages 305–314. Springer, 2020.

[391] Viergever M. A., Maintz J. A., Klein S., Murphy K., Staring M., and Pluim J. P. A survey of medical image registration. *Medical Image Analysis*, 33:140–144, 2016.

[392] Recht M. P., Dewey M., Dreyer K., Langlotz C., Niessen W., Prainsack B., and Smith J. J. Integrating artificial intelligence into the clinical practice of radiology: Challenges and recommendations. *European Radiology*, 30(6):3576–3584, 2020.

[393] Strohm L., Hehakaya C., Ranschaert E. R., Boon W. P. C., and Moors E. H. M. Implementation of artificial intelligence (AI) applications in radiology: Hindering and facilitating factors. *European Radiology*, 30(10):5525–5532, 2020.

[394] Chokshi F. H., Flanders A. E., Prevedello L. M., and Langlotz C. P. Fostering a healthy AI ecosystem for radiology: Conclusions of the 2018 RSNA summit on AI in radiology. *Radiology: Artificial Intelligence*, 1(2):190021, 2019.

[395] Grand-challenge. Grand challenge: AI for radiology. `https://grand-challenge.org/aiforradiology/`, 2021.

[396] van Leeuwen K. G., Schalekamp S., Rutten M. J. C. M., van Ginneken B., and de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European Radiology*, 31(6):3797–3804, 2021.

[397] Fischer A. M., Varga-Szemes A., Martin S. S., Sperl J. I., Sahbaee P., Neumann D., Gawlitza J., Henzler T., Johnson C. M., Nance J. W., Schoenberg S. O., and Schoepf U. J. Artificial Intelligence-based Fully Automated Per Lobe Segmentation and Emphysema-quantification Based on Chest Computed Tomography Compared With Global Initiative for Chronic Obstructive Lung Disease Severity of Smokers. *Journal of Thoracic Imaging*, 35(3):S28–S34, 2020.

[398] Sim Y., Chung M. J., Kotter E., Yune S., Kim M., Do S., Han K., Kim H., Yang S., Lee D.-J., and Choi B. W. Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. *Radiology*, 294(1):199–209, 2019.

[399] Murphy K., Smits H., Knoops A. J. G., Korst M. B. J. M., Samson T., Scholten E. T., Schalekamp S., Schaefer-Prokop C. M., Phi lipsen R. H. H. M., Meijers A., Melendez J., van Ginneken B., and Rutten M. COVID-19 on Chest Radiographs: A Multireader Evaluation of an Artificial Intelligence System. *Radiology*, 296(3):E166–E172, 2020.

[400] Murphy K., Habib S. S., Zaidi S. M. A., Khowaja S., Khan A., Melendez J., Scholten E. T., Amad F., Schalekamp S., Verhagen M., Philipsen R. H. H. M., Meijers A., and van Ginneken B. Computer aided detection of tuberculosis on chest radiographs: An evaluation of the CAD4TB v6 system. *Scientific Reports*, 10(1):5492, 2020.

[401] Habib S. S., Rafiq S., Zaidi S. M. A., Ferrand R. A., Creswell J., Van Ginneken B., Jamal W. Z., Azeemi K. S., Khowaja S., and Khan A. Evaluation of computer aided detection of tuberculosis on chest radiography among people with diabetes in Karachi Pakistan. *Scientific Reports*, 10(1): 6276, 2020.

[402] Qin Z. Z., Sander M. S., Rai B., Titahong C. N., Sudrungrot S., Laah S. N., Adhikari L. M., Carter

E. J., Puri L., Codlin A. J., and Creswell J. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Scientific Reports*, 9(1):15000, 2019.

[403] Santos A. d. S., Oliveira R. D. d., Lemos E. F., Lima F., Cohen T., Cords O., Martinez L., Gonçalves C., Ko A., Andrews J. R., and Croda J. Yield, Efficiency and Costs of Mass Screening Algorithms for Tuberculosis in Brazilian Prisons. *Clinical Infectious Diseases*, 72(5):771–777, 2020.

[404] Liang C.-H., Liu Y.-C., Wu M.-T., Garcia-Castro F., Alberich-Bayarri A., and Wu F.-Z. Identifying pulmonary nodules or masses on chest radiography using deep learning: External validation and strategies to improve clinical practice. *Clinical Radiology*, 75(1):38–45, 2020.

[405] Hwang E. J., Nam J. G., Lim W. H., Park S. J., Jeong Y. S., Kang J. H., Hong E. K., Kim T. M., Goo J. M., Park S., Kim K. H., and Park C. M. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology*, 293(3):573–580, 2019.

[406] Singh R., Kalra M. K., Nitiwarangkul C., Patti J. A., Homayounieh F., Padole A., Rao P., Putha P., Muse V. V., Sharma A., and Digumarthy S. R. Deep learning in chest radiography: Detection of findings and presence of change. *PLOS One*, 13(10):e0204155, 2018.

[407] Nash M., Kadavigere R., Andrade J., Sukumar C. A., Chawla K., Shenoy V. P., Pande T., Huddart S., Pai M., and Saravu K. Deep learning, computer-aided radiography reading for tuberculosis: A diagnostic accuracy study from a tertiary hospital in India. *Scientific Reports*, 10(1):210, 2020.

[408] Engle E., Gabrielian A., Long A., Hurt D. E., and Rosenthal A. Performance of Qure.ai automatic classifiers against a large annotated database of patients with diverse forms of tuberculosis. *PLOS One*, 15(1):e0224445, 2020.

[409] Kim J. R., Shim W. H., Yoon H. M., Hong S. H., Lee J. S., Cho Y. A., and Kim S. Computerized Bone Age Estimation Using Deep Learning Based Program: Evaluation of the Accuracy and Efficiency. *American Journal of Roentgenology*, 209(6):1374–1380, 2017.

[410] Homayounieh F., Digumarthy S. R., Febbo J. A., Garrana S., Nitiwarangkul C., Singh R., Khera R. D., Gilman M., and Kalra M. K. Comparison of Baseline, Bone-Subtracted, and Enhanced Chest Radiographs for Detection of Pneumothorax. *Canadian Association of Radiologists Journal*, 72(3):519–524, 2020.

[411] Dellios N., Teichgraeber U., Chelaru R., Malich A., and Papageorgiou I. E. Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. *Journal of Clinical Imaging Science*, 7:8, 2017.

[412] Schalekamp S., Karssemeijer N., Cats A. M., De Hoop B., Geurts B. H. J., Berger-Hartog O., van Ginneken B., and Schaefer-Prokop C. M. The Effect of Supplementary Bone-Suppressed Chest Radiographs on the Assessment of a Variety of Common Pulmonary Abnormalities: Results of an Observer Study. *Journal of Thoracic Imaging*, 31(2):119–125, 2016.

[413] Schalekamp S., Ginneken B. v., Berk I. A. H. v. d., Hartmann I. J. C., Snoeren M. M., Odink A. E., Lankeren W. v., Pegge S. A. H., Schijf L. J., Karssemeijer N., and Schaefer-Prokop C. M. Bone Suppression Increases the Visibility of Invasive Pulmonary Aspergillosis in Chest Radiographs. *PLOS One*, 9(10):e108551, 2014.

[414] Schalekamp S., van Ginneken B., Koedam E., Snoeren M. M., Tiehuis A. M., Wittenberg R.,

Karssemeijer N., and Schaefer-Prokop C. M. Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images. *Radiology*, 272(1):252–261, 2014.

[415] Szucs-Farkas Z., Schick A., Cullmann J. L., Ebner L., Megyeri B., Vock P., and Christe A. Comparison of dual-energy subtraction and electronic bone suppression combined with computer-aided detection on chest radiographs: Effect on human observers' performance in nodule detection. *American Journal of Roentgenology*, 200(5):1006–1013, 2013.

[416] López-Cabrera J. D., Orozco-Morales R., Portal-Diaz J. A., Lovelle-Enríquez O., and Pérez-Díaz M. Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging. *Health and Technology*, 11(2):411–424, 2021.

[417] DeGrave A. J., Janizek J. D., and Lee S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2020.

[418] Garcia Santa Cruz B., Bossa M. N., Sölter J., and Husch A. D. Public covid-19 x-ray datasets and their impact on model bias – a systematic review of a significant problem. *Medical Image Analysis*, 74:102225, 2021.

[419] Maguolo G. and Nanni L. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 76:1–7, 2021.

[420] Isensee F., Jaeger P. F., Kohl S. A. A., Petersen J., and Maier-Hein K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[421] Sheller M. J., Reina G. A., Edwards B., Martin J., and Bakas S. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 92–104, 2019.

[422] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

[423] Simonyan K., Vedaldi A., and Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv*, 2014.

[424] Demner-Fushman D., Kohli M. D., Rosenman M. B., Shooshan S. E., Rodriguez L., Antani S. K., Thoma G. R., and McDonald C. J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[425] MacMahon H., Doi K., Chan H.-P., Giger M. L., Katsuragawa S., and Nakamori N. Computer-aided diagnosis in chest radiology. *Journal of Thoracic Imaging*, 5(1):67–76, 1990.

[426] Dimopoulos K., Giannakoulas G., Bendayan I., Liodakis E., Petraco R., Diller G.-P., Piepoli M. F., Swan L., Mullen M., Best N., Poole-Wilson P. A., Francis D. P., Rubens M. B., and Gatzoulis M. A. Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. *International Journal of Cardiology*, 166(2):453–457, 2013.

[427] Yao L., Prosky J., Poblenz E., Covington B., and Lyman K. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

[428] Gündel S., Grbic S., Georgescu B., Zhou S. K., Ritschl L., Meier A., and Comaniciu D. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. *arXiv:1803.04565*, 2018.

[429] Yao L., Poblenz E., Dagunts D., Covington B., Bernard D., and Lyman K. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017.

[430] van Ginneken B., Stegmann M., and Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006.

[431] Candemir S., Jaeger S., Lin W., Xue Z., Antani S. K., and Thoma G. R. Automatic heart localization and radiographic index computation in chest x-rays. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 9785, pages 302 – 309, 2016.

[432] Dallal A. H., Agarwal C., Arbabshirani M. R., Patel A., and Moore G. Automatic estimation of heart boundaries and cardiothoracic ratio from chest x-ray images. In *SPIE Medical Imaging: Computer-Aided Diagnosis*, volume 10134, pages 134 – 143, 2017.

[433] Buda M., Maki A., and Mazurowski M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[434] Bergstra J., Komer B., Eliasmith C., Yamins D., and Cox D. D. Hyperopt: A python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1): 014008, 2015.

[435] Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., Kudlur M., Levenberg J., Monga R., Moore S., Murray D. G., Steiner B., Tucker P., Vasudevan V., Warden P., Wicke M., Yu Y., and Zheng X. TensorFlow: A System for Large-Scale Machine Learning. In *USENIX Conference on Operating Systems Design and Implementation*, pages 265–283, 2016.

[436] Chollet F. et al. Keras. `https://keras.io`, 2015. Accessed: 2024-05-15.

[437] Chlebus G., Schenk A., Hendrik Moltz J., van Ginneken B., Meine H., and Karl Hahn H. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Scientific Reports*, 8(1):15497, 2018.

[438] Clevert D., Unterthiner T., and Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv:1511.07289*, 2015.

[439] Duchi J., Hazan E., and Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[440] Klambauer G., Unterthiner T., Mayr A., and Hochreiter S. Self-normalizing neural networks. In *International Conference on Neural Information Processing Systems*, pages 972–981, 2017.

[441] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

[442] Ioffe S. and Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, volume 37, pages 448–456, 2015.

[443] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[444] Efron B. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158, 1981.

[445] Newcombe R. G. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998.

[446] Rubin J., Sanghavi D., Zhao C., Lee K., Qadir A., and Xu-Wilson M. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*, 2018.

[447] Candemir S., Rajaraman S., Thoma G., and Antani S. Deep learning for grading cardiomegaly severity in chest x-rays: An investigation. In *Life Sciences Conference*, pages 109–113, 2018.

[448] Que Q., Tang Z., Wang R., Zeng Z., Wang J., Chua M., Gee T. S., Yang X., and Veeravalli B. CardioXNet: Automated detection for cardiomegaly based on deep learning. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 612–615. IEEE, 2018.

[449] Rolnick D., Veit A., Belongie S. J., and Shavit N. Deep learning is robust to massive label noise. *arXiv:1705.10694*, 2017.

[450] Quanjer P. H., Tammeling G. J., Cotes J. E., Pedersen O. F., Peslin R., and Yernault J.-C. Lung volumes and forced ventilatory flows. *European Respiratory Journal*, 6(Suppl 16):5–40, mar 1993.

[451] Flesch J. D. and Dine C. J. Lung volumes. *Chest*, 142(2):506–510, aug 2012.

[452] Pedone C., Scarlata S., Chiurco D., Conte M. E., Forastiere F., and Antonelli-Incalzi R. Association of reduced total lung capacity with mortality and use of health services. *Chest*, 141(4): 1025–1030, apr 2012.

[453] Tantucci C., Bottone D., Borghesi A., Guerini M., Quadri F., and Pini L. Methods for measuring lung volumes: Is there a better one? *Respiration*, 91(4):273–280, 2016.

[454] Coxson H. O., Fauerbach P. V. N., Storness-Bliss C., Muller N. L., Cogswell S., Dillard D. H., Finger C. L., and Springmeyer S. C. Computed tomography assessment of lung volume changes after bronchial valve treatment. *European Respiratory Journal*, 32(6):1443–1450, dec 2008.

[455] Iwano S., Okada T., Satake H., and Naganawa S. 3d-CT volumetry of the lung using multidetector row CT. *Academic Radiology*, 16(3):250–256, mar 2009.

[456] Lundsgaard C. and Slyke D. D. V. STUDIES OF LUNG VOLUME. i. *Journal of Experimental Medicine*, 27(1):65–86, jan 1918.

[457] Ries A. Measurement of lung volumes. *Clinics in chest medicine*, 10(2):177—186, June 1989.

[458] Harris T. R., Pratt P. C., and Kilburn K. H. Total lung capacity measured by roentgenograms. *The American Journal of Medicine*, 50(6):756–763, jun 1971.

[459] Cobb S., Blodgett D. J., Olson K. B., and Stranahan A. Determination of total lung capacity in disease from routine chest roentgenograms. *The American Journal of Medicine*, 16(1):39–54, jan 1954.

[460] Schlesinger A. E., White D. K., Mallory G. B., Hildeboldt C. F., and Huddleston C. B. Estimation of total lung capacity from chest radiography and chest CT in children: comparison with body

plethysmography. *American Journal of Roentgenology*, 165(1):151–154, jul 1995.

[461] Barnhard H. J., Pierce J. A., Joyce J. W., and Bates J. H. Roentgenographic determination of total lung capacity. *The American Journal of Medicine*, 28(1):51–60, jan 1960.

[462] Pierce R. J., Brown D. J., Holmes M., Cumming G., and Denison D. M. Estimation of lung volumes from chest radiographs using shape information. *Thorax*, 34(6):726–734, dec 1979.

[463] Loyd H. M., String S. T., and DuBois A. B. Radiographie and plethysmographie determination of total lung capacity. *Radiology*, 86(1):7–30, jan 1966.

[464] Regan E. A., Hokanson J. E., Murphy J. R., Make B., Lynch D. A., Beaty T. H., Curran-Everett D., Silverman E. K., and Crapo J. D. Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, mar 2010.

[465] Xie Y., Chen M., Kao D., Gao G., and Chen X. A. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, 2020.

[466] Wanger J., Clausen J. L., Coates A., Pedersen O. F., Brusasco V., Burgos F., Casaburi R., Crapo R., Enright P., van der Grinten C. P. M., Gustafsson P., Hankinson J., Jensen R., Johnson D., MacIntyre N., McKay R., Miller M. R., Navajas D., Pellegrino R., and Viegi G. Standardisation of the measurement of lung volumes. *European Respiratory Journal*, 26(3):511–522, sep 2005.

[467] Simonyan K. and Zisserman A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[468] Shapiro S. S. and Wilk M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591, dec 1965.

[469] Wilk M. B. and Gnanadesikan R. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1, mar 1968.

[470] Haas M., Hamm B., and Niehues S. M. Automated lung volumetry from routine thoracic CT scans. *Academic Radiology*, 21(5):633–638, may 2014.

[471] Daghighi A. and Tropp H. Computed tomography lung volume estimation and its relation to lung capacities and spine deformation. *Journal of Spine Surgery*, 5(1):132–141, mar 2019.

[472] Park C. H., Haam S. J., Lee S., Han K. H., and Kim T. H. Prediction of anatomical lung volume using planimetric measurements on chest radiographs. *Acta Radiologica*, 57(9):1066–1071, jul 2016.

[473] Wade O. L. and Gilson J. C. The effect of posture on diaphragmatic movement and vital capacity in normal subjects. *Thorax*, 6(2):103–126, jun 1951.

[474] Bray F., Ferlay J., Soerjomataram I., Siegel R. L., Torre L. A., and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, sep 2018.

[475] Howlader N., Noone A., Krapcho M., Miller D., Brest A., Yu M., Ruhl J., Tatalovich Z., Mariotto A., Lewis D., Chen H., Feuer E., and Cronin K. Seer cancer statistics review, 1975-2017. *SEER Cancer Statistics Review, 1975-2017*, 2020.

[476] Team N. L. S. T. R. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, aug 2011.

[477] de Koning H. J., van der Aalst C. M., de Jong P. A., Scholten E. T., Nackaerts K., Heuvelmans M. A., Lammers J.-W. J., Weenink C., Yousaf-Khan U., Horeweg N., van 't Westeinde S., Prokop M., Mali W. P., Hoesein F. A. M., van Ooijen P. M., Aerts J. G., den Bakker M. A., Thunnissen E., Verschakelen J., Vliegenthart R., Walter J. E., ten Haaf K., Groen H. J., and Oudkerk M. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *New England Journal of Medicine*, 382(6):503–513, feb 2020.

[478] Toyoda Y., Nakayama T., Kusunoki Y., Iso H., and Suzuki T. Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography. *British Journal of Cancer*, 98(10): 1602–1607, may 2008.

[479] Homayounieh F., Digumarthy S., Ebrahimian S., Rueckel J., Hoppe B. F., Sabel B. O., Conjeti S., Ridder K., Sistermanns M., Wang L., Preuhs A., Ghesu F., Mansoor A., Moghbel M., Botwin A., Singh R., Cartmell S., Patti J., Huemmer C., Fieselmann A., Joerger C., Mirshahzadeh N., Muse V., and Kalra M. An artificial intelligence–based chest x-ray model on human nodule detection accuracy from a multicenter study. *JAMA Network Open*, 4(12):e2141096, dec 2021.

[480] Litjens G. J. S., Hogeweg L., Schilham A. M. R., de Jong P. A., Viergever M. A., and van Ginneken B. Simulation of nodules and diffuse infiltrates in chest radiographs using CT templates. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, pages 396–403. Springer Berlin Heidelberg, 2010.

[481] Schultheiss M., Schmette P., Bodden J., Aichele J., Müller-Leisse C., Gassert F. G., Gassert F. T., Gawlitza J. F., Hofmann F. C., Sasse D., et al. Lung nodule detection in chest x-rays using synthetic ground-truth data comparing cnn-based diagnosis to human performance. *Scientific Reports*, 11(1):1–13, 2021.

[482] https://grand-challenge.org/, 2017.

[483] https://node21.grand-challenge.org/, 2021.

[484] https://github.com/DIAGNijmegen/opencxr, 2020.

[485] Philipsen R. H. H. M., Maduskar P., Hogeweg L., Melendez J., Sanchez C. I., and van Ginneken B. Localized energy-based normalization of medical images: Application to chest radiography. *IEEE Transactions on Medical Imaging*, 34(9):1965–1975, sep 2015.

[486] https://zenodo.org/record/5548363, 2021.

[487] Demner-Fushman D., Kohli M. D., Rosenman M. B., Shooshan S. E., Rodriguez L., Antani S. K., Thoma G. R., and McDonald C. J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[488] https://node21.grand-challenge.org/Annotation_Process/, 2021.

[489] Setio A. A. A., Traverso A., de Bel T., Berens M. S., van den Bogaard C., Cerello P., Chen H., Dou Q., Fantacci M. E., Geurts B., van der Gugten R., Heng P. A., Jansen B., de Kaste M. M., Kotov V., Lin J. Y.-H., Manders J. T., Sóñora-Mengana A., García-Naranjo J. C., Papavasileiou E., Prokop M., Saletta M., Schaefer-Prokop C. M., Scholten E. T., Scholten L., Snoeren M. M., Torres E. L., Vandemeulebroucke J., Walasek N., Zuidhof G. C., van Ginneken B., and Jacobs C. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42: 1–13, dec 2017.

[490] Setio A. A. A., Ciompi F., Litjens G., Gerke P., Jacobs C., van Riel S. J., Wille M. M. W., Naqibullah M., Sanchez C. I., and van Ginneken B. Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, may 2016.

[491] Nguyen H. Q., Lam K., Le L. T., Pham H. H., Tran D. Q., Nguyen D. B., Le D. D., Pham C. M., Tong H. T. T., Dinh D. H., Do C. D., Doan L. T., Nguyen C. N., Nguyen B. T., Nguyen Q. V., Hoang A. D., Phan H. N., Nguyen A. T., Ho P. H., Ngo D. T., Nguyen N. T., Nguyen N. T., Dao M., and Vu V. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020.

[492] Dwibedi D., Misra I., and Hebert M. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[493] Pérez P., Gangnet M., and Blake A. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*. ACM, jul 2003.

[494] https://node21.grand-challenge.org/evaluation/detection-track-final-test-set/leaderboard/, 2022.

[495] Behrendt F., Bengs M., Bhattacharya D., Krüger J., Opfer R., and Schlaefer A. A systematic approach to deep learning-based nodule detection in chest radiographs. *Scientific Reports*, 13 (1), jun 2023.

[496] Tan M., Pang R., and Le Q. V. EfficientDet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[497] Solovyev R., Wang W., and Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, mar 2021.

[498] Yan K., Wang X., Lu L., and Summers R. M. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(03): 1, jul 2018.

[499] https://node21.grand-challenge.org/evaluation/challenge-2/leaderboard/, 2022.

[500] Zeng Y., Lin Z., Lu H., and Patel V. M. CR-fill: Generative image inpainting with auxiliary contextual reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2021.

[501] DeLong E. R., DeLong D. M., and Clarke-Pearson D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44 (3):837–845, 1988.

Acknowledgements

In the immortal words of renowned astronomer Carl Sagan, "To make an apple pie from scratch, one must first invent the universe". Such wisdom holds true in the context of accomplishments as well - each achievement is not an isolated event, but rather the fruit of collective efforts made by those who came before us, as well as those who have supported us in every high and low of our lives. With this understanding deeply ingrained in me, I wish to recognize the numerous individuals whose contributions have been instrumental in the realization of this thesis.

I would like to extend my appreciation to my supervisor, **Bram van Ginneken**, for their guidance, patience, and support during my doctoral journey. His expertise and insightful suggestions have shaped my research and writing, elevating them to new levels. Bram's constructive criticism and patience have refined my work. I have consistently been inspired by Bram's intellectual acumen and his ability to generate sharp and insightful ideas. His understanding of academia and the subject matter at hand is truly remarkable.

I would also like to express my gratitude to my daily supervisor, **Keelin Murphy**, who has been an exceptional mentor and a remarkable person throughout my doctoral journey. Keelin's support and guidance have played an important role in my academic development. Her critical insights and expertise in our field have shaped my research, and her assistance with writing has been invaluable, as she has provided suggestions that have greatly enhanced the quality of my work. However, Keelin's impact extends far beyond academics. She has consistently shown genuine care and compassion, providing support not only as a supervisor but also as a friend. I feel incredibly grateful and privileged to have had Keelin by my side, and I am profoundly thankful for the impact she has had on both my personal and academic growth.

I am very grateful for the friendship and collaboration I had with **Erdi Calli** throughout my PhD journey. Erdi was not only a close friend but also a valued colleague who played a significant role in my academic and personal growth. We worked together on numerous research papers, and his insightful contributions and dedication greatly enriched our collaborations. We were there for each other during challenging times, providing a listening ear, sharing encouragement, and offering understanding. His support and empathy helped me navigate the highs and lows of the PhD journey, and I am grateful for his presence by my side. And thank you Erdi for letting me hold **Ada** in my arms, her pictures always make my days :)

I am extremely grateful for the collaboration and guidance of **Ernst Scholten** throughout my PhD journey. As a radiologist, Ernst has been instrumental in my research, providing essential expertise and insights that have enhanced the quality and impact of my work. His willingness to share his knowledge, along with his exceptional helpfulness, has been a constant source of support. Thank you Ernst, it has been an absolute pleasure to work with you. You have created a fun, productive, and inclusive environment.

I would like to convey my gratitude to **Mathias Prokop**, with whom I had the privilege of collaborating during my work. I am deeply impressed by his knowledge, remarkable patience, and invaluable feedback and insights. Mathias exemplifies the qualities of an extraordinary scientist, demonstrating a remarkable ability to simplify and explain complex concepts with utmost clarity. His contributions have left a lasting impact on me, and I am truly happy for the opportunity to work alongside such a distinguished researcher and mentor.

I am thankful for the friendships I have built with **Anton Schreuder, Gabriel Humpire, Matin Hosseinzadeh, Patrick Brand, and Weiyi Xie** during my doctoral journey. These individuals have not only been my colleagues but have also become cherished friends with whom I have shared the most enjoyable times during my PhD. Our lunch breaks, after-work gatherings, and engaging conversations spanning various topics have served as a welcome escape from the demands of research.

Gabriel, thank you for always being there to help and for being a listening ear (yes, you do deserve to be a second co-author as always, and i forgive you for the biscuit joke:)). I am grateful for our friendship. **Anton**, thank you for your empathetic ear and encouraging words. I appreciate you.
I would like to thank my colleagues **Steven Schaelkamp, Luuk Boulogne, Nils Hendrix, Ward Hendrix, Hans Pinckaers, Henkjan Huisman, Colin Jacobs, Kicky van Leeuwen, and Paul Konstantin Gerke, David Tellez, Bart Liefers, Coen de Vente, James Meakin, John-Melle Bokhorst, Péter Bándi, Marta Pinto, Olga Sliwicka, and many other DIAG members** who I had an opportunity to work with. I would like to extend special thanks to **Cristina González-Gonzalo** for our heartfelt conversations, which have brought warmth to my doctoral journey. Our girls night out are always something special, and I am happy we still make time for them :) Additionally, I am very grateful to **Kiran Vaidhya Venkadesh and his wife Ann** for their kindness in taking care of our beloved dogs during our absences. Kiran and Ann, thank you for being so kind and warm. I understood the depth of our friendship when you met me even when my dogs were not with me. I am very thankful for the friendship we have built.
I would like to express my appreciation to my colleagues at **Adyen, namely Joao, Sean, Rik, Aakansha, Rhys, Rolf, Kris, Sai, Sertac, Vivian, Thijs and Mo**, and others. Meeting individuals like you has been an absolute privilege, you have all been extraordinary colleagues. A special thanks to **Joao and Sean**— I genuinely worry I will never meet people like you again. You make Gen(ie)Squad truly special.
Thank you to my Spanish family — my beloved suegros **Amable and Suso, my cuñadas Maria and Nati, Fani, and our cherished little ones, Eire and Alvaro**. I am deeply grateful to each one of you for your love and support. You make me feel home, os quiero a todos.
I would like to take a moment to express my deep appreciation for my beloved mother **Tuba** and my grandparents, **Guler and Orhan Kanturvardar**. Being your daughter and granddaughter have been an immense blessing in my life. You have consistently stood by my side through the ups and downs, offering unwavering support and celebrating every achievement with joy. I am forever grateful for your unconditional love and presence in my life. Thank you from the bottom of my heart for being such remarkable family and for shaping me into the person I am today. I would like to extend my gratitude to my cousin, **Fulden**, who has been more than just a family member — she has been like another sister to me. I am incredibly grateful for her empathetic ear, understanding, and boundless compassion. Her caring nature and genuine understanding have made a profound impact, and I am fortunate to have her by my side. I also want to extend a special thank you to my aunt, **Fulya**, for her support and compassion. Her kindness and compassion have been a source of strength and comfort throughout this journey. Annem, ananem, dedem, Gizem, Fulden, teyzem ve enişte m, beni koşulsuz sevdiğiniz için ve her an yanımda olduğunuz için size minnettarım. Siz benim evimsiniz, ve ben İzmir'deki küçük apartman dairemizde yemek masamızda oturduğumuz halimizi hep yanımda taşıyorum. Sizleri çok seviyorum. Bu doktora tezi siz olmadan yazılamazdı.
Thank you to my childhood friends **Funda and Simge**. I feel so lucky to have you in my life. Even when we meet after a long time, it always feels like we were only apart for a few minutes, picking up right where we left off. You both are so special to me. I am thankful for the friendship I built in Amsterdam, my Turkish friends **Ipek, Cansu, Onur, Cagri**. Spending time with you made me feel close to home.
I would like to thank to **Amsterdam**, a city that gave me a fresh and beautiful start, surrounded by a diverse and vibrant international community. Strolling with my dogs in Westerpark and taking in the stunning canals and lights on my daily walks home from work filled me with inspiration and a sense of belonging.

Figure 7.1: Our office in DIAG where it all started (Thank you Gabriel for sharing this picture with me)

I would like to express my deepest gratitude to the most critical pillars of my life, my family - affectionately known as the **Quadros**. This group comprises my loving husband **Jesus**, my dear sister **Gizem**, her supportive husband **Zeki**, and of course, myself. Together, we create a vibrant and loving family unit that has been my sanctuary, inspiration, and support system.
A special nod goes to our canine family members - our delightful dogs, **Oslo and Frida**, and the charming **Mars** who is Gizem and Zeki's sweet companion. Their joyful presence, innocent affection, and unconditional love have often been the stress reliever amidst the pressures of this work. They remind me of life's simple pleasures even in the most complex times.
To my husband - You have believed in me even when I doubted myself, always there to listen, offer advice, and patiently help me through it all. And honestly, the little things—like baking a cake when I needed a break—meant the world to me. Your support has been my rock, and I could not have done this without you. To my sister and her husband – your encouragement, advice, and constant support have meant everything. You have always been there to help carry my worries, celebrate the wins, and give me a fresh perspective when I needed it most. To all of you, I am profoundly grateful. The accomplishment of this thesis is not merely my own; it is a testament to our collective effort, love, and resilience. Thank you for being my universe as I invented this 'apple pie'.

Curriculum Vitae

Ecem Sogancioglu was born in Izmir, Turkiye on March 15, 1990. She received her Bachelor's degree in Computer Science at Hacettepe University. She received her Master's degree in Computer Science at the University of Freiburg with specialization in Artificial Intelligence on March 2017. She joined the Diagnostic Image Analysis Group as a PhD candidate on October 2017. Her research is focused on deep learning algorithms for Chest X-Rays, under the supervision of Bram van Ginneken and Keelin Murphy. The results of her work is in your hands right now (or in your screen).

PhD Portfolio

Name: Ecem Sogancioglu
Graduate school: Radboud Institute for Health Sciences (RIHS)
PhD period: 02-10-2017 until 30-11-2021

Courses & workshops	Year(s)	ECTS
Radboudumc International Introduction Day	2017	0.25
Radboudumc Work Safety	2017	0.5
Coursera Deep Learning Specialization	2017-2018	5
Radboud University Mindfulness-based Stress Reduction	2018	2
Scientific writing for PhD candidates	2018	3
Achieving your goals and performing more successfully in your PhD	2019	1
Scientific integrity	2020	1
Teaching Activities		
Teaching assistant at Intelligent Systems in Medical Imaging (RU)	2019	1.5
Seminars & lectures		
Radiology research meeting	2017-2021	3
Deep learning Nijmegen meetup	2017-2021	0.5
DIAG discussion hour	2017-2021	5
Deep learning journal club	2017-2021	4
DLMedia regular meetup	2017-2021	8
Symposia & congresses		
NFBIA Summer School	2017	3
International Summer School on Deep Learning	2018	3
Medical Imaging with Deep Learning	2018	2
Medical Imaging with Deep Learning	2019	2
Radboud Frontiers	2019	0.25
Other		
Grand Challenge Website	2018-2020	2
Diag/Axti Weekend	2020	1

Research Data Management

This thesis is based on the results of medical-scientific research with human participants. The research project described in this PhD thesis makes use of an extensive amount of data with the purpose of training and evaluation several machine learning algorithms. This data consists of three main components: (1) digitized Chest X-ray images of patient internals, (2) labels that describe these images at patient level, and (3) test results indicating various patient health markers.

Regarding the origin, ownership, and permission to use this data, we strictly follow the regulations of the Radboudumc. The medical ethical committee Radboud CMO, Nijmegen, the Netherlands has given approval to conduct these studies. Informed consent was obtained from research participants. Technical and organizational measures were followed to safeguard the availability, integrity and confidentiality of the data (these measures include the use of independent monitoring, pseudonymization, access authorization and secure data storage).

Data for Chapter 4 and 5 was collected and securely stored within the Radboudumc storage system. More generally, all scientific experiments conducted within the context of this research project have been executed exclusively within the Radboudumc IT infrastructure. In order to protect patients' privacy rights, all data used within the context of this research project has been subject to pseudonymization. This process ensures that personally identifiable information is replaced by artificial identifiers, or pseudonyms, before conducting any of the experiments described within this thesis. We adhere to the FAIR data principles (findable, accessible, interoperable and re-usable) whenever possible. This data is only accessible by project members working at the Radboudumc. The data will be archived for 15 years after termination of the study. Reusing the data for future research is only possible after a renewed permission by the participants. The anonymous datasets that were used for analysis are available from the corresponding author upon reasonable request.

The Chapter 5 additionally contains labels of publicly available data, and this labels were shared under Zenodo[1], an open repository operated by CERN, in fully anonymized form for public use. Data were made reusable by adding sufficient documentation (research protocol, codebook and a readme file), by using preferred and sustainable data formats and by publishing under the CC.BY.4.0 license. The data not suitable for reuse will be archived for 15 years after termination of the study.

[1] `https://zenodo.org/records/5548363`

9 789493 296770

Radboudumc university medical center

Radboud University