Deep Learning for Chest X-ray Analysis: A Survey

Ecem Sogancioglu*, Erdi Çalli*, Bram van Ginneken, Kicky G. van Leeuwen, Keelin Murphy

*Radboud University Medical Center, Institute for Health Sciences, Department of Medical Imaging, Nijmegen, The Netherlands

Abstract

Recent advances in deep learning have led to a promising performance in many medical image analysis tasks. As the most commonly performed radiological exam, chest radiographs are a particularly important modality for which a variety of applications have been researched. The release of multiple, large, publicly available chest X-ray datasets in recent years has encouraged research interest and boosted the number of publications. In this paper, we review all studies using deep learning on chest radiographs, categorizing works by task: image-level prediction (classification and regression), segmentation, localization, image generation and domain adaptation. Commercially available applications are detailed, and a comprehensive discussion of the current state of the art and potential future directions are provided.

Keywords:
Deep Learning, chest radiograph, chest X-ray analysis, survey

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 (United Nations, 2008). In 2006, it is estimated that 129 million CXR images were acquired in the United States alone (Mettler et al., 2009). 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 pathological patterns. The CXR is often the first imaging study acquired and remains central to screening, diagnosis, and management of a broad range of conditions (Raoof et al., 2012).

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 [1]

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 (Quekel et al., 2001; Balabanova et al., 2005; Young, 1994).

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 (Lodwick et al., 1963; Becker et al., 1964; Meyers et al., 1964; Kruger et al., 1972; Toriwaki et al., 1973). 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 (Litjens et al., 2017). 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 (Wang et al., 2017b). In 2019 alone, more than 755,000 images were released in 3 labeled databases (CheXpert (Irvin et al., 2019), MIMIC-CXR (Johnson et al., 2019), PadChest (Bustos et al., 2020)). 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 (Litjens et al., 2017; van Ginneken et al., 2018). A cornerstone of radiological imaging for many decades...

*Corresponding author: E-mail: erdi-calli@radboudumc.nl

Preprint submitted to Medical Image Analysis
Figure 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 (Irvin et al., 2019), patient 184.

The remainder of this work is structured as follows: Section 2 provides a brief introduction to the concept of deep learning and the main network architectures encountered in the current literature. In Section 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 4 categorized according to the major themes identified. Commercial systems available for chest radiograph analysis are described in Section 5. The paper concludes in Section 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. 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 (Litjens et al., 2017). 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 2.3 to 2.6 describe commonly used architectures categorized by types of application in the CXR literature.

2.1. Convolutional Neural Networks

In the 1980s, networks using convolutional layers were first introduced for image analysis (Fukushima and Miyake, 1982), and the idea was formalized over the following years (LeCun and Bengio, 1998). 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. 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 (Yosinski et al., 2014). 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) (Baltruschat et al., 2019b) 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 2.7.

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 (Krizhevsky et al., 2012), 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 (Deng et al., 2009) by a large margin in 2012. Since then many deep convolutional neural network architectures have been proposed. The VGG family of models (Simonyan and Zisserman, 2014) use 8 to 19 convolutional layers followed by 3 fully-connected layers. The Inception architecture was first introduced in 2015 (Szegedy et al., 2015) using multiple convolutional filter sizes within layered blocks known as Inception modules. In 2016, the ResNet family of models (He et al., 2016) 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 (Huang et al., 2017), 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) (Szegedy et al., 2017). The Xception network architecture (Chollet, 2017) 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.

2.4. 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 (Chen et al., 2018b). 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 (Shelhamer et al., 2017). 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 (Ronneberger et al., 2015) and this work has become the most cited paper in the history of medical image analysis. The U-Net consists of several convolutional layers in a contracting (downsampling) path, followed by further convolutional layers in an expanding (upsampling) path which restore the result to the input resolution. It additionally uses skip connections 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.
2.5. 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 (Girshick et al., 2014), 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 (Girshick, 2015) and subsequently by faster-RCNN (Ren et al., 2017) 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 (He et al., 2017). 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 (Redmon et al., 2016) as a single-stage object detection method, and improved in subsequent versions in 2017 and 2018 (Redmon and Farhadi, 2017, 2018). 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 (Lin et al., 2017). 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.

2.6. Image Generation Networks

Image generation was first popularized with the introduction of the generative adversarial network (GAN) in 2014 (Goodfellow et al., 2014). 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 (Salimans et al., 2016; Heusel et al., 2017; Karras et al., 2018; Arjovsky et al., 2017). GANs have also been adapted to conditional data generation (Chen et al., 2016; Odena et al., 2017) by incorporating class labels, image-to-image translation (conditioned on an image in this case) (Isola et al., 2017), and unpaired image-to-image translation (CycleGAN, Zhu et al., 2017).

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 (Yi et al., 2019b). Many of the image generation works identified in this review employed GAN based architectures.

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 Wang and Deng, 2018): 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.

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 other available to the best of our knowledge. Details are provided in Table 1. Each dataset is given an acronym which is used in the literature review tables (Tables 2 to 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 [Wang et al. 2017]. 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 [Irvin et al. 2019]. 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 [Johnson et al. 2019]. 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 [Bustos et al. 2020]. 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) [Bodenreider 2004].

5. PLCO (PL) is a screening trial for prostate, lung, colorectal and ovarian (PLCO) cancer [Zhu et al. 2013]. 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 [Demner-Fushman et al. 2012]. The CXRs are collected from the Indiana Network for Patient Care [McDonald et al. 2005]. The images are distributed as anonymized DICOMs. The radiological findings obtained by radiologist interpretation are available in MeSH format.

7. Ped-Pneumonia (PP) is a dataset consisting of 5, 856 pediatric CXRs [Kermany 2018]. 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 [Shiraishi et al. 2000]. 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 [van Ginneken et al. 2006] and we group these datasets together in this work.

9. RSNA-Pneumonia (RP) is a dataset consisting of 30, 000 CXRs with pneumonia annotations [RSNA 2018]. 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.

10. Shenzhen (S) is a dataset consisting of 662 CXRs [Jaeger et al. 2014]. The CXRs are collected at Shenzhen No.3 Hospital in 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.

https://www.nlm.nih.gov/mesh/meshhome.html
Table 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) | View Positions | Annotation Types | Image Format | Labeling method | Gold Standard Data |
|-------------|----------------|------------------|--------------|-----------------|-------------------|
| **ChestX-ray14** (C) | | | | | |
| Wang et al. (2017b) | | | | | |
| P: 31K | I: 112K | PA: 67K | BB | 14 | 112K | PNG | RP | 984 |
| | | AP: 45K | | 8 | 983 | |

| **CheXpert (X)** | | | | | |
| Irvin et al. (2019) | | | | | |
| P: 65K | S: 189K | I: 224K | PA: 29K | BB | 14 | 224K | JPEG | RI | 235 |
| | | | AP: 162K | | | | | | |
| | | | LL: 32K | | | | | | |

| **MIMIC-CXR** (M) | | | | | |
| Johnson et al. (2019) | | | | | |
| P: 65K | S: 224K | I: 372K | PA:AP 250K | R (V1) | 14 | 372K | JPEG (V1) | RP | |
| | | | LL: 122K | (V2) | | | | | |
| | | | | | | | | | |

| **PadChest** (P) | | | | | |
| Biston et al. (2020) | | | | | |
| P: 67K | S: 110K | I: 160K | PA: 96K | R | 193 | 110K | DICOM | RIR | 27593 |
| | | | AP: 20K | | | | | | |
| | | | LL: 51K | | | | | | |

| **PLCO (PL)** | | | | | |
| Zhu et al. (2013) | | | | | |
| P: 25K | S: 89K | I: 89K | PA: 89K | R | 22 | 89K | TIFF | RI | All |
| | | | | | | | | | |

| **Open-i (O)** | | | | | |
| Demner-Fushman et al. (2012) | | | | | |
| I: 5,856 | | | | | |
| | | | | | | | | | |

| **Ped-pneumonia (PP)** | | | | | |
| Kermani [2018] | | | | | |
| I: 5,910 | | | | | |
| | | | | | | | | | |

| **JSRT+SCR (J)** | | | | | |
| Shirazi [2007] | | | | | |
| I: 247 | | | | | |
| | | | | | | | | | |

| **Shenzhen (S)** | | | | | |
| Jaeger et al. (2014) | | | | | |
| I: 340 | | | | | |
| | | | | | | | | | |

| **Montgomery (MO)** | | | | | |
| Jaeger et al. (2014) | | | | | |
| I: 138 | | | | | |
| | | | | | | | | | |

| **BIMCV (B)** | | | | | |
| Vayá et al. (2020) | | | | | |
| I: 1,305 | | | | | |
| | | | | | | | | | |

| **COVIDDSL (CD)** | | | | | |
| HMHospitales | | | | | |
| I: 1,725 | | | | | |
| | | | | | | | | | |

| **COVIDGR (CG)** | | | | | |
| Tabik et al. (2020) | | | | | |
| I: 852 | | | | | |
| | | | | | | | | | |

| **SIIM-ACR (SE)** | | | | | |
| ACR [2019] | | | | | |
| I: 16K | | | | | |
| | | | | | | | | | |

| **CXR14-Rad-Labels (CR)** | | | | | |
| Majewska et al. (2019) | | | | | |
| I: 1,709 | | | | | |
| | | | | | | | | | |

| **COVID-CXR (CC)** | | | | | |
| Cohen et al. (2020c) | | | | | |
| I: 1,866 | | | | | |
| | | | | | | | | | |

| **NLST (N)** | | | | | |
| National Lung Screening Trial Research Team et al. [2011] | | | | | |
| I: 5,493 | | | | | |
| | | | | | | | | | |

| **Object-CXR (OB)** | | | | | |
| Belarus (BL) | | | | | |
| I: 300 | | | | | |
| | | | | | | | | | |

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 [Jaeger et al., 2014]. 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 [Vayá et al., 2020]. 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 [HMHospitales]. 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 [Tabik et al., 2020]. 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 [ACR, 2019]. 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 [Majkowska et al., 2019]. 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 (CV) is a dataset consisting of 930 CXRs at the time of writing (the dataset remains in continuous development) [Cohen et al., 2020c]. 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 [National Lung Screening Trial Research Team et al., 2011]. 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 (https://jifhealthcare.github.io/object-CXR/) is no longer available at the time of writing. 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 illustrates the cumulative number of publicly available CXR images and the number of publications on deep learning with CXR per year.

Figure 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.

### 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 [Irvin et al., 2019; Oakden-Rayner, 2020]. 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 [Olatunji et al., 2019]. 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 [Olatunji et al., 2019]. Acknowledging some of these issues, Irvin et al. [2019] 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’ [Calli et al., 2019; Oakden-Rayner, 2020].

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 [Calli et al., 2019; Rolnick et al., 2018]. 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 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.

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 3 shows the number of studies for each of the tasks.

4.1. 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.

We classified 187 studies, fully detailed in Table 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 4.

The most commonly studied image-level prediction task is predicting the labels of the ChestX-ray14 dataset (31 studies). For example, [Baltruschat et al. (2019a)] compares the performance of various approaches to classify the 14 disease labels provided by the ChestX-ray14 dataset. [Rajpurkar et al. (2018)] 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 ex-
Table 2: Image-Level Prediction Studies (Section 4.1).

| Citation                        | Method                                                                 | Other Tasks       | Labels            |
|---------------------------------|------------------------------------------------------------------------|-------------------|-------------------|
| Liu et al. (2019)               | Combines lung cropped CXR model and a CXR model to improve model performance | SE, LC, PR        | C, LJ, CJ         |
| Sogancioglu et al. (2020)       | Comparison of image-level prediction and segmentation models for cardiology | SE                 | CM                |
| Que et al. (2018)               | A network with DenseNet and U-Net for classification of cardiology       | SE                 | CM                |
| Li et al. (2019)                | U-Net based model for heart and lung segmentation for cardiothoracic ratio | SE                 | CM, PR            |
| Mohan et al. (2020)             | Combines lung cropped CXR model and a CXR model using the segmentation quality | SE                 | E, LO, PE, PT, Z M |
| Li et al. (2019)                | Pneumonia detection is improved by use of lung segmentation              | SE                 | PM                |
| Hurt et al. (2020)              | U-Net based model to segment pneumonia                                   | SE                 | PM, RP            |
| Wang et al. (2020)              | Multi-scale DenseNet based model for pneumonia segmentation              | SE                 | PT                |
| Liu et al. (2020b)              | DenseNet based U-Net for segmentation of the left and right humerus of the infant | SE                 | Z                 |
| Ouali et al. (2020)             | Uses a database of the intermediate ResNet-50 features to find similar studies | OT, IR            | TB, M, OS        |
| Nguyen et al. (2019)            | Uses activation and gradient based attention for localization and classification | LC                 | C                |
| Samala et al. (2019)            | Detects and localizes COVID-19 using various networks and ensembling     | LC                 | C                |
| Park et al. (2020)              | GoogleNet trained with CXR patches, correlates with COVID-19 severity score | LC                 | CV, PM           |
| Nham et al. (2019)              | Proposes a segmentation and classification model compares with radiologist cohort | LC                 | ND, PE, PT       |
| Perez et al. (2019)             | Trains a semisupervised network on a large CXR dataset with CT-confirmed nodule cases | LC                 | ND, PR           |
| Tugrul et al. (2019)            | Defines a loss that minimizes the saliency map errors to improve model performance | LC                 | PM, C            |
| Li et al. (2019)                | A weakly supervised localization with variational model, leverages attention maps | LC                 | PM, RP           |
| Wang et al. (2019)              | Attention guided CNN for pneumonia detection with bounding boxes         | LC                 | T                 |
| Nasri et al. (2020)             | Introduces a visualization method to identify regions of interest from classification | LC                 | T                 |
| Park et al. (2020)              | Weakly supervised framework jointly trained with localization and classification | LC                 | TB, PR           |
| Liu et al. (2019)               | Combines classification loss and autencoder reconstruction loss           | IG, SE            | T                 |
| Wang et al. (2019b)             | Wasserman GAN to permute diseased radiographs to appear healthy          | IG, LC, Z         |
| Wience et al. (2019)            | Novel GAN model trained with healthy and abnormal CXR to predict difference map | IG                 | PE                |
| Tang et al. (2019)              |GANs with U-Net autoencoder and CNN discriminator and encoder for one-class learning | IG                 | T                 |
| Miao et al. (2019)              | Autencoder uses uncertainty for reconstruction in one-class learning setting | IG                 | T                 |
| Longa et al. (2019)             | Continual learning methods to classify data from new domains             | DA                 | C, M             |
| Ling et al. (2019)              | CycleGAN model to adapt adult to pediatric CXR for pneumonia classification | DA                 | PP, PP           |
| Oyayama et al. (2019)           | Trains a Variational Autencoder, uses encoded features to train models   | WS                 | X                 |
| Gwyd et al. (2020)              | Predicts labels for unlabeled data using latent space similarity for semisupervision | WS                 | X                 |
| McMang et al. (2020)            | Y-Net to normalize image geometry for preprocessing                       | SE, PR            |
| Huan et al. (2019)              | COVID-19 opacity localization and severity detection on CXRs             | SE, LC, CV, PR    |
| Or et al. (2019)                | ResNet-18 backbone for Covid-19 classification with limited data availability | SE, LC, CV, PP    |
| Tabak et al. (2020)             | Proposes a new dataset COVIDOR and a novel method using transformations GANs | SEL, GC, CV       |
| Hummel et al. (2020)            | DenseNet for cardiology detection given lung cropped CXR                   | SE                 | CM, O, P         |
| Tirtatung et al. (2020)         | Multiple models and combinations of CXR datasets for COVID-19 detection   | SE                 | CV                |
| Kusakumuran et al. (2021)       | ResNet-101 trained for COVID-19, heatmaps are generated for lung-segmented regions | SE                 | CV, PM           |
| Narayan et al. (2020)           | Multiple architectures considered for two-stage classification of pediatric pneumonia | SE                 | PM, PP            |
| Orange et al. (2020)            | Compares visualization methods for pneumonia localization                 | SE                 | PP                |
| Buchmuller et al. (2018b)       | Classifies patches and uses the positive area size to classify the image  | SE                 | PT, PR            |
| Rajaram et al. (2018b)          | Feature extraction from CNN and segmenting methods                         | SE                 | TB, MO, PR, S     |
| Subramanian et al. (2019)       | Detection of central venous catheters using segmentation shape analysis   | SE                 | TU                |
| Mathieu et al. (2019)           | Detection of air-trapping in pediatric CXRs using Stacked Autoencoders   | SE                 | Z, PR             |
| Wang et al. (2019)              | Pneumococci detection using Inception-v3 and evaluation against two radiologists | SE                 | Z, PR             |
| Liu et al. (2019)               | Introduces CheXpert dataset and model performance on radiologist labeled test set | XL                 | X                 |
| Oly et al. (2019)               | Curates data for interval change detection, proposes method comparing local features | X                  |
| Oh et al. (2019)                | Parses reports to define a topic model and predicts those using CXRs      | RP                 | CM, PE, Z         |
| Ohm et al. (2020)               | Trains model using image and reports to improve image only performance    | RP                 | E                 |
| Chaitra et al. (2019)           | Extracts ambiguity of labels from reports, proposes model that uses this information | RP                 | E, PT, Z          |
| Karak et al. (2019)             | Creates and parses reports for ChestX-ray14 AF data to obtain 73 labels for training | RP                 | MN                 |
| Syeda-Mahmood et al. (2019)     | Obtains findings by tagging common report sentences to train models       | RP                 | MN, PR            |
| Hessel et al. (2019)            | An ensemble of two CNNs to predict priority level for CXR queue management | RP                 | T, PR             |
| Amarasinghe et al. (2019)       | Evaluates bone suppression and lung segmentation, detection of 8 abnormalities | PR, SE            |
| Hulirschal et al. (2019a)       | Classification of pediatric pneumonia types using adapted VGG-16 architecture | PR, SE            |
| Pfeiffer et al. (2019a)         | Evaluates various image preprocessing algorithms on performance of DenseNet-121 | PR                 | T, CM, PR, S      |
| Nie et al. (2020)               | Detects 8 findings and analyzes how these can improve workflow prioritization | OT                 |
| Hulirischal et al. (2020)       | Proposes a model for weakly supervised classification and localization     | LC, WS             |
| Hermonzi et al. (2020)          | Proposes a recurrent attention mechanism to improve model performance      | LC                 |
| Sheine et al. (2020)            | Evaluates the use of various model configurations for classification       | LC                 |
| Hulischal et al. (2020)         | Attention mining and knowledge preservation for classification with localization | LC                 |
| Cat et al. (2018)               | Attention based model compared with well-known architectures              | LC                 |
| Wang et al. (2018b)             | Minimizes the encoding differences of a CXR from multiple models         | LC                 |
| Wei et al. (2019)               | DenseNet used to predict COVID-19 severity as scored by radiologists      | LC                 |
| Wang et al. (2018b)             | Uses a ResNet-50 backed segmentation model to detect healthy, pneumonia, COVID-19 | LC                 |
| Niu et al. (2018)               | Uses multi instance learning for classification with localization          | LC                 |
| Yedla et al. (2018)             | Lung cancer and nodule prediction using ResNet                              | LC                 |
| Wang et al. (2019)              | GradCam based attention mining loss, compared with labels extracted from reports | LC                 |
| Nijhoffs et al. (2020)          | Trains Xception using 750k CXR, compares results with radiologist labels | LC                 |

continued on the next page
Table 2: Image-Level Prediction Studies (Section 4.1).

| Citation | Method | Other Tasks | Labels | Datasets |
|----------|--------|-------------|--------|----------|
| Hosch et al. (2020) | ResNet and VGG used to distinguish AP from PA images | LC | OR | PR,RP |
| Ritter et al. (2020) | DenseNet-121 trained on public data evaluated using CT-based labels | LC | PE,PM | C,PL |
| Ureta et al. (2020) | Evaluates the performance of various models trained on pediatric CXRs on adult CXRs | LC | PM | PP |
| Chen et al. (2020b) | Evaluates ensembling methods and visualization on pediatric CXRs | LC | PM,PT,Z | PR |
| Ye et al. (2020) | Compares a ResNet-152 against radiologists and shows the statistical significance | LC | PT | C |
| Crosby et al. (2020a) | Compares GradCAM with radiologist segmentations for evaluation of VGG-19 | LC | PT | C,PR |
| Crosby et al. (2020b) | Apical regions and patches from them extracted to detect pneumonia | LC | PT | PR |
| Tang et al. (2020) | Detection of abnormality, various networks compared with radiologist labeling | LC | T | C,OP,RP |
| performed in 2020a | Evaluates pre-training on ImageNet and CheXpert on various models/settings | LC | T | RP |
| Nakao et al. (2020) | Proposes a GAN-based model trained only with healthy images for anomaly detection | LC | T | RP |
| Pisa et al. (2019) | Proposes a new model for faster classification of TB | LC | TB | BL,MO,LS |
| Hwang et al. (2020) | Evaluates the use of a ResNet-based model on a large gold standard dataset | LC | TB | PR |
| Nataraj et al. (2020) | Evaluates multiple models for detection of feeding tube malpositioning | LC | TU | PR |
| Chakraborty et al. (2020) | Graph CNN solution with ensembling which models which diseases dependences | LC | X | X |
| Matsunoto et al. (2020) | Curates a dataset of heart failure cases and evaluates VGG-16 on it | LC | Z | C |
| Shi et al. (2021) | CNN for identifying the presence of subpleural free air from CXR | LC | Z | PR |
| Zhai et al. (2021) | Evaluates several models to predict hypertension and artery systemic pressure | LC | Z | PR |
| Zhang et al. (2021a) | Uses ResNet-18 to measure the Brachial Score, predicts Cystic Fibrosis based on it | LC | Z | PR |
| Zaidan et al. (2020) | Simulates CXRs from CT scans and predicts emphysema scores | LC | Z | PR |
| Campos et al. (2018) | Inception network to predict pulmonary to systemic flow ratio from pediatric CXR | LC | Z | PR |
| Tobin et al. (2018) | Predicts COVID-19 severity by comparing CXRs to previous ones | LC | T | C,OP,RP |
| Liao et al. (2020) | Addresses domain and label discrepancies in multi-dataset training | LC | T | RP |
| Lui et al. (2019) | Method to increase robustness of CNN classifiers to adversarial samples | LC | A,IG | PM | RP |
| Li and Zhao (2019) | Uses the features extracted from the training set to detect adversarial CXRs | LC | A | C |
| Akhtar et al. (2020) | Self-supervision and adversarial training improves on transfer learning | LC | A | C |
| Khatibi et al. (2021) | Claims 0.99 AUC for predicting TB, uses complex feature engineering and ensembling | LC | AA | C |
| Neuböcker et al. (2021) | ResNet model trained with frontal and lateral images to predict COPD with PFT results | LC | AA | PP |
| Zhang et al. (2020) | One-class identification of viral pneumonia cases compared with binary classification | LC | AA | C |
| Boscaglia et al. (2020) | A distributed learning method that overcomes problems of multi-institutional settings | LC | C | C |
| Bouteiller et al. (2020) | Geometric deep learning including metadata with graph structure. Application to CXR | LC | C | C |
| Negotio et al. (2020) | Proposes a new weighting scheme to improve abnormality classification | LC | C | C |
| Dey et al. (2019) | Proposes data augmentation on CheXNet v2 | LC | C | C |
| Irazzimbio et al. (2019) | Investigates effect of data augmentations on CheXNet v2 | LC | C | C |
| Miao et al. (2018) | Proposes a variational/generative architecture, demonstrates performance on CXRs | LC | C | C |
| Raperk et al. (2019) | Evaluates the performance of an ensemble against many radiologists | LC | C | C |
| Amoura et al. (2019) | Novel method for multi-label classification, application to CXR | LC | C | C |
| Pal et al. (2020) | Defines a few-shot learning method by extracting features from autoencoders | LC | C | C |
| Umkehrishan et al. (2020) | Mean teacher inspired a probabilistic graphical model with a novel loss | LC | C | C |
| Michael and Voon (2020) | Examines the effect of denoising on pathology classification using DenseNet-121 | LC | C | C |
| Wang et al. (2020) | Proposes integrating three attention mechanisms that work at different levels | LC | C | C |
| Patal et al. (2020a) | Step-wise trained CNN and saliency-based autoencoder for few shot learning | LC | C | C |
| Patal et al. (2020b) | Uses CT and CXR reports with CXR images during training to diagnose unseen diseases | LC | C | C |
| Bistoun et al. (2020) | Proposes a new model using CXR images during training to diagnose unseen diseases | LC | C | C |
| Li et al. (2020) | Lesion detection network used to improve image-level classification | LC | C | C |
| Ghosh et al. (2019) | Method to produce confidence measure alongside probability, uses DenseNet-121 | LC | C | C |
| Haglindh et al. (2020) | Uses self-supervised learning for pretraining, compares with ImageNet pretraining | LC | C | PT |
| Zhou et al. (2020) | Proposes a new model for COVID-19 | LC | C | PT |
| Zhou et al. (2020a) | Proposes a new model for COVID-19 | LC | C | PT |
| Zhou et al. (2019) | Compares several models for the detection of cardiomegaly | LC | C | C |
| Bogus et al. (2020) | Tests four off-the-shelf networks for prediction of cardiomegaly | LC | C | C |
| Bisesi et al. (2019) | Inception v3 trained to detect 4 abnormalities and compared with expert observers | LC | C | C |
| Canny et al. (2020) | googLeNet to classify normal and 5 abnormalities on a large proprietary dataset | LC | C | C |
| Bu et al. (2019a) | Compares the performance of deep learning with traditional feature extraction methods | LC | C | C |
| Bu et al. (2019b) | ImageNet pre-training and feature extraction methods for pathology detection | LC | C | C |
| Zhao et al. (2020) | An ensemble of DenseNet-121 models used for COVID-19 classification | LC | C | C |
| Hua et al. (2020) | Investigates the value of soft tissue CXR for training DenseNet-121 for COVID-19 | LC | C | C |
| Zhu et al. (2020) | Labels and predicts COVID-19 severity using CNN | LC | C | C,P |
| Fricks et al. (2020) | Uses a model trained on COVID-19 cases to evaluate the effect of an imaging parameter | LC | C | PR |
| Wu et al. (2020) | COVID-19 detection based on RT-PCR labels, evaluates an ensemble against radiologists | LC | C | PR |
| Zheng et al. (2021) | Ensemble of ResNet models for COVID-19 detection | LC | C | C |
| Zhang et al. (2021c) | Compares the performance of a DenseNet-121 ensemble to radiologists | LC | C | C |
| Zhang et al. (2021d) | Various models and use of semi-supervised labels for edema severity estimation | LC | C | C |
| Wang et al. (2020) | Age prediction on PA or AP images using DenseNet-121 | LC | C | C |
| Arafat et al. (2019a) | Gender prediction using features from deep-learning models in traditional classifiers | LC | C | C |
| Arafat et al. (2019b) | Age prediction on PA images using DenseNet-121 and ResNet-50 | LC | C | C |
| Li et al. (2020) | Combines the CXR with age/sex/smoking history to predict the lung cancer risk | LC | C | C |
| Thompson et al. (2020) | Densnet-121 pre-trained with public data used to identify 6 classes | LC | C | C |
| Kuo et al. (2020) | Evaluates deep learning on pictures of CXRs captured with mobile phones | LC | C | C |

continued on the next page
Table 2: Image-Level Prediction Studies (Section 4.1).

| Citation                  | Method                                      | Other Tasks | Labels | Datasets |
|---------------------------|---------------------------------------------|-------------|--------|----------|
| Cohen et al. (2020a)      | Investigates the domain and label shift across publicly available CXR datasets | MN | C.M.O.F.R.P.R.X. |          |
| Anand et al. (2020)       | Explores the use of the lateral view CXR for classification of 64 different labels | MN, P | P |          |
| Rajkomar et al. (2017)    | Classification of CXRs as Frontal or Lateral using GoogLeNet architecture | OR | PR |          |
| Crootsy et al. (2019)     | Assesses the effect of imprinted labels on AP/PA classification | OR | PR |          |
| Crootsy et al. (2018)     | Distinguishes the CXR orientation, bone CXRs and soft tissue CXRs from dual energy | OR, Z, PR |          |          |
| Betram et al. (2019)      | Compares PA and Lateral images for pathology detection with DenseNet | P | P |          |
| Chen et al. (2019)        | Introduces a loss term that uses the label hierarchy to improve model performance | PL | PL |          |
| Shah et al. (2020)        | Methods to mitigate imbalanced class sizes. Applied to CXR using ResNet-18 | PM | PP |          |
| Yue et al. (2020)         | Evaluation of MobileNet to detect pneumonia on pediatric CXRs | PM | PP |          |
| Elshennawy and Ibrahim (2020) | Evaluates various capsule network architectures for pneumonia detection | PM | PP |          |
| Mittal et al. (2020)      | Uses ResNet-50 to classify pediatric pneumonia | PM | PP |          |
| Deshpande et al. (2020)   | Compares traditional and generative data augmentation techniques on CXRs | PM | RP |          |
| Ganiesan et al. (2019)    | Addresses catastrophic forgetting, application to pneumothorax detection using VGG-13 | PT | C |          |
| Ravishankar et al. (2019) | Construction of large dataset, multiple architectures and hyperparameters optimized | PT | PR |          |
| Talyade et al. (2019)     | Model pre-trained with public data and fine-tuned for pneumothorax detection | PT | PR |          |
| Xu et al. (2019a)         | Distinguishes pneumonia from different foreign object types and detects using various architectures | Q | C |          |
| Moradi et al. (2019a)     | GoogleNet combined with rule-based approach to determine the image quality | Q | PR |          |
| Takaki et al. (2020)      | Detects abnormal CXRs using several models. Evaluates on independent private data | T | C |          |
| Lu et al. (2019b)         | Defines a model on top of features extracted from Inception-ResNet-v2 for triaging | T | C |          |
| Moradi et al. (2019b)     | Collects features from pretrained models and adds a CNN on top for triaging | T | C.M |          |
| Wong et al. (2020a)       | Studies the effect of various label noise levels on classification with DenseNet-121 | T | C.P.R.X |          |
| Zhang et al. (2020)       | Various models for detection of abnormal CXRs, effect of different train set sizes | T | PR |          |
| Dimou et al. (2019a)      | Defines 10 abnormalities to define a triaging model and uses CT based test labels | T | PR |          |
| Dyer et al. (2021)        | Ensemble of DenseNet and EfficientNet for identification of normal CXR | T | PR |          |
| Ogawa et al. (2019)       | Examines the use of data augmentation in small data setting | T | PR |          |
| Ellis et al. (2020)       | Evaluation of extra supervision in the form of localized region of interest | T | PR |          |
| Rajaraman et al. (2019a)  | Evaluates various models and ensembling methods for the triage task | T | RP |          |
| Rajaraman et al. (2019b)  | Evaluates deep learning approaches for tuberculosis detection | TB | BL, MO, P.R.S |          |
| Lakhdhri and Elnoufana (2017) | Evaluates the use of transfer learning for tuberculosis detection | TB | MO, P.R.S |          |
| Hwang et al. (2020a)      | Extracts features using off-the-shelf models and trains a model using those | TB | MO, P.R.S |          |
| Mittal et al. (2020a)     | Combines hand-crafted features and CNN for tuberculosis diagnosis | TB | MO, S |          |
| Ayas et al. (2021)        | Evaluates a Bayesian-based CNN for detection of TB | TB | MO, S |          |
| Friese and Ackermann (2020) | Evaluates assisting clinicians with an AI based system to improve diagnosis of TB | TB | PR |          |
| Rajparkar et al. (2020)   | Various architectures, inclusion of patient demographics in model considered | TB | PR |          |
| Haroon et al. (2021)      | Addresses preservation of learned data, application to TB detection using ResNet-21 | TB | PR |          |
| Kim et al. (2018)         | Pre-training using CXR pathology and metadata labels, application to TB detection | TB | S |          |
| Gooes and Greenspan (2019) | Pre-training using CXR pathology and metadata labels, application to TB detection | TB | S |          |
| Rajaraman and Anjali (2020) | Compares various models using various pretraining and ensembling strategies | TB | S |          |
| Lakhdhri and Elnoufana (2020) | Evaluates models on detecting the position of feeding tube in abdominal and CXRs | TB | S |          |
| Arafa et al. (2020)       | Comparison of seven architectures and ensembling for detection of nine pathologies | TU | PR |          |
| Pham et al. (2020)        | A method to incorporate label dependencies and uncertainty data during classification | X | X |          |
| Rajan et al. (2020)       | Proposes self-training and student-teacher model for sample efficiency | X | X |          |
| Cai et al. (2019)         | Analyses the effect of label noise in training and test datasets | Z | C |          |
| Deshpande et al. (2020)   | Labels 6 different foreign object types and detects using various architectures | Z | M |          |
| Liu et al. (2019)         | Evaluates the use of CXRs to predict long term mortality using Inception-v4 | Z | PL |          |
| Zhang et al. (2019a)      | Low-res segmentation is used to crop high-res lung areas and predict pneumocissusis | Z | PR |          |
| Devine et al. (2020)      | Pneumocissusis prediction with DenseNet-121 and SVMs applied to extracted features | Z | PR |          |
| Liu et al. (2017)         | Detection of coronary artery calcification using various CNN architectures | Z | PR |          |
| Hirata et al. (2021)      | ResNet-50 for detection of the presence of elevated pulmonary arterial wedge pressure | Z | PR |          |
| Kosanose et al. (2020)    | A network is designed to identify subjects with elevated pulmonary artery pressure | Z | PR, RP |          |

pert observers in most of the 14 labels provided by ChestX-ray 14. 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 (Rajaraman et al. [2018a], Yue et al. [2020], Li et al. [2020b], Behzadi-khormouji et al. [2020], Elshennawy and Ibrahim [2020], Ureta et al. [2020], Mittal et al. [2020], Shah et al. [2020], Qu et al. [2020], Ferreira et al. [2020], Anand et al. [2020]). 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, Annarumma et al. (2019) develops a triaging pipeline based on the urgency of exams. Similarly, Tang et al. (2020) 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, Taylor et al. (2018) 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). Tasks that need to be done for chest X-ray datasets include: Image Level Prediction, Adversarial Attack, Domain Adaptation, Interval Change, Image Generation, Image Retrieval, Localization, Other, Preprocessing, Report Parsing, Segmentation, Weak Supervision.
of attention mining models are [Cai et al., 2018; Saednia et al., 2020]. Multiple-instance learning (multi-instance learning or MIL) [Xu et al., 2014] 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 [Crosby et al., 2020c; Schwab et al., 2020]. Other topics within the literature include model uncertainty [Ul Abideen et al., 2020; Ghesu et al., 2019], quality of the CXR [Moradi et al., 2020; McManagle et al., 2020; Moradi et al., 2019a; Takaki et al., 2020; McManagle et al., 2020] and defence against adversarial attack [Li and Zhu, 2020; Anand et al., 2020; Xue et al., 2019].

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 [Pham et al., 2020; Chen et al., 2020b; Chakravarty et al., 2020]. 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 [Luo et al., 2020; Cohen et al., 2020b]. The effect of dataset sizes is evaluated by Dunnmon et al. (2019). Semi-supervised learning methods combine a small set of labeled and a large set of unlabeled data to train a model [Gyawali et al., 2019; Wang et al., 2019; Unnikrishnan et al., 2020].

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 [Bertrand et al., 2019; Hashir et al., 2020; Bayat et al., 2020].

4.2. 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 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 [Baltruschat et al., 2019b; Wang et al., 2020c; Rajaraman et al., 2019b; Heo et al., 2019; Liu et al., 2019; Mansoor et al., 2016]. 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 [Sogancioglu et al., 2020; Li et al., 2019]. Organ segmentation has, for these reasons, become one of the most commonly studied subjects among CXR segmentation tasks as seen in Figure 5.

Another application found in the CXR literature is foreign object segmentation, i.e. catheter, tubes, lines, for which high
Table 3: Segmentation Studies (Section 1.2).

| Tasks                          | Method                                                                 | Other Tasks | Labels | Datasets  |
|-------------------------------|------------------------------------------------------------------------|-------------|--------|-----------|
| DA                           | A model based on U-Net and Faster R-CNN to detect PICC catheter and its tip | LC          | PR     | PR        |
| IL                           | Tailored Mask R-CNN for simultaneous detection and segmentation         | LC          | L      | PR        |
| IL                           | Uses Mask R-CNN iteratively to segment and detect ribs.                | LC          | R      | PR        |
| PE                           | Combines lung cropped CXR model and a CXR model to improve model performance | IL, LC, PR  | C      | C         |
| PE                           | Comparison of image-level precision and prediction models for cardiomegaly | IL          | CM     | C         |
| Other                        | A network with DenseNet and U-Net for classification of cardiomegaly    | IL          | CM     | C         |
| PE                           | U-Net based model for heart and lung segmentation for cardiothoracic ratio | IL          | CM     | PR        |
| Other                        | Combines lung cropped CXR model and a CXR model using the segmentation quality | IL          | E, L, O, P, E, T, Z | M         |
| PE                           | Pneumonia detection is improved by use of lung segmentation            | IL          | PM     | J, MO, PPCPR |
| Other                        | U-Net based model to segment pneumonia                                 | IL          | PM     | RP        |
| Other                        | Multi-scale DenseNet based model for pneumothorax segmentation          | IL          | PT     | PR        |
| PE                           | DenseNet based U-Net for segmentation of the left and right humerus of the infant | IL          | Z      | J, PR     |
| Other                        | Attention-based network and CXR synthesis process for data augmentation | IG, JG     | L      | J, MO, PR |
| Other                        | Conditional GANs for multi-class segmentation of heart,clavicles and lungs | IG          | CL, HL | J         |
| Other                        | Processing method to produce scatter-corrected CXRs and segments masses with U-Net | IG          | LO     | SM        |
| BL                           | MUNIT based DA model for lung segmentation                              | DA          | CL, HL | J         |
| BL                           | Adversarial training of lung and heart segmentation for DA              | DA          | CM     | J, PR     |
| Other                        | CycleGAN guided by a segmentation module to convert CXR to CT projection images | DA          | H, L, Z | PR        |
| Other                        | CycleGAN based DA model with semantic aware loss for lung segmentation  | DA          | L      | MO        |
| Other                        | Conditional GANs based DA for bone segmentation                         | DA          | R      | SM        |
| Other                        | FCN based novel model incorporating weak landmarks and bounding boxes annotations | WS          | CL, HL | J         |
| Other                        | U-Net segmentation model integrating unlabeled data through consistency loss | WS          | CL, HL | J         |
| Other                        | Attention masks derived from classification model to guide the segmentation model | IL          | PT     | PR        |
| Other                        | U-Net based network for classification and segmentation with simulated data | IL          | Z      | C, J      |
| Other                        | U-Net based model for segmentation and a classification for existence of lines | IL          | Z      | C, J      |
| Other                        | U-Net for bone suppression given lung-segmented CXR image with patches | PR          |        |           |
| Other                        | Proposes teacher-student based learning with noisy segmentations        | CL, HL     | J       |           |
| Other                        | Various FCN based models explored for simultaneous pixel and contour segmentation | CL, HL     | J       |           |
| Other                        | Investigates various FCN type architecture including U-Net for organ segmentation | CL, HL     | J       |           |
| Other                        | Capsule networks adapted for multi-class organ segmentation             | CL, HL     | J       |           |
| Other                        | U-Net based architecture with residual connections for organ segmentation | CL, HL     | J       | MO, S     |
| Other                        | U-Net based architecture based on dense connections                     | CL, R      | PR      |           |
| Other                        | CNN trained with CT projection images for quantification of airspace disease | CV          | PR      |           |
| Other                        | Denoising autoencoder as post-processing to improve segmentations        | H, L       | J       |           |
| Other                        | Evaluates U-Net performance with various loss functions, and data augmentation | H, L       | PR      |           |
| Other                        | Stacked denoising autoencoder model for space and shape parameter estimation | L          | BL, J, PR|           |
| Other                        | Investigates the effect of fine-tuning different layers for U-Net based model | L          | J       |           |
| Other                        | Proposes a human-in-the-loop one shot anatomy segmentor                | L          | J       |           |
| Other                        | U-Net with conditional random field post processing for lung segmentation | L          | J       |           |
| Other                        | Investigates U-Net with different optimizer and dropout                 | L          | J       |           |
| Other                        | U-Net with dense connections for reducing network parameters for lung segmentation | L          | J, L, MO |           |
| Other                        | Yahilatyabe et al. (2020)                                              | L          | J, L, MO|           |
| Other                        | Yu and Lee (2021)                                                      | L          | J, L, MO|           |
| Other                        | Arbitbashari et al. (2017)                                             | L          | J, PR   |           |
| Other                        | Multi-scale and patch-based CNN to segment lungs                        | L          | J, PR   |           |
| Other                        | Investigates 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 (Arbabshirani et al., 2017) Souza performance levels have been reported using deep learning (Lee et al., 2018; Frid-Adar et al., 2019; Sullivan et al., 2020). Interestingly, only a small number of works addressed segmentation of abnormalities. Hurt et al. (2020) focused on segmentation of pneumonia, and Toktachev et al. (2020) 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 (Arbabshirani et al., 2017) Souza.
However, this approach is generally considered inefficient for segmentation and most works use fully convolutional networks (FCN) (Shelhamer et al., 2017), 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 (Ronneberger et al., 2015), 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 (Novikov et al., 2018; Furutani et al., 2019; Kitahara et al., 2019), pneumonia (Hurt et al., 2020) and foreign objects (Sullivan et al., 2020; Frid-Adar et al., 2019). For example, Novikov et al. (2018) compared three U-Net variants 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 overall predictions with an optimized binarization threshold and a dilution technique for pneumothorax segmentation. Some authors proposed to learn post-processing by training an independent network, inputting segmentation predictions for refinement, rather than using conventional methods. For example, Larrazabal et al. (2020) used denoising autoencoders, trained to produce anatomically plausible segmentations from the initial predictions. Similarly, Souza et al. (2019) 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 (Wessel et al., 2019; Souza et al., 2019; Xue et al., 2018d; 2020). For example, Xue et al. (2018d) 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, Souza et al. (2019) employed two networks, where the second network received the predictions of the first to refine the segmentation results. Wessel et al. (2019) 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 (Yi et al., 2019a; Milletari et al., 2018; Mathai et al., 2019) and report good performance. Milletari et al. (2018) 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, Yi et al. (2019b) 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 systems which incorporate weak-labels or simulated datasets with the aim of reducing annotation costs (Frid-Adar et al., 2019; Ouyang et al., 2019; Lu et al., 2020b; Yi et al., 2019a). Several works addressed this using weakly supervised learning approaches (Lu et al., 2020b; Ouyang et al., 2019; Lu et al., 2020b) 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. Ouyang et al. (2019) 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.

### 4.3. 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.

| Lesion or Opacity | Catheter or Tube | Pneumothorax | Cardiomegaly |
|-------------------|------------------|--------------|--------------|
| Lung              | Heart            | Clavicle     | Other        |
| Pneumonia         | Edema            | COVID-19     | Rib          |
| Lesion or Opacity | Edema            | Effusion     | Other        |

Figure 5: Number of studies for the Segmentation labels. 58 papers are included, each may study more than one label.
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 quantitatively evaluated and such works are not included here. Table 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 6.

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 (Zhao et al., 2019). Such state of the art architectures have been rapidly adapted for CXR analysis and shown to achieve high-level performance. For example, Park et al. (2019) demonstrated that the (original) YOLO architecture was successful at identifying the location of pleural effusion 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 (Kim et al., 2019a; Li et al., 2020c) 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.

Figure 6: Number of studies for the Localization labels. 30 papers are included, each may study more than one label.

4.4. Image Generation

There are 35 studies identified in this work whose main focus is Image Generation, as detailed in Table 5. Image generation techniques have been harnessed for a wide variety of purposes including data augmentation (Salehinejad et al., 2019), visualization (Bigolin Lanfredi et al., 2019; Seah et al., 2019), abnormality detection through reconstruction (Tang et al., 2019c; Wolleb et al., 2020), domain adaptation (Zhang et al., 2018) or image enhancement techniques (Lee et al., 2019).

The generative adversarial network (GAN) (Goodfellow et al., 2014; Yi et al., 2019b) has become 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 (Moradi et al., 2018b; Zhang et al., 2019a; Salehinejad et al., 2019) by using unconditional GANs which synthesize images from random noise. For example, Salehinejad et al. (2019) trained a DCGAN model, similar to Moradi et al. (2018b), 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.
Table 4: Localization Studies (Section 4.3).

| Citation          | Method                                                                 |
|-------------------|------------------------------------------------------------------------|
| Yu et al. (2020)   | A model based on U-Net and Faster R-CNN to detect PCC cathe-er and its tip |
| Zhang et al. (2019b) | Tailored Mask R-CNN for simultaneous detection and segmentation        |
| Wessely et al. (2020) | Uses Mask R-CNN iteratively to segment and detect ribs.               |
| Ouyang et al. (2020) | Uses activation and gradient based attention for localization and classification |
| Rajpurkar et al. (2019) | Detects and localizes COVID-19 using various networks and ensembling   |
| Simaitis et al. (2020) | GoogleNet trained with CXR patches, correlates with COVID-19 severity score |
| Park et al. (2020)   | Proposes a segmentation and classification model compares with radiologist cohort |
| Nam et al. (2019)    | Trains a semisupervised network on a large CXR dataset with CT-confirmed nodule cases |
| Ponce et al. (2019)  | Defines a loss that minimizes the saliency map errors to improve model performance |
| Huang and Kim (2020) | A weakly supervised localization with variational model, leverages attention maps |
| Li et al. (2019)     | Attention guided CNN for pneumonia detection with bounding boxes       |
| Huang et al. (2019)  | A CNN for identification of abnormal CXRs and localization of abnormalities |
| Tang et al. (2019c)  | Introduces a visualization method to identify regions of interest from classification |
| Hwang and Kim (2016) | Weakly supervised framework jointly trained with localization and classification |
| Chen et al. (2019a)  | Extracts nucleate candidates using traditional methods and trains GoogleNet |
| Schirmer et al. (2020) | RetinaNet for detecting nodules incorporating lung segmentation         |
| Huang et al. (2019c) | Combines reports and CXRs for weakly supervised localization and classification |
| Moraitis et al. (2018a) | Proposes a model using LSTM and CNN, combining reports and images as inputs |
| Khakzar et al. (2019) | Adversarially trained weakly supervised localization framework for interpretability |
| Kim et al. (2019)    | Evaluates the effect of image size for nodule detection with Mask R-CNN and RetinaNet |
| Cheh et al. (2019b)  | Evaluates the reproducibility of various detection architectures in follow up exams |
| Kim et al. (2019)    | Evaluates the reproducibility of various detection architectures in follow up exams |
| Cha et al. (2019a)   | ResNet model using CT and surgery based annotations for lung cancer prediction |
| Takemura et al. (2019) | R-CNN for localization of lung nodules                                 |
| Wang et al. (2019)   | Fuses AlexNet and hand-crafted features to improve random forest performance |
| Li et al. (2019)     | Patch-based nodule detect, combines features from different resolutions |
| Park et al. (2019)   | Evaluates the detection of pneumothorax before, 3h and 1d after biopsy |
| Souda et al. (2019)  | Proposes a U-Net based model for localizing and labeling individual ribs |
| Xie et al. (2019)    | AlexNet for localizing tuberculosis with patch-based approach           |
| von Berg et al. (2020) | Localizes anatomical features for image quality check                  |

Another work (Zhang et al. 2019a) proposed a novel GAN architecture to improve the quality of generated CXRs 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 (Tang et al. 2019c; Mao et al. 2020). 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. Tang et al. (2019c) 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, Mao et al. (2020) 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.

The most widely studied subject in the image generation literature is image enhancement. Several researchers investigated bone suppression (Liu et al. 2020a; Matsubara et al. 2020; Zarshenas et al. 2019; Gozes and Greenspan 2020; Lin et al. 2020; Zhou et al. 2020a) and lung enhancement (Li et al. 2019; Gozes and Greenspan 2020) techniques to improve image interpretability. A number of works (Liu et al. 2020a; Zhou et al. 2020b) employed GANs to generate bone-suppressed images. For example, Liu et al. (2020a) 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 (Li et al. 2019; Gozes and Greenspan 2020) addressed this by leveraging digitally reconstructed radiographs for enhancing the lungs and bones in CXR. For instance, Li et al. (2019) trained an autoencoder for generating CXR with bone suppression and lung enhancement, and the knowledge obtained from DRR images were integrated through the encoder.

4.5. 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 (Kim et al. 2019).
Table 6: Image Generation Studies (Section 4.4).

| Citation                  | Method                                                                 | Other Tasks | Labels |
|---------------------------|----------------------------------------------------------------------|-------------|--------|
| Tang et al. (2019)        | Attention-based network and CXR synthesis process for data augmentation | SE,IG       | L, JMO, PR |
| Yamasaki et al. (2020)    | Conditional GANs for multi-class segmentation of heart, clavicles and lungs | SE          | CL, IL, J |
| Ohodera et al. (2020)     | Processing method to produce scatter-corrected CXRs and segments masses with U-Net | SE          | LO, SM |
| Wang et al. (2018)        | Combines classification loss and autoencoder reconstruction loss       | IL, SE      | T, JMO, O.S |
| Nazer et al. (2021)       | Wasserstein GAN to permute diseased radiographs to appear healthy    | IL, LC, Z   | PR |
| Wolkfe (2020)             | Novel GAN model trained with healthy and abnormal CXR to predict difference map | IL          | PE, SM, X |
| Tang et al. (2019b)       | GANs with U-Net autoencoder and CNN discriminator and encoder for one-class learning | IL          | T, C |
| Man et al. (2020)         | Autoencoder uses uncertainty for reconstruction error in one-class learning setting | IL          | T, FPR |
| Taghanaki et al. (2019a)  | Adversarial based method adapting new domains for abnormality classification | DA, IL      | CM, PL |
| Lin et al. (2020)         | Proposes a patch-based CNN super resolution method                    | SR, J       | Z, J |
| Zhang et al. (2019a)      | Generates high resolution CXRs using multi-scale, patch based GANs    | SR          | Z, O |
| Uzunova et al. (2019a)    | Novel GAN model with sketch guidance module for high resolution CXR generation | SR          | Z, PP |
| Prevedello et al. (2019)  | AutoEncoder for bone suppression and segmentation with statistical similarity losses | SE, PR      | BS, J |
| Dong et al. (2019)        | Uses neural architecture search to find a discriminator network for GANs | SE          | H.L, JPR |
| Prevedello et al. (2019a) | Proposes an iterative gradient based input preprocessing for improved performance | SE          | S |
| Tang et al. (2019c)       | Learns transformations to register two CXRs, uses the difference for interval change | REL, IC     | Z, PR |
| Yang et al. (2019b)       | Generates bone and soft tissue (dual energy) images from CXRs         | PR          | BS, PR |
| Zarneshes et al. (2019)   | Proposes an CNN with multi-resolution decomposition for bone suppression images | PR          | BS, PR |
| Ameen et al. (2020)       | U-Net for bone generation with CT projection images, used for CXR enhancement | PR          | BS, SM |
| Lee et al. (2019)         | U-Net based network to generate dual energy CXR                        | PR          | Z, PR |
| Liu et al. (2020a)        | GAN integrates edges of ribs and clavicles to guide DES-like images generation | PR          | Z, PR |
| Xing et al. (2019)        | Generates diseased CXRs, evaluates their realism with radiologists and trains models | LC          | C, C |
| Li et al. (2019)          | Novel CycleGAN model to decompose CXR images incorporating CT projection images | IL           | C, C, CPR, SM |
| Salehinejad et al. (2019) | Uses DCGAN model to generate CXR with abnormalities for data augmentation | IL          | CME, PE, PT, PR |
| Albagouni et al. (2017)   | U-Net based architecture to decompose CXR structures, application to TB detection | IL          | TB, PR |
| Yi et al. (2018)          | Two DCGAN trained with normal and abnormal images for data augmentation | IL          | Z, PL |
| Hossaini et al. (2019)    | Novel conditional GAN using lung function test results to visualize COPD progression | IL          | Z, PR |
| Zarei et al. (2021)       | Conditional GAN and two variational autoencoders designed for CXR generation | IL          | Z, PR |
| Omran et al. (2020)       | Novel reconstruction algorithm for CXR enhancement                     | PR          | |
| Zhou et al. (2020)        | Bone shadow suppression using conditional GANs with diluted U-Net variant | BS          | J |
| Matynin et al. (2020)     | Generates CXRs from CT to train CNN for bone suppression               | BS          | PR |
| Zoumaar and Hamza (2021)  | Generates COVID-19 CXR images to improve network training and performance | CV          | CC, PR |
| Bilyat et al. (2020)      | 2D-to-3D encoder-decoder network for generating 3D spine models from CXR studies | Z           | PR |
| Bigolin Lanfredi et al. (2020) | Generates normal from abnormal CXRs, uses the deformations as disease evidence | Z           | PR |

Prevedello et al. (2019) and highlights an important concern: most of the performance levels reported in the literature might not generalize well to data from other domains (Zech et al. 2018). Several studies (Yao et al., 2019; Zech et al., 2018; Cohen et al., 2020b) 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, Yao et al. (2019) 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. Zech et al. (2018) 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 investigated 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 6.

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, Dong et al. (2018) 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 (Chen et al., 2018a; Zhang et al., 2018) Oliveira and dos Santos (2018) addressed unsupervised DA using CycleGAN-based models to transform source images to resemble those from the target domain. For example, Zhang et al. (2018) 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 simultaneous-
Table 6: Domain Adaptation Studies (Section 4.5).

| Citation                       | Method                                                                 | Other Tasks | Labels | Datasets |
|-------------------------------|------------------------------------------------------------------------|-------------|--------|----------|
| Oliveira et al. (2020b)       | MUNIT based DA model for lung segmentation                              | SE          | CL,HL  | J        |
| Abou et al. (2019)            | Adversarial training of lung and heart segmentation for DA             | SE          | CM     | JPR      |
| Zhang et al. (2020)           | CycleGAN guided by a segmentation module to convert CXR to CT projection images | SE          | H,L,Z  | PR       |
| Chen et al. (2018a)           | CycleGAN based DA model with semantic aware loss for lung segmentation | SE          | L      | MO       |
| Oliveira et al. (2020a)       | Conditional GANs based DA for bone segmentation                        | SE          | R      | SM       |
| Lenga et al. (2020)           | Continual learning methods to classify data from new domains           | IL          | C,M    | CM       |
| Tang et al. (2019)            | CycleGAN model to adapt adult to pediatric CXR for pneumonia classification | IL          | PM     | PPRP     |
| Mahapatra and Ghosh (2019)    | Conditional GAN based DA for image registration using segmentation guidance | TG,RE,SE   | L      | C        |
| Madani et al. (2018)          | Adversarial based model adapting new domains for abnormality classification | IG,II      | CM     | PL       |
| Zec et al. (2019)             | Assessment of generalization to data from different institutes         | SE          | PM     | C-O      |
| Sathrittanaichaweewin et al.  (2020) | Demonstrates the effect of training and test on data from different domains | IL          | TB     | S        |

Another task that did not belong to previously defined categories is out-of-distribution detection. Studies working on this (Marquez-Neila and Sznitman, 2019; Calì et al., 2019; Bozorgtabar et al., 2020) 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. Calì et al. (2019) 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. Bozorgtabar et al. (2020) 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 studies (Li et al. 2020b; Yuan et al. 2019; Syeda-Mahmood et al. 2020; Xue et al. 2018a). These studies aim to partially automate the radiology workflow by evaluating the chest X-ray and producing a text radiology report. For example, Syeda-Mahmood et al. (2020) 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 (Mansilla et al., 2020). 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. Detecting new findings, tracking the course of a disease, or evaluating the efficacy of a treatment are among the many uses of image registration (Viergever et al., 2016). To that end, Mansilla et al. (2020) aims to create an anatomically plausible registration by using the heart and lung segmentations to guide the registration process.

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 analysis 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 (Recht et al., 2020; Strohm et al., 2020), however there is growing acceptance that
In this work we have detailed datasets, literature and commercial products relevant to deep learning in CXR analysis. It is clear that this area of research has thrived on the release of multiple large, public, labeled datasets in recent years, with 209 of 295 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 3. However, although these datasets are extremely valuable, there are multiple caveats to be considered in relation to their use, as described in Section 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 papers include offer limited novel contributions although 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 content.

### 6. Discussion

The commercial products include applications for a wide range of abnormalities, with 6 of them reporting results for more than 209 of 295 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 3. However, although these datasets are extremely valuable, there are multiple caveats to be considered in relation to their use, as described in Section 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.
sons 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 presentation of results optimized on a validation set (without a validation set), whereby an algorithm can be trained using data from multiple sources 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 the use of AI in medical decision making (Recht et al., 2020; Strohm et al., 2020), 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 (Sheller et al., 2019). Further research is required to determine how this system will work in clinical practice.

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 or data-selection influence are rarely considered. Many such architectures 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 or data-selection influence are rarely considered.

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 (Dietterich, 2000). At the time of writing, most of the top-10 submissions from the public challenges (CheXpert (Irvin et al., 2019), SIIM-ACR (ACR, 2019), and RSNA-Pneumonia (RSNA, 2018)) consist of network ensembles. There is also promise in the development of self-adapting frameworks such as the nnU-Net (Isensee et al., 2021) 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 (Recht et al., 2020; Strohm et al., 2020), 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 (Sheller et al., 2019). Further research is required to determine how this type of system will work in clinical practice.

### Table 8: Commercial Products for CXR analysis. (Section 5)

| Company                  | Product                              | Literature (4 most recent)                                                                 | Labels (Total number) | Output |
|--------------------------|--------------------------------------|------------------------------------------------------------------------------------------|------------------------|--------|
| Siemens Healthineers     | All-Rad Companion Chest X-Ray        | [Fischer et al., 2020]                                                                  | LO PE PT Z (5)         | LOC, SCOR, REP |
| Samsung Healthcare       | Auto Lung Nodule Detection           | [Sim et al., 2019]                                                                     | ND (1)                 | LOC    |
| Thirona                  | CADCOVID-XRay                        | [Murphy et al., 2020]                                                                  | CV (1)                 | LOC, SCOR |
|                         | CAD4TB                               | [Habib et al., 2020]                                                                  | TB (1)                 | LOC, SCOR |
| Oxitut                   | ChestEye CAD                         | [Lang et al., 2020]                                                                   | T (1)                  | REP(healthy) |
| Arttry                   | Chest                                | [Sim et al., 2019]                                                                    | LO, ND, PE, PT (4)     | LOC, SCOR, PRI |
| Quibim                   | Chest X-Ray Classifier               | [Jiang et al., 2020]                                                                  | 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-DXR                              |                                                                                         | LC Z (16)              | LOC, SCOR |
| Limit                    | Lunit INSIGHT CXR                    | [Hwang et al., 2019; Qin et al., 2019]                                                | TB CM ND PE PT Z (11) | LOC, SCOR, PRI, REP |
| qure.ai                  | qXR                                  | [Singh et al., 2018; Nash et al., 2020; Engle et al., 2020]                            | T CV TB Z (30)         | LOC, SCOR, PRI, REP |
| Digitec                  | TIRESYA                              |                                                                                         | BS (1)                 | Bone Suppressed Image |
| VUNO                     | VUNO Med-CHEST X-Ray                 |                                                                                         | LO ND PE PT Z (5)      | LOC, SCOR |
| Riverain Technologies    | ClearRead Xray - Bone Suppress       | [Homayouni et al., 2020; Deliols et al., 2017]                                        | BS(1)                  | Bone Suppressed Image |
| Riverain Technologies    | ClearRead Xray - Compare             | [Schalekamp et al., 2018; 2018]                                                       | LC(1)                  | Subtraction Image |
| Riverain Technologies    | ClearRead Xray - Confirm             |                                                                                         | TU(1)                  | LOC    |
| Riverain Technologies    | ClearRead Xray - Detect              |                                                                                         | 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 |
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.

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 (Dyer et al., 2021; Dunnmon et al., 2019; Baltruschat et al., 2020).

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 bedside/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 (Hashir et al., 2020). 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.

Machine learning will almost certainly improve sensitivity and specificity of data from multiple sources along with the CXR image 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.

7. 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’).

References

ACR, 2019. SIIM-ACR Pneumothorax Segmentation. URL: https://kaggle.com/c/siim-acr-pneumothorax-segmentation
Albarqouni, S., Fotouhi, J., Navab, N., 2017. X-Ray In-Depth Decomposition: Revealing the Latent Structures, in: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. Springer. volume 10435, pp. 444–452, doi:10.1007/978-3-319-66179-7_51
Amiri, M., Brooks, R., Rivaz, H. 2020. Fine-Tuning U-Net for Ultrasound Image Segmentation: Different Layers, Different Outcomes. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 67, 2510–2518, doi:10.1109/TUFFC.2020.3015081
sisted Intervention – MICCAI 2018. Springer, volume 11071, pp. 589–598. doi:10.1007/978-3-030-00934-2_66
Çalli, E., Murphy, K., Sogancioglu, E., van Ginneken, B., 2019. (FRODO): Free rejection of out-of-distribution samples: application to chest x-ray analysis, in: International Conference on Medical Imaging with Deep Learning – Extended Abstract Track, pp. 1–4.
Çalli, E., Scholten, E.T., Murphy, K., van Ginneken, B., Sogancioglu, E., 2019. Predicting label noise through model confidence and uncertainty: application to chest radiograph classification, in: Medical Imaging 2019: Computer-Aided Diagnosis, SPIE. p. 41. doi:10.1117/12.2542290
Campo, M.I., Pascau, J., Estep, R.S.J., 2018. Emphysema quantification on simulated x-rays through deep learning techniques, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, pp. 273–276. doi:10.1109/ISBI.2018.8365752
Cardenas, D.A.C., Jr, J.R.F., Moreno, R.A., Rebelo, M.d.F.d.S., Krieger, J.E., Youssef, M., 2021. Computer-aided detection selects shortcuts over signal. medRxiv doi:10.1101/2021.01.13.21251775
Cohen, J.P., Hashir, M., Brooks, R., Bertrand, H., 2020b. On the limits of cross-domain generalization in automated x-ray prediction. Proceedings of the Third Conference on Medical Imaging with Deep Learning, PMLR, 121:136–155. doi:10.1186/s41747-020-00203-z
Cohen, J.P., Morrison, P., Dao, L., 2020c. Covid-19 image data collection: Prospective predictions are the future. arXiv preprint arXiv:2006.11988
Conjeti, S., Roy, A.G., Katouzian, A., Navab, N., 2017. Hashing with Residual Networks for Image Retrieval, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Springer. volume 10435, pp. 541–549. doi:10.1007/978-3-319-66179-7_62
Crosby, J., Chen, S., Li, F., MacMahon, H., Giger, M., 2020b. Network output visualization to uncover limitations of deep learning detection of pneumothorax, in: Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, SPIE. p. 22. doi:10.1117/12.2550066
Crosby, J., Rhines, T., Duan, C., Li, F., MacMahon, H., Giger, M., 2019. Impact of imprinted labels on deep learning classification of PA and AP thoracic radiographs, in: Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications, SPIE. p. 13. doi:10.1117/12.2513026
Crosby, J., Rhines, T., Li, F., MacMahon, H., Giger, M., 2020a. 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, 1. doi:10.1117/1.JMI.7.1.014001
Crosby, J., Rhines, T., Li, F., MacMahon, H., Giger, M., 2020c. Deep learning for pneumothorax detection and localization using networks fine-tuned with multiple institutional datasets, in: Medical Imaging 2020: Computer-Aided Diagnosis, SPIE. p. 11. doi:10.1117/12.2549709
Cruz, B.G.S., Bossa, M.N., Söltzer, J., Husch, A.D., 2021. Public covid-19 x-ray datasets and their impact on model bias - a systematic review of a significant problem. medRxiv doi:10.1101/2021.02.15.21251775
Daniels, Z.A., Metaxas, D.N., 2019. Exploiting Visual and Report-Based Information for Chest X-RAY Analysis by Jointly Learning Visual Classifiers and Topic Models, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, pp. 1270–1274. doi:10.1109/ISBI.2019.8759548
DeGrave, A.J., Janizek, J.D., Lee, S.I., 2020. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv doi:10.1101/2020.09.13.20193565
Dellino, N., Teichgraber, U., Chelaru, R., Malich, A., Papageorgiou, I.E., 2017. Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiographs. Journal of Clinical Imaging Science 7. doi:10.4103/jcis.JCIS_75_16
Denner-Fushman, D., Antani, S., Simpson, M., Thoma, G.R., 2012. Design and Development of a Multimodal Biomedical Information Retrieval System. Journal of Computing Science and Engineering 6, 168–177. doi:10.4103/jcis.2012.6.2.168
DeKosky, S.T., Webster, J., 2015. Beyond the amyloid hypothesis: Alzheimer’s disease as a disorder of neurovascular unit dysfunction. Neurology 85, 1660–1668. doi:10.1212/WNL.0000000000002494
Deng, J., Dong, W., Socher, R., Li, L., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. doi:10.1109/CVPR.2009.5206884
Deshpande, H., Harder, T., Saalbach, A., Sawarkar, A., Buelow, T., 2020. Detection Of Foreign Objects In Chest Radiographs Using Deep Learning, in: 2020 IEEE 17th International Symposium on Biomedical Imaging this article, all use of the data in this article is subject to the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).
cardiothoracic ratio calculation with deep learning. IEEE Access 7, 37749–37756. doi: 10.1109/ACCESS.2019.2900083

Li, Z., Li, H., Han, H., Shi, G., Wang, J., Zhou, S.K., 2019. Encoding CT Anatomy Knowledge for Unpaired Chest X-Ray Image Decomposition, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer, volume 11769, pp. 275–283. doi: 10.1007/978-3-030-32266-7_31

Liang, C.H., Liu, Y.C., Wang, M.T., Garcia-Castro, F., Alberich-Bayarri, A., Wu, F.Z., 2020. Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. Clinical Radiology 75, 38–45. doi: 10.1016/j.crad.2019.08.005

Liang, G., Zheng, L., 2020. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Computer Methods and Programs in Biomedicine 187, 104964. doi: 10.1016/j.cmpb.2019.06.023

Lin, C., Tai, R., Lin, D.D., Liu, L., Lu, F., Chen, Y., Gao, D., Zhou, J., 2020. Deep Feature Disentanglement Learning for Bone Suppression in Chest Radiographs, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 795–798. doi: 10.1109/ISBI45749.2020.9098389

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 2980–2988. doi: 10.1109/iccv.2017.324

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., et al., 2018. A Survey on Deep Learning in Medical Image Analysis, Medical Image Analysis 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., Pu, J., 2019. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest x-rays. Computerized Medical Imaging and Graphics 75, 66–73. doi: 10.1016/j.compmedimag.2018.05.005

Liu, X., Wang, X., Deng, Y., Chen, K., 2017. Coronary artery calcification (CAC) classification with deep convolutional neural networks, in: Medical Imaging: Computer-Aided Diagnosis, SPIE. p. 101340M. doi: 10.1117/12.2283974

Liu, Y., Liu, M., Yi, Y., Qin, G., Shen, D., Yang, W., 2020a. Generating Dual-Energy Subtraction Soft-Tissue Images from Chest Radiographs via Bone Edge-Guided GAN, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer. volume 12266, pp. 709–713. doi: 10.1007/978-3-030-59713-9_85

Liu, Y.C., Lin, Y.C., Tsai, P.Y., Iwata, O., Chuang, C.C., Huang, Y.H., Tsai, Y.S., Sun, Y.N., 2020b. Convolutional Neural Network-Based Humerus Segmentation and Application to Bone Mineral Density Estimation from Chest X-ray Images of Critical Infants. Diagnostics 10, 1028. doi: 10.3390/diagnostics10121028

Lodwick, G.S., Keats, T.E., Dorst, J.P., 1963. The coding of roentgen images for computer analysis as applied to lung cancer. Radiology 81, 185–200. doi: 10.1148/81.2.185

Longjiang, E., Baisong Zhao, Liu, H., Zheng, C., Song, X., Cai, Y., Liang, H., Liang, C.H., Liu, Y.C., Wu, M.T., Garcia-Castro, F., Alberich-Bayarri, A., Wu, F.Z., 2020. Segmenting chest radiographs for deep space and shape learning. IEEE Transactions on Biomedical Engineering 67, 1206–1220. doi: 10.1109/TBME.2019.2935058

Lumsden, A.O., von Berg, J., Fabritius, A., Lorenz, C., Meyer, C., 2018. Localization and Labeling of Posterior Ribs in Chest Radiographs Using a CRF-regularized FCN with Local Refinement, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer. volume 11071, pp. 562–570. doi: 10.1007/978-3-030-00934-2_63

Maguolo, G., Nanni, L., 2020. A critic evaluation of methods for covid-19 automatic detection from x-ray images. arXiv preprint arXiv:2004.12823

Marañón, U., 2019. Deep Learning to Assess Long-term Mortality From Cardiovascular Disease, in: 2019 IEEE International Symposium on Biomedical Imaging (ISBI 2019), IEEE. pp. 1038–1042. doi: 10.1109/ISBI.2019.8536759

Mader, A.O., von Berg, J., Fabritius, A., Lorenz, C., Meyer, C., 2018. Localization and Labeling of Posterior Ribs in Chest Radiographs Using a CRF-regularized FCN with Local Refinement, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer. volume 11071, pp. 562–570. doi: 10.1007/978-3-030-00934-2_63

Mansoor, A., Cerrolazaj, J.P., Perez, G., Biggs, E., Okada, K., Nino, G., Lingurar, M.G., 2020. A Generic Approach to Lung Field Segmentation From Chest Radiographs Using Deep Space and Shape Learning. IEEE Transactions on Biomedical Engineering 67, 1206–1220. doi: 10.1109/TBME.2019.2935058

Mansoor, A., Perez, G., Nino, G., Lingurar, M.G., 2016. Automatic tissue characterization of air trapping in chest radiographs using deep neural networks, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 97–100. doi: 10.1109/EMBC.2016.7590649

Mao, C., Yao, L., Pan, Y., Luo, Y., Zeng, Z., 2018. Deep Generative Classifiers for Thoracic Disease Diagnosis with Chest X-Rays, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 1209–1214. doi: 10.1109/BIBM.2018.8621107

Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H., 2020. Abnormality Detection in Chest X-Ray Images Using Uncertainty Prediction Autoencoders, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer. volume 12266, pp. 529–538. doi: 10.1007/978-3-030-59725-2_51

Mathai, T.S., Gorantla, V., Galeotti, J., 2019. Segmentation of Vessels for Thoracic Disease Diagnosis with Chest X-ray Images, in: 2018 IEEE International Conference on Computer Vision Workshops, IEEE. pp. 4109–4117. doi: 10.1109/ICCVW.2018.8759247

Matsubara, N., Teramoto, A., Saito, K., Fujita, H., 2020. Bone suppression for chest x-ray image using a convolutional neural filter. Physical and Engineering Sciences in Medicine 43, 97–108. doi: 10.1007/s10488-019-00822-v

Matsumoto, T., Kodaka, S., Shinohara, H., Ieki, H., Yamaguchi, T., Higashikuni, Y., Kiyosue, A., Ito, K., Ando, J., Takimoto, E., Akazawa, H., 2020. Detecting Heart Failure from Chest X-Ray Images Using Deep Learning. International Heart Journal 61, 781–786. doi: 10.1016/j.ihj.2019.07.014

McDonald, C.J., Overhage, J.M., Barnes, M., Schadow, G., Blevins, L., Dexter, P.R., Mamlin, B., 2005. The Indiana Network For Patient Care: A Working Local Health Information Infrastructure. Health Affairs 24, 1214–1220.
McManigle, J.E., Bartz, R.R., Carin, L., 2020. Y-Net for Chest X-Ray Preprocessing: Simultaneous Classification of Geometry and Segmentation of Annotates, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 1266–1269. doi: 10.1109/EMBC44109.2020.9176354

Mettler, F.A., Bhargavan, M., Faulkner, K., Gilley, D.B., Gray, J.E., Ibbott, G.S., Lipoti, I.A., Mahesh, M., Mcloone, J.L., Stabin, M.G., Thomsen, B.R., Yoshizumi, T.T., 2009. Radiologic and nuclear medicine studies in the United States and worldwide: Frequency, radiation dose, and comparison with other radiation sources—1950–2007. Radiology 253, 520–531. doi: 10.1148/radiol.2532082101

Meyers, P.H., Nice, C.M., Becker, H.C., Nettleton, W.J., Sweeney, J.W., Meckstroth, G.R., 1964. Automated computer analysis of radiographic images. Radiology 83, 1029–1034. doi: 10.1148/83.6.1029

Michael, P., Yoon, H.J., 2020. Survey of image denoising methods for medical images. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 1225–1228. doi: 10.1109/EMBC44109.2020.9175246

Moradi, M., Madani, A., Karagyris, A., Syeda-Mahmood, T.F., 2018b. Chest x-ray generation and data augmentation for cardiovascular abnormality classification, in: Medical Imaging 2018: Image Processing, SPIE. p. 57. doi: 10.1117/12.2293971

Moradi, M., Madani, A., Karagyris, A., Syeda-Mahmood, T., 2018. Development and validation of a deep learning algorithm for detecting 10 common abnormalities on chest radiographs. Eur. Respir. J. 53, 1020–1029. doi: 10.1183/13993003.1020-2019

Nam, J.G., Park, J., Hwang, E.J., Lee, J.H., Jin, K.N., Lim, K.Y., Yu, T.H., Sohn, J.H., Hwang, S., Goo, J.M., Park, C.M., 2019. Development and validation of Deep Learning—Based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 290, 218–228. doi: 10.1148/radiol.2018182037

Novikov, A.A., Lenis, D., Major, D., Hladuvka, J., Wimmer, M., Buhler, K., 2018. Fully convolutional architectures for multiclass segmentation in chest radiographs. IEEE Transactions on Medical Imaging 37, 1865–1876. doi: 10.1109/tmi.2018.2805086

Oh, D.Y., Kim, J., Lee, K.J., 2019. Longitudinal Change Detection on Chest X-Rays Using Geometric Correlation Maps, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer. volume 11769, pp. 748–756. doi: 10.1007/978-3-030-32226-7_83

Oh, Y., Park, S., Ye, J.C., 2020. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. IEEE Transactions on Medical Imaging 39, 2688–2700. doi: 10.1109/tmi.2020.2993291

Onodera, S., Lee, Y., Tanaka, Y., 2020. Evaluation of dose reduction potential in scatter-corrected bedside chest radiography using U-Net. Radiological Physics and Technology 13, 336–347. doi: 10.1007/s12194-020-00586-z

Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., 2020. Deep Anomaly Detection in Chest Radiographs. J. Digit. Imaging doi: 10.1007/s10278-020-00413-2

Oyama, T., Sato, T., Takahashi, K., 2018. Computerized chest radiography. In: Radiology 253, 520–531. doi: 10.1148/radiol.2018180237

Ozsoylu, E., Bozdag, I., 2019. Evaluating transfer knowledge for rib segmentation in chest x-rays. Pattern Recognition 97, 30–30–32251–9_36

Park, C.M., 2020. Development and validation of a deep learning algorithm for detecting 10 common abnormalities on chest radiographs. Eur. Respir. J. 53, 1020–1029. doi: 10.1183/13993003.1020-2019

Nakao, T., Hanaoka, S., Nomura, Y., Murata, M., Takenaga, T., Miki, S., Watanadi, T., Yoshikawa, T., Hayashi, N., Abe, O., 2021. Unsupervised Deep Anomaly Detection in Chest Radiographs. J. Digit. Imaging doi: 10.1007/s10278-020-00413-2

Nam, J.G., Park, J., Hwang, E.J., Lee, J.H., Hong, J.H., Goo, J.M., Park, C.M., 2020. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. Eur. Respir. J. 53, 1020–1029. doi: 10.1183/13993003.1020-2019

Nam, J.G., Park, J., Hwang, E.J., Lee, J.H., Jin, K.N., Lim, K.Y., Yu, T.H., Sohn, J.H., Hwang, S., Goo, J.M., Park, C.M., 2019. Development and validation of Deep Learning-Based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 290, 218–228. doi: 10.1148/radiol.2018182037

Oliveira, H.N., Ferreira, E., Santos, J.A.D., 2020b. Truly generalizable radiographic image classification potential in scatter-corrected bedside chest radiography using U-Net. European Journal of Radiology 121, Ahead of Print. doi: 10.1016/j.ejrad.2020.105898

Oliveira, H., dos Santos, J., 2018. Deep transfer learning for segmentation of anatomical structures in chest radiographs, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE. pp. 264–2651. doi: 10.1109/sibgrapi.2018.00033

Oliveira, V.P., Pande, T., Huddart, S., Pai, M., Saravu, K., 2020. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. Scientific Reports 10, 210. doi: 10.1038/s41598-019-56589-3

Ogawa, R., Kido, T., Kido, T., Mochizuki, T., 2019. Effect of augmented datasets on deep convolutional neural networks applied to chest radiographs. Clinical Radiology 74, 697–701. doi: 10.1106/j.crad.2019.04.025

Oktay, H., 2020. Deep Transfer learning for Segmentation of Chest Radiographs Using Semantic Segmentation. In: Deep Learning for Medical Image Analysis (2020). pp. 264–2651. doi: 10.1109/tmi.2020.2993291

Oh, D.Y., Kim, J., Lee, K.J., 2019. Longitudinal Change Detection on Chest X-Rays Using Geometric Correlation Maps, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer. volume 11769, pp. 748–756. doi: 10.1007/978-3-030-32226-7_83

Oh, Y., Park, S., Ye, J.C., 2020. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. IEEE Transactions on Medical Imaging 39, 2688–2700. doi: 10.1109/TMI.2020.2993291

Ouahabi, M., Bouchoucha, H., Bouzidi, M., 2020. Transfer learning knowledge for scattering chest x-rays. Pattern Recognition Letters 140, 10–17. doi: 10.1016/j.patrec.2020.09.021

Oliveira, H., dos Santos, J., 2018. Deep transfer learning for segmentation of anatomical structures in chest radiographs, in: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE. pp. 204–211. doi: 10.1109/sibgrapi.2018.00003

Onodera, S., Lee, Y., Tanaka, Y., 2020. Evaluation of dose reduction potential in scatter-corrected bedside chest radiography using U-Net. Radiological Physics and Technology 13, 336–347. doi: 10.1007/s12194-020-00586-z

Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., 2020. Deep Anomaly Detection in Chest Radiographs. J. Digit. Imaging doi: 10.1007/s10278-020-00413-2

Park, C.M., 2020. Development and validation of a deep learning algorithm for detecting 10 common abnormalities on chest radiographs. Eur. Respir. J. 53, 1020–1029. doi: 10.1183/13993003.1020-2019
Wang, W., Feng, H., Bu, Q., Cui, L., Xie, Y., Zhang, A., Feng, J., Zhu, Z., Wehbe, R.M., Sheng, J., Dutta, S., Chai, S., Dravid, A., Barutcuh, S., Wu, Y., Wang, X., Yu, J., Zhu, Q., Li, S., Zhao, Z., Yang, B., Pu, J., 2020. Po-
Wang, Q., Liu, Q., Luo, G., Liu, Z., Huang, J., Zhou, Y., Xu, W., Wang, M., Deng, W., 2018. Deep visual domain adaptation: A survey. Neuro-
Wang, H., Wang, S., Qin, Z., Zhang, Y., Li, R., Xia, Y., 2021a. Triple attention
Wang, C., Elazab, A., Wu, J., Hu, Q., 2017a. Lung nodule classification using
Wang, C., Elazab, A., Jia, F., Wu, J., Hu, Q., 2018. Automated chest
Vayá, M.d.l.I., Saborit, J.M., Montell, J.A., Pertusa, A., Bustos, A., Cazorla, M., 2020, 1–9. doi:
Chen, Z., 2020d. MDU-Net: A Convolutional Network for Clavicle and Rib
Cantrell, D.R., Xiao, N., Allen, B.D., MacNealy, G.A., Savas, H., Agrawal, R., Parekh, N., Katsaggelos, A.K., 2020. DeepCOVID-X: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large US Clinical Dataset. Radiology, 205311doi 10.1148/radiol.2020203511
Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S., 2018. Object Reg-
Wessel, J., Heinrich, M.P., von Berg, J., Franz, A., Saalbach, A., 2019. Sequential
Wolleb, J., Sandkühler, R., Cattin, P.C., 2020. DeScarGAN: Disease-Specific
Ying, S., Ge, Z., Zeng, R., Mahapatra, D., Sea, J., Law, M., Drummond, T., 2019. Adversarial Pulmonary Pathology Translation for Pairwise Chest X-Ray Data Augmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer. volume 11769, pp. 757–765. doi:
Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Chang, E.L., 2014. Deep learning of feature representation with multiple instance learning for medical image analysis, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1626–1630. doi:
Xie, C., Deng, Q., Li, X., Dou, Q., Heng, P.A., 2020. Cascaded Robust Learning at Imperfect Labels for Chest X-ray Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer. volume 12266, pp. 579–588. doi:
Xue, F.F., Peng, J., Wang, R., Zhang, Q., Zheng, W.S., 2019. Improving Robustness of Medical Image Diagnosis with Denoising Convolutional Neural Networks, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer. volume 11769, pp. 846–854. doi:
Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X., 2018a. Multimodal Recurrent Model with Attention for Automated Radiology Report Generation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer. volume 11070, pp. 457–466. doi:
Xue, Z., Antani, S., Long, R., Thoma, G.R., 2018b. Using deep learning for detecting gender in adult chest radiographs, in: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, SPIE. p. 10. doi:
Xue, Z., Jaeger, S., Antani, S., Long, R., Karagyris, A., Siegelman, J., Folio, L.R., Thoma, G.R., 2018c. Localizing tuberculosis in chest radiographs with deep learning, in: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, SPIE. p. 28. doi:
Wang, X., Xue, Z., Long, R., Jaeger, S., Folio, L., George Thoma, R., Antani, A.S., 2018d. Extraction of Aortic Knuckle Contour in Chest Radiographs Using Deep Learning, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. pp. 5890–5893. doi:
Yahyatabar, M., Jouvet, P., Cheriet, F., 2020. Dense-Unet: A light model for lung fields segmentation in chest X-ray images, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. pp. 985–988. doi:
Yang, W., Chen, Y., Liu, Y., Zhong, L., Qin, G., Lu, Z., Feng, Q., Chen, W., 2017. Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. Medical Image Analysis 35, 421–433. doi:
Yao, L., Proksy, J., Covington, B., Lyman, K., 2019. A strong baseline for domain adaptation and generalization in medical imaging, in: International Conference on Medical Imaging with Deep Learning – Extended Abstract Track, pp. 1–4.
Yi, P.H., Kim, T.K., Yu, A.C., Bennett, B., Eng, J., Lin, C.T., 2020. Can AI out-perform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. Emergency Radiology 27, 367–375. doi:
Xi, Y., Adams, S., Babyn, P., Elnahal, A., 2019a. Automatic Catheter and Tube
Detection in Pediatric X-ray Images Using a Scale-Reccurrent Network and Synthetic Data. Journal of Digital Imaging 33, 181–190. doi:10.1007/s10278-019-00201-7

Yi, X., Wallia, E., Babyn, P., 2019b. Generative adversarial network in medical imaging: A review. Medical Image Analysis 58, 101552. doi:10.1016/j.media.2019.101552

Yoo, H., Kim, K.H., Singh, R., Dignamurthy, S.R., Kalra, M.K., 2020. Validation of a Deep Learning Algorithm for the Detection of Malignant Pulmonary Nodules in Chest Radiographs. JAMA Network Open 3, e2017135. doi:10.1001/jamanetworkopen.2020.17135

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks?, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 1–9.

Young, M., 1994. Interobserver variability in the interpretation of chest roentgenograms of patients with possible pneumonia. Archives of Internal Medicine 154, 2729. doi:10.1001/archinte.1994.0402003122014

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., Zhao, R., 2020. 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. doi:10.1016/j.cmpb.2020.105674

Yuan, J., Liao, H., Luo, K., Luo, J., 2019. Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer, volume 11769, pp. 721–729. doi:10.1007/978-3-030-32226-7_89

Yue, Z., Ma, L., Zhang, R., 2020. Comparison and Validation of Deep Learning Models for the Diagnosis of Pneumonia. Computational Intelligence and Neuroscience 2020, 1–8. doi:10.1155/2020/8876798

Zarei, M., Abadi, E., Fricks, R., Segars, W.P., Samei, E., 2021. A probabilistic conditional adversarial neural network to reduce imaging variation in radiography, in: Medical Imaging 2021: Physics of Medical Imaging, International Conference on Computer Vision (ICCV), IEEE. pp. 2223–2232. doi:10.1109/ICCV49210.2016.779224

Zarchenas, A., Liu, J., Forti, P., Suzeiki, K., 2019. Separation of bones from soft tissue in chest radiographs: Anatomy-specific orientation-frequency-specific deep neural network convolution. Medical Physics 46, 2232–2242. doi:10.1002/mp.13468

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2019. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS ONE 15, e0236378. doi:10.1371/journal.pone.0236378

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., Chen, G.H., 2021c. Diagnosis of Coronavirus Disease 2019 Pneumonia by Using Chest Radiographs: Value of Artificial Intelligence. Radiology 298, E88–E97. doi:10.1148/radiol.2020202944

Zhang, T., Fu, H., Zhao, Y., Cheng, J., Guo, M., Gu, Z., Yang, B., Xiao, Y., Gao, S., Liu, J., 2019a. SkrGAN: Sketching-Rendering Unconditional Generative Adversarial Networks for Medical Image Synthesis, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer, volume 11769, pp. 779–788. doi:10.1007/978-3-030-32251-9_85

Zhang, W., Li, G., Wang, F., Li, L., Lin, L., Liang, H., 2019c. Simultaneous Lung Field Detection and Segmentation for Pediatric Chest Radiographs, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer, volume 11769, pp. 594–602. doi:10.1007/978-3-030-32226-7_68

Zhang, Y., Miao, S., Mansi, T., Liao, R., 2018. Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-ray Image Segmentation, in: Medical Image Computing and Computer Assist-