U.Porto Journal of Engineering, 7:4 (2021) 16-32
ISSN 2183-6493
DOI: 10.24840/2183-6493_007.004_0002

Received: 27 November, 2020
Accepted: 14 December, 2020
Published: 26 November, 2021

A Review on Deep Learning Methods for Chest X-Ray based Abnormality Detection and Thoracic Pathology Classification

Joana Rocha[1], Ana Maria Mendonça[2], Aurélio Campilho[3]

[1]INESC TEC-Institute for Systems and Computer Engineering and Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (joana.m.rocha@inesctec.pt) ORCID 0000-0002-4856-138X; [2]INESC TEC-Institute for Systems and Computer Engineering and Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (amendon@fe.up.pt) ORCID 0000-0002-4319-738X; [1]INESC TEC-Institute for Systems and Computer Engineering and Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal (campilho@fe.up.pt) ORCID 0000-0002-5317-6275

Abstract
Backed by more powerful computational resources and optimized training routines, Deep Learning models have proven unprecedented performance and several benefits to extract information from chest X-ray data. This is one of the most common imaging exams, whose increasing demand is reflected in the aggravated radiologists' workload. Consequently, healthcare would benefit from computer-aided diagnosis systems to prioritize certain exams and further identify possible pathologies. Pioneering work in chest X-ray analysis has focused on the identification of specific diseases, but to the best of the authors' knowledge no paper has specifically reviewed relevant work on abnormality detection and multi-label thoracic pathology classification. This paper focuses on those issues, selecting the leading chest X-ray based deep learning strategies for comparison. In addition, the paper discloses the current annotated public chest X-ray databases, covering the common thorax diseases.

Author Keywords. Computer-aided Diagnosis, Deep Neural Network, Medical Imaging, Radiology, Thorax.

Type: Research Article
Open Access Peer Reviewed CC BY

1. Introduction

Among the popular medical imaging exams, the Chest X-Ray (CXR) is frequently requested by healthcare professionals to assess the presence of thoracic diseases, due to its low-cost non-invasive nature. Nevertheless, the thorough analysis of CXR images is time-consuming and their interpretation may be dubious even for expert radiologists (Shaw, Hendry, and Eden 1990). For this reason, the incorporation of computer-aided diagnosis systems in the hospitals is an attractive solution to increase the productivity and efficiency in the interpretation of these exams, by providing a second opinion. Considering the recurring need to assess several types of thoracic pathologies, Deep Learning (DL) based systems have been preferred over traditional machine learning approaches, following the advances in computational capabilities and the increasing availability of medical datasets. This way, the data-driven nature of DL has proved to achieve great performance for multi-disease detection and classification tasks, being a great preliminary diagnostic tool that reduces the physicians' workload. A CXR-based computer-aided diagnosis system encompasses several steps to reach a diagnosis, and perhaps one of the most important is abnormality detection, focused on the prioritization of more urgent abnormal cases. As mentioned in Yasaka and Abe (2018), this would be highly valuable for the clinicians to manage their time and resources, considering that cardiothoracic and pulmonary abnormalities are one of the leading causes of morbidity and mortality, according to Wang et al. (2016). This step could be further complemented with another important task, which is the identification of the pathologies present in the exam at hand. Here, one must consider a multi-label thoracic pathology classification approach, minding that it is possible to have more than one in the same image/patient.

The detection and classification of cardiothoracic and pulmonary abnormalities often resorts to Convolutional Neural Networks (CNNs), due to their great ability to handle data with strong spatial relationship. In fact, these networks have been capable of matching or even exceeding human performance in other medical-related tasks, namely the diabetic retinopathy detection (Ting et al. 2017), and skin cancer classification (Esteva et al. 2017). Yet, to the best of the authors' knowledge, no paper addresses a state-of-the-art review based exclusively on deep learning approaches to solve both the thoracic abnormality detection and classification tasks. For this reason, the present publication intends to gather and describe all public annotated CXR databases and analyse how they have been used in the most relevant papers to tackle abnormality detection and pathology classification. Consequently, certain criteria were defined to select the papers from arXiv, IEEE Xplore, PubMed, and Scopus: employing a 100% DL based methodology, published after 2015 to ensure the novelty of the work, which exclusively extracts information from images (and not radiology reports), and presents a relevant study in the field. It was also established that the comparison between the selected papers would be done based on the most frequently observed evaluation metric, the Area Under Curve (AUC). In summary, besides this introduction, the paper includes a review of the CXR datasets in Section 2, abnormality detection in Section 3, and multi-label thoracic pathology classification in Section 4. Finally, Section 5 presents the main conclusions.

2. Chest X-ray Datasets

Although DL approaches have proven to significantly improve the performance of computer-aided diagnosis systems, it is also noticeable that their distinctive data-hungry nature impairs further achievements. In fact, any achievements made in the recent past were only enabled by the publication of larger public CXR datasets. For this reason, it is still ambitious to say that these systems will soon have a truly large-scale high precision implementation in a real-life clinical domain, considering the challenges tied to the collection and annotation of CXR datasets. For example, Shin, Lu, and Summers (2017) state that it is not clear how to annotate the large amount of CXR images needed for DL methods, particularly ensuring their required precision. Besides, there are multiple approaches to even define the labels themselves, or the criteria to follow during the annotation process.

In spite of all these difficulties, several CXR datasets have been published, which can be split into two main groups - the ones which tackle a specific thoracic pathology, and the ones which annotate multiple pathologies. While this paper will briefly describe the first group, the focus of this work are the datasets which encompass more than a single pathology. For instance, the JRST dataset presented in Shiraishi et al. (2000) contains 247 frontal CXR images with and without lung nodules, from 14 medical centers, being one of the first available collections. Jaeger et al. (2014)

provided two datasets centered in tuberculosis, named MC and Shenzhen sets. These were collected in the United States and Shenzhen, and contain 138 and 662 frontal CXR images respectively, presenting both normal and tuberculosis cases. Additionally, Ryoo and Kim (2014) also introduced a total of 10848 observations from the Korean Institute of Tuberculosis (KIT). Considering CXR datasets which tackle several pathologies, Gohagan et al. (2000) proposed the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial which resulted in a 13-label partially public set of 185 421 CXR images from 56 071 patients. The Open-I Indiana University dataset was published in Demner-Fushman et al. (2016), collecting 8 121 associated frontal images from two large hospitals in the Indiana Network for Patient Care, and addressing the 10 most prevalent conditions observed in 3 996 subjects.

Later on, the National Institutes of Health (NIH) released the ChestX-ray8 in Wang et al. (2017a) and compiled 108 948 frontal views belonging to 32 717 unique patients and a total of 8 associated pathologies (Figure 1) extracted from radiological reports using natural language processing. This dataset evolved to include 6 more categories, increasing the overall number of frontal CXR images and resulting in ChestX-ray14. It is argued that this version is more representative of the patients' distributions and diagnosis in comparison to the previously mentioned set (Wang et al. 2017b). This way, ChestX-ray14 comprises a total of 112 120 images from 30 805 patients and 14 pathologies, and is by far the most popular dataset being used in today's research. Another staple among the most popular CXR datasets is the CheXpert, as seen in Irvin et al. (2019), counting with 224 316 frontal and lateral images and 65 240 patients from the Standford Hospital. CheXpert is distinctive because it not only recognizes the presence of 12 pathology-related classes, but also the presence of medical support devices and fractures, all described in radiology reports that were released along with the images.

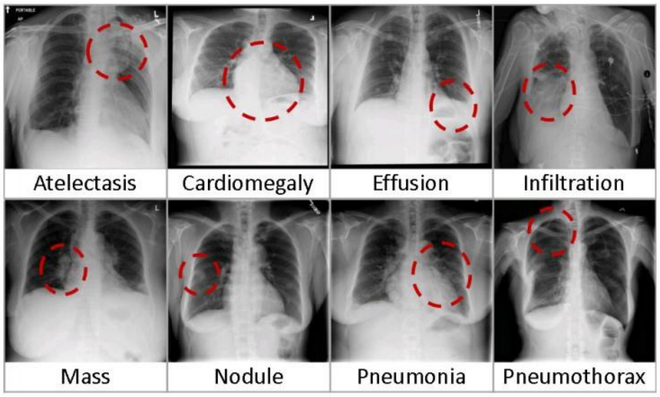

Figure 1: Eight common thoracic diseases observed in ChestX-ray8 and ChestX-ray14 (Wang et al. 2017a)

After that, Johnson et al. (2019a) released the MIMIC-CXR, making available 377 110 frontal and lateral views of 65 379 patients from the Beth Israel Deaconess Medical Center Emergency Department, along with text radiology reports. Another version of the same dataset was published, the MIMIC-CXR-JPG, where the images are in a JPG format instead of the original DICOM (Johnson et al. 2019b). The MIMIC-CXR was annotated with the same labels as CheXpert. Finally, PadChest became very recently available in Bustos et al. (2020), containing a total of 193 labels applied to 160 868 frontal and lateral CXR images of 67 625 patients. It was collected from the San Juan Hospital, considering radiology reports written in Spanish. Table 1 focuses on the multiple pathology datasets presented above and summarizes their content.

As will be made clear in the following sections, the ChestX-ray14 is undoubtedly the benchmark dataset for CXR-based computer-aided diagnosis. For this reason, there are some important considerations to be made about its limitations. Firstly, the ChestX-ray14 labels were text-mined from radiology reports through natural language processing techniques, and no expert validation was performed to confirm if the final annotations match the image content. Instead, the annotated images were validated using the Open-I Indiana University dataset (F1 score of 0.90). This raises some questions regarding the accuracy of the annotations, namely how accurately these labels reflect the pathology(ies) present in each image. The lack of manual and expert-based verification does not ensure that the positive predictive values of the text-mined ground-truth match the positive predictive values one would achieve with the visual queues. In addition to that, the established labels are not detailed, in the sense that they do not provide information on the expected range of abnormalities beside those 14 pathologies (e.g. pacemakers and invasive lines), and that the "no finding" hypothesis does not guarantee a healthy observation - it simply ensures the absence of those 14 diseases (Yates, Yates, and Harvey 2018). Other issues can be addressed, such as the class imbalance among pathologies, or the relevance between the CXR images and some of the proposed annotations.

To conclude, and while the overall lack of diversity impairs the ChestX-ray14's generalization ability in heterogeneous real-world settings, this dataset has fuelled innovation and research and is considered highly valuable. Table 2 shows the labels and label distributions of ChestX-ray8, ChestX-ray14, and CheXpert, presenting these labels according to the groups defined in Irvin et al. (2019) for easier comparison.

Year	Dataset	Patients	Images	Format	CXR view	Non-normal labels
2000	PLCO	56 071	185 421	TIFF	frontal	12
2015	Open-I Indiana	3996	8 121	DICOM	frontal	10
2017	ChestX-ray8	32 717	108 948	DICOM	frontal	8
2017	ChestX-ray14	30 805	112 120	PNG	frontal	14
2019	CheXpert	65 240	224 316	PNG	frontal and lateral	13
2019	MIMIC-CXR MIMIC-CXR-JPG	65 379	377 110	DICOM JPG	frontal and lateral	13
2020	PadChest	67 625	160 868	DICOM	frontal and lateral	193

Table 1: Description of multiple pathology CXR datasets

Label group	**CheXpert**	
	No finding	16 627
Enlarged Cardiomediastinum	Enlarged Cardiomediastinum	9 020
	Cardiomegaly	23 002
Lung Opacity	Lung Opacity	92 669
	Atelectasis	29 333
	Lung Lesion	6 856
	Pneumonia	4 576
	Consolidation	12 730
	Edema	48 905
Pleural	Pleural Other	2 441
	Effusion	75 696
	Pneumothorax	17 313
Others	Fracture	7 270
	Support devices	105 831

Label group	**ChestX-ray8**		**ChestX-ray14**	
	No finding	84 312	No finding	60 412
Enlarged Cardiomediastinum	Cardiomegaly	1 010	Cardiomegaly	2 772
Lung Opacity	Atelactasis	5 789	Atelactasis	11 535
	Pneumonia	1 062	Pneumonia	703
			Consolidation	4 667
			Edema	2 303
Pleural	Effusion	6 331	Effusion	13 307
	Pneumothorax	2 793	Pneumothorax	5 298
Not contemplated in CheXpert	Infiltration	10 317	Infiltration	19 871
	Mass	6 046	Mass	5 746
	Nodule	1 971	Nodule	6 323
			Emphysema	2 516
			Fibrosis	1 686
			Pleural thickening	3 385
			Hernia	227

Table 2: Comparison of the ChestX-ray8, ChestX-ray14, and CheXpert annotations, with the respective number of samples in each class (adapted from Wang et al. (2017a) and Irvin et al. (2019))

3. Abnormality Detection

Published work in this field has typically favored pathology classification rather than abnormality detection; yet, such detection task can have a high impact when it comes to building a triage system for the CXR images being analyzed. Several approaches can be established to define the automated triage criteria of the patients' images, i.e. which labels to consider. While Tataru et al. (2017) suggest a more elaborate three label system (normal, abnormal. and emergent), the most common annotations are simply normal and abnormal. It is also possible to address the detection of a specific pathology, as tuberculosis (Sivaramakrishnan et al. 2018), pneumonia (Chouhan et al. 2020), or cardiomegaly (Islam et al. 2017), in which case the abnormal label stands for the presence of the considered condition. However, in this section only the generic normal and abnormal annotations will be considered, and so assuming a binary classification exercise.

Current standard off-the-shelf CNN-based methods are frequently applied to detect abnormalities in CXR, and there are several papers which establish a comparison between well-known architectures, as illustrated in Tang et al. (2020) and Dunnmon et al. (2019). In the first work, the authors consider the AlexNet (Krizhevsky, Sutskever, and Hinton 2017), VGG (Simonyan and Zisserman 2015), GoogLeNet (Szegedy et al. 2014), ResNet (He et al. 2015) and DenseNet (Huang et al. 2018), and the ChestX-ray14 dataset. Using transfer learning with pre-trained ImageNet weights (Deng et al. 2009), all CNNs achieved good results, with the DenseNet slightly outperforming the remaining methods. Regarding Dunnmon et al. (2019), which exploits a private database but also uses ImageNet weights, only the AlexNet, ResNet, and DenseNet were assessed for the automated binary triage, where the DenseNet surpassed the other networks. Yates, Yates, and Harvey (2018) used transfer learning on the Inception CNN (Szegedy et al. 2014), retraining its final layer to execute abnormality detection on a mixed of frontal CXR data from the ChestX-ray14 and Open-I Indiana datasets. Besides using transfer learning to reduce the needed computational resources, the authors advise to skip data augmentation, arguing that it is unlikely to result in a reliable representation of any collected real datasets. This way, they gathered the normal Open-I Indiana CXR images (which unlike in ChestX-ray14 guarantee normality) and the 14 pathology examples from the latter as abnormality-positive samples. The CXNet-m1 is presented in Xu, Wu, and Bie (2019) and has a reduced number of convolutional layers in comparison to VGG, ResNet, and DenseNet. Unlike the previous publications, the authors argue against transfer learning in this context due to the dissimilarity between medical images and ImageNet's. Instead, they suggest that ChestX-ray14 is large enough to train a smaller CNN from scratch without time or memory limitations, proposing a hierarchical shorter CNN structure with an improved loss function (sin-loss) to address the information present in indistinguishable features and misclassified images.

All these methodologies look at the task at hand as a binary classification problem, but there are alternatives to approach abnormality detection. One of them is considering it a one-class exercise, where the goal is to classify a specific category of data amongst all observations, by primarily learning from a training set containing merely the objects of that class. Tang et al. (2019) adopt this research line and suggest an end-to-end architecture for abnormality detection using generative adversarial one-class learning and ChestX-ray14 (Figure 2). For this reason, the network only takes a normal CXR as input, which go through three main modules: a U-Net autoencoder (Ronneberger, Fischer, and Brox 2015), a CNN discriminator, and an encoder, which compete during the learning task while collaborating for the target task. Considering the model is trained exclusively on normal observations, the adversarial generative model is able to reconstruct a normal CXR, but performs poorly on an abnormal image, thus gaining the ability to distinguish both situations based on the reconstruction differentiation. A one-class autoencoder-based approach is also implemented in Mao et al. (2020), taking normal samples and outputting the reconstructed normal version of the images with an associated pixel-wise uncertainty. This way, abnormal observations in ChestX-ray14 can be identified considering the uncertainty-weighted reconstruction error as a measurement for abnormality presence. Both these publications are valuable in cases where annotating all abnormalities is impractical for large scale training or cannot be obtained (e.g. rare forms of abnormality that are difficult to collect).

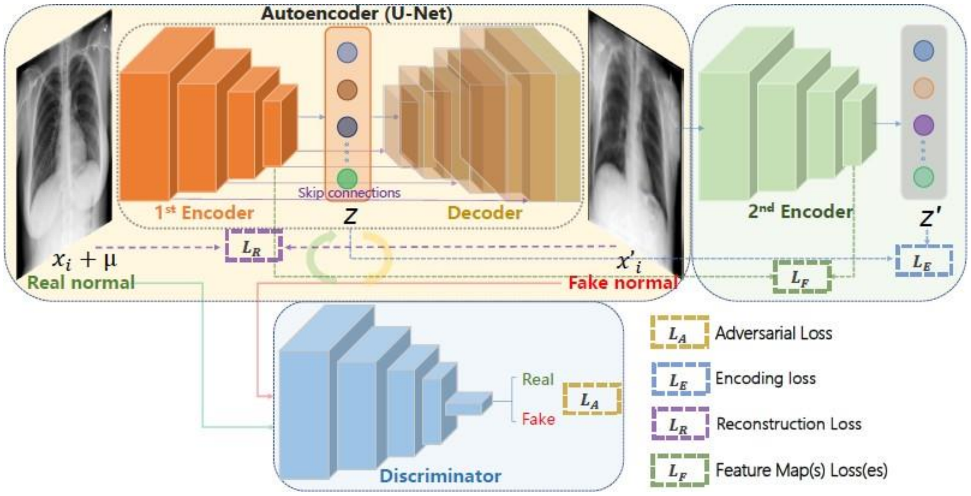

Figure 2: Architecture of a deep adversarial one-class learning model for abnormality detection (Tang et al. 2019)

While Shvetsova et al. (2020) agree that the autoencoders' implicit modelling of more complex data distribution is great for medical abnormality detection, the authors suggest to soften the one-class assumption. In other words, the authors skip an unsupervised detection where no abnormal observations are taken into account during the model's training, and instead use a limited subset of abnormal images to initiate hyperparameter search and grant the model a more flexible understanding of normality. Consequently, the deep perceptual autoencoder is capable of learning common patterns between normal observations and so accurately restore them, using the perceptual loss function to measure pattern dissimilarity. This works by minimizing the difference between the normalized features of the original and reconstructed images. Also evaluated on ChestX-ray14, the overall framework is represented in Figure 3.

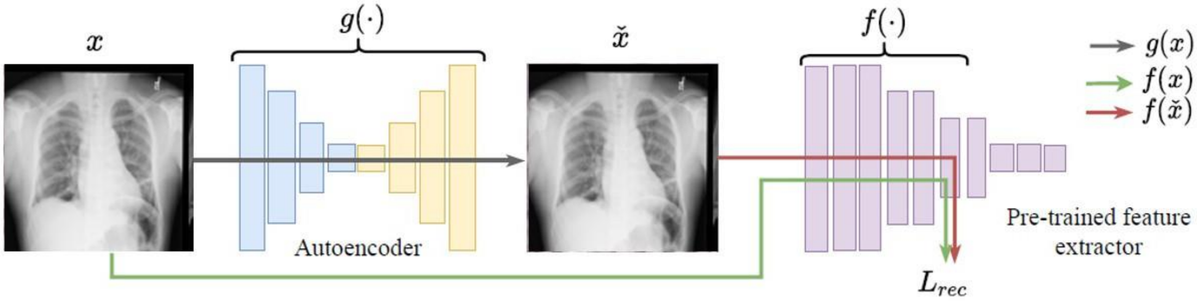

Figure 3: The proposed deep perceptual autoencoder for image anomaly detection: g denotes the autoencoder network, f denotes a feature extractor, x is an image, and $\hat{x}= g(x)$ is a reconstructed "image". Reconstruction loss $Lrec$ calculates difference between deep features $f(x)$ and $f(\hat{x})$ (Shvetsova et al. 2020)

Finally, a different approach is proposed in Kieu et al. (2018) to tackle this decision, in which a private dataset goes through three different CNNs simultaneously (Multi-CNNs), represented in Figure 4. One of the networks takes the full CXR image, while the other two take either the left or right half of the same image, to ensure both sides are equally analysed. They all output the probability of normality and abnormality, which are then combined in a fusion rule to compute the final decision.

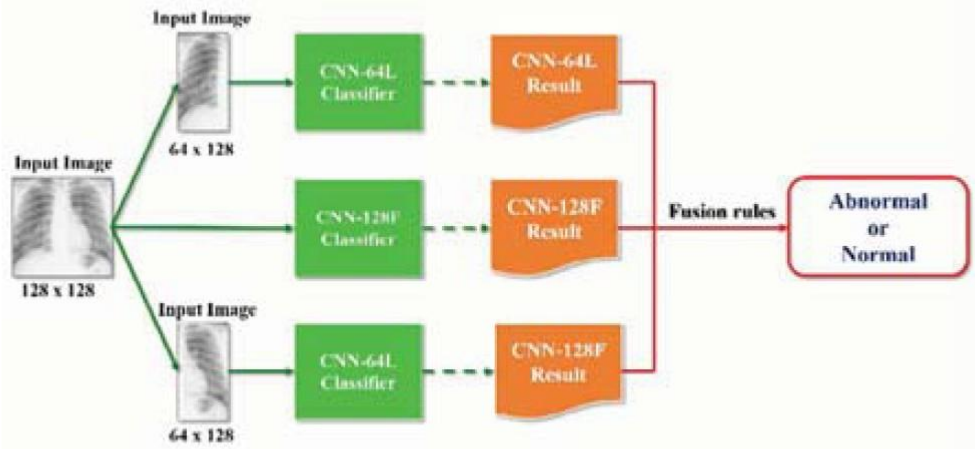

Figure 4: Architecture of the Multi-CNNs model (Kieu et al. 2018)

Table 3 summarizes the results of all the previously mentioned papers which evaluate their methodologies on publicly available datasets, highlighting the highest scores. While Yates, Yates, and Harvey (2018) and Tang et al. (2020) employ the standard off-the-shelf Inception and DenseNet CNNs respectively, Shvetsova et al. (2020) propose to train an autoencoder with a limited number of abnormal samples. Note that these results correspond to the best detection experiment in each paper and cannot be directly compared, as they may consider different databases or subsets of the same database. Further information on the data splits for validation and testing of the models can be found in the original publications.

Publication	AUC	Publication	AUC
Yates, Yates, and Harvey (2018)	**0.980**	Mao et al. (2020)	0.780
Xu, Wu, and Bie (2019)	0.795	Shvetsova et al. (2020)	**0.926**
Tang et al. (2019)	0.841	Tang et al. (2020)	**0.980**

Table 3: AUC scores for the mentioned CXR abnormality detection publications that are evaluated on public datasets. The highest performances are in bold.

4. Multi-label Thoracic Pathology Classification

The automatic identification of multiple pathologies in CXR is a much more common exercise in comparison to general abnormality detection. Consequently, there is a higher number of published articles with this particular aim, which often ally the classification with a location task. In such cases, the goal is not only to identify the pathologies present in the image, but also where they appear to be. While most papers presented in this section combine the two aspects, the focus of analysis will be the methodology and performance of their classification task. Nonetheless, it is still relevant to address that several articles seek to interpret their results with heat maps (frequently achieved with class activation mapping) to highlight class-specific regions of images and better demonstrate what the network considered relevant for pathology identification. Additionally, it is also common practice to use transfer learning with pre-trained ImageNet weights to speed the convergence of the classification models. All mentioned papers follow this procedure, unless stated otherwise.

As previously introduced, the work presented in this section tackles a multi-label classification exercise, meaning multiple pathologies can be identified in the same image. Perhaps one of the most popular publications for such purpose is the CheXNet's Rajpurkar et al. (2017), which is a classical example of a simple DenseNet implementation. Urinbayev et al. (2020) follow a similar approach to the CheXNet's, incorporating it in a more comprehensive end-to-end diagnosis framework, and claiming to outperform the state-of-the-art by using a more robust version of the Adam optimizer, known as RAdam. This is a variation that provides an automated, dynamic adjustment to the adaptive learning rate. Furthermore, Kumar, Grewal, and Srivastava (2018) apply the DenseNet in a boosted cascaded context without any transfer learning. The authors argue it is able to model complex dependencies among class labels, whilst taking advantage of the boosting strategy during training compared to single classifiers. Gündel et al. (2019) go a step further and use a DenseNet variant to propose the location-aware DNetLoc, which opposes class imbalance with additional weights within the loss function. These weights are tuned based on the label frequency per batch.

The ResNet is also a frequent option for image classification, as exemplified in Li et al. (2018). Here, the authors attempt to classify and locate the pathologies with limited supervision and a single model. More specifically, by slicing the image into a patch grid, the model is able to capture local information on each disease, while at the same time considering information present in the whole image. Alternatively, one can combine the ResNet and DenseNet to build the DualCheXNet by Chen et al. (2019). Its novel dual asymmetric architecture, i.e. with two asymmetric networks depicted in Figure 5, adaptively captures more discriminative features of several pathologies. In other words, since the DenseNet and ResNet capture different and unique features, the network is able to learn complementary details, thus increasing its performance. The two asymmetric feature streams are later combined with a fusion classifier, and evaluated based on a unified loss function, which is a variation of the weighted cross entropy loss with a modulating factor to deal with class imbalance.

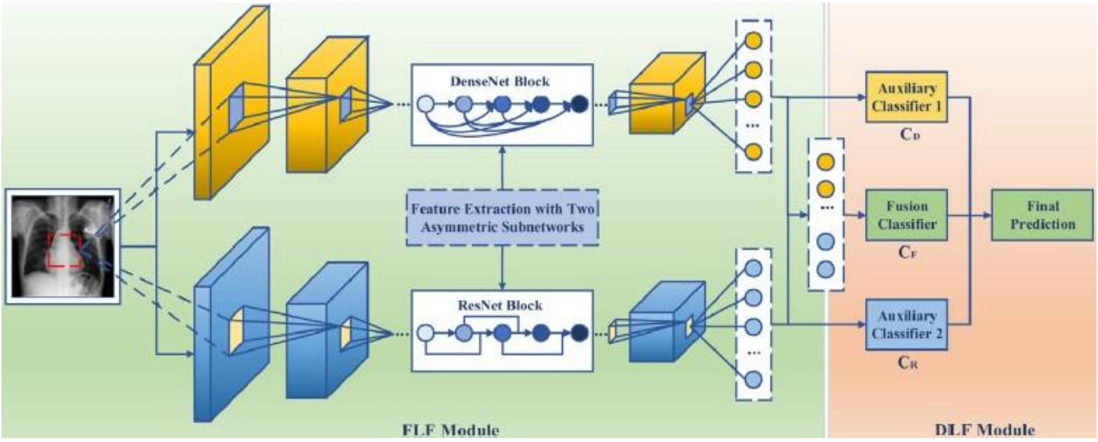

Figure 5: Schematic of the DualCheXNet (Chen et al. 2019)

Li et al. (2018) suggested limited supervision, but in fact several approaches have been implemented with weakly supervised networks. For instance, the original ChestX-ray8 publication, Wang et al. (2017a), highlight a ResNet-based approach for classification and location within a unified weakly supervised setting, considering different loss functions and pooling strategies. Simultaneously, Yan et al. (2018) tackle the same goals and context by enhancing the DenseNet with squeeze-and-excitation blocks and multi-map transfer. These contribute to boost the model's sensitivity to subtle differences between normal and abnormal regions, and the learning process of disease-specific features, respectively. Zhang, Chen, and Chen (2020) suggest a weakly supervised distance learning framework which, by learning discriminative features among triplets of images, is able to discriminate subtle disease characteristics. As shown in Figure 6, the network considers a pair of images that share the class annotation, and another image which does not. By comparing the unannotated observation (anchor) with images whose pathology is known (positive/negative), the network is able to differentiate the classes by imposing a similarity metric to be lower when image pairs share a similar disease, and higher when there is nothing in common. In addition, the approach also trains a different classifier on region features to verify if the attentive regions contain information indicative of any disease.

This leads to another trend called attention learning, where the approaches selectively focus on relevant image regions to assess the presence of the pathologies. Guan et al. (2018) support these methods and defend that irrelevant noisy areas are present during global image training. In Figure 7, the authors provide an example of an attention-guided DenseNet (AG-CNN) with three branches, to learn from both the disease- specific regions, solving the noise issue, and the global image information, avoiding the loss of discriminative clues. Another example is given in the A3Net's triple attention learning strategy (Wang et al. 2021). Here, a model with a DenseNet backbone encompasses three learning modules with channel-wise, element-wise, and scale-wise attention. Each of these grants information on the most discriminative feature channels, regions of interest, and scales, respectively. Moving on to Guan and Huang (2020), category-wise residual attention learning (CRAL) embraces the classification exercise with a class-specific attentive view. This means that the relevance of the features is endorsed by weights based on each category and region, and that these scores are then embedded into a DenseNet's attention blocks to output a final classification. To conclude, Liu et al. (2019) present a contrast-induced attention network (CIA-Net) for disease classification and location based on the contrastive learning of positive and negative observations. In detail, the framework starts by adjusting all images in terms of scale and angles, to take advantage of a highly structured input and so compute a distance between corresponding pixel coordinates in the positive and negative samples. The distances act as an indication of the lesion areas, thus assisting the contrast induced attention branch of the CIA-Net in the final prediction. Note that this particular branch generates attention for every label when analysing a pair of negative and positive images.

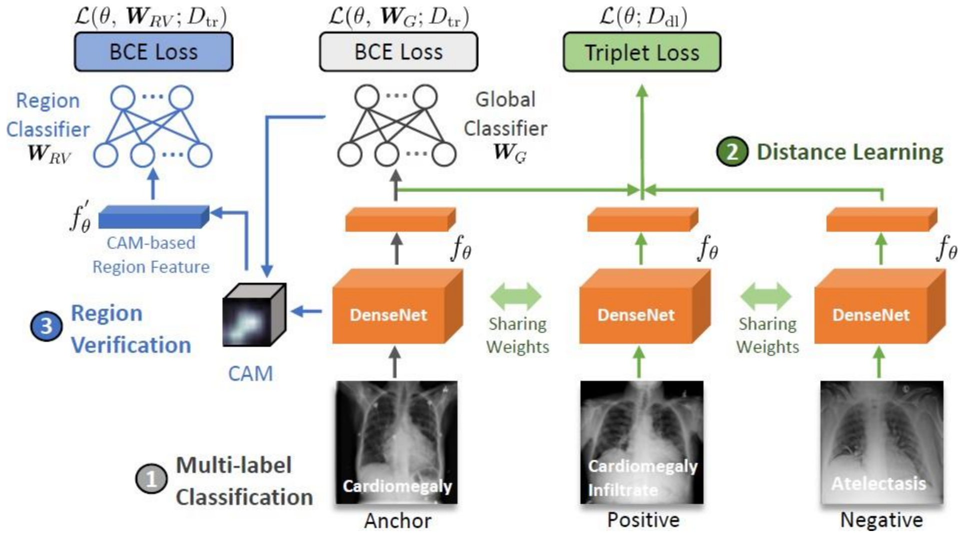

Figure 6: A distance learning based model for thoracic pathology classification and location (Zhang, Chen, and Chen 2020)

Considering that all the papers mentioned in this section evaluated their thoracic pathology classification models on the ChestX-ray14, Table 4 presents the 14 labels established for this dataset, along with the metrics achieved by each publication, i.e. their AUC scores per class and mean AUC scores. The mean values are present in the last column, which highlights the three highest scores. All highlighted publications focus on capturing more discriminative characteristics of each pathology present in the images. Guan et al. (2018) identify those subtle features by implementing an attention-guided CNN with three branches, while Chen et al. (2019) do that by combining the ResNet and the DenseNet. Finally, Zhang, Chen, and Chen (2020) opted for a weakly supervised distance learning approach to spot the same indicative attributes. It is important to once again remind that the same data split is not guaranteed, and so it is not possible to establish a direct comparison of the publications. However, one can perceive that there is an overall consistency between the performance values.

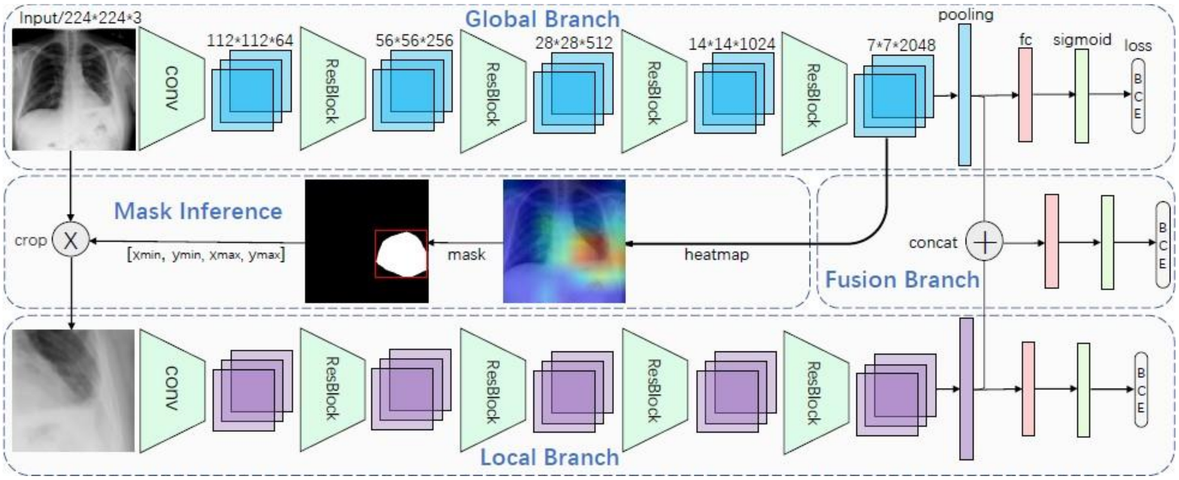

Figure 7: Overall structure of the Attention Guided Convolutional Neural Network (AG-CNN) (Guan et al. 2018)

Publication	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Consolidation	Edema	Emphysema	Fibrosis	Pleural Thick.	Hernia	AUC
Rajpurkar et al. (2017)	0.809	0.925	0.864	0.735	0.868	0.780	0.768	0.889	0.790	0.888	0.937	0.805	0.806	0.916	0.837
Wang et al. (2017a)	0.700	0.810	0.759	0.661	0.693	0.669	0.658	0.799	0.703	0.805	0.833	0.786	0.684	0.872	0.731
Kumar, Grewal, and Srivastava (2018)	0.743	0.893	0.862	0.675	0.789	0.704	0.638	0.853	0.760	0.882	0.916	0.752	0.757	0.864	0.775
Li et al. (2018)	0.700	0.870	0.870	0.700	0.830	0.750	0.670	0.870	0.800	0.880	0.910	0.780	0.790	0.770	0.795
Guan et al. (2018)	0.853	0.939	0.903	0.754	0.902	0.828	0.774	0.921	0.842	0.924	0.932	0.864	0.837	0.921	**0.883**
Yan et al. (2018)	0.792	0.881	0.842	0.710	0.847	0.811	0.740	0.876	0.760	0.848	0.942	0.833	0.808	0.934	0.837
Gündel et al. (2019)	0.826	0.911	0.885	0.716	0.854	0.774	0.765	0.872	0.806	0.892	0.925	0.820	0.785	0.941	0.840
Chen et al. (2019)	0.836	0.917	0.889	0.717	0.863	0.824	0.783	0.893	0.815	0.901	0.948	0.862	0.806	0.947	**0.863**
Liu et al. (2019)	0.790	0.870	0.880	0.690	0.810	0.730	0.750	0.890	0.790	0.910	0.930	0.800	0.800	0.920	0.805
Guan and Huang (2020)	0.781	0.883	0.831	0.697	0.830	0.764	0.725	0.866	0.758	0.853	0.911	0.826	0.780	0.918	0.828
Urinbayev et al. (2020)	0.810	0.910	0.870	0.720	0.850	0.780	0.740	0.900	0.790	0.910	0.920	0.810	0.790	0.990	0.830
Zhang, Chen, and Chen (2020)	0.845	0.905	0.877	0.817	0.859	0.824	0.804	0.871	0.810	0.862	0.896	0.849	0.829	0.927	**0.854**
Wang et al. (2021)	0.779	0.895	0.836	0.710	0.834	0.777	0.737	0.878	0.759	0.855	0.933	0.838	0.791	0.938	0.835

Table 4: Performance comparison for thoracic pathology classification, minding the AUC score per ChestX- ray14 class. The highest average AUC scores are in bold.

5. Conclusions

Computer-aided diagnosis seeks to provide a second opinion to healthcare professionals, reducing their workload and promoting a more accurate early diagnosis. These systems are particularly important to analyse CXR images containing complex information on a variety of pathologies that affect vital organs. Recent advances in DL strategies and computational resources have led to a steep performance increase in CXR-based computer-aided diagnosis algorithms, which also escalated due to the availability of larger annotated public CXR datasets. The present publication grants a description of the most relevant public annotated CXR datasets, as well as a comprehensive state-of-the-art review on two particular tasks - abnormality detection and thoracic pathology classification. One may notice that all selected papers were published in or after 2017, which is expected because they follow the recent release of the most popular datasets and the prominent DL trend. It is also noticeable that the results published for each task show no significant disparity, i.e. similar performance. In terms of abnormality detection, the leading publications concentrate mainly on standard off-the-shelf CNNs, which can be combined with one-class learning or fusion rule-based classification. In thoracic pathology classification, besides the same common CNNs, special attention is given to weakly supervised approaches and attention learning. To conclude, this publication provides an overview on the current knowledge on abnormality detection and thoracic pathology identification by describing and comparing a selected set of papers, considered by the authors as the most relevant in the field, in order to promote future research in this area.

References

Bustos, A., A. Pertusa, J. M. Salinas, and M. de la Iglesia-Vayá. 2020. "PadChest: A large chest x-ray image dataset with multi-label annotated reports". *Medical Image Analysis* 66 (december): Article number 101797. https://doi.org/10.1016/j.media.2020.101797.

Chen, B., J. Li, X. Guo, and G. Lu. 2019. "DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays". *Biomedical Signal Processing and Control* 53 (august): Article number 101554. https://doi.org/10.1016/j.bspc.2019.04.031.

Chouhan, V., S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius, and V. H. C. de Albuquerque. 2020. "A novel transfer learning based approach for pneumonia detection in chest X-ray images". *Applied Sciences* 10, no. 2 (january): Article number 559. https://doi.org/10.3390/app10020559.

Demner-Fushman, D., M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. 2016. "Preparing a collection of radiology examinations for distribution and retrieval". *Journal of the American Medical Informatics Association* 23, no. 2 (march): 304-10. https://doi.org/10.1093/jamia/ocv080.

Deng, J., W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. 2009. "ImageNet: A large-scale hierarchical image database". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-55. IEEE. https://doi.org/10.1109/CVPR.2009.5206848.

Dunnmon, J. A., D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren. 2019. "Assessment of convolutional neural networks for automated classification of chest radiographs". *Radiology* 290, no. 3 (march): 537-44. https://doi.org/10.1148/radiol.2018181422.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. "Dermatologist-level classification of skin cancer with deep neural networks". *Nature* 542, no. 7639 (february): 115-18. https://doi.org/10.1038/nature21056.

Gohagan, J. K., P. C. Prorok, R. B. Hayes, and B. S. Kramer. 2000. "The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the National Cancer Institute: History, organization, and status". *Controlled Clinical Trials* 21, no. 6 Suppl (december): 251S-72S. https://doi.org/10.1016/s0197-2456(00)00097-0.

Guan, Q., and Y. Huang. 2020. "Multi-label chest X-ray image classification via category-wise residual attention learning". *Pattern Recognition Letters* 130 (february): 259-66. https://doi.org/10.1016/j.patrec.2018.10.027.

Guan, Q., Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang. 2018. "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification". Preprint, submitted January 30, 2018. https://arxiv.org/abs/1801.09927.

Gündel, S., S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu. 2019. "Learning to recognize abnormalities in chest X-rays with location-aware dense networks". In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, edited by R. Vera-Rodriguez, J. Fierrez, and A. Morales, 757-65. Lecture Notes in Computer Science, vol. 11401. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-13469-3_88.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. "Deep residual learning for image recognition". Preprint, submitted December 10, 2015. https://arxiv.org/abs/1512.03385.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger. 2018. "Densely connected convolutional networks". Preprint, submitted August 25, 2016. https://arxiv.org/abs/1608.06993.

Irvin, J., P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. 2019. "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison". *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 1: 590-97. https://doi.org/10.1609/aaai.v33i01.3301590.

Islam, M. T., M. A. Aowal, A. T. Minhaz, and K. Ashraf. 2017. "Abnormality detection and localization in chest X-rays using deep convolutional neural networks". Preprint, submitted May 27, 2017. https://arxiv.org/abs/1705.09850.

Jaeger, S., S. Candemir, S. Antani, Y. X. Wang, P. X. Lu, and G. Thoma. 2014. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases". *Quantitative Imaging in Medicine and Surgery* 4, no. 6 (december): 475-77. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.

Johnson, A. E. W., T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. 2019a. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". *Scientific Data* 6, no. 1 (december): Article number 317. https://doi.org/10.1038/s41597-019-0322-0.

Johnson, A. E. W., T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. 2019b. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". Preprint, submitted January 21, 2019. https://arxiv.org/abs/1901.07042.

Kieu, P. N., H. S. Tran, T. H. Le, T. Le, and T. T. Nguyen. 2018. "Applying Multi-CNNs model for detecting abnormal problem on chest x-ray images". In *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering*, 300-05. IEEE. https://doi.org/10.1109/KSE.2018.8573404.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2017. "ImageNet classification with deep convolutional neural networks". *Communications of the ACM* 60, no. 6 (june): 84-90. https://doi.org/10.1145/3065386.

Kumar, P., M. Grewal, and M. M. Srivastava. 2018. "Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs". In *Image Analysis and Recognition. ICIAR 2018*, edited by A. Campilho, F. Karray, and B. ter Haar Romeny, 546-52. Lecture Notes in Computer Science, vol. 10882. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93000-8_62.

Li, Z., C. Wang, M. Han, Y. Xue, W. Wei, L. Li, and L. Fei-Fei. 2018. "Thoracic disease identification and localization with limited supervision". In *Proceedings 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8290-99. IEEE. https://doi.org/10.1109/CVPR.2018.00865.

Liu, J., G. Zhao, Y. Fei, M. Zhang, Y. Wang, and Y. Yu. 2019. "Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision". In *Proceedings 2019 International Conference on Computer Vision - ICCV 2019*, 10631-40. IEEE. https://doi.org/10.1109/ICCV.2019.01073.

Mao, Y., F.-F. Xue, R. Wang, J. Zhang, W.-S. Zheng, and H. Liu. 2020. "Abnormality detection in chest X-ray images using uncertainty prediction autoencoders". In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, edited by A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, 529-38. Lecture Notes in Computer Science, vol. 12266. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-59725-2_51.

Rajpurkar, P., J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, et al. 2017. "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning". Preprint, last revised December 25, 2017. https://arxiv.org/abs/1711.05225.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional networks for biomedical image segmentation". Preprint, submitted May 18, 2015. https://arxiv.org/abs/1505.04597.

Ryoo, S., and H. J. Kim. 2014. "Activities of the Korean Institute of Tuberculosis". *Osong Public Health and Research Perspectives* 5, no. S (december): S43-S49. https://doi.org/10.1016/j.phrp.2014.10.007.

Shaw, N. J., M. Hendry, and O. B. Eden. 1990. "Inter-observer variation in interpretation of chest X-rays". *Scottish Medical Journal* 35, no. 5: 140-41. https://doi.org/10.1177/003693309003500505.

Shin, H.-C., L. Lu, and R. M. Summers. 2017. "Natural language processing for large-scale medical image analysis using deep learning". In *Deep Learning for Medical Image Analysis*, edited by S. K. Zhou, H. Greenspan, and D. Shen, 405-21. Academic Press. https://doi.org/10.1016/B978-0-12-810408-8.00023-7.

Shiraishi, J., S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. I. Komatsu, M. Matsui, et al. 2000. "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules". *American Journal of Roentgenology* 174, no. 1 (january): 71-74. https://doi.org/10.2214/ajr.174.1.1740071.

Shvetsova, N., B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov. 2020. "Anomaly detection in medical imaging with deep perceptual autoencoders". Preprint, last revised October 15, 2020. https://arxiv.org/abs/2006.13265.

Simonyan, K. and A. Zisserman. 2015. "Very deep convolutional networks for large-scale image recognition". Preprint, submitted April 10, 2015. https://arxiv.org/abs/1409.1556.

Sivaramakrishnan, R., S. Antani, S. Candemir, Z. Xue, J. Abuya, M. Kohli, P. Alderson, and G. Thoma. 2018. "Comparing deep learning models for population screening using chest radiography". In

Progress in Biomedical Optics and Imaging - Proceedings of SPIE, Article number 105751E. https://doi.org/10.1117/12.2293140.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2014. "Going deeper with convolutions". Preprint, submitted September 17, 2014. https://arxiv.org/abs/1409.4842.

Tang, Y., Y. Tang, M. Han, J. Xiao, and R. M. Summers. 2019. "Abnormal chest X-ray identification with generative adversarial one-class classifier". In *Symposium Proceedings - 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1358-61. IEEE. https://doi.org/10.1109/ISBI.2019.8759442.

Tang, Y. X., Y. B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, et al. 2020. "Automated abnormality classification of chest radiographs using deep convolutional neural networks". *npj Digital Medicine* 3, no. 1 (december): Article number 70. https://doi.org/10.1038/s41746-020-0273-z.

Tataru, C., D. Yi, A. Shenoyas, and A. Ma. 2017. "Deep Learning for abnormality detection in chest X-ray images". In *IEEE Conference on Deep Learning*. http://cs231n.stanford.edu/reports/2017/pdfs/527.pdf

Ting, D. S. W., C. Y. L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, et al. 2017. "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes". *JAMA - Journal of the American Medical Association* 318, no. 22 (december): 2211-23. https://doi.org/10.1001/jama.2017.18152.

Urinbayev, K., Y. Orazbek, Y. Nurambek, A. Mirzakhmetov, and H. A. Varol. 2020. "End-to-end deep diagnosis of X-ray images". Preprint, submitted March 19, 2020. https://arxiv.org/abs/2003.08605.

Wang, H., M. Naghavi, C. Allen, R. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, et al. 2016. "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015". *The Lancet* 388, no. 10053: 1459-544. https://doi.org/10.1016/s0140-6736(16)31012-1.

Wang, H., S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia. 2021. "Triple attention learning for classification of 14 thoracic diseases using chest radiography". *Medical Image Analysis* 67 (january): Article number 101846. https://doi.org/10.1016/j.media.2020.101846.

Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 2017a. "ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition CVPR 2017*, 3462-71. IEEE. https://doi.org/10.1109/CVPR.2017.369.

———. 2017b. "NIH Chest X-rays". https://www.kaggle.com/nih-chest-xrays/data.

Xu, S., H. Wu, and R. Bie. 2019. "CXNet-m1: Anomaly detection on chest X-rays with image-based deep learning". *IEEE Access* 7: 4466-77. https://doi.org/10.1109/ACCESS.2018.2885997.

Yan, C., J. Yao, R. Li, Z. Xu, and J. Huang. 2018. "Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays". In *BCB '18: Proceedings of the 2018*

ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 103-10. Association for Computing Machinery. https://doi.org/10.1145/3233547.3233573.

Yasaka, K., and O. Abe. 2018. "Deep learning and artificial intelligence in radiology: Current applications and future directions". *PLoS Medicine* 15, no. 11 (november): Article number e1002707. https://doi.org/10.1371/journal.pmed.1002707.

Yates, E. J., L. C. Yates, and H. Harvey. 2018. "Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification". *Clinical Radiology* 73, no. 9 (september): 827-31. https://doi.org/10.1016/j.crad.2018.05.015.

Zhang, C., F. Chen, and Y.-Y. Chen. 2020. "Thoracic disease identification and localization using distance learning and region verification". Preprint, submitted August 11, 2020. https://arxiv.org/abs/2006.04203.