PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

EDUARDO HENRIQUE PAIS POOCH

PATHOLOGY LOCALIZATION ON CHEST RADIOGRAPHS WITH LIMITED SUPERVISION VIA SEMI-SUPERVISED MULTIPLE INSTANCE LEARNING

Porto Alegre
2021

AD VERVM DVCIT

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

PATHOLOGY LOCALIZATION ON CHEST RADIOGRAPHS WITH LIMITED SUPERVISION VIA SEMI-SUPERVISED MULTIPLE INSTANCE LEARNING

EDUARDO HENRIQUE PAIS POOCH

Master Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Rodrigo Coelho Barros

Porto Alegre
2021

Ficha Catalográfica

P821p   Pooch, Eduardo Henrique Pais

Pathology localization on chest radiographs with limited supervision via semi-supervised multiple instance learning / Eduardo Henrique Pais Pooch. – 2021.

70 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Rodrigo Coelho Barros.

1. Deep learning. 2. Medical Imaging. 3. Semi-supervised learning. 4. Multiple instance learning. I. Barros, Rodrigo Coelho. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

EDUARDO HENRIQUE PAIS POOCH

PATHOLOGY LOCALIZATION ON CHEST RADIOGRAPHS WITH LIMITED SUPERVISION VIA SEMI-SUPERVISED MULTIPLE INSTANCE LEARNING

This Master Thesis has been submitted in partial fulfillment of the requirements for the degree of Master in Computer Science, of the Computer Science Graduate Program, School of Technology of the Pontifical Catholic University of Rio Grande do Sul

Sanctioned on March 25, 2021.

COMMITTEE MEMBERS:

Prof. Dr. Julien Cohen-Adad (DEE/PolyMtl)

Prof. Dr. Felipe Rech Meneguzzi (PPGCC/PUCRS)

Prof. Rodrigo Coelho Barros (PPGCC/PUCRS - Advisor)

ACKNOWLEDGMENTS

This research was conducted within a project supported by the Brazilian Informatics Law (Law nº 8.248 of 1991) and was developed over Agreement 001/2015 between Pontifícia Universidade Católica do Rio Grande do Sul and HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda.

Research supported by HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda. using financial incentives of IPI refund reggarding the Law (Law nº 8.248 of 1991)

LOCALIZAÇÃO DE PATOLOGIAS EM RADIOGRAFIAS DE TÓRAX COM SUPERVISÃO LIMITADA VIA APRENDIZADO DE MÚLTIPLAS INSTÂNCIAS SEMI-SUPERVISIONADO

RESUMO

Radiografias são exames primários para a avaliação das condições do tórax. Na prática clínica, vem se popularizando a utilização de abordagens de aprendizado profundo para apoiar radiologistas no processo de tomada de decisão visando aumentar a acurácia diagnóstica. Para dar suporte adequado aos radiologistas, é insuficiente um modelo que simplesmente infere um rótulo diagnóstico. Idealmente, o modelo deve fornecer mais informações para apoiar o resultado da classificação, como a localização espacial do achado radiológico. Para treinar adequadamente modelos de aprendizado profundo, geralmente é necessário utilizar muitos dados anotados. Há uma grande quantidade de imagens de radiografias de tórax disponíveis publicamente, anotadas de acordo com a presença de achados radiológicos, mas poucas contêm uma anotação com a localização desses achados. O objetivo deste trabalho é utilizar a quantia limitada de dados anotados e a vasta quantia de dados não anotados para melhorar o desempenho de métodos de localização automática de patologias em radiografias de tórax. Identificamos o estado-da-arte de métodos semi-supervisionados e avaliamos seu desempenho em um cenário de classificação. Em seguida, estendemos o melhor método, Mean Teacher, para realizar a tarefa de localização em um framework de aprendizado de múltiplas instâncias, introduzindo nosso método `C-MIL`. Nesse paradigma, existem dois tipos de rótulos: um rótulo geral que é conhecido, e um rótulo mais específico e desconhecido mas que é relacionado ao conhecido, no caso, a presença de patologia e sua localização. Os resultados mostram melhorias na aplicação de regularização de consistência em um cenário de localização por meio de aprendizado de múltiplas instâncias e demonstram que os métodos de aprendizado semi-supervisionado são promissores para o avanço do desempenho de métodos de localização automática de patologias em imagens médicas.

Palavras-Chave: aprendizado profundo, imagens médicas, aprendizado semi-supervisionado, aprendizado de múltiplas instâncias.

PATHOLOGY LOCALIZATION ON CHEST RADIOGRAPHS WITH LIMITED SUPERVISION VIA SEMI-SUPERVISED MULTIPLE INSTANCE LEARNING

ABSTRACT

Radiographs are the primary examination for diagnosing chest conditions, and yet they are frequently misread/misdiagnosed due to human-observer confusion. In clinical practice, there is an increase of deep learning approaches to support radiologists on the decision-making process to improve diagnostic accuracy. To properly support radiologists, it is insufficient for the system to simply output a diagnosis label. Ideally, the model should provide more information to support the classification result, such as the spatial localization of the finding. To properly train deep learning models, we usually need lots of annotated data. There is a vast amount of publicly-available chest radiographs labeled according to their radiological findings (labels for classification), but very few contain a location annotation. Our goal is to extend the use of unlabeled data to improve pathology localization in chest radiographs in a scenario with limited labeled data. We identify state-of-the-art semi-supervised methods and evaluated their performance on a classification scenario. Next, we extend the best method, Mean Teacher, to perform localization within a multiple instance learning framework, introducing our method `C-MIL`. Multiple instance learning is a paradigm with two types of labels: a general label that is known, and a more specific and unknown label but related to the one known, in our case, pathology presence and its localization. Our results show improvements of applying consistency regularization over a multiple instance localization framework and demonstrate that semi-supervised learning methods are promising to advance the state-of-the-art performance of pathology localization methods.

Keywords: deep learning, medical imaging, semi-supervised learning, multiple instance learning.

LIST OF FIGURES

1.1 Chest radiographs annotated with 8 different radiological findings. Images and annotations from the ChestX-ray14 dataset [63]. 22

2.1 Chest pathology localization using three different learning paradigms: supervised, semi-supervised, and multiple instance learning. 25

4.1 Example of a chest radiograph (positive for consolidation) randomly sampled from each of the four analyzed datasets: ChestX-ray14, CheXpert, MIMIC-CXR, and PadChest. 35

4.2 Dataset pixel intensity probability density function (PDF) of the four datasets. 37

4.3 Average image of each of the four datasets. Last image contains 1/4 of each average image to better visualize the pixel intensity differences (I - ChestX-ray14, II - CheXpert, III - MIMIC-CXR, IV - PadChest). 37

4.4 Performance of a model trained on ChestX-ray14 (a), CheXpert (b), MIMIC-CXR (c), and PadChest (d) on each of the four test sets. 39

6.1 Architecture of the developed multiple instance learning localization model based on [33]. The model comprises a ResNet-50 encoder, which extracts features from the input image, an upsampling layer, and a sequence of convolution, batch normalization, ReLU, and another convolution, outputting a patch score matrix. A pooling function δ converts the patch scores into class scores and the model is optimized to reduce the classification loss $\mathcal{L}_{cls}$ between the prediction and the ground-truth y_{cls}. When the localization label is available, a localization loss $\mathcal{L}_loc$ is also computed using the patch scores and the y_{loc} ground-truth. 53

6.2 Architecture of `C-MIL`. A training dataset $\mathcal{D}$ contains images x, classification labels y_{cls} for each image, and localization labels y_{loc} for some images. The image is fed to a student model m_s that outputs patch scores, which are converted into class scores by a pooling function δ and compared to the ground-truth class labels y_cls to compute the classification loss $\mathcal{L}_{cls}$. If the localization labels are available for that image, a localization loss $\mathcal{L}_{loc}$ is computed using the patch scores and the ground-truth. And for all images, a teacher model m_t with an identical architecture receives a flipped version of the input image $\phi_t(x)$ and also generates patch scores. Then, the teacher output is flipped and compared against the student output to compute the consistency loss $\mathcal{L}_{con}$. The student is optimized to reduce these loss functions and the teacher is updated through an EMA of the student's weights. 55

6.3 Visual comparison of both baseline and C-MIL regarding their pathology localization predictions (test set images). The white bounding box represent the annotated ground-truth and the heatmap colors represent the patch score intensity for the specified pathology. Images and annotations from the ChestX-ray14 dataset [63]. 60

LIST OF TABLES

3.1 Reported results of four approaches for radiological finding localization on the ChestX-ray14 dataset [63]. Missing values (-) were not reported on the original papers. 33

3.2 Summary of the contributions of the related work and a comparison to our proposal. 34

4.1 Positive label frequency (in number of radiographs) in the training and test split for each dataset. 38

4.2 Test AUC for the 8 radiological findings common to the four datasets. Best results for each test set are in bold. 38

5.1 Error rate on CIFAR-10 from four different semi-supervised learning methods and a supervised baseline under different amounts of labeled data during training. Values reported on the original papers. CIFAR-10 consists of 50000 training samples and 10000 test samples. 41

5.2 Supervised baselines using all available training data of ChestX-ray14. 46

5.3 Average AUCs of our proposed approaches and baselines using different amounts of labeled samples on ChestX-ray14. 48

5.4 AUC results for the implemented methods using 25 labels per class during training. 48

5.5 AUC results for the implemented methods using 100 labels per class during training. 49

5.6 AUC results for the implemented methods using 400 labels per class during training. 49

5.7 Average AUC of our best approach with two previous approaches for semi-supervised classification in ChestX-ray14. 50

6.1 Comparison of IoR accuracy in all 8 pathologies of the re-implemented baseline and the metrics reported by Li et al. [33], in which we based our baseline, and Wang et al. [63], which introduced the dataset. The "Avg" column presents the value of a macro-average of the accuracy results on the 8 pathologies. 53

6.2 Frequency of bounding boxes for each pathology in each data fold. 56

6.3 Pathology localization accuracy based on the Intersection over the detected Region (IoR), where $T(IoR) = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. The reported methods are our re-implementation of the method from Li et al. [33] (baseline) and our semi-supervised proposed approach `C-MIL` with ResNet-50 and DenseNet-121 as backbones. 58

6.4 Point localization accuracy of each pathology. The reported methods are our re-implementation of the method from Li et al. [33] (baseline) and our semi-supervised proposed approach C-MIL with ResNet-50 and DenseNet-121 as backbones. 58

6.5 Classification results (AUC) of our implementations of the multiple instance models on the ChestX-ray14 test set with the classification model designed by Wang et al. [63] and the CheXNet [49] network. Results from related work as reported in the original papers. 59

6.6 Ablation study to assess the impact of keeping an exponential moving average (EMA) model as teacher, compared to using a single model to compute consistency. Reported values are point accuracy performance for each pathology. The reported methods are the original C-MIL and C-MIL without the EMA model. Experiments include both ResNet-50 and DenseNet-121 as backbones. 61

6.7 Ablation study to include the use of CheXpert data [22] in the computation of the consistency loss. Reported values are point accuracy performance for each pathology. We performed this ablation with C-MIL without an EMA model and using only DenseNet-121 as backbone. 62

LIST OF SYMBOLS

x_i – A labeled sample's feature vector 26

y_i – Label vector of the input sample x_i 26

s – Size of input vector x 26

n – Total number of instances 26

$\mathcal{X}$ – An m-dimensional feature space 26

C – Number of classes 26

$\mathcal{Y}$ – A C-dimensional label space 26

$f(\cdot)$ – A real function that models $\mathcal{X} \rightarrow \mathcal{Y}$ 26

$m(\cdot)$ – An approximate function that models $\mathcal{X} \rightarrow \mathcal{Y}$ 26

$\mathbb{L}$ – Set of labeled samples 26

$\mathbb{U}$ – Set of unlabeled samples 26

u_i – The feature vector of an unlabeled sample 26

X_i – A bag of samples which contain the same bag label 26

Y_i – Label of the bag X_i 26

L – Number of layers on the neural network 27

$l^{(i)}$ – A layer on the neural network 27

$\Theta^{(i)}$ – The weight matrix connecting layers $l^{(i)}$ e $l^{(i+1)}$ 27

$a^{(i)}$ – Activation values on layer $l^{(i)}$ 27

$z^{(i)}$ – Output values on layer $l^{(i)}$ before activation function 27

$\sigma(\cdot)$ – Logistic sigmoid activation function 27

$g(\cdot)$ – A generic activation function 27

I – Image byte matrix 28

h – Number of rows in the matrix 28

w – Number of columns in the matrix 28

$\phi(\cdot)$ – A data augmentation policy 28

$T(\cdot)$ – A threshold 30

k – Number of folds in cross validation 30

$\mathbb{1}(.)$ – Indicator function, returns 1 if the condition is true and 0 otherwise 30

Θ^S – Student model's weights 42

Θ^T – Teacher model's weights 42

e – Traning step 42

ρ – Hyperparameter controlling the exponential moving average decay 42

$\mathcal{L}_{comb}$ – Combined loss 42

$\mathcal{L}_{task}$ – Task loss 42

$\mathcal{L}_{cons}$ – Consistency loss 42

λ_{con} – Hyperparameter weighting the consistency loss 42

$m_s(\cdot)$ – Student model 42

$m_t(\cdot)$ – Teacher model 42

$\phi_s(\cdot)$ – Student augmentation policy 42

$\phi_t(\cdot)$ – Teacher augmentation policy 42

Q – Number of augmentations 44

$\hat{y}$ – Pseudo-label 44

$\tilde{x}$ – Mixed up labeled inputs 44

$\tilde{y}$ – Mixed up labels 44

$\tilde{u}$ – Mixed up unlabeled inputs 44

$\tilde{\hat{y}}$ – Mixed up pseudo-labels 44

τ – Hyperparameter controlling sharpen temperature 44

γ – A random value sampled from a $\beta(\alpha, \alpha)$ distribution on MixUp 44

Φ – A strong augmentation policy 45

c – One specific class 53

λ_{loc} – Hyperparameter weighting the localization loss 53

$\delta(\cdot)$ – Multiple instance pooling function to convert patch scores into class scores 53

y_{loc} – Localization labels 53

y_{cls} – Classification labels 53

$\mathcal{L}_{loc}$ – Localization loss 53

$\mathcal{L}_{cls}$ – Classification loss 53

p_i^c – Patch score for patch number i of class c 53

$\mathcal{M}$ – Set of $P \times P$ patches of an image 53

P – Total number of patches 53

$\mathcal{L}_{pcons}$ – Patch consistency loss 54

CONTENTS

1 INTRODUCTION . . . 21

1.1 RADIOGRAPHS . . . 21

1.2 AUTOMATED DIAGNOSIS . . . 21

1.3 RESEARCH PROBLEM . . . 22

1.3.1 RESEARCH QUESTION AND HYPOTHESIS . . . 23

1.3.2 GOALS . . . 23

2 BACKGROUND . . . 25

2.1 MACHINE LEARNING . . . 25

2.1.1 SUPERVISED LEARNING . . . 26

2.1.2 SEMI-SUPERVISED LEARNING . . . 26

2.1.3 MULTIPLE INSTANCE LEARNING . . . 26

2.2 DEEP LEARNING . . . 27

2.2.1 ARTIFICIAL NEURAL NETWORKS . . . 27

2.2.2 CONVOLUTIONAL NEURAL NETWORKS . . . 28

2.2.3 REGULARIZATION . . . 28

2.3 OBJECT DETECTION . . . 29

2.3.1 LOCALIZATION AS REGRESSION . . . 29

2.3.2 ACTIVATION MAPS . . . 29

2.3.3 ONE-STAGE APPROACHES . . . 30

2.3.4 TWO-STAGE APPROACHES . . . 30

2.4 EVALUATION . . . 30

2.4.1 METRICS . . . 30

2.4.2 CROSS-VALIDATION . . . 31

3 RELATED WORK . . . 33

4 DOMAIN SHIFT ANALYSIS . . . 35

4.1 CONTEXT . . . 35

4.2 CHEST RADIOGRAPHS DATASETS . . . 36

4.3 EXPERIMENT DESIGN . . . 36

4.4 EXPERIMENT RESULTS . . . 38

4.5 DISCUSSION . . . 40

5 SEMI-SUPERVISED PATHOLOGY CLASSIFICATION . . . 41
5.1 MOTIVATION . . . 41
5.2 SEMI-SUPERVISED LEARNING METHODS . . . 42
5.2.1 MEAN TEACHER . . . 42
5.2.2 UNSUPERVISED DATA AUGMENTATION . . . 43
5.2.3 MIXMATCH . . . 44
5.2.4 FIXMATCH . . . 45
5.3 EXPERIMENTAL DESIGN . . . 45
5.3.1 METHODS . . . 46
5.3.2 HYPERPARAMETER SEARCH . . . 47
5.4 RESULTS . . . 47
5.4.1 COMPARISON WITH PREVIOUS APPROACHES . . . 49
5.5 DISCUSSION . . . 50

6 SEMI-SUPERVISED PATHOLOGY LOCALIZATION . . . 51
6.1 METHODOLOGY . . . 52
6.1.1 DATA DESCRIPTION . . . 52
6.1.2 BASELINE . . . 52
6.1.3 PROPOSED APPROACH . . . 54
6.1.4 EXPERIMENTAL SETTINGS . . . 56
6.2 RESULTS . . . 57
6.2.1 QUANTITATIVE ANALYSIS . . . 57
6.2.2 QUALITATIVE ANALYSIS . . . 59
6.2.3 ABLATION STUDY . . . 61

7 CONCLUSIONS . . . 63
7.1 LIMITATIONS . . . 64

REFERENCES . . . 65

1. INTRODUCTION

Lung cancer is the first cause of cancer death in several countries [2], affecting both developed and emerging nations. Global 5-year survival rates vary between 10% and 20% [2]. The lack of effective early-detection methods is one of the main reasons for its poor prognosis [20]. Lung cancer signs are mostly identified through imaging exams, which are examined by the doctor specialist on medical imaging analysis, the radiologist.

Since the prognosis is better in earlier stages [41], missing to diagnose lung cancer in imaging exams is a great concern among radiologists. In 90% of the times, the misdiagnosis occurs on chest radiographs, mostly due to observer error [13]. There are stated cases [57] in which the radiography presented early-stage cancer signs that were overlooked when the cancer was still resectable. One way to overcome this issue is by automated medical image analysis methods, which might improve diagnostic accuracy and early-detection rates of lung cancer, leading to a better prognosis [41].

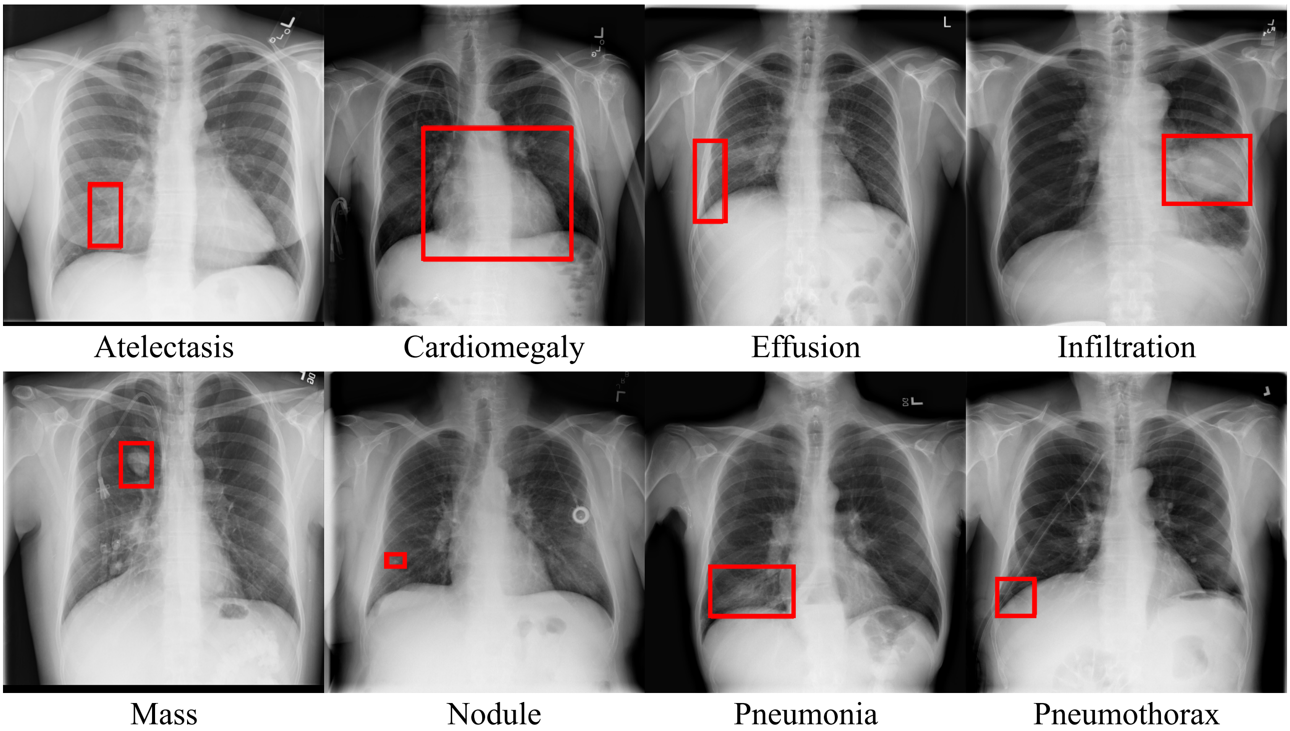
1.1 Radiographs

Radiography is a common exam to diagnose chest conditions since it is a low-cost, fast, and widely available imaging modality [14]. A Computed Tomography is an improvement over X-rays, providing more rich information to the radiologist, though exposing the patient to about 350 times more ionizing radiation [38], besides being costly, and of lower availability, especially considering third-world public health systems and low-income regions. Therefore, extracting as much information as possible from the radiography is vital.

Abnormalities identified on radiographs are called radiological findings. The radiologist reports the identified radiological findings on a text-based radiological report. In a chest radiograph, the radiological findings manifest as areas of high density, which appear on the radiograph as lighter shades, or as areas of low density, which appear as darker shades. The reported findings usually indicate a known pathology or condition. For instance, the appearance of lung lesions, consolidations, or atelectases, might indicate lung cancer [12]. We exemplify some chest radiological findings in Figure 1.1.

1.2 Automated Diagnosis

With the digitization of radiology, computer-aided diagnosis systems can be integrated into the radiological practice workflow, providing support via automated diagnosis tools. The development of automated diagnosis methods involves knowledge from software development, digital image processing, and machine learning. These consist of the main areas that form


Atelectasis Cardiomegaly Effusion Infiltration

Mass Nodule Pneumonia Pneumothorax

Figure 1.1: Chest radiographs annotated with 8 different radiological findings. Images and annotations from the ChestX-ray14 dataset [63].

computer vision, an area of science that seeks to provide machines the ability to describe and interpret digital images automatically.

Automated diagnosis tools might deal with classic computer vision problems, such as image classification, object detection, and segmentation, which are usually solved by image feature extraction and classification algorithms. Some of the methods used for medical image classification are decision trees, linear classifiers, and artificial neural networks [1]. Convolutional neural networks and other deep learning methods are becoming the method of choice for most medical imaging applications in recent years [35], mostly due to its high performance in image classification when a large amount of data is available for training, achieving radiologist-level performance on some tasks such as pneumonia detection [49].

1.3 Research Problem

To train deep learning models in a supervised fashion, we need a significant amount of training data. However, as it happens in most medical imaging scenarios [35], there is a lack of available annotated data. Although public datasets provide over 700, 000 chest radiographs labeled with radiological findings [63, 22, 24], only a small amount of those (880) are annotated with the findings localization. Classification labels are becoming increasingly easier to obtain since we can automatically extract them using natural language processing algorithms on radiological reports. In contrast, the localization label needs to be manually annotated by an expert, which is an expensive and time-consuming task. Our problem thus consists of training deep learning models to locate pathology patterns on radiographs using limited localization supervision but abundant samples with classification labels (presence or absence of radiological findings).

In this dissertation, we propose to extend a deep neural architecture for pathology localization on chest radiographs trained with limited annotated data to better use the available non-annotated samples through semi-supervised learning methods. Multiple instance learning is a paradigm that learns from two types of labels: a general bag label that is known, and a more specific and scarce instance label related to the bag label. Since we have a large number of samples labeled regarding pathology presence, we can replicate a multiple instance learning scenario in which the bag labels are pathology presence, and the instance labels are pathology localization. The instance labels are mostly unknown, as the available public datasets contain only 984 bounding box annotations. The bag labels, however, are widely available ($\approx 700,000 samples$). We intend to adapt a state-of-the-art semi-supervised approach to a multiple instance learning framework by assuming that each bag X_k comprises instance-labeled samples $\mathbb{L}_k$ and bag-labeled samples $\mathbb{U}_k$. Then, to predict instance-level labels, we can train a model that learns both from instance-labeled samples and is leveraged by the remaining unlabeled samples of each bag X_k.

1.3.1 Research Question and Hypothesis

Can state-of-the-art semi-supervised methods be adapted to a multiple instance learning scenario to perform pathology localization on chest radiographs? We hypothesize that the proposed approach will improve the performance of chest radiograph pathology localization methods since there is a limited amount of annotated data available for supervised training. We also expect that the proposed approach will be extendable to similar problems, such as localization of pathologies in other medical imaging modalities and general object detection.

1.3.2 Goals

Our general goal is to improve automated pathology localization on chest radiographs using limited localization supervision by leveraging state-of-the-art semi-supervised learning methods to perform multiple instance learning. Our three specific goals are:

1. To explore the available public datasets annotated for chest radiograph classification, and how they relate to each other in terms of generalizability and representativeness. The experiments and findings on this matter are described in Chapter 4.

2. To identify the state-of-the-art approaches regarding pathology localization on chest radiographs and semi-supervised learning. Following, to benchmark the identified semi-supervised learning methods on a chest radiograph classification scenario. The experiments and findings on this matter are described in Chapters 3 and 5.

3. Finally, to investigate how the top semi-supervised learning method can be adapted for the task of localization and to a multiple instance learning scenario. Then, implement, evaluate and report the results of the proposed approach. The experiments and findings on this matter are described in Chapter 6.

2. BACKGROUND

2.1 Machine Learning

Machine learning is a subfield of Artificial Intelligence focused in researching how to develop computer programs that automatically improve based on experience [40]. Machine learning algorithms are mathematical models that learn to represent a certain distribution based on available data. A dataset is composed of instances; in supervised learning, each instance has a set of attributes or features, which are the input of the model, and an associated label, which is the output. To validate machine learning models, usually, we split the dataset into train, validation, and test sets. We use the train set to fit the model so the approximated function can map the feature distribution to the associated label; the validation set is used to tune hyperparameters based on the model's performance; and the test set is used as proxy to unseen data, so we can check the model's capability to generalize to new input data.

Concerning learning paradigms, we traditionally divide machine learning algorithms into three categories: supervised, unsupervised and reinforcement learning. In supervised learning, each training sample has an associated known label, and the goal is to infer this label on unknown inputs. In unsupervised learning, the training samples are unlabeled, and the rely on the data distribution alone to infer the underlying structure of the data. Usually, a large amount of training samples increase the model performance and generalization ability since the learned function is a better approximation of the real distribution we aim to represent. In scenarios where there is a lack of labeled data, which is the case in medical imaging, to achieve better results we must exploit approaches that go beyond traditional supervised learning, such as semi-supervised learning, and multiple instance learning [9]. Figure 2.1 illustrates the three different learning paradigms.

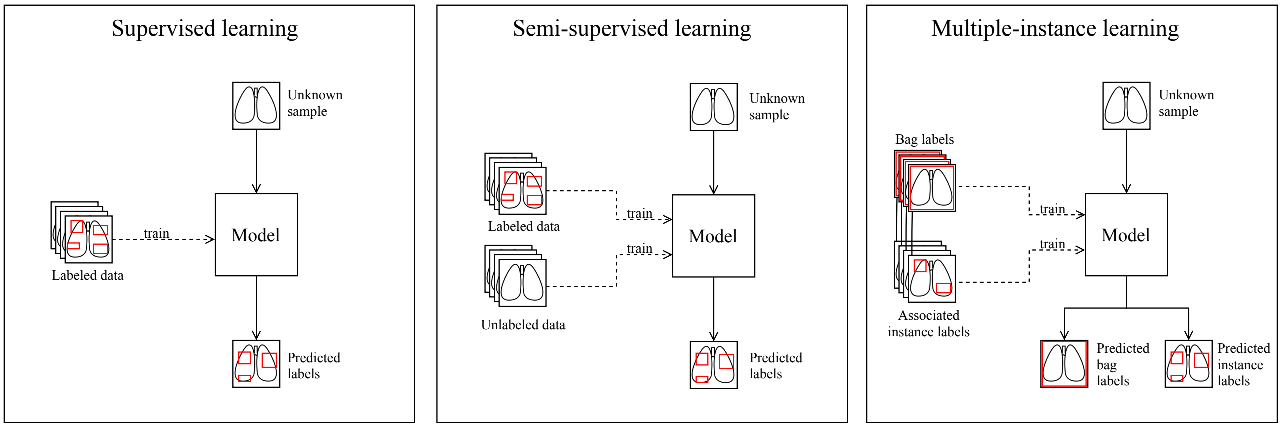


Figure 2.1: Chest pathology localization using three different learning paradigms: supervised, semi-supervised, and multiple instance learning.

2.1.1 Supervised Learning

In the supervised learning approach, the model learns based on known annotated examples. As the system is presented with input and output variables in the training set, it seeks to create a model that represents this data distribution. Then, this model is extrapolated to infer the output variable of an unseen input sample. Formally, the training data comprises samples $\{x_1, x_2 \ldots, x_n\}, x_i \in \mathcal{X}$ along with their corresponding labels $\{y_1, y_2 \ldots, y_n\}, y_i \in \mathcal{Y}$. We use the training set in order to model function $f(x) : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the s-dimensional feature space and $\mathcal{Y}$ is the C-dimensional label space. Through function $f(x)$ (sometimes referred to as $m(x)$) we can predict the labels of previously unseen samples.

2.1.2 Semi-Supervised Learning

Semi-supervised learning is a learning paradigm intersecting supervised and unsupervised learning. In this scenario, besides labeled samples $\mathbb{L} = \{x_1, x_2 \ldots, x_n\}$ we also have unlabeled samples $\mathbb{U} = \{u_1, u_2 \ldots, u_n\}$ that are also within the feature space $\mathcal{X}$ but whose output labels within label space $\mathcal{Y}$ are unknown [67]. We can use $\mathbb{U}$ in the training set alongside $\mathbb{L}$ in order to improve the modeling of the function $m(x) : \mathcal{X} \to \mathcal{Y}$. Intuitively, the unlabeled samples provide important clues on the data distribution based on sample similarity and they help to add robustness to the model by exploring this distribution [48].

Semi-supervised learning methods are mainly based on three assumptions: smoothness, low-density, and manifold [61]. The smoothness assumption states that if two samples x_1 and x_2 are close in the feature space, their labels y_1 and y_2 are probably the same. The low-density assumption states that the decision boundary of a classifier probably does not pass through high-density areas of the feature space. Finally, the manifold assumption says that samples located on the same low-dimensional feature space manifold probably have the same labels.

2.1.3 Multiple Instance Learning

In a multiple instance learning scenario, the training set consists of bags of samples $\{X_1, X_2 \ldots, X_c\}$ along with bag labels $\{Y_1, Y_2 \ldots, Y_c\}$. The samples from X_i $\{x_{i1}, x_{i2} \ldots, x_{in}\}$ have associated labels $\{y_{i1}, y_{i2} \ldots, y_{in}\}$ that are somehow related to the bag label Y_i [4]. For instance, in a chest radiograph scenario we can assume that bag labels Y_i indicate pathology presence and sample labels y_{ij} indicate pathology location. Bag X_i contains all images positive for a particular pathology. Considering there are 14 pathologies in the dataset ($c = 14$), there are 14 different sets of bags X_i and 14 possible bag labels Y_i. In our dataset [63], all samples have bag labels Y_i, but only a few of them have a known label y_{ij}. Therefore, we can exploit this scenario as a special case of semi-supervised learning.

2.2 Deep Learning

One important aspect of machine learning projects is defining how to represent previous experience as information within a dataset. Conventional machine learning methods usually require a domain expert to define the features that must be extracted to represent the desired target in an initial step of the project called feature engineering. With deep learning methods, models learn not only the underlying function that maps data to desired output, but also the data representation with multiple levels of abstraction [29]. Deep learning models have processing layers that automatically extract features from raw data, advancing the state-of-the-art in many data processing tasks such as image recognition [27], speech recognition [19], genomic data analysis [32], and machine translation [62].

2.2.1 Artificial Neural Networks

Artificial neural networks are an example of biologically-inspired machine learning method. McCulloch and Pitts [39] first introduced the model of an artificial neuron in 1943. Current neural networks have multiple layers of artificial neurons and a large number of connections between them, being able to model any computable function. In fully-connected feedforward networks with L layers, each layer $l^{(i)}$ has a weight matrix $\Theta^{(i)}$ which is multiplied with an input vector $a^{(i)}$ resulting in $z^{(i+1)}$. Then, its activation values are computed through an *activation function* $g(z^{(i+1)})$ outputting $a^{(i+1)}$, which is used as input to the following layer or as the model's result y for the case case $i + 1 = L$.

Activation functions allow the existence of non-linearities in the model, which ultimately gives the neural networks expressiveness to approximate complex non-linear functions. A common activation function is the logistic sigmoid (Equation 2.1). The sigmoid output values are in the range $[0, 1]$, saturating to 0 when the input becomes very negative and to 1 when it becomes very positive. Its output value can represent a firing rate of a neuron or a class probability.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.1}$$

Current deep learning architectures mostly implement the Rectified Linear Unit (ReLU) activation function (Equation 2.2) or its variations, mostly because it does not involve computationally expensive operations and also accelerates the optimization convergence time [27].

$$ReLU(z) = max(0, z) \tag{2.2}$$

The learning process of artificial neural networks is done by optimizing the weights to minimize a loss function computed over training data. The weights are randomly initialized and iteratively updated in the negative gradient direction of the loss function with a pre-defined step size called learning rate. For classification problems, a common loss function is the binary cross-entropy loss (Equation 2.3). It is computed over all n samples, for samples with positive ground-truth labels ($y_i = 1$), it adds the log of the model output probability $m(x_i)$, and for samples with negative labels ($y_i = 0$), it adds $\log(1 - m(x_i))$, returning a high output for predictions far from its true value, and a low output for predictions close to the ground-truth.

$$BCE(m(x_i), y_i) = -\frac{1}{n}\sum_{i=0}^{n} y_i \log\left(m(x_i)\right) + (1 - y_i)\log\left(1 - m(x_i)\right) \tag{2.3}$$

2.2.2 Convolutional Neural Networks

A digital image is represented through a matrix of bytes I. We call each element of this matrix a *pixel*. In a grayscale image, the value in each pixel represents its gray-level intensity. The most common representation approach is to assume that the lowest intensity value represents the black color, and the highest one represents the white color. The image resolution is the size of matrix I; let h be the number of rows and w the number of columns, the image has a total of $h \times w$ pixels.

Convolutional neural networks are one of the most successful deep learning methods [29], achieving state-of-the-art results in several medical image analysis tasks [35]. Instead of performing matrix multiplication between inputs and weights, convolutional neural networks exploit spatially-local correlations by using a mathematical operation called *convolution*, which leads to local connections between neurons of adjacent layers and shared weights [30]. Each convolutional layer has matrices of weights, also called filters or kernels, that are convolved with the inputs. Each resulting matrix is called a feature map. Convolutional layers' filters are optimized during the training process to learn the best features to represent the desired output.

2.2.3 Regularization

Overfitting is a common problem with machine learning algorithms. It happens when models perform well on training data but do not generalize well to unseen inputs like the validation or test sets. Regularization strategies are modifications made in a learning algorithm in order to reduce its generalization error [17]. One regularization strategy is adding extra terms in the objective (loss) function in order to apply constraints or penalties over the parameter values (magnitude). Weight decay regularization adds the sum of all weights to the loss function, making the optimizer deal with the trade-off of fitting the training data and keeping the weights with low values, resulting in a smoother (potentially less-complex) decision boundary.

Dataset augmentation is another popular regularization technique that has been proved highly effective on computer vision tasks [17]. It consists of creating fake data points by applying a transformation $\phi(\cdot)$ over training instances x_i and training the model with $\phi(x_i)$ and their original label y_i. When working with images, we can apply random image transformations like translation, rotation, scaling, and horizontal flips, or apply small random noises (jittering) by changing pixel values.

2.3 Object Detection

2.3.1 Localization as Regression

Considering a single-object scenario, localization can be seen as a merge of both classification and regression. Instead of outputting a class label y, a regression model can output four values indicating the bounding-box coordinate that contains the desired object. We can implement this approach by making an output layer with four parameters and training the model with the ground-truth bounding box using a standard regression loss (e.g., mean squared error).

A limitation of this approach is that it becomes impractical when we have a scenario with multiple objects that can appear multiple times within a single image. Since the number of objects vary, it is not possible to define a fixed output layer, making this approach only convenient when the problem has a fixed number of objects.

2.3.2 Activation maps

To provide localization on a classification framework, we can explore visual explanation methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) [56]. Grad-CAM generates heat maps using the gradients flowing from the final convolutional layer, which shows how significant each part of the input image is for the classification result. With this approach, we do not need the bounding box annotations, since we perform localization in a weakly-supervised manner by using only the classification labels.

One can use the highest activation region for a certain class and design a bounding box assuming it to be the object localization. We can implement it in a very straightforward way with only classification labels, as Grad-CAM does not use localization labels. However, it can be misleading since wrong classes can still produce high activation values on image regions, as shows [53], which states that these approaches mostly explain where the network is looking, even if the class used to generate the activation map is not correct.

2.3.3 One-stage Approaches

Currently, computer vision research has two main lines of state-of-the-art approaches on object detection: one-stage, and two-stage detectors. One-stage detectors compute objectness score, classification, and location regression in the same stage. Popular one-stage detectors are SDD [34] and YOLO [50]. For instance, YOLO detects scene objects by dividing the image into a grid and predicting for each cell a score of conditional class probability. Then, places a set of anchor boxes, which have their sizes learned based on the training set annotations, in each grid cell, and a final bounding box to maximize the object classification confidence suppressing similar boxes.

2.3.4 Two-stage Approaches

Two-stage detectors are mainly based on the original R-CNN approach [16], like Fast R-CNN [15], and Faster R-CNN [51]. These approaches first generate regions of interest (region proposals), and then compute the class score probabilities for each proposal, adjusting the bounding boxes using regression. Faster R-CNN [51] introduces a neural network to propose the regions of interest to perform the first stage, reducing the inference time of the detector. The region proposal network outputs image regions and preserves the top candidates based on objectness scores. It makes the region proposal also differentiable, being trained end-to-end with the object classifier and the bounding box regressor.

Usually, two-stage approaches perform better on object detection. However, one-stage detectors are easier to train and have lower inference time, being more suited to real-time applications. Usually, both approaches need to be trained with a large amount of annotated data to perform well.

2.4 Evaluation

2.4.1 Metrics

The evaluation of localization approaches relies on the intersection between the prediction $m(x_i)$ with the ground-truth y_i to decide whether a prediction is correct. The intersection over union (IoU) metric is shown in Equation 2.4. IoU is a ratio computed by dividing the area of overlap $|m(x_i) \cap y_i|$ with the area of union $|m(x_i) \cup y_i|$ of the two bounding boxes. An

IoU of 1.0 indicates that the predicted region is exactly the same as the ground-truth region. The Intersection over the detected region (IoR), which may also be called Intersection over the detected bounding box (IoBB), is defined in Equation 2.5. The IoR ratio is also a value between 0 and 1, and it is computed by dividing the area of overlap $|m(x_i) \cap y_i|$ with the area of the detected region or bounding box $|m(x_i)|$. An IoR of 1.0 means that the predicted region is completely inside the ground-truth region.

If a prediction's IoU or IoR with the ground-truth is above a given threshold $T(\cdot)$, the prediction is considered correct, and we can compute the model's accuracy by dividing the number of correct predictions with the total number of instances n, as shows Equation 2.6, considering each instance has one object.

$$\mathrm{IoU}(m(x_i), y_i) = \frac{|m(x_i) \cap y_i|}{|m(x_i) \cup y_i|} \tag{2.4}$$

$$\mathrm{IoR}(m(x_i), y_i) = \frac{|m(x_i) \cap y_i|}{|m(x_i)|} \tag{2.5}$$

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left(\mathrm{IoU} \left(m\left(x_i \right), y_i \right) \geqslant T(IoU) \right) \tag{2.6}$$

Another evaluation measure proposed to be used in weakly-supervised scenarios [66] is the point localization accuracy. This measure is computed for each class of the dataset, being a ratio of the hits and misses of the proposed heatmap. A hit is counted if the highest scoring pixel is inside of the bounding box, otherwise that prediction was a miss. The point localization accuracy value for that class is then computed, as shown in Equation 2.7, by dividing the total number of hits by the sum of hits and misses.

$$\text{Point localization accuracy} = \frac{\#Hits}{\#Hits + \#Misses} \tag{2.7}$$

2.4.2 Cross-validation

When limited labeled data is available, which is the problem we are facing, a common approach to validate model performance is to perform k-fold cross-validation. The general procedure is to split the data into k different groups and then train and test the model k times, each time with a different group as test set. Finally, we compute the k tests average accuracy (or simply add the total amount of true positives, false positives, true negatives and false negatives).

3. RELATED WORK

Current methods that exceed average radiologist performance [5, 33, 49] are multi-label classification approaches using convolutional neural networks trained with a large amount of data to output diagnosis labels. Network architectures that have the best reported performance on chest radiographs are ResNet-50 and DenseNet-121 [5]. Though performing radiologist-level classification, their reported localization performance is still unsatisfactory [63]. Table 3.1 summarizes the top accuracies with different IoU thresholds from four different approaches of pathology localization on the ChestX-ray14 dataset [63].

Table 3.1: Reported results of four approaches for radiological finding localization on the ChestX-ray14 dataset [63]. Missing values (-) were not reported on the original papers.

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.1	Wang et al. [63]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.57
	Li et al. [33]	0.71	0.98	**0.87**	0.92	**0.71**	0.40	0.60	0.63	0.73
	Liu et al. [36]	-	-	-	-	-	-	-	-	-
	Rozenberg et al. [52]	**0.78**	**1.00**	0.84	**0.95**	**0.71**	**0.44**	**0.92**	**0.73**	**0.80**
0.3	Wang et al. [63]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	Li et al. [33]	0.36	**0.94**	0.56	0.66	0.45	**0.17**	0.39	**0.44**	0.49
	Liu et al. [36]	**0.53**	0.88	**0.57**	**0.73**	**0.48**	0.10	**0.49**	0.40	**0.53**
	Rozenberg et al. [52]	-	-	-	-	-	-	-	-	-
0.5	Wang et al. [63]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	Li et al. [33]	0.14	**0.84**	0.22	0.30	0.22	**0.07**	0.17	0.19	0.27
	Liu et al. [36]	**0.32**	0.78	**0.40**	**0.61**	**0.33**	0.05	**0.37**	**0.23**	**0.39**
	Rozenberg et al. [52]	-	-	-	-	-	-	-	-	-
0.7	Wang et al. [63]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	Li et al. [33]	0.04	0.52	0.07	0.09	0.11	0.01	0.05	0.05	0.12
	Liu et al. [36]	**0.18**	**0.70**	**0.28**	**0.41**	**0.27**	**0.04**	**0.25**	**0.10**	**0.29**
	Rozenberg et al. [52]	-	-	-	-	-	-	-	-	-

Wang et al. [63] introduce the ChestX-ray14 dataset and propose a method to predict bounding boxes based on activation heat maps of a convolutional neural network trained for classification. The top activation values on the heatmap are considered to be the pathology location. The use of only activation heat maps for localization is not reliable and might be misleading [53]. The predictions are compared with manually-labeled samples with different threshold values for considering a correct prediction. With a threshold $T(IoU) = 0.1$, it achieves a mean accuracy of 0.57, the lowest one being 0.14 for nodule location and the highest one 0.94 for cardiomegaly. Using a threshold $T(IoU) = 0.7$, the mean accuracy decreases to 0.01.

Li et al. [33] improve the work in [63] by exploiting bounding-box supervision. To locate the pathologies, they handle images as groups of patches, treating each patch as a classification target. To train a model with limited annotated data, they propose a multiple instance learning framework that assumes during training that if an image is labeled positive for a pathology, then there is at least one positive patch on the image. They achieve a mean accuracy of 0.73 with $T(IoU) = 0.1$, ranging from 0.40 for nodule location to 0.98 for cardiomegaly location. When testing with $T(IoU) = 0.7$ the mean accuracy falls down to 0.12.

Liu et al. [36] propose a novel approach called Contrast-Induced Attention Network (CIA-Net), which is trained on a multiple instance learning framework to perform pathology localization. The images are geometrically aligned via a learnable alignment module to maintain radiographs' structural consistency, and an attention branch generates attention for every class based on paired positive and negative images for that finding. Their reported results overcome [33] in most findings and mean accuracy, achieving a mean of 0.29 using $T(IoU) = 0.7$ with 80% of the annotated images and 50% of the unannotated images, showing an evident improvement over previous studies at a higher $T(IoU)$.

Rozenberg et al. [52] perform localization with limited annotation with two contributions. A novel loss function to combine labeled and weakly labeled data, and the incorporation of conditional random field layers and anti-aliasing filters on the network architecture to account for patch dependency and shift-invariance. The reported results outperform [33] with $T(IoU) = 0.1$, achieving a mean accuracy of 0.80, but the authors do not report accuracies on other IoU thresholds, so it is not fully comparable to Li et al. [33] and Liu et al. [36] approaches.

Table 3.2 summarizes the related work and contextualize our proposed approach among the previous studies. We propose to extend previous work by introducing a new paradigm to the multiple instance localization scenario. By introducing semi-supervised learning we can make more use of the classification labeled samples and extend the model to learn from unlabeled samples from other sources. In the next chapters, we will discuss our work and present our experimental analyses for validating our research hypothesis.

Table 3.2: Summary of the contributions of the related work and a comparison to our proposal.

Method	Paradigm	Contribution
Wang et al. [63]	Weakly-supervised learning	Introduces the ChestX-ray14 dataset and proposes a baseline solution on localization by using activation heatmaps of a model trained only on classification labels.
Li et al. [33]	Multiple instance learning	Introduces a multiple instance architecture for the localization problem that uses both classification and localization labels during training.
Liu et al. [36]	Multiple instance learning	Extends the architecture proposed by [33] adding pathology attention maps and an alignment module to standardize the positioning of input samples.
Rozenberg et al. [52]	Multiple instance learning	Extends the architecture proposed [33] adding anti-aliasing filters and conditional random field layers and proposes a different loss function.
C-MIL (our approach)	Semi-supervised multiple instance learning	Extends the architecture proposed by [33] introducing semi-supervised learning mechanisms that make use of unlabeled data during training.

4. DOMAIN SHIFT ANALYSIS

In this chapter, we analyze the available chest radiograph classification datasets in order to evaluate which of the datasets are the most representative of the others and to assess the generalization ability of a deep learning model outside its domain of training in a medical imaging scenario. The experiments of this chapter resulted in the work "*Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification*" published in MICCAI's International Workshop on Thoracic Image Analysis [45].

4.1 Context

Considering chest radiographs, deep learning approaches are usually developed within a multi-label classification scenario, predicting radiological findings to assist physicians with the diagnosis process. Recent work in the field achieved near radiologist-level accuracy at identifying radiological findings using convolutional neural networks [49].

One assumption underlying deep learning models is that training and test data are independent and identically distributed ($i.i.d$). This assumption often does not hold when data come from different settings. This is a common case for medical imaging, a scenario in which image acquisition protocols and machines may vary among diagnostic centers, being defined by the quality of the machine, its parameters, and the acquisition protocol. Another aspect of medical imaging is the epidemiological variation among different populations, which may change the label distribution in different datasets. This difference in data distribution from the same task is called *domain shift*. The domain from where training data is sampled is the source domain, with distribution $p(X_s)$, and the one where the model is applied to is the target domain, with distribution $p(X_t)$. When $p(X_s) \sim p(X_t)$, it means that the model will most likely handle test data the same way as it did in training. As $p(X_s)$ diverges from $p(X_t)$, trained models tend to yield poor results, failing to effectively handle the input data [60]. Figure 4.1 shows a sample image labeled for "*Consolidation*" in four different chest radiograph datasets.

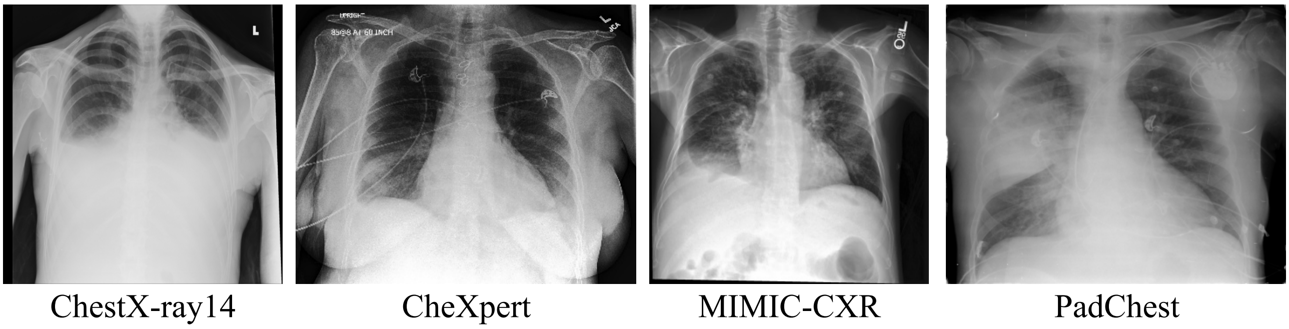

ChestX-ray14 CheXpert MIMIC-CXR PadChest

Figure 4.1: Example of a chest radiograph (positive for consolidation) randomly sampled from each of the four analyzed datasets: ChestX-ray14, CheXpert, MIMIC-CXR, and PadChest.

In the analysis of this chapter, we evaluate how well models trained on a hospital-scale database generalize to unseen data from other hospitals or diagnostic centers by analyzing the degree of domain shift among four large datasets of chest radiographs. We train a state-of-the-art convolutional neural network for multi-label classification on each of the four datasets and evaluate each model's performance in predicting labels on the other three datasets.

4.2 Chest Radiographs Datasets

Four large datasets of chest radiographs are available to this date. ChestX-ray14 [63] from the National Institute of Health contains 112, 120 frontal-view chest radiographs from 32, 717 different patients labeled with 14 radiological findings and with 984 manually annotated bounding boxes on 880 different images for 8 of the 14 findings. CheXpert [22] from the Stanford Hospital contains 224, 316 frontal and lateral chest radiographs of 65, 240 patients. MIMIC-CXR [24] from Massachusetts Institute of Technology presents 371, 920 chest X-rays associated with 227, 943 imaging studies from 65, 079 patients. Both CheXpert and MIMIC-CXR are labeled with the same 14 observations. PadChest [8] contains 160, 000 images obtained from 67, 000 patients of San Juan Hospital in Spain. The radiographs are labeled with 174 different findings. Most labels from all four datasets are automatically extracted using natural language processing algorithms on the radiological reports.

We show the pixel intensity distribution of each dataset in Figure 4.2. We see a spike at low intensities (especially 0) for most centers. However, the distribution for higher intensities is somewhat different for every center, which may imply in a decrease of the models' predictive performance, except for CheXpert and MIMIC-CXR, which show similar distributions. Figure 4.3 shows the average radiograph of each dataset (computed using 10, 000 random samples), in which we can see small differences in pixel intensity and that a common artifact appears on the top left corner of PadChest radiographs. Another difference that might cause domain shift is that PadChest labels are extracted from reports in Spanish, while the other three are extracted from reports in English.

4.3 Experiment design

We employ a multi-label classification approach reproducing CheXNet [49], which achieved state-of-the-art results in classification of multiple pathologies using a DenseNet121 convolutional neural network architecture [21]. The model is pre-trained on the ImageNet dataset, and the images are resized to 224 × 224 pixels and normalized using ImageNet mean and standard deviation. We train four models, one for each dataset, and subsequently evaluate our model at the other three. Each model is trained with the training set and evaluated on its own test set and the other three test sets. The four datasets have the same train, test, and

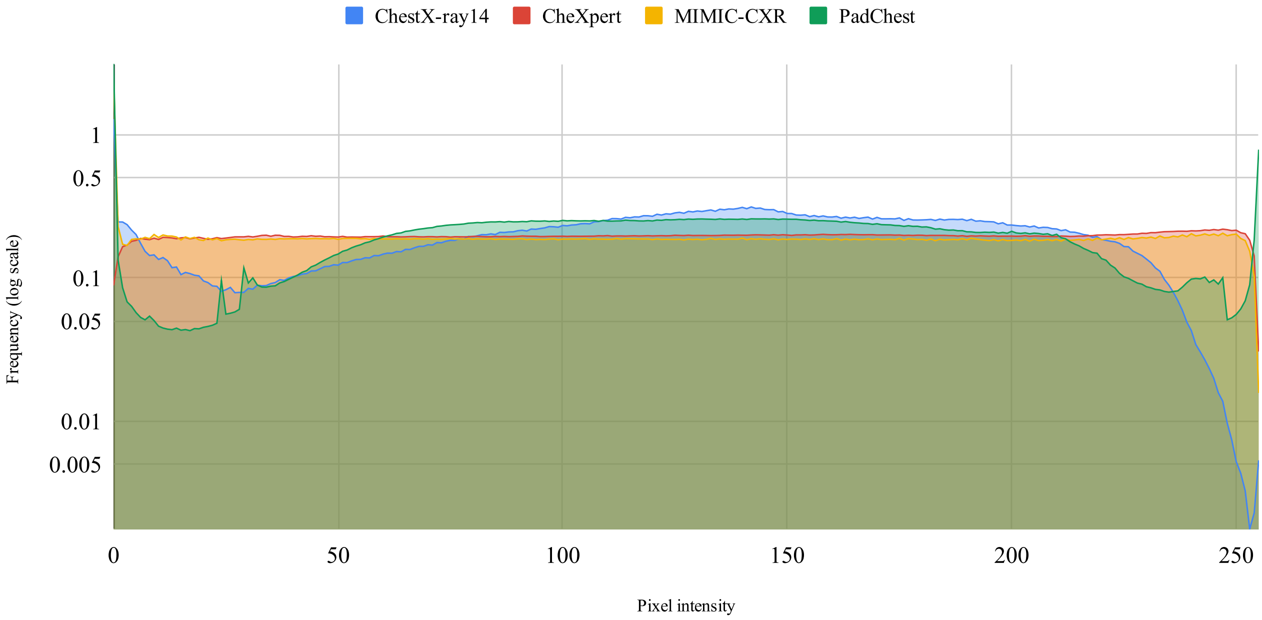


Figure 4.2: Dataset pixel intensity probability density function (PDF) of the four datasets.

validation sets across the experiments. For the ChestX-ray14 dataset, we use the original split, but since CheXpert and MIMIC-CXR test sets are not publicly available and PadChest does not have an original split, we randomly re-split their data, keeping ChestX-ray14 split ratio (70% train, 20% test, and 10% validation) and no patient overlap between the sets. Table 4.1 shows the frequency of the labels in each training and test split. As both CheXpert and MIMIC-CXR have labels for uncertainty, we assumed these labels as negatives (U-Zeros approach in [22]).

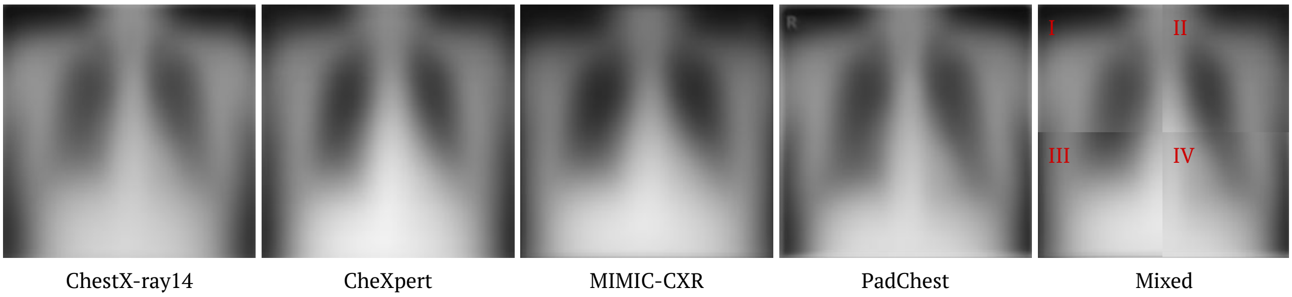


Figure 4.3: Average image of each of the four datasets. Last image contains 1/4 of each average image to better visualize the pixel intensity differences (I - ChestX-ray14, II - CheXpert, III - MIMIC-CXR, IV - PadChest).

One limitation we encountered is that the datasets have distinct sets of labels between each other. We fix this by training each model with all labels available, but reporting the results only on the common labels for all four (Atelectasis, Cardiomegaly, Consolidation, Edema, Lesion, Pneumonia, Pneumothorax, and No Finding). We create a "Lesion" label on ChestX-ray14 by joining the samples annotated as "Nodule" or "Mass". For PadChest, we joined labels that can fit into the 8 common findings, (i.e. "Atelectasis Basal", "Total Atelectasis", "Lobar Atelectasis", and "Round Atelectasis" were merged into "Atelectasis"). Another limitation is that ChestX-ray14 has only frontal X-rays. Therefore, we only use the frontal samples from the

Table 4.1: Positive label frequency (in number of radiographs) in the training and test split for each dataset.

	Atelectasis		Cardiomegaly		Consolidation		Edema		Lesion		Pneumonia		Pneumothorax		No Finding	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
ChestX-ray14	7,996	2,420	1,950	582	3,263	957	1,690	413	7,758	2,280	978	242	3,705	1,089	42,405	11,928
CheXpert	20,630	6,132	15,885	5,044	9,063	2,713	34,066	10,501	4,976	1,411	3,274	935	12,583	3,476	12,010	3,293
MIMIC-CXR	34,653	10,071	34,097	9,879	8,097	2,430	20,499	5,954	5,025	1,341	12,736	3,711	8,243	2,231	58,135	16,670
PadChest	1,841	574	3283	953	664	210	127	44	878	261	678	194	163	33	25,268	7,200

other three datasets, which means $191,229$ samples on CheXpert, $249,995$ on MIMIC-CXR, and $111,176$ on PadChest.

To evaluate domain shift, we use a standard performance metric in multi-label classification, the Area Under the Receiver Operating Characteristic curve (AUC), to report both individual radiological findings results and their average for an overall view of model performance. Both the true positive rate and the false positive rate are considered for computing the AUC. Higher AUC values indicate better performance.

4.4 Experiment results

We train the same neural network architecture with the same hyperparameters on each of the four datasets individually. When training and testing on ChestX-ray14, we achieve results similar to the ones reported by CheXnet [49], which exceeded radiologists' performance in detecting pneumonia. After training, we evaluate each of our models with images from the remaining three datasets.

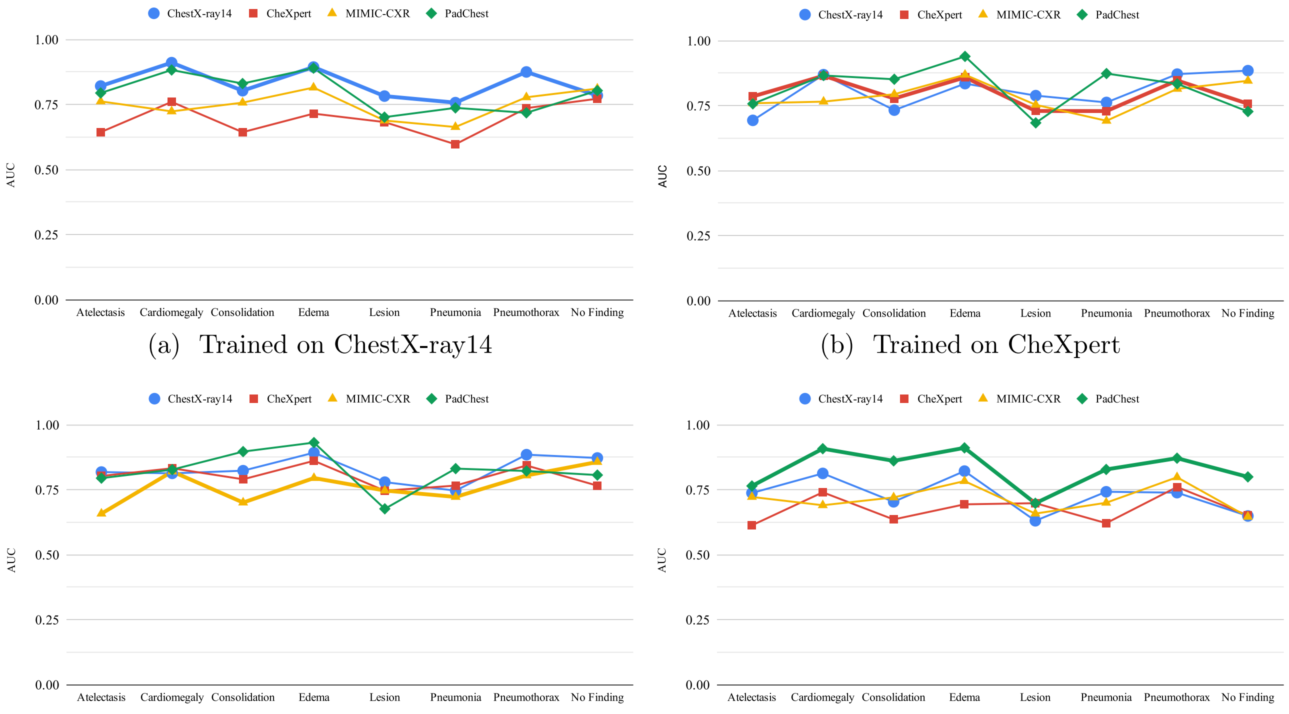
Table 4.2: Test AUC for the 8 radiological findings common to the four datasets. Best results for each test set are in bold.

Test set	Training set	Atelectasis	Cardiomegaly	Consolidation	Edema	Lesion	Pneumonia	Pneumothorax	No Finding	Mean
ChestX-ray14	**ChestX-ray14**	**0.8205**	**0.9104**	**0.8026**	**0.8935**	**0.7819**	0.7567	**0.8746**	**0.7842**	**0.8343**
	CheXpert	0.7850	0.8646	0.7771	0.8584	0.7291	0.7287	0.8464	0.7569	0.7933
	MIMIC-CXR	0.8024	0.8322	0.7898	0.8609	0.7457	**0.7656**	0.8429	0.7652	0.8006
	PadChest	0.7371	0.8124	0.7031	0.8213	0.6301	0.6487	0.7417	0.7384	0.7291
CheXpert	ChestX-ray14	0.6433	0.7596	0.6431	0.7145	0.6821	0.5967	0.7356	0.7717	0.6821
	CheXpert	**0.6930**	**0.8687**	**0.7323**	**0.8344**	**0.7882**	**0.7619**	**0.8709**	**0.8842**	**0.8042**
	MIMIC-CXR	0.6576	0.8197	0.7002	0.7946	0.7465	0.7219	0.8046	0.8564	0.7627
	PadChest	0.6127	0.7397	0.6352	0.6934	0.6978	0.6510	0.6209	0.7600	0.6764
MIMIC-CXR	ChestX-ray14	0.7616	0.7230	0.7567	0.8146	0.6880	0.6630	0.7773	0.8106	0.7406
	CheXpert	0.7587	0.7650	0.7936	0.8685	0.7527	0.6913	0.8142	0.8452	0.7861
	MIMIC-CXR	**0.8177**	**0.8126**	**0.8229**	**0.8922**	**0.7788**	**0.7461**	**0.8845**	**0.8718**	**0.8283**
	PadChest	0.7218	0.6899	0.7200	0.7828	0.6577	0.6454	0.6995	0.7976	0.7143
PadChest	ChestX-ray14	0.7938	0.8822	0.8300	0.8893	**0.7010**	0.7366	0.7176	0.8028	0.7929
	CheXpert	0.7566	0.8656	0.8511	**0.9390**	0.6833	0.7269	**0.8731**	0.8335	0.8161
	MIMIC-CXR	**0.7942**	0.8270	**0.8963**	0.9310	0.6761	**0.8060**	0.8308	0.8217	0.8229
	PadChest	0.7641	**0.9075**	0.8607	0.9107	0.6975	0.7990	0.8276	**0.8710**	**0.8298**

We summarize the results in Table 4.2. We can see that the best average result for each test set appears when the training set is from the same dataset. This shows that clinicians should expect a decrease in the reported performance of machine learning models when applying them in real-world scenarios. The decrease may vary according to the dataset distribution in which the model was trained on. For instance, running a model trained on MIMIC-CXR over

CheXpert's test set reduces the mean AUC in 0.04, while the model trained on ChestX-ray14 reduces it by 0.12. On MIMIC-CXR's test set, a model trained on CheXpert shows almost the same decrease in mean AUC (0.04), reducing the AUC in all of the findings. The model trained on ChestX-ray14 has the highest average AUC when testing on its own test set, but when testing in other datasets, it shows a significant performance drop, lowering CheXpert's mean AUC in 0.12, MIMIC-CXR's in 0.08 and PadChest in 0.04. Both the models trained on CheXpert and MIMIC-CXR mostly preserve the ChestX-ray14 baseline mean AUC, while the model trained on PadChest drops the average performance in 0.10. PadChest presented some variations on the best AUC for each disease, probably due to the smaller number of training instances. The models trained on CheXpert and MIMIC-CXR got very close results to PadChest's baseline.

Figure 4.4 shows the performance on the test set of the four trained models, represented as lines to better visualize AUC variations. The CheXpert (4.4b) and MIMIC-CXR (4.4c) models show smaller variations on the AUCs of the findings compared to their own test sets, presenting close lines, while PadChest (4.4d) and ChestX-ray14 (4.4a) shows the line of their own test set mostly on top and a drop in performance on the other test sets.



(a) Trained on ChestX-ray14

(b) Trained on CheXpert

(c) Trained on MIMIC-CXR

(d) Trained on PadChest

Figure 4.4: Performance of a model trained on ChestX-ray14 (a), CheXpert (b), MIMIC-CXR (c), and PadChest (d) on each of the four test sets.

Clear evidence of the impact of domain shift over model performance may be measured by how frequently the best AUC for each radiological finding comes from the same dataset. In the ChestX-ray14 test set, the best AUC appears 7 out of 8 times when training with the same set. The same phenomenon happens on both CheXpert (8 out of 8) and MIMIC-CXR (8 out of 8). Furthermore, in all four test sets, the best average AUC comes from their respective training set. One possible cause of domain shift is the label extraction method. CheXpert and MIMIC-CXR used the same labeler, while ChestX-ray14 has its own.

ChestX-ray14 labeler has raised some questions concerning its reliability. A visual inspection study [43] states that its labels do not accurately reflect the content of the images. Estimated label accuracies are $10 - 30\%$ lower than the values originally reported. It also might be that ChestX-ray14 and PadChest do not have representative training sets since models trained on CheXpert and MIMIC-CXR perform well on ChestX-ray14 and PadChest test sets, but the models trained on ChestX-ray14 and PadChest do not perform well on CheXpert and MIMIC-CXR's test sets.

4.5 Discussion

Our experiments showed that a model with reported radiologist-level performance [49] had a significant drop in performance outside its source dataset, pointing the existence of domain shift in chest X-rays datasets. Despite recent efforts for the creation of large radiograph datasets in the hope of training generalizable models, it seems that the data acquisition methodology of some of the available datasets does not capture the required heterogeneity for this purpose.

Among the analyzed datasets, CheXpert and MIMIC-CXR seem to be the most representative of the other datasets, as the models trained on them show a smaller performance drop when comparing to the baseline. Therefore, these two datasets should be preferred by researchers when developing models for chest radiograph analysis. The least representative dataset seems to be ChestX-ray14, whose model did not perform as well outside its own test set, while the models trained on the other datasets performed well when testing on ChestX-ray14. Models trained on PadChest also show a significant performance drop in other test sets, but it might be because of the smaller amount of available data for each finding.

5. SEMI-SUPERVISED PATHOLOGY CLASSIFICATION

In this chapter, we describe the experiment analysis that we have designed to compare different semi-supervised classification methods in a chest radiograph classification scenario. The experiments of this chapter resulted in the work "Semi-supervised classification of chest radiographs" published in MICCAI's International Workshop on Medical Image Learning with Less Labels and Imperfect Data [46].

5.1 Motivation

Public datasets of chest radiographs provide over $100,000$ chest radiographs labeled with the most common findings [63].These datasets have automatically-extracted labels obtained via natural language processing algorithms on radiological reports and have been used to build radiologist-level models [49]. However, in most medical imaging scenarios, there is a lack of annotated data available [35], since, for most tasks, the samples need to be manually annotated by an expert, which is an expensive and time-consuming task, like annotating chest radiographs for pathology localization.

Recently, research in semi-supervised learning for image classification had some considerable progress [48]. Methods based on consistency regularization strategies such as Mean Teacher [59], Unsupervised Data Augmentation [64], MixMatch [7], and FixMatch [58] achieve results comparable to supervised training but with only a fraction of the training samples. For instance, training a model on the SVHN dataset [42] in a supervised fashion using all training data ($73,257$ labeled samples) results in an error rate of 2.59%, whereas training the same model with the MixMatch approach and only 250 labeled samples achieves an error of 3.78% [7]. However, the best-performing method can vary when comparing them in different datasets and tasks.

Table 5.1: Error rate on CIFAR-10 from four different semi-supervised learning methods and a supervised baseline under different amounts of labeled data during training. Values reported on the original papers. CIFAR-10 consists of 50000 training samples and 10000 test samples.

Method / Labels	40	250	500	1000	2000	4000
Supervised-only [59]	-	-	-	46.43	33.94	20.66
Mean Teacher [59]	-	47.32	42.01	17.32	12.17	10.36
UDA [64]	-	8.76	**6.68**	**5.87**	**5.51**	5.29
MixMatch [7]	-	11.08	9.65	7.75	7.03	6.24
FixMatch [58]	**11.39**	**5.07**	-	-	-	**4.31**

These methods were developed considering natural images and popular computer vision benchmark datasets like CIFAR-10[26], therefore, they were not thoroughly validated and compared in a medical imaging scenario. Table 5.1 shows the results of some state-of-the-art methods on the CIFAR-10 dataset. Our goal in this experimental analysis is to compare state-of-the-art semi-supervised classification methods in a medical imaging classification scenario.

We adapt the semi-supervised classification methods to a multi-label scenario and compare them to a strong supervised baseline in chest radiograph classification, the CheXNet architecture [49].

5.2 Semi-Supervised Learning Methods

The methods we investigate implement both perturbation-based and entropy minimization techniques. Perturbation-based approaches rest on the smoothness assumption, which implies that small perturbations on the input should not alter the model's prediction. This behavior does not depend on knowing the ground-truth label. Therefore, we can apply noise to input data points, and use the distance between the output of clean and noisy input samples on the loss function, adjusting the model based on unlabeled data [61]. These methods take advantage of artificial neural networks because of their straightforward incorporation of additional terms on the objective optimization function. Entropy minimization approaches rest on the low-density assumption and encourage the model to make confident predictions even on unlabeled data in order to keep the decision boundary far from high-density areas.

5.2.1 Mean Teacher

The Mean Teacher approach [59] is based on a previous state-of-the-art semi-supervised learning method called Temporal Ensembling [28], which first proposed using an exponential moving average to combine prediction scores from models on different epochs and different regularization conditions to predict unknown labels. Mean Teacher [59] consists of using two models with the same architecture, which are called *student* and *teacher.* The student and the teacher receive the same inputs with different augmentation policies. Then, a consistency loss is computed based on the distance between both models predictions. Finally, the student weights Θ^S are updated via loss optimization, and the teacher weights Θ^T are updated via an exponential moving average of the student weights after each training step e. A hyperparameter ρ controls the exponential moving average to update the teacher's weights, as shown in Equation 5.1.

$$\Theta_e^T = \rho\Theta_{e-1}^T + (1-\rho)\Theta_e^S \tag{5.1}$$

The loss function $\mathcal{L}_{comb}$ used to update the student's weights is defined in Equation 5.2. The combined loss is the sum of the task loss $\mathcal{L}_{task}$ with the consistency loss $\mathcal{L}_{cons}$. The task loss (Equation 5.3) is a regular binary cross-entropy loss between the ground-truth labels y and the labels predicted by the student model $m_s(x)$, which is only computed on labeled instances. The consistency loss (Equation 5.4) is a mean-squared error of the predictions from the student and the teacher on unlabeled data u when submitted to two augmentation policies ϕ_s e ϕ_t. A hyperparameter λ_{con} defines the weight of the consistency loss on the combined loss.

$$\mathcal{L}_{comb} = \mathcal{L}_{task} + \lambda_{con}\mathcal{L}_{cons} \tag{5.2}$$

$$\mathcal{L}_{task} = BCE(m_s(x), y) \tag{5.3}$$

$$\mathcal{L}_{cons} = ||m_s(\phi_s(u)) - m_t(\phi_t(u))||^2 \tag{5.4}$$

5.2.2 Unsupervised Data Augmentation

Xie et al. [64] argue that the quality of the input noise plays a crucial role in perturbation-based semi-supervised learning methods. They propose Unsupervised Data Augmentation (UDA), which uses advanced data augmentation techniques to input noise on the training data. For image classification tasks, the authors propose using RandAugment [11] as the data augmentation technique. RandAugment presents an improvement over AutoAugment [10], a search method to test multiple image processing procedures to find a good augmentation policy using reinforcement learning. AutoAugment was first proposed by the authors and was further replaced by RandAugment, a simpler technique that does not require to be learned ahead with labeled data. RandAugment randomly selects transformations for each sample from a collection of transformations. A global magnitude parameter controls the distortions. This hyperparameter is optimized via grid search on a validation set.

UDA uses only one model $m(\cdot)$, which is updated by a combined loss just like the one defined in Equation 5.2, except that the consistency loss $\mathcal{L}_{cons}$ is a KL divergence (Equation 5.6) between augmented ($\phi(u)$) and non-augmented (u) unlabeled data, and that the task loss $\mathcal{L}_{task}$ (Equation 5.5) is computed on the same model m.

$$\mathcal{L}_{task} = BCE(m(x), y) \tag{5.5}$$

$$\mathcal{L}_{cons} = log\frac{m(\phi(u))}{m(u)} \tag{5.6}$$

Since there is usually a limited amount of labeled data, for preventing overfitting the labeled data and underfitting the unlabeled data, the authors propose a technique called training signal annealing (TSA). It consists of defining a confidence threshold for the model's predictions to use the training signals of the labeled sample, gradually increasing the threshold η_t from $1/n$ to 1 (n being the number of classes) according to a schedule. This technique prevents over-training on easy samples and focuses the initial stage of the training on complex samples.

5.2.3 MixMatch

MixMatch [7] is an algorithm that combines techniques from different semi-supervised learning regularization approaches. It starts by sampling and augmenting labeled and unlabeled samples. Each unlabeled sample is augmented Q times, and the model computes predictions for each augmented sample. These predictions are averaged and sharpened (Equation 5.9) to become pseudo-labels $\hat{y}$. Then, the augmented labeled and unlabeled data form a batch with their respective labels and pseudo-labels. This batch is shuffled and regularized using the MixUp regularizer [65], which interpolates data points to create a smoother data distribution, and the model training set becomes the interpolated points $\tilde{x}$ and their labels $\tilde{y}$ and $\tilde{\hat{y}}$.

The loss function is a combination of the losses on labeled and unlabeled data controlled by a hyperparameter λ_{con} like the previous approaches (Equation 5.2). The loss for labeled data (Equation 5.7) is a binary cross-entropy as the one in UDA, except that it uses the data points $\tilde{x}$ and labels $\tilde{y}$ generated by MixUp. The loss for unlabeled data (Equation 5.8) is a mean-squared error between generated pseudo-labels $\tilde{\hat{y}}$ and the predictions on mixed-up unlabeled inputs $\tilde{u}$.

$$\mathcal{L}_{task} = BCE(m(\tilde{x}), \tilde{y}) \tag{5.7}$$

$$\mathcal{L}_{cons} = ||m(\tilde{u}) - \tilde{\hat{y}}||^2 \tag{5.8}$$

The sharpen function is based on the entropy minimization concept and is a softmax function adjusted by a temperature hyperparameter τ, which reduces the softness of the result in order to keep predictions more close to a one-hot distribution.

$$\text{sharpen}(p, \tau) = \frac{p^{\frac{1}{\tau}}}{\sum_{i=1}^{c} p_i^{\frac{1}{\tau}}} \tag{5.9}$$

MixUp regularization computes a new input $\tilde{x}$ and a new target $\tilde{y}$ by interpolating two data points. This interpolation is detailed in Equation 5.10 using two points x_1 and x_2 and their labels y_1 and y_2. The γ value is randomly sampled from a $\beta(\alpha, \alpha)$ distribution. As we increase the value of α, the interpolated points become farther from the real points and closer to the center of the two points.

$$\begin{aligned} \tilde{x} &= \gamma x_1 + (1-\gamma)x_2 \\ \tilde{y} &= \gamma y_1 + (1-\gamma)y_2 \end{aligned} \tag{5.10}$$

5.2.4 FixMatch

FixMatch [58] is a simple yet effective approach that combines consistency regularization with pseudo-labeling. It leverages strong and weak augmentation policies. At first, an input sample u_i is weakly augmented with a policy ϕ and fed to a model $m(\cdot)$. Its output becomes a pseudo-label for u_i using $\hat{y}_i = \arg\max(m(\phi(u_i)))$. Then, the input u_i is strongly augmented with a policy Φ, and the model is trained with a regular cross-entropy loss using the previously generated pseudo-label $\hat{y}_i$.

FixMatch optimizes a combined loss of labeled and unlabeled data controlled by a hyperparameter λ_{con} like the one defined in Equation 5.2 used in previous methods. The task loss $\mathcal{L}_{task}$ (Equation 5.11) is a binary cross-entropy between weakly augmented inputs $\phi(x_i)$ and their ground-truth labels y_i. The consistency loss $\mathcal{L}_{cons}$ (Equation 5.12) is also a binary cross-entropy, but between the strongly-augmented $\Phi(u)$ and the pseudo-labels $\hat{y}$ that have a confidence score $max(m(\phi(x_i)))$ higher than a threshold T.

$$\mathcal{L}_{task} = BCE(m(\phi(x)), y) \tag{5.11}$$

$$\mathcal{L}_{cons} = \mathbb{1}(max(m(\phi(u)) \geqslant T) BCE(m(\Phi(u), \hat{y}) \tag{5.12}$$

The weak augmentations are random flips and image translations. The strong augmentations are two approaches based on AutoAugment[10], RandAugment[11], the same used in UDA [6] and explained in Section 5.2.2, and CTAugment [6], which learns during training the best augmentation policy by separating the possible augmentation values in bins and assigning weights for each bin. To update the weights, it uses labeled data. Two bins are randomly sampled, and the weights are updated according to how close the model's prediction is to the true label.

5.3 Experimental Design

We employ a multi-label classification approach reproducing the CheXNet model [49], a popular approach that achieved state-of-the-art results in classifying multiple pathologies using a DenseNet-121 convolutional neural network architecture [21]. We use it as our supervised baseline and also as a backbone for the semi-supervised methods. Table 5.2 shows the performance of the CheXNet model [49] compared with the initial baseline proposed with the launch of the ChestX-ray14 dataset [63].

The model is pre-trained on the ImageNet dataset, and the images are resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation. We use a learning rate of 0.01, a cosine learning rate schedule, and a Stochastic Gradient Descent optimizer with

Table 5.2: Supervised baselines using all available training data of ChestX-ray14.

Finding	Wang et al. [63]	CheXNet [49]
Atelectasis	0.716	**0.8094**
Cardiomegaly	0.807	**0.9248**
Consolidation	0.708	**0.7901**
Edema	0.835	**0.8878**
Effusion	0.784	**0.8638**
Emphysema	0.815	**0.9371**
Fibrosis	0.769	**0.8047**
Hernia	0.767	**0.9164**
Infiltration	0.609	**0.7345**
Mass	0.706	**0.8676**
Nodule	0.671	**0.7802**
Pleural Thickening	0.708	**0.8062**
Pneumonia	0.633	**0.7680**
Pneumothorax	0.806	**0.8887**
Average	0.738	**0.8414**

0.9 momentum, a weight decay of 0.001 and a mini-batch size of 16. In the semi-supervised methods, we use 8 labeled and 8 unlabeled samples for each batch. The weak augmentations are the same ones performed in the supervised baseline [49], the strong augmentations are done by RandAugment [11] with $n = 2$ and $m = 10$. Every method is trained for 20 epochs, as we empirically observed that a longer training does not show improvement.

We use the same model hyperparameters for supervised training in all methods, varying only the hyperparameters referring to the semi-supervised training. We use subsets containing 25, 100, and 400 labeled samples per class for each method and leave the rest of the training set as unlabeled samples, which is a common setup for semi-supervised evaluation. We have three different subsets with different samples used as labeled for each labeled amount, and we report the mean and standard deviation of the top performance on the three experiments. We evaluate the models' performance computing the area under the receiver operating characteristic curve (AUC) for each label.

5.3.1 Methods

We evaluate the following state-of-the-art semi-supervised learning methods: Mean Teacher [59], Unsupervised Data Augmentation [64], MixMatch [7], and FixMatch [58], as well as a supervised baseline. We also compared these methods with a simple semi-supervised baseline, called Pseudo-labeling. In that approach, the model is trained with a regular cross-entropy loss on labeled data, and we also take the top prediction made in unlabeled data and use it as a pseudo-label to compute the unsupervised loss, and add it to the combined loss that optimizes the model. Our approach was based on the work of Lee [31], and since our task is a multi-label scenario, we tested a soft label approach, in which the pseudo-label is the classes score prediction, and a hard label approach, in which the pseudo-label is a one-hot vector with the top prediction as one and the rest as zero.

We replaced the original softmax output and categorical cross-entropy loss with a sigmoid output and a multi-label binary cross-entropy loss to adapt to a multi-label scenario. FixMatch [58], MixMatch [7] and Pseudo-labeling [31] make use of pseudo-labels. To adapt these approaches, we use a hard pseudo-label through a one-hot vector with the top prediction as one and the rest as zero, which means that the pseudo-label is not multi-label. In methods that use a score threshold (FixMatch [58] and UDA [64]), we compute the threshold based only on the top prediction.

5.3.2 Hyperparameter Search

To select the best set of hyperparameters for our objective task, we performed a random hyperparameter search for each method using a 25 labels subset. We trained the model with different hyperparameters for 20 epochs and selected the ones that achieved a higher AUC on the validation set. In all methods, we searched for a consistency weight between 0.5 and 100.

For **Pseudo-labeling**, we selected 1 as the unsupervised weight and also searched for two different pseudo-labeling strategies using soft and hard pseudo-labels. Hard pseudo-labels had the best performance.

In **Mean Teacher**, we selected a consistency weight of 100, with an exponential consistency rampup as proposed by [59] with a length of 10 epochs. We also searched for an EMA decay rate ρ for the teacher model between 0.8 and 0.99, and selected 0.99.

In **UDA**, we selected a consistency weight of 2 and searched for a traning signal annealing schedule among the three proposed by [64], which are linear, exponential, and logarithmic. The logarithmic schedule showed the best validation results, but not using a threshold was still better, so we did not use TSA.

For **MixMatch**, we selected a consistency weight of 10 and also searched for the α of the $\beta(\alpha, \alpha)$ distribution between 0.1 and 50, selecting 0.1, which indicates a more conservative interpolation.

For **FixMatch**, we selected a consistency weight of 1, and also searched for a threshold between 0.7 and 0.95, selecting 0.8. Since in the original paper the authors reported that a larger ratio of unlabeled samples increased the model performance, we also searched for a ratio of 2,3, and 4, but the ratio of 1 still presented the best results.

5.4 Results

We summarize the average results for each method and label subset in Table 5.3. Our strongest baseline is the fully-supervised CheXNet [49], which achieves an average AUC of 0.8414. Its results and a comparison with the baseline proposed by Wang et al. [63] is shown in Table 5.2. The results of all the semi-supervised approaches are similar, with the most gain being obtained by Mean Teacher using 400 labels (See Table 5.6), achieving an average AUC 9% higher than the one obtained by supervised training. With 25 labels (See Table 5.4), the highest average result was obtained by UDA, improving supervised training in 5%. Using 100 labels (shown in Table 5.5), the best performance was achieved with Pseudo-label, improving the baseline in 6%.

Table 5.3: Average AUCs of our proposed approaches and baselines using different amounts of labeled samples on ChestX-ray14.

	25 labels	100 labels	400 labels
Supervised	0.6142 ± 0.0291	0.6596 ± 0.0300	0.6805 ± 0.0675
Pseudo-labeling	0.6675 ± 0.0155	**0.7232 ± 0.0014**	0.7565 ± 0.0050
Mean Teacher	0.6677 ± 0.0155	0.7223 ± 0.0102	**0.7708 ± 0.0013**
MixMatch	0.6627 ± 0.0195	0.7139 ± 0.0045	0.7612 ± 0.0048
FixMatch	0.6643 ± 0.0186	0.7129 ± 0.0110	0.7634 ± 0.0029
UDA	**0.6691 ± 0.0176**	0.7225 ± 0.0125	0.7612 ± 0.0065

The proposed baseline method presented by Wang et al. [63] in ChestX-ray14's release achieved an average AUC of 0.738. Comparing it with the original results in [63], our implemented approaches were capable of outperforming their fully-supervised model using only 400 labeled samples per class, in which we achieved 0.7708 AUC, and achieved similar results when using 100 labels, which scored an average AUC of 0.7232. Based on the average results of each method shown in Table 5.3, we computed an overall gain for each method with respect to the supervised baseline. Mean teacher had the best performance with an overall gain of 0.2065, followed by UDA with 0.1985, Pseudo-labeling with 0.1929, FixMatch with 0.1863 and MixMatch with 0.1839.

Table 5.4: AUC results for the implemented methods using 25 labels per class during training.

			25 labels			
	Supervised	Pseudo-label	Mean Teacher	MixMatch	FixMatch	UDA
Atelectasis	0.5987	0.6637	0.6609	0.6665	**0.6775**	0.6736
Cardiomegaly	0.5682	0.6421	0.6283	0.6270	0.6646	**0.6671**
Consolidation	0.6753	0.6816	0.6882	0.6771	0.6797	**0.6945**
Edema	0.7748	0.8082	0.7953	**0.8086**	0.7995	0.8043
Effusion	0.6587	**0.7635**	0.7591	0.7464	0.7545	0.7482
Emphysema	0.5447	0.6579	**0.6864**	0.6761	0.6600	0.6535
Fibrosis	0.6594	0.6586	**0.6687**	0.6474	0.6336	0.6478
Hernia	0.6992	0.8005	**0.8192**	0.7851	0.7811	0.8141
Infiltration	**0.6099**	0.6028	0.6031	0.6042	0.6080	0.5943
Mass	0.5243	0.5362	0.5495	0.5148	0.5590	**0.5647**
Nodule	0.5512	0.5752	0.5640	**0.5826**	0.5713	0.5537
Pleural Thickening	0.5650	0.6231	**0.6353**	0.6248	0.6071	0.6233
Pneumonia	0.6088	**0.6311**	0.6144	0.6180	0.6267	0.6261
Pneumothorax	0.5799	0.6997	0.6915	**0.7018**	0.6870	0.7035
No Finding	0.5953	0.6681	0.6521	0.6598	0.6553	**0.6683**
Average	0.6142	0.6675	0.6677	0.6627	0.6643	**0.6691**

Table 5.5: AUC results for the implemented methods using 100 labels per class during training.

	100 labels					
	Supervised	Pseudo-label	Mean Teacher	MixMatch	FixMatch	UDA
Atelectasis	0.6464	**0.7170**	0.7112	0.6833	0.6966	0.7121
Cardiomegaly	0.6489	**0.7986**	0.7622	0.7791	0.7690	0.7773
Consolidation	0.7117	0.7292	0.7230	0.7222	0.7315	**0.7360**
Edema	0.8130	0.8366	**0.8378**	0.8350	0.8236	0.8289
Effusion	0.7225	0.8055	**0.8068**	0.7870	0.7989	0.8032
Emphysema	0.6381	0.7628	0.7717	0.7663	0.7515	**0.7850**
Fibrosis	0.6603	0.6829	**0.7035**	0.6777	0.6941	0.6976
Hernia	0.7834	**0.8945**	0.8844	0.8923	0.8692	0.8743
Infiltration	0.6268	0.6298	**0.6393**	0.6314	0.6309	0.6255
Mass	0.5613	0.6221	**0.6283**	0.6192	0.6238	0.6235
Nodule	0.5648	0.5881	**0.5880**	0.5828	0.5733	0.5814
Pleural Thickening	0.5797	0.6515	**0.6544**	0.6258	0.6273	0.6466
Pneumonia	0.6356	**0.6657**	0.6495	0.6577	0.6616	0.6552
Pneumothorax	0.6561	0.7609	0.7661	0.7512	0.7457	**0.7780**
No Finding	0.6450	0.7030	0.7089	0.6970	0.6968	**0.7129**
Average	0.6596	**0.7232**	0.7223	0.7139	0.7129	0.7225

Table 5.6: AUC results for the implemented methods using 400 labels per class during training.

	400 labels					
	Supervised	Pseudo-label	Mean Teacher	MixMatch	FixMatch	UDA
Atelectasis	0.6739	0.7448	**0.7489**	0.7404	0.7417	0.7343
Cardiomegaly	0.6884	0.8516	**0.8642**	0.8639	0.8547	0.8501
Consolidation	0.7220	0.7367	0.7319	0.7430	**0.7613**	0.7430
Edema	0.8196	0.8536	**0.8674**	0.8626	0.8613	0.8563
Effusion	0.7282	0.8286	**0.8365**	0.8349	0.8317	0.8261
Emphysema	0.6761	0.8415	**0.8706**	0.8496	0.8382	0.8564
Fibrosis	0.6873	0.7249	**0.7575**	0.7555	0.7385	0.7503
Hernia	0.8125	0.8934	0.8908	0.8841	**0.9089**	0.8863
Infiltration	0.6345	0.6506	0.6562	0.6438	**0.6580**	0.6410
Mass	0.5910	0.6909	**0.7226**	0.7056	0.6987	0.7040
Nodule	0.5893	0.6328	**0.6529**	0.6439	0.6505	0.6463
Pleural Thickening	0.6073	0.6859	0.6956	0.6772	0.7042	**0.7060**
Pneumonia	0.6373	0.6651	**0.7071**	0.6721	0.6919	0.6660
Pneumothorax	0.6889	0.8113	**0.8178**	0.8120	0.7767	0.8177
No Finding	0.6518	0.7353	**0.7418**	0.7298	0.7342	0.7338
Average	0.6805	0.7565	**0.7708**	0.7612	0.7634	0.7612

5.4.1 Comparison with Previous Approaches

The work of Rivero et al. [3] aims at reducing the need for annotated data in medical imaging. They propose GraphX^{NET}, a graph-based semi-supervised learning approach for X-ray data classification. It is a graph model that contains all the training samples and only a limited amount of them are labeled. They tested the approach in the ChestX-ray14 dataset. When using only 20% of the data, they achieve results close to a fully-supervised model. However, under extreme minimal supervision (2% labeled data), the model does not perform well, having an average AUC of 0.53.

Tanan et al. [37] perform semi-supervised classification in skin lesion classification and thoracic image analysis. The proposed method is called SRC-MT. It is a semi-supervised classifier based on Mean Teacher [59] and introduces a sample relation consistency term to the optimization function. This enforces the consistency based on the relationship information among different samples instead of individual predictions. They achieve similar results to GraphX-NET when using 20% of ChestX-ray14, but when using only 2% of labeled data, they achieve an average AUC of 0.67.

These two previous studies [3, 37] had addressed the problem of reducing the need for annotated data in medical imaging and evaluated their results on the ChestX-ray14 dataset. They use subsets of the available training data as labeled data and the rest as unlabeled data. We selected our Mean Teacher approach to compare with their results. When using 2% and 5% of the labeled data, our approach outperforms the previous results by Rivero et al. [3] and the results by Tanan et al. [37], as presented in Table 5.7.

Table 5.7: Average AUC of our best approach with two previous approaches for semi-supervised classification in ChestX-ray14.

	2%	5%
GraphX^{NET} [3]	53	58
SRC-MT [37]	66.95	72.29
Ours (Mean Teacher)	**71.82**	**74.82**

5.5 Discussion

In this chapter, we evaluated different semi-supervised learning methods performing multi-label classification in a medical imaging scenario and achieved state-of-the-art results on semi-supervised classification on ChestX-ray14. Most of the trained methods showed similar results, with Mean Teacher having a slightly better gain in overall performance when compared to a supervised baseline. The improvement over a supervised baseline is not as high as the ones reported by the original methods in common computer vision datasets like CIFAR-10 [26], highlighting that we might need to make some adaptations to these methods specifically designed for the context of medical imaging scenarios.

6. SEMI-SUPERVISED PATHOLOGY LOCALIZATION

In this chapter, we propose a novel training procedure that combines multiple instance learning with semi-supervised learning for pathology localization on chest radiographs.

As public datasets provide over $700,000$ chest radiographs labeled with radiological findings [63, 22, 24], but only a small amount of those (880) are annotated with bounding boxes, approaches for pathology localization have been developed to make use of the classification labels by performing weakly-supervised learning [63] and multiple instance learning [33].

In the weakly-supervised learning scenario [63], a model is trained for classification using the available labels. Then, saliency maps like Grad-CAM [56] are used to explain the classification output and the pixels with the higher contribution to the selected output are used as a localization inference. This approach is not ideal since wrong classes can still produce high activation values on image regions [53], and it performs poorly on locating diagnostically-relevant regions for medical image interpretation, as shown by Saporta et al. [55].

To work within a multiple instance learning scenario, Li at al. [33] use an architecture that divides the image into patches and predicts the pathology probability in each region by outputting a $P \times P \times C$ matrix of scores (C is the number of classes/pathologies) and training in a supervised manner using the ground-truth localization labels. Then, they use the classification labels by formulating a pooling mechanism that transforms a $P \times P \times C$ matrix into a $1 \times C$ vector containing the score for each class. Assuming that an image is positive for a particular pathology, then at least one patch needs to be positive. The pooled vector is compared with the ground-truth and the difference is added to the loss function and used during training to optimize the network's weights. They make use of both the classification and localization labels available on the dataset.

Our problem thus consists of extending a semi-supervised learning method to a model that locates pathology patterns on radiographs using limited localization supervision but abundant classification labels. Since we have a large amount of samples labeled regarding a pathology presence, we can propose a multiple instance learning scenario based on [33], in which the bag labels are pathology presence, and the sample labels are pathology localization. To extend the use of the unannotated data, we propose to adapt Mean Teacher, the we identified as being the best semi-supervised approach in the experiments of Chapter 5, to this multiple instance learning framework.

6.1 Methodology

6.1.1 Data description

We use the ChestX-ray14 dataset in our experiments. ChestX-ray14 [63] is a public dataset from the National Institute of Health containing $112,120$ frontal-view chest radiographs from $32,717$ different patients labeled with 14 radiological findings and with 984 manually annotated bounding boxes on 880 different images for 8 of the 14 pathologies. The pathologies containing location annotation are: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. Based on the results of the analysis of Chapter 4, we also performed an experiment using the CheXpert dataset as unlabeled data. CheXpert [22] is a dataset from the Stanford Hospital containing $224,316$ frontal and lateral chest radiographs of $65,240$ patients.

In our experiments, we use the official training, validation, and test split of the ChestX-ray14 dataset, which keeps the images from the same patient on the same set to avoid data leakage. The training set contains $78,484$ samples, the validation set has $8,040$ samples, and the test set has $25,596$ samples with classification labels. Since the samples annotated with bounding boxes are limited (880), we perform a 4-fold split cross-validation to evaluate the method's performance. In each fold, there are 660 images used for training and 220 used for testing. The training and validation set of unannotated data remains the same in all 4 folds. We also perform one experiment using CheXpert data as unlabeled data. In this experiment we select one random frontal sample for each patient, adding $65,240$ samples to the training set.

6.1.2 Baseline

For our baseline, we implemented the multiple instance learning architecture proposed by Li et al. [33]. The first component is a ResNet-50 [18] network without the final linear layer, which works as a feature extractor of the $h \times w$ input image, producing a $h/32 \times w/32 \times 2048$ feature vector. Then, an upsampling layer rescales the feature vectors to the desired patch setting of $P \times P \times 2048$ using bilinear interpolation. P here is the number of patches, an adjustable hyperparameter which we set to 20. The upscaling layer is followed by a convolution layer with 512 filters with a 3×3 size, a batch normalization layer, and a ReLU activation function. Then, a final convolution layer with C filters with 1×1 size and a sigmoid activation function outputs a $P \times P \times C$ score matrix. A visual depiction of the method is presented in Figure 6.1.

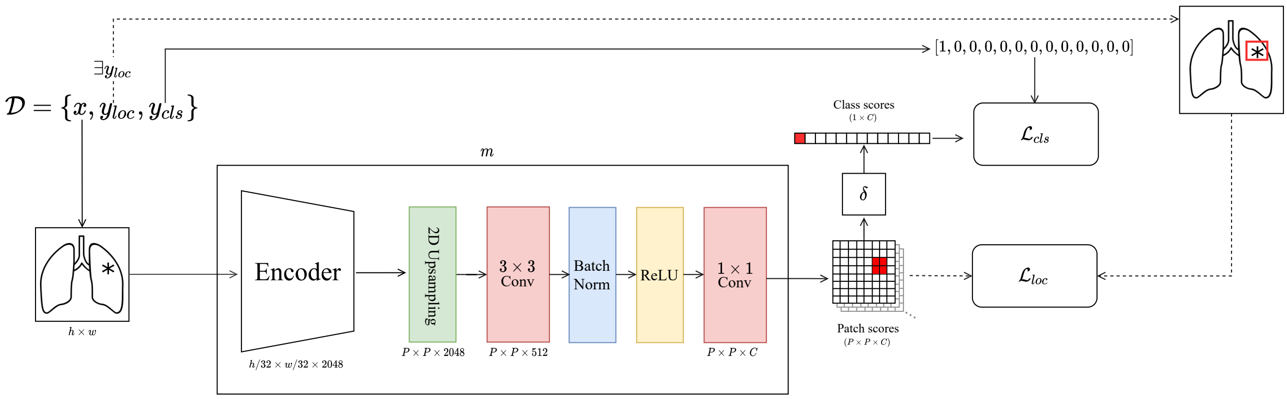


Figure 6.1: Architecture of the developed multiple instance learning localization model based on [33]. The model comprises a ResNet-50 encoder, which extracts features from the input image, an upsampling layer, and a sequence of convolution, batch normalization, ReLU, and another convolution, outputting a patch score matrix. A pooling function δ converts the patch scores into class scores and the model is optimized to reduce the classification loss $\mathcal{L}_{cls}$ between the prediction and the ground-truth y_{cls}. When the localization label is available, a localization loss $\mathcal{L}_loc$ is also computed using the patch scores and the y_{loc} ground-truth.

We did our implementation by adapting the re-implementation code made available by the work of Preechakul et al. [47] on weakly-supervised pathology detection, since the original authors of [33] did not publicly release their code. Table 6.1 compares the performance of our re-implemented baseline with the original reported results. Our baseline did not match the expected results reported by Li et al. [33].

Table 6.1: Comparison of IoR accuracy in all 8 pathologies of the re-implemented baseline and the metrics reported by Li et al. [33], in which we based our baseline, and Wang et al. [63], which introduced the dataset. The "Avg" column presents the value of a macro-average of the accuracy results on the 8 pathologies.

T(IoR)	Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Avg
0.1	Wang et al. [63]	0.62	**1.00**	0.80	0.91	0.59	0.15	**0.86**	0.52	0.68
	Li et al. [33]	**0.77** ± 0.06	0.99 ± 0.01	**0.91** ± 0.04	**0.95** ± 0.05	**0.75** ± 0.08	**0.40** ± 0.11	0.69 ± 0.09	**0.68** ± 0.10	**0.75**
	Our baseline	0.59 ± 0.06	0.96 ± 0.09	0.76 ± 0.08	0.83 ± 0.09	0.60 ± 0.13	0.19 ± 0.14	0.69 ± 0.06	0.44 ± 0.10	0.63
0.25	Wang et al. [63]	0.39	0.99	0.63	0.80	0.46	0.05	**0.71**	0.34	0.55
	Li et al. [33]	**0.57** ± 0.09	**0.99** ± 0.01	**0.79** ± 0.02	**0.88** ± 0.06	**0.57** ± 0.07	**0.25** ± 0.10	0.62 ± 0.05	**0.61** ± 0.07	**0.66**
	Our baseline	0.33 ± 0.08	0.93 ± 0.12	0.56 ± 0.16	0.68 ± 0.09	0.36 ± 0.09	0.03 ± 0.03	0.58 ± 0.13	0.28 ± 0.11	0.47
0.5	Wang et al. [63]	0.19	0.95	0.42	**0.65**	0.31	0.00	0.48	0.27	0.41
	Li et al. [33]	**0.35** ± 0.04	**0.98** ± 0.02	**0.52** ± 0.03	0.62 ± 0.08	**0.40** ± 0.06	**0.11** ± 0.04	**0.49** ± 0.08	**0.43** ± 0.10	**0.49**
	Our baseline	0.12 ± 0.06	0.88 ± 0.15	0.28 ± 0.15	0.39 ± 0.10	0.21 ± 0.09	0.00 ± 0.00	0.37 ± 0.06	0.11 ± 0.07	0.30
0.75	Wang et al. [63]	0.09	0.82	0.23	0.44	0.16	0.00	0.29	0.17	0.28
	Li et al. [33]	**0.20** ± 0.04	**0.87** ± 0.05	**0.34** ± 0.06	**0.46** ± 0.07	**0.29** ± 0.06	**0.07** ± 0.04	**0.43** ± 0.06	**0.30** ± 0.07	**0.37**
	Our baseline	0.04 ± 0.02	0.70 ± 0.16	0.12 ± 0.07	0.18 ± 0.07	0.08 ± 0.05	0.00 ± 0.00	0.18 ± 0.08	0.05 ± 0.05	0.17
0.9	Wang et al. [63]	0.07	**0.65**	0.14	**0.36**	0.09	0.00	0.23	0.12	0.21
	Li et al. [33]	**0.15** ± 0.03	0.59 ± 0.04	**0.23** ± 0.05	0.32 ± 0.07	**0.22** ± 0.05	**0.06** ± 0.03	**0.34** ± 0.04	**0.22** ± 0.05	**0.27**
	Our baseline	0.01 ± 0.01	0.56 ± 0.12	0.07 ± 0.06	0.07 ± 0.03	0.05 ± 0.04	0.00 ± 0.00	0.12 ± 0.08	0.03 ± 0.04	0.11

If an image is positive for class c, then, at least one of the patch scores p_i^c must be classified as positive. The probability of an image being positive, which can also be seen as the classification scores, are obtained through a pooling function $\delta(\cdot)$. Considering a set of generated patches $\mathcal{M}$, where $|\mathcal{M}| = P \times P$, the function $\delta(\cdot)$ multiplies the score p_i^c outputted by $m(x)$ for every patch in $\mathcal{M}$, as shown in Equation 6.1. To avoid the numerical underflow that happens when multiplying small numbers, the patch scores $1 - p_{ij}^k$ are rescaled to $[0.98, 1.00]$.

$$\delta(m(x)^c) = 1 - \prod_{i \in \mathcal{M}} (1 - p_i^c) \tag{6.1}$$

The loss function is calculated over the patch scores for images with annotated bounding boxes and over the classification scores provided by δ when there is no localization annotation. The localization loss (Equation 6.2) is computed using a binary cross entropy $BCE(\cdot)$ between the predicted patch scores for each class c and the ground-truth. The ground-truth y_{loc} is obtained by converting the bounding boxes information into a $h \times w \times C$ binary matrix, containing a value of 1 if the pixel is inside the bounding box of the pathology class c and 0 otherwise. Then, this matrix is downscaled to $P \times P \times C$ using nearest neighbor interpolation. In the classification loss term (Equation 6.3), we use the pooled scores $\delta(m(x))$ and compute a BCE loss with the ground-truth class scores y_{cls}. Both losses are added to form a combined loss function that is used to optimize the network's parameters (Equation 6.4). Since there are far less samples with localization annotations, the localization loss has a weight hyperparameter λ_{loc} to increase its importance on the combined function.

$$\mathcal{L}_{loc} = BCE(m(x), y_{loc}) \tag{6.2}$$

$$\mathcal{L}_{cls} = BCE(\delta(m(x)), y_{cls}) \tag{6.3}$$

$$\mathcal{L}_{comb} = \lambda_{loc}\mathcal{L}_{loc} + \mathcal{L}_{cls} \tag{6.4}$$

6.1.3 Proposed approach

To extend the use of the non-annotated data, we propose to introduce consistency regularization, a semi-supervised training mechanism to the baseline multiple instance learning framework. We call our method *Consistent Multiple Instance Localization* (`C-MIL`). Since Mean Teacher [59] was identified as the best semi-supervised approach for classifying chest radiographs as shown in the experiments of Chapter 5, we decided to base `C-MIL` on the key aspects of Mean Teacher.

The original Mean Teacher framework [59] consists of using two models with identical architecture, which are called the *student* m_s and the *teacher* m_t. At every training iteration, both models are fed the same inputs with different augmentation policies, then a consistency loss is computed based on the distance between both models predictions. Following the smoothness assumption, these perturbations on the input should not alter the model's prediction. The

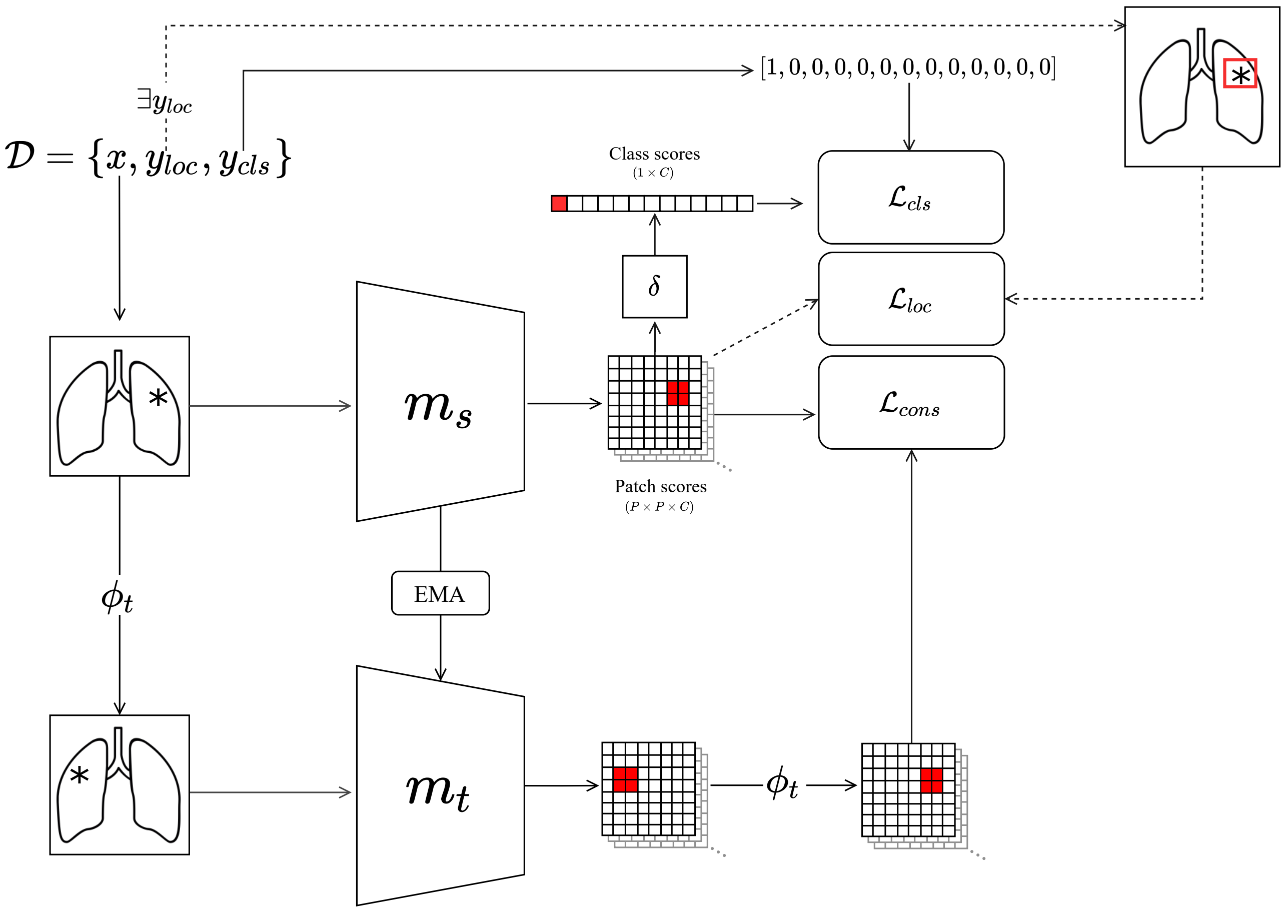


Figure 6.2: Architecture of C-MIL. A training dataset $\mathcal{D}$ contains images x, classification labels y_{cls} for each image, and localization labels y_{loc} for some images. The image is fed to a student model m_s that outputs patch scores, which are converted into class scores by a pooling function δ and compared to the ground-truth class labels y_cls to compute the classification loss $\mathcal{L}_{cls}$. If the localization labels are available for that image, a localization loss $\mathcal{L}_{loc}$ is computed using the patch scores and the ground-truth. And for all images, a teacher model m_t with an identical architecture receives a flipped version of the input image $\phi_t(x)$ and also generates patch scores. Then, the teacher output is flipped and compared against the student output to compute the consistency loss $\mathcal{L}_{con}$. The student is optimized to reduce these loss functions and the teacher is updated through an EMA of the student's weights.

student weights Θ^s are updated via loss optimization, and the teacher weights Θ^t are updated via an exponential moving average (EMA) of the student weights after each training step e. A hyperparameter ρ controls the EMA decay rate to update the teacher's weights, as in $\Theta^t_e = \rho\Theta^t_{e-1} + (1-\rho)\Theta^s_e$. A combined loss function $\mathcal{L}_{comb}$ is used to update the student's weights. This loss is the sum of the task loss $\mathcal{L}_{task}$ with the consistency loss $\mathcal{L}_{cons}$ controlled by a consistency weight hyperparameter λ_{con} as in $\mathcal{L}_{comb} = \mathcal{L}_{task} + \lambda_{con}\mathcal{L}_{cons}$. The task loss $\mathcal{L}_{task}$ is a regular cross-entropy loss between the ground-truth labels y and the predictions of the student model $m_s(x)$, which is only computed on labeled instances. The consistency loss is a mean-squared error of the predictions from the student and the teacher on unlabeled data u when submitted to two different augmentation policies ϕ_s and ϕ_t, defined as $||m_s(\phi_s(u)) - m_t(\phi_t(u))||^2$.

As shown in Equation 6.4, the multiple instance loss already contains classification and localization terms, so to enforce consistency regularization, we add a consistency loss term to the combined function. Since most of the samples do not provide localization annotation, we design the consistency loss on patch-level scores. Image augmentation pipelines can include rotation, translation and image distortion functions. In a classification scenario, the labels remain the same if the image is rotated or shifted, however in a localization scenario these functions can alter the output prediction, being a concern when computing the consistency metric. Based on the work of Jeong et al. [23], we design an augmentation strategy $\phi(\cdot)$ that flips the input image x horizontally, and thus the patch-level output of the model is also flipped. As shown in Equation 6.5, to compute the patch consistency loss $\mathcal{L}_{pcon}$ we also apply the transformation $\phi_t(\cdot)$ to flip the output of the teacher model in order to correctly enforce consistency. The mean-squared error is computed over every patch score on the $P \times P \times C$ matrix predicted by both the student and the teacher.

$$\mathcal{L}_{pco} = ||m_s(x) - \phi(m_t(\phi(x))||^2 \tag{6.5}$$

We use the following combined loss shown on Equation 6.6 to update the weights of `C-MIL`'s student model. The teacher's weights are updated through an EMA of the student's weights as it is done in the original mean teacher framework. Our proposed framework is illustrated on Figure 6.2.

$$\mathcal{L}_{comb} = \lambda_{loc}\mathcal{L}_{loc} + \mathcal{L}_{cls} + \lambda_{con}\mathcal{L}_{cons} \tag{6.6}$$

6.1.4 Experimental Settings

Throughout this experimental analysis, we use the official training, validation, and test sets of the ChestX-ray14 dataset. We perform a 4-fold cross-validation with the images with bounding box annotation to evaluate the method performance regarding localization. In each split, there are 660 images used for training and 220 used for testing. The reported metrics are the mean and standard deviation of the best model trained on each of the 4 folds. The best model is selected based on its performance in the validation set. The training and validation sets of unannotated data remain the same in all 4 folds. Table 6.2 shows pathology frequency in each of the 4 folds.

Table 6.2: Frequency of bounding boxes for each pathology in each data fold.

Fold	Atelectasis		Cardiomegaly		Effusion		Infiltration		Mass		Nodule		Pneumonia		Pneumothorax	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0	124	56	117	29	109	44	87	36	64	21	62	17	97	23	74	24
1	135	45	107	39	116	37	95	28	63	22	62	17	77	43	78	20
2	136	44	112	34	117	36	100	23	64	21	57	22	86	34	74	24
3	145	35	102	44	117	36	87	36	64	21	56	23	100	20	68	30

The feature extractor proposed in [33] is a ResNet architecture, so we make use of a ResNet-50 backbone in our experiments. We also use a DenseNet-121 as a feature extractor in some experiments, based on its superior performance in chest radiograph classification [49]. Both network architectures are initialized with pre-trained weights on the ImageNet Dataset [54]. We use a batch size of 64 when using a ResNet-50 backbone and a batch size of 48 for DenseNet-121. Each batch randomly contains annotated and unannotated samples, with a fixed random seed for each fold in order to standardize comparisons. Due to the low number of annotated samples, some batches might only have non-annotated samples. The baseline and proposed methods were executed with the same hyperparameters and under the same conditions.

We use a learning rate value of 10^{-4}, with a scheduled reduction of the learning rate by a factor of 0.2 based on a plateau of the validation loss. The training stops when the learning rate becomes lower than 10^{-6}, which happens usually in about 15 epochs. For the loss hyperparameters, the weight of the localization term λ_{loc} is set to 5, the patch consistency weight λ_{con} is set to 10 and we employ a consistency rampup lenght of 15 epochs. The EMA decay rate ρ is set to 0.99. The weights are optimized using the Adam optimizer [25].

An augmentation pipeline is applied to the unannotated data, containing random resized crops, 90° rotation, horizontal flip, brightness, and contrast variations. Each augmentation function has a 0.5 probability of happening. The images are pre-processed by resizing them to 256×256 pixels, and normalized based on ChestX-ray14 mean and standard deviation.

We use the Python programming language and the Pytorch framework [44] to implement our experiments. The experiments were executed in a NVIDIA Geforce GTX 1080 TI GPU, with 12GB of VRAM.

6.2 Results

6.2.1 Quantitative analysis

`C-MIL` performed better than the baselines in almost every pathology. Table 6.3 shows the localization accuracy under multiple thresholds of IoR. The improvement over higher thresholds like $T(IoR) = 0.75$ or $T(IoR) = 0.9$ is more significant than on smaller thresholds as $T(IoR) = 0.1$, which shows that the consistency regularization is helping the model to increase the confidence of the predictions closer to the ground-truth. To compute the accuracy metrics we apply a score threshold of 0.5 to define positive or negative patches. In Table 6.3, we also compare `C-MIL` using different backbones, the ResNet-50 and the DenseNet-121 architectures.

The largest improvement over the baseline was on the Cardiomegaly localization metrics, with the accuracy values going from 0.56 to 0.85 in a stronger threshold $T(IoR) = 0.9$. The smaller pathologies like Mass and Nodule did not improve significantly with `C-MIL`. The

Table 6.3: Pathology localization accuracy based on the Intersection over the detected Region (IoR), where $T(IoR) = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. The reported methods are our reimplementation of the method from Li et al. [33] (baseline) and our semi-supervised proposed approach `C-MIL` with ResNet-50 and DenseNet-121 as backbones.

T(IoR)	Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Micro-Avg.
0.1	Baseline	0.59 ± 0.06	0.96 ± 0.09	0.76 ± 0.08	0.83 ± 0.09	0.60 ± 0.13	**0.19** ± 0.14	0.69 ± 0.06	0.44 ± 0.10	0.67 ± 0.03
	`C-MIL` (ResNet)	0.59 ± 0.15	**1.00** ± 0.00	0.84 ± 0.04	0.85 ± 0.05	0.51 ± 0.13	0.03 ± 0.04	0.83 ± 0.13	0.54 ± 0.07	0.69 ± 0.04
	`C-MIL` (DenseNet)	**0.67** ± 0.08	**1.00** ± 0.00	**0.85** ± 0.03	**0.90** ± 0.05	**0.61** ± 0.06	0.01 ± 0.02	**0.92** ± 0.08	**0.60** ± 0.08	**0.74** ± 0.01
0.25	Baseline	0.33 ± 0.08	0.93 ± 0.12	0.56 ± 0.16	0.68 ± 0.09	0.36 ± 0.09	**0.03** ± 0.03	0.58 ± 0.13	0.28 ± 0.11	0.51 ± 0.02
	`C-MIL` (ResNet)	0.40 ± 0.12	**1.00** ± 0.00	**0.74** ± 0.04	0.67 ± 0.08	**0.40** ± 0.18	0.00 ± 0.00	0.66 ± 0.23	0.31 ± 0.09	0.56 ± 0.06
	`C-MIL` (DenseNet)	**0.44** ± 0.02	**1.00** ± 0.00	0.73 ± 0.10	**0.73** ± 0.09	0.37 ± 0.07	0.00 ± 0.00	**0.79** ± 0.09	**0.44** ± 0.12	**0.61** ± 0.02
0.5	Baseline	0.12 ± 0.06	0.88 ± 0.15	0.28 ± 0.15	0.39 ± 0.10	0.21 ± 0.09	0.00 ± 0.00	0.37 ± 0.06	0.11 ± 0.07	0.32 ± 0.02
	`C-MIL` (ResNet)	**0.26** ± 0.14	0.99 ± 0.01	0.44 ± 0.05	**0.52** ± 0.11	**0.24** ± 0.12	0.00 ± 0.00	0.45 ± 0.24	0.15 ± 0.08	0.42 ± 0.08
	`C-MIL` (DenseNet)	0.25 ± 0.04	**1.00** ± 0.00	**0.53** ± 0.12	**0.52** ± 0.14	0.21 ± 0.02	0.00 ± 0.00	**0.51** ± 0.07	**0.30** ± 0.12	**0.45** ± 0.04
0.75	Baseline	0.04 ± 0.02	0.70 ± 0.16	0.12 ± 0.07	0.18 ± 0.07	0.08 ± 0.05	0.00 ± 0.00	0.18 ± 0.08	0.05 ± 0.05	0.19 ± 0.02
	`C-MIL` (ResNet)	**0.12** ± 0.11	0.95 ± 0.04	0.20 ± 0.05	0.38 ± 0.12	**0.14** ± 0.12	0.00 ± 0.00	**0.36** ± 0.19	0.10 ± 0.07	0.31 ± 0.08
	`C-MIL` (DenseNet)	0.11 ± 0.01	**0.99** ± 0.02	**0.29** ± 0.05	**0.43** ± 0.15	0.07 ± 0.06	0.00 ± 0.00	0.30 ± 0.06	**0.18** ± 0.08	**0.33** ± 0.02
0.9	Baseline	0.01 ± 0.01	0.56 ± 0.12	0.07 ± 0.06	0.07 ± 0.03	0.05 ± 0.04	0.00 ± 0.00	0.12 ± 0.08	0.03 ± 0.04	0.13 ± 0.02
	`C-MIL` (ResNet)	**0.05** ± 0.05	0.78 ± 0.06	0.12 ± 0.05	**0.26** ± 0.07	**0.08** ± 0.08	0.00 ± 0.00	**0.25** ± 0.15	0.10 ± 0.07	**0.22** ± 0.06
	`C-MIL` (DenseNet)	0.04 ± 0.01	**0.85** ± 0.08	**0.14** ± 0.04	0.24 ± 0.11	0.06 ± 0.04	0.00 ± 0.00	0.19 ± 0.03	**0.11** ± 0.04	**0.22** ± 0.01

performance of "Nodule" localization was the only one that dropped when training with `C-MIL`. This might be happening due to the fact that most nodule bounding boxes are very small, and since it is smaller than the other pathologies the generated consistency values are not significant enough to help improving the model, and the optimization method focuses on the bigger pathologies instead.

The ResNet-50 backbone shows better results on pathologies with smaller bounding boxes. On $T(IoR) \geq 0.5$, ResNet-50 performed better than the DenseNet-121 on localizing "Atelectasis", and with $T(IoR) \geq 0.1$ on localizing "Mass". The "Nodule" localization metric was also better using ResNet-50 on $T(IoR) = 0.1$.

Point localization accuracy values reported on Table 6.4 show how often the highest scoring patch is inside of the bounding box. The results of the point accuracy values seem consistent with the $T(IoR)$ evaluations. The performance is also higher than the baseline when training with `C-MIL` with the exception of the "Nodule" class, and it also presents higher values for the smaller pathologies like Mass and Atelectasis for ResNet-50 over the Densenet-121 backbone.

Table 6.4: Point localization accuracy of each pathology. The reported methods are our reimplementation of the method from Li et al. [33] (baseline) and our semi-supervised proposed approach `C-MIL` with ResNet-50 and DenseNet-121 as backbones.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Micro-Avg.
Baseline	0.28 ± 0.10	0.90 ± 0.14	0.53 ± 0.07	0.55 ± 0.10	0.39 ± 0.16	**0.05** ± 0.05	0.38 ± 0.06	0.21 ± 0.08	0.44 ± 0.04
`C-MIL` (ResNet)	**0.45** ± 0.13	0.99 ± 0.03	0.66 ± 0.05	0.63 ± 0.11	**0.43** ± 0.12	0.01 ± 0.02	0.62 ± 0.14	0.32 ± 0.05	0.55 ± 0.05
`C-MIL` (DenseNet)	0.39 ± 0.05	**1.00** ± 0.00	**0.69** ± 0.03	**0.69** ± 0.04	0.40 ± 0.06	0.00 ± 0.00	**0.66** ± 0.16	**0.40** ± 0.12	**0.57** ± 0.02

Table 6.5 shows the classification performance of `C-MIL` when using the multiple instance learning pooling mechanism defined in Equation 6.1 to generate classification scores. We also compare `C-MIL` with our baseline and with classification-only methods [63, 49]. `C-MIL` decreases the classification performance for every pathology, which is an evidence that the consis-

tency loss on the patch-level labels ends up prioritizing localization performance. However, the most significant drop happens on the non-annotated pathologies. The last 6 of the 14 pathologies (Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia) do not provide bounding box annotation, and the consistency loss for localization becomes completely unsupervised. It seems that without bounding box supervision, the consistency regularization affects the classification performance much more harshly. The classification weight in the combined loss of C-MIL is lower than the localization and classification ones, becoming a smaller part of the training objective.

Table 6.5: Classification results (AUC) of our implementations of the multiple instance models on the ChestX-ray14 test set with the classification model designed by Wang et al. [63] and the CheXNet [49] network. Results from related work as reported in the original papers.

Finding	Wang et al. [63]	CheXNet [49]	Baseline	C-MIL (ResNet)	C-MIL (DenseNet)
Atelectasis	0.716	**0.809**	0.762 ± 0.06	0.702 ± 0.04	0.710 ± 0.02
Cardiomegaly	0.807	**0.925**	0.856 ± 0.08	0.834 ± 0.06	0.852 ± 0.02
Effusion	0.784	**0.864**	0.818 ± 0.04	0.786 ± 0.02	0.771 ± 0.02
Infiltration	0.609	**0.734**	0.698 ± 0.02	0.657 ± 0.00	0.665 ± 0.02
Mass	0.706	**0.868**	0.791 ± 0.12	0.661 ± 0.08	0.715 ± 0.04
Nodule	0.671	**0.780**	0.753 ± 0.08	0.620 ± 0.04	0.650 ± 0.02
Pneumonia	0.633	**0.768**	0.734 ± 0.06	0.684 ± 0.04	0.692 ± 0.02
Pneumothorax	0.806	**0.889**	0.855 ± 0.06	0.761 ± 0.07	0.812 ± 0.03
Consolidation	0.708	**0.790**	0.739 ± 0.04	0.671 ± 0.02	0.663 ± 0.01
Edema	0.835	**0.888**	0.830 ± 0.05	0.715 ± 0.03	0.728 ± 0.01
Emphysema	0.815	**0.937**	0.856 ± 0.15	0.623 ± 0.03	0.664 ± 0.06
Fibrosis	0.769	**0.805**	0.783 ± 0.11	0.590 ± 0.02	0.612 ± 0.05
Pleural Thickening	0.708	**0.806**	0.755 ± 0.08	0.612 ± 0.04	0.612 ± 0.03
Hernia	0.767	**0.916**	0.809 ± 0.21	0.456 ± 0.01	0.538 ± 0.04
Average (14 classes)	0.738	**0.841**	0.788	0.669	0.692
Average (8 classes)	0.716	**0.828**	0.783	0.713	0.733

6.2.2 Qualitative analysis

Figure 6.3 shows some visual examples of the output score prediction of the baseline compared to C-MIL with a DenseNet-121 backbone. The patch score matrices were interpolated to match the image size. As stated before based on Table 6.3, the impact of the consistency regularization seems larger on higher threshold values, meaning that even though the baseline presents small intersection with the ground-truth, the consistency is helping to bring the larger and more confident prediction scores closer to the pathology. In some cases, C-MIL seems to learn beyond the labels as we can see that some inferences follow anatomical lines and shapes, predicting a better localization than the ground-truth bounding box, which is limited to being in a square shape. We also show two examples of classes that did not perform as well with C-MIL , “Nodule” and “Mass”. On the “Mass” example, there are multiple masses inside one bounding box, which is not very common on the dataset and might have confused the model.

The "Nodule" class performance is the weakest of C-MIL, which usually predicts high scores randomly over the lung area.

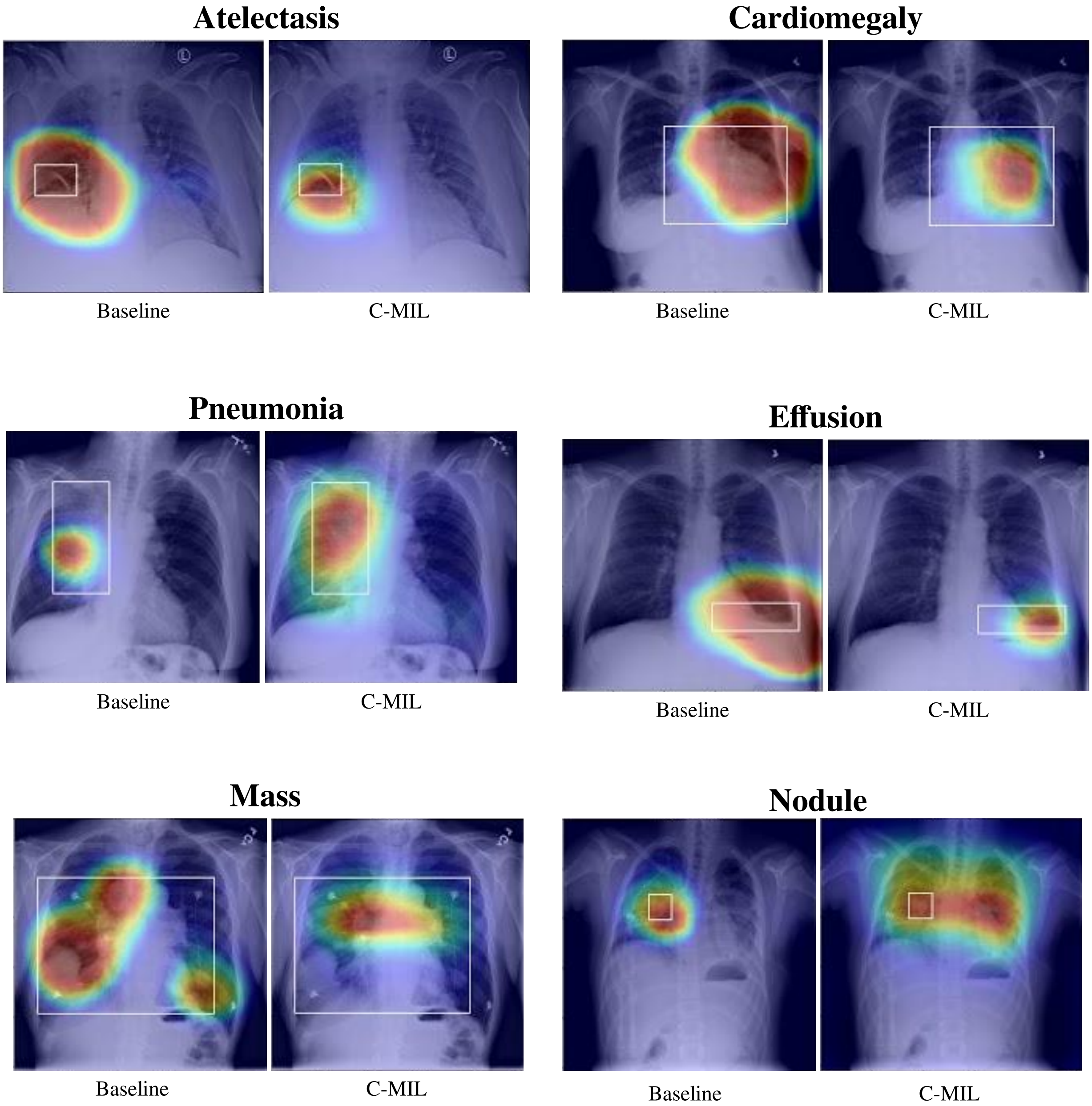


Figure 6.3: Visual comparison of both baseline and C-MIL regarding their pathology localization predictions (test set images). The white bounding box represent the annotated ground-truth and the heatmap colors represent the patch score intensity for the specified pathology. Images and annotations from the ChestX-ray14 dataset [63].

6.2.3 Ablation Study

Exponential Moving Average

We perform an ablation study to evaluate the impact of different aspects of `C-MIL` on its final performance. The first experiment removes the teacher model, which is updated via an EMA of the optimized model, and using the same model for computing the consistency, using only the transform function ϕ_t to apply some noise to the input image. We do that by replacing m_t with m_s on the original `C-MIL` consistency function. Equation 6.7 shows how the consistency loss term is computed in this case.

$$\mathcal{L}_{pcons} = ||m_s(x) - \phi(m_s(\phi(x))||^2 \tag{6.7}$$

As shown in Table 6.6, both training procedures have similar performance, presenting the same average result with both backbones. Using an EMA of the optimized model to compute consistency does not seem to improve the final result with the used hyperparameters. We speculate that using custom hyperparameters and schedulers for the EMA model could maybe boost the method's performance.

Table 6.6: Ablation study to assess the impact of keeping an exponential moving average (EMA) model as teacher, compared to using a single model to compute consistency. Reported values are point accuracy performance for each pathology. The reported methods are the original `C-MIL` and `C-MIL` without the EMA model. Experiments include both ResNet-50 and DenseNet-121 as backbones.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Micro-Avg.
`C-MIL` (ResNet) with EMA	**0.45** ± 0.13	**0.99** ± 0.03	**0.66** ± 0.05	**0.63** ± 0.11	**0.43** ± 0.12	0.01 ± 0.02	0.62 ± 0.14	**0.32** ± 0.05	**0.55** ± 0.05
`C-MIL` (ResNet) no EMA	0.43 ± 0.13	**0.99** ± 0.01	0.62 ± 0.06	**0.63** ± 0.05	0.31 ± 0.09	**0.03** ± 0.04	**0.71** ± 0.11	**0.32** ± 0.12	**0.55** ± 0.05
`C-MIL` (DenseNet) with EMA	0.39 ± 0.05	**1.00** ± 0.00	0.69 ± 0.03	**0.69** ± 0.04	**0.40** ± 0.06	0.00 ± 0.00	0.66 ± 0.16	**0.40** ± 0.12	**0.57** ± 0.02
`C-MIL` (DenseNet) no EMA	**0.40** ± 0.08	0.98 ± 0.03	**0.70** ± 0.09	0.67 ± 0.11	0.38 ± 0.08	**0.01** ± 0.02	**0.72** ± 0.15	0.33 ± 0.07	**0.57** ± 0.02

Adding CheXpert Data

In the experiments of Chapter 4, classification models trained on the CheXpert dataset generalized better to other datasets. Based on this result, we assume that CheXpert data is more representative of other datasets, and thus could improve the performance of the consistency regularization when used as unlabeled data. We make use of a subset of the CheXpert dataset, selecting one frontal image per patient, in a total of $65,000$ images to balance the number of samples from each dataset. Therefore, we add CheXpert data to our pipeline, using it to compute the consistency loss, discarding the original labels and using both NIH and CheXpert samples in the consistency loss, and only NIH data for the classification and localization loss terms. We also perform an experiment using NIH for classification and localization only, and using CheXpert unlabeled samples in the consistency loss. Table 6.7 shows the results of these experiments.

Table 6.7: Ablation study to include the use of CheXpert data [22] in the computation of the consistency loss. Reported values are point accuracy performance for each pathology. We performed this ablation with `C-MIL` without an EMA model and using only DenseNet-121 as backbone.

Method (Unlabeled data)	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Micro-Avg.
`C-MIL` (NIH Only)	0.40 ± 0.08	0.98 ± 0.03	**0.70** ± 0.09	**0.67** ± 0.11	0.38 ± 0.08	0.01 ± 0.02	**0.72** ± 0.15	**0.33** ± 0.07	**0.57** ± 0.02
`C-MIL` (NIH + CheXpert)	**0.41** ± 0.09	**0.99** ± 0.01	0.66 ± 0.05	0.58 ± 0.12	**0.48** ± 0.04	**0.10** ± 0.02	0.65 ± 0.22	0.30 ± 0.18	0.55 ± 0.06
`C-MIL` (CheXpert only)	0.38 ± 0.07	**0.99** ± 0.01	0.67 ± 0.03	0.66 ± 0.09	0.41 ± 0.05	0.05 ± 0.04	0.65 ± 0.10	0.14 ± 0.07	0.53 ± 0.03

The results of this experiment show that despite having a lower average performance, adding CheXpert data helps increasing the performance of the pathologies that `C-MIL` performed worse, “Mass” and “Nodule”. We speculate that maybe the increase in the frequency of nodule positive samples might have increased the importance of the “Nodule” class in the consistency loss, boosting the localization performance of this pathology and achieving better results than our baseline (Table 6.4). Using only CheXpert data still presented a slight improvement in those metrics, but had a lower performance on the remaining pathologies.

7. CONCLUSIONS

In this dissertation, we argued on how semi-supervised learning methods can make use of unlabeled data when training deep learning models with limited annotations available. We explained the task of pathology localization on chest radiographs and used it as an example of a relevant medical imaging task with abundant publicly available data but limited available annotations.

First, we evaluated the available large public datasets, showing how a model trained over one large-scale dataset generalizes to another. We showed that radiologist-level performing models can display performance drop on unseen data if the training set is not representative enough. In this experiment, we also found that some large public datasets were more representative of others. For instance, classification models trained over the CheXpert dataset performed well on unseen data from the ChestX-ray14 dataset, but we experienced performance drops when doing it the other way around.

Next, we defined and compared state-of-the art semi-supervised methods in a chest radiograph classification scenario. These methods were originally developed to perform multi-class classification and were only benchmarked on natural images. In our work, we extended those methods to also perform multi-label classification and compared them in a medical image classification problem. Based on the experiments, we identified Mean Teacher as the best-performing method for semi-supervised medical imaging classification. With Mean Teacher, we achieved state-of-the-art performance when comparing to previous work that developed methods for classification of chest radiographs with limited supervision.

Finally, we introduced C-MIL by extending the Mean Teacher method to a multiple instance localization framework. Our method uses the key concepts in the Mean Teacher approach, mainly consistency regularization and self-ensembling, to extend the use of non-annotated data on a multiple instance learning scenario. Comparing to a supervised baseline, our experimental analysis showed superior performance of C-MIL for almost all pathologies when evaluating localization metrics and, therefore, evidencing that semi-supervised mechanisms such as consistency regularization, can improve the results of a multiple instance learning scenario.

Notwithstanding, C-MIL had a negative impact on classification performance. However, when considering a clinical setting, using C-MIL as a classification model might provide better insights to justify the prediction of the model, presenting an advantage regarding its explainability power, since the classification scores are computed directly on patch scores, indicating exactly which portions of the image contributed to the score of that particular class.

As future work, we want to explore C-MIL in other deep learning scenarios with limited annotated data, such as localization and segmentation of pathologies in other medical imaging modalities and common object detection in natural images. We also want to explore how to mitigate the weaknesses of C-MIL and whether extending mechanisms from other state-of-the-art semi-supervised methods can improve its performance, such as pseudo-labels, MixUp regularization, and stronger augmentation policies as well as proposing different ways to convert the localization labels into classification labels.

7.1 Limitations

Reproducing the baseline performance was our greatest limitation. We did not achieve the results reported by the original work in our re-implementation, and therefore we did not directly compare `C-MIL`'s performance improvement to the state-of-the-art pathology localization methods. This limitation could be solved with a public release of the original code, since their work was not entirely reproducible using the information provided by the paper.

In addition, we believe a more thorough hyperparameter search could improve `C-MIL` performance, but this is very computationally demanding and due to hardware limitations we could not perform a more extensive hyperparameter search. The results we have achieved in this dissertation may not be a true display of our proposed approach's real performance due to our inability in properly optimizing its hyperparameters.

REFERENCES

[1] Amir, G. J.; Lehmann, H. P. “After detection: The improved accuracy of lung cancer assessment using radiologic computer-aided diagnosis”, *Academic Radiology*, vol. 23–2, Feb 2016, pp. 186–191.

[2] Araujo, L. H.; Baldotto, C.; de Castro Jr, G.; Katz, A.; Ferreira, C. G.; Mathias, C.; Mascarenhas, E.; de Lima Lopes, G.; Carvalho, H.; Tabacof, J.; Martínez-Mesa, J.; de Souza Viana, L.; de Souza Cruz, M.; Zukin, M.; Marchi, P. D.; Terra, R. M.; Ribeiro, R. A.; de Lima, V. C. C.; Werutsky, G.; Barrios, C. H. “Lung cancer in Brazil”, *Jornal Brasileiro de Pneumologia*, vol. 44–1, Feb 2018, pp. 55–64.

[3] Aviles-Rivero, A. I.; Papadakis, N.; Li, R.; Sellars, P.; Fan, Q.; Tan, R. T.; Schönlieb, C.-B. “GraphXNET chest X-ray classification under extreme minimal supervision”. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 504–512.

[4] Babenko, B. “Multiple instance learning: Algorithms and applications”, Technical Report, University of California, 2008, 19p.

[5] Baltruschat, I. M.; Nickisch, H.; Grass, M.; Knopp, T.; Saalbach, A. “Comparison of deep learning approaches for multi-label chest X-ray classification”, *Scientific Reports*, vol. 9–1, Dec 2019, pp. 6381.

[6] Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. “Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring”, *CoRR*, vol. 1911.09785, Nov 2019, pp. 13.

[7] Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; Raffel, C. “Mixmatch: A holistic approach to semi-supervised learning”. In: Annual Conference on Neural Information Processing Systems, 2019, pp. 5050–5060.

[8] Bustos, A.; Pertusa, A.; Salinas, J. M.; de la Iglesia-Vayá, M. “Padchest: A large chest X-ray image dataset with multi-label annotated reports”, *CoRR*, vol. 1901.07441, Dec 2019, pp. 35.

[9] Cheplygina, V.; de Bruijne, M.; Pluim, J. P. W. “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis”, *Medical Image Analysis*, vol. 54, May 2019, pp. 280–296.

[10] Cubuk, E. D.; Zoph, B.; Mané, D.; Vasudevan, V.; Le, Q. V. “Autoaugment: Learning augmentation policies from data”, *CoRR*, vol. 1805.09501, Aug 2018, pp. 14.

[11] Cubuk, E. D.; Zoph, B.; Shlens, J.; Le, Q. V. “Randaugment: Practical data augmentation with no separate search”, *CoRR*, vol. 1909.13719, Oct 2019, pp. 13.

[12] De Lacey, G.; Morley, S.; Berman, L. “The Chest X-ray: A Survival Guide E-Book”. Elsevier Health Sciences, 2012, 384p.

[13] del Ciello, A.; Franchi, P.; Contegiacomo, A.; Cicchetti, G.; Bonomo, L.; Larici, A. R. “Missed lung cancer: when, where, and why?”, *Diagnostic and Interventional Radiology*, vol. 23–2, Mar 2017, pp. 118–126.

[14] Gibbs, J. M.; Chandrasekhar, C. A.; Ferguson, E. C.; Oldham, S. A. “Lines and stripes: where did they go? From conventional radiography to CT”, *Radiographics*, vol. 27–1, Jan 2007, pp. 33–48.

[15] Girshick, R. “Fast R-CNN”. In: IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[16] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[17] Goodfellow, I.; Bengio, Y.; Courville, A. “Deep Learning”. MIT Press, 2016, 800p.

[18] He, K.; Zhang, X.; Ren, S.; Sun, J. “Deep residual learning for image recognition”. In: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[19] Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al.. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”, *IEEE Signal Processing Magazine*, vol. 29–6, October 2012, pp. 82–97.

[20] Hirsch, F. R.; Franklin, W. A.; Gazdar, A. F.; Bunn, P. A. “Early detection of lung cancer: clinical perspectives of recent advances in biology and radiology”, *Clinical Cancer Research*, vol. 7–1, Jan 2001, pp. 5–22.

[21] Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. “Densely connected convolutional networks”. In: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[22] Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R. L.; Shpanskaya, K. S.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; Ng, A. Y. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: AAAI Conference on Artificial Intelligence, 2019, pp. 590–597.

[23] Jeong, J.; Lee, S.; Kim, J.; Kwak, N. “Consistency-based semi-supervised learning for object detection”. In: Annual Conference on Neural Information Processing Systems, 2019, pp. 10758–10767.

[24] Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.; Mark, R. G.; Horng, S. “MIMIC-CXR: A large publicly available database of labeled chest radiographs”, *CoRR*, vol. 1901.07042, Feb 2019, pp. 7.

[25] Kingma, D. P.; Ba, J. “Adam: A method for stochastic optimization”. In: International Conference on Learning Representations, 2015, pp. 15.

[26] Krizhevsky, A.; Hinton, G.; et al.. “Learning multiple layers of features from tiny images”, Technical Report, Citeseer, 2009, 60p.

[27] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. “Imagenet classification with deep convolutional neural networks”. In: Annual Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[28] Laine, S.; Aila, T. “Temporal ensembling for semi-supervised learning”. In: International Conference on Learning Representations, 2017, pp. 13.

[29] LeCun, Y.; Bengio, Y.; Hinton, G. “Deep learning”, *Nature*, vol. 521–7553, May 2015, pp. 436–444.

[30] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, vol. 86–11, November 1998, pp. 2278–2324.

[31] Lee, D.-H. “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks”. In: ICML Workshop on challenges in representation learning, 2013, pp. 1–6.

[32] Leung, M. K. K.; Xiong, H. Y.; Lee, L. J.; Frey, B. J. “Deep learning of the tissue-regulated splicing code”, *Bioinformatics*, vol. 30–12, Jun 2014, pp. 121–129.

[33] Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.-J.; Fei-Fei, L. “Thoracic disease identification and localization with limited supervision”. In: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8290–8299.

[34] Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. “Focal loss for dense object detection”. In: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

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

[36] Liu, J.; Zhao, G.; Fei, Y.; Zhang, M.; Wang, Y.; Yu, Y. “Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision”. In: IEEE International Conference on Computer Vision, 2019, pp. 10632–10641.

[37] Liu, Q.; Yu, L.; Luo, L.; Dou, Q.; Heng, P. "Semi-supervised medical image classification with relation-driven self-ensembling model", *IEEE Trans. Medical Imaging*, vol. 39–11, May 2020, pp. 3429–3440.

[38] McCollough, C. H.; Bushberg, J. T.; Fletcher, J. G.; Eckel, L. J. "Answers to common questions about the use and safety of ct scans". In: Mayo Clinic Proceedings, 2015, pp. 1380–92.

[39] McCulloch, W. S.; Pitts, W. H. "A logical calculus of the ideas immanent in nervous activity". In: *The Philosophy of Artificial Intelligence*, Oxford University Press, 1990, pp. 22–39.

[40] Mitchell, T. M. "Machine learning, International Edition". McGraw-Hill, 1997, 432p.

[41] Mountain, C. F. "Revisions in the international system for staging lung cancer", *Chest*, vol. 111–6, Jun 1997, pp. 1710–1717.

[42] Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y. "Reading digits in natural images with unsupervised feature learning". In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011, pp. 9.

[43] Oakden-Rayner, L. "Exploring large scale public medical image datasets", *CoRR*, vol. 1907.12720, Aug 2019, pp. 9.

[44] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. "Pytorch: An imperative style, high-performance deep learning library". In: Annual Conference on Neural Information Processing Systems, 2019, pp. 8024–8035.

[45] Pooch, E. H. P.; Ballester, P.; Barros, R. C. "Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification". In: Thoracic Image Analysis, 2020, pp. 74–83.

[46] Pooch, E. H. P.; Ballester, P.; Barros, R. C. "Semi-supervised classification of chest radiographs". In: Interpretable and Annotation-Efficient Learning for Medical Image Computing, 2020, pp. 172–179.

[47] Preechakul, K.; Sriswasdi, S.; Kijsirikul, B.; Chuangsuwanich, E. "High resolution weakly supervised localization architectures for medical images", *CoRR*, vol. 2010.11475, Oct 2020, pp. 6.

[48] Qi, G.; Luo, J. "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods", *CoRR*, vol. 1903.11260, Mar 2019, pp. 24.

[49] Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D. Y.; Bagul, A.; Langlotz, C.; Shpanskaya, K. S.; Lungren, M. P.; Ng, A. Y. "Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning", *CoRR*, vol. 1711.05225, Jul 2017, pp. 7.

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

[51] Ren, S.; He, K.; Girshick, R. B.; Sun, J. "Faster R-CNN: towards real-time object detection with region proposal networks". In: Annual Conference on Neural Information Processing Systems, 2015, pp. 91–99.

[52] Rozenberg, E.; Freedman, D.; Bronstein, A. "Localization with limited annotation for chest X-rays". In: NeurIPS Workshop on Machine Learning for Health, 2020, pp. 52–65.

[53] Rudin, C. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, vol. 1–5, May 2019, pp. 206–215.

[54] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Fei-Fei, L. "Imagenet large scale visual recognition challenge", *International Journal of Computer Vision*, vol. 115–3, Apr 2015, pp. 211–252.

[55] Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S. Q.; Nguyen, C. D.; Ngo, V.-D.; Seekins, J.; Blankenberg, F. G.; Ng, A. Y.; Lungren, M. P.; Rajpurkar, P. "Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation", *medRxiv*, vol. 2021.02.28.21252634, Feb 2021, pp. 34.

[56] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[57] Shah, P. K.; Austin, J. H. M.; White, C. S.; Patel, P.; Haramati, L. B.; Pearson, G. D. N.; Shiau, M. C.; Berkmen, Y. M. "Missed non–small cell lung cancer: Radiographic findings of potentially resectable lesions evident only in retrospect", *Radiology*, vol. 226, Jan 2003, pp. 235–41.

[58] Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; Li, C. "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: Annual Conference on Neural Information Processing Systems, 2020, pp. 13.

[59] Tarvainen, A.; Valpola, H. “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”. In: Annual Conference on Neural Information Processing Systems, 2017, pp. 1195–1204.

[60] Torralba, A.; Efros, A. A.; et al.. “Unbiased look at dataset bias.” In: Conference on Computer Vision and Pattern Recognition, 2011, pp. 7.

[61] van Engelen, J. E.; Hoos, H. H. “A survey on semi-supervised learning”, *Machine Learning*, vol. 109–2, Nov 2019, pp. 373–440.

[62] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. “Attention is all you need”. In: Annual Conference on Neural Information Processing Systems, 2017, pp. 5998–6008.

[63] Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R. M. “Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 3462–3471.

[64] Xie, Q.; Dai, Z.; Hovy, E. H.; Luong, T.; Le, Q. “Unsupervised data augmentation for consistency training”. In: Annual Conference on Neural Information Processing Systems, 2020, pp. 6256–6268.

[65] Zhang, H.; Cissé, M.; Dauphin, Y. N.; Lopez-Paz, D. “Mixup: Beyond empirical risk minimization”. In: International Conference on Learning Representations, 2018, pp. 13.

[66] Zhang, J.; Lin, Z. L.; Brandt, J.; Shen, X.; Sclaroff, S. “Top-down neural attention by excitation backprop”. In: European Conference on Computer Vision, 2016, pp. 543–559.

[67] Zhu, X.; Goldberg, A. B. “Introduction to semi-supervised learning”, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3–1, Jun 2009, pp. 1–130.