

BRIDGING THE GAP: LEVERAGING MACHINE LEARNING FOR IMPROVED DATA INTERPRETATION AND DECISION MAKING UNDER UNCERTAINTY

by

Artin Majdi

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY

In the Graduate College
UNIVERSITY OF ARIZONA

2023

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Artin Majdi, titled *Bridging the Gap: Leveraging Machine Learning for Improved Data Interpretation and Decision Making under Uncertainty*, and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Jeffrey J. Rodriguez

Date

Carlos Alsua

Date

Abhijit Mahalanobis

Date

Janet M. Roveda

Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Jeffrey J. Rodriguez
Dissertation Director

Date

STATEMENT by AUTHOR

This dissertation *Bridging the Gap: Leveraging Machine Learning for Improved Data Interpretation and Decision Making under Uncertainty* prepared by Artin Majdi has been submitted in partial fulfillment of requirements for a doctoral degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under the rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by

SIGNED: _____

ACKNOWLEDGMENTS

As I reflect upon the completion of this invigorating journey, my heart brims with profound gratitude towards those who journeyed with me, imparting their wisdom, support, and guidance at every step.

First and foremost, my deepest appreciation is directed towards my mentor, Dr. Jeffrey J. Rodriguez, whose steadfast guidance and support throughout the project has been unparalleled. His expertise, dedication, and continual encouragement have been priceless to me, for which I am eternally thankful.

My sincere gratitude goes out to Mr. Nirav Merchant, Ms. Maliaca Oxnam, and Drs. Carlos Alsua, Manoj Saranathan, and Mahesh Keerthivasan, whose unceasing support has been invaluable throughout this endeavor. Additionally, I wish to convey my gratitude to Drs. Janet Roveda and Abhijit Mahalanobis for their insightful contributions and feedback which significantly enhanced the depth and quality of my work. Your invaluable support and mentorship have played a crucial role in my personal and academic evolution.

An acknowledgment wouldn't be complete without a heartfelt tribute to my family. Your unwavering faith in my capabilities, your patience during the tough times, and your consistent encouragement have served as my strongest pillars of support.

Contents

List of Figures	6
List of Tables	7
1 Leveraging Disease Taxonomy for Enhanced Multi-Label Classification in Chest Radiography	8
1.1 Introduction	9
1.2 Related Work	10
1.3 Methods	12
1.3.1 Problem Formulation	12
1.3.2 Label Taxonomy Structure	16
1.3.3 Approach 1: Conditional Predicted Probability	17
1.3.4 Approach 2: Conditional Loss	18
1.3.5 Updating Loss Values and Predicted Probabilities	23
1.3.6 Experimental Setup	26
1.4 Results	29
1.5 Discussion and Conclusion	38
References	41

List of Figures

1.1	Taxonomy Structure of Lung Pathologies in Chest Radiographs	30
1.2	Comparative Analysis of ROC Curves for Nine Thoracic Pathologies: 'Logit', 'Loss', and Baseline Techniques	34
1.3	Heatmap Visualization of Model Performance Metrics (ACC, AUC, F1) for Different Techniques across Pathologies	38

List of Tables

1.1	Representation of Pathologies Across Datasets	31
1.2	Sample Distribution Per Pathology in Evaluated Datasets (CheX, NIH, and PC)	32
1.3	Statistical Performance Comparison of “Logit”, “Loss”, and “Baseline” Techniques Across Various Pathologies	36

Chapter 1

Leveraging Disease Taxonomy for Enhanced Multi-Label Classification in Chest Radiography

Accurate diagnosis of thoracic diseases from chest radiographs is a challenging task that can lead to diagnostic errors and negative patient outcomes. This study introduces two novel hierarchical multi-label classification methods, leveraging the taxonomy of pathologies to boost both the accuracy and interpretability of disease classifications. These methods cater to scenarios where ground truth is accessible (termed “loss”) as well as when it isn’t (termed “logit”). By utilizing disease taxonomy, the proposed methods acknowledge the interrelationships among diseases, thereby enhancing their adaptability to new tasks. The “logit” method offers ease of integration with existing pre-trained models, eliminating the need for re-optimization and ensuring extensive applicability. Conversely, the “loss” method modifies the loss function during the training phase, thus providing an avenue for integration into the current training process. The proposed techniques were evaluated on three publicly accessible, diverse chest radiograph datasets, namely CheXpert, PadChest, and NIH Chest-Xray14 as well as various statistical tests. The results underpin the significant enhancement in accuracy and interpretability these methods provide in the diagnosis of thoracic diseases in chest radiography. This approach has the potential to promote an accurate and efficient diagnosis by providing an additional layer of decision support to radiologists, ultimately leading to better patient outcomes.

KEYWORDS: Chest radiography, hierarchical classification, disease taxonomy, multilabel classification, conditional loss function, diagnostic errors, machine learning, medical imaging

1.1 Introduction

Chest X-ray (CXR) is a prevalent radiological examination for diagnosing lung and heart disorders, constituting a significant share of ordered imaging studies. Fast and accurate detection of different thoracic diseases, such as pneumothorax, is crucial for optimal patient care [1]. However, interpreting CXRs can be challenging due to similarities between different thoracic diseases, which may result in misinterpretation even by experienced radiologists [2]. Consequently, devising an accurate system to identify and localize common thoracic diseases can aid radiologists in minimizing diagnostic errors [3,4]. Progress in natural language processing (NLP) has enabled the collection of extensive annotated datasets such as ChestX-ray8 [5], PADCHEST [6], and CheXpert [7], allowing researchers to develop more efficient and robust supervised learning algorithms. Convolutional neural networks (CNNs) exhibit potential for learning intricate relationships between image objects. However, their training necessitates vast amounts of labeled data, which can be both expensive and time-consuming to acquire. Despite these challenges, deep learning techniques have become increasingly popular in medical imaging, especially in radiology, due to their ability to perform complex tasks with minimal human intervention [8]. The timely diagnosis and effective treatment of diseases depend on the fast and accurate detection of anomalies in medical images. Deep learning techniques have made substantial progress in the medical imaging domain, exhibiting impressive success across various applications [9,10]. Although recent advances in deep learning have facilitated the creation of CAD systems capable of classifying and localizing prevalent thoracic diseases using CXR images, most of these techniques have concentrated on specific diseases [11,12,13,14], leaving ample opportunities to investigate a unified deep learning framework that can efficiently detect a broad spectrum of common thoracic diseases. Further, conventional classification methods are primarily designed for single-label predictions and struggle with multi-label classification, which requires predicting multiple labels for each input sample. In multi-label classification, common methods like the One-vs-All (OVA) approach exhibit limitations, including high computational complexity and an inability to capture intricate label relationships [15]. This paper aims to tackle the challenges of multi-label classification by introducing a hierarchical framework that incorporates the relationships between different classes to provide a more accurate classification framework. We propose one approach termed as “loss-based” for scenarios where ground truth is available, in which the proposed technique is applied to the loss function of a network (e.g., a

classification or segmentation network such as DenseNet121 [16] or U-Net [17]). For scenarios where ground truth is not available, we propose an alternative approach termed as “logit-based”, where the hierarchical framework is applied to the logit values of an existing pre-trained network. Logits are the output of the last layer of a neural network before applying the activation function. For multi-class problems with K classes, the number of logits is K , and the value of each logit represents the model’s confidence in the k -th class being positive. For example, consider a binary classification problem where one needs to determine if an email is spam. In that case, the logit will be a single value representing the confidence that the email is spam. The higher the value of the logit, the more confident the model is that the email is spam. The logit-based technique provides a transfer learning approach that improves classification accuracy without necessitating an extensive computational investment. The rest of this paper is structured as follows. Section 1.2 discusses related work on multi-label classification and hierarchical loss functions; Section 1.3 describes the proposed techniques for integrating label hierarchy into multi-label classification techniques; Section 1.4 presents experimental results using the chest radiograph dataset; and Section 1.5 concludes the paper and outlines future research directions.

1.2 Related Work

The introduction of the ChestX-ray8 dataset and its associated model [5] marked a significant advancement in large-scale CXR classification, leading to numerous improvements in both modeling and dataset collection. These enhancements include the integration of ensemble methods [18], attention mechanisms [19, 20], and localization techniques [21, 22, 23, 24]. Most early approaches use “binary relevance” (BR) learning, which reduces the multi-label classification problem to binary classification by training a binary classifier for each class [25]. However, BR-based techniques do not account for label dependence—either conditional (instance-specific label dependence) where in a given instance the presence or absence of one label may impact another, or marginal (dataset-specific label dependence) where certain labels may co-occur more frequently [26]. Multi-label classification, unlike multi-class methods, classifies instances into multiple categories simultaneously. For example, a single chest radiograph image can have both edema and cardiomegaly [15, 27]. Significant research on integrating taxonomies through hierarchical classification was conducted prior to the advent of deep learning by extracting a set of binary hierarchical multi-label classification (HMLC) labels from pseudo-probability

predictions [28]. Early methods used hierarchical and multi-label generalizations of traditional algorithms, such as nearest-neighbor or multi-layer perceptron [29] and decision trees [30]. With the rise of deep learning, the adaptation of convolutional neural networks (CNN) for hierarchical classification has gained increasing attention [31, 32, 33, 34].

Hierarchical Multi-Label Classification Technique: In many cases, the diagnosis or observation of a particular condition on a CXR (or other medical imaging data) is dependent on the presence or absence of the parent class [35]. For example, if a radiologist is trying to diagnose pneumonia in a patient, they may first look for evidence of lung consolidation (parent label) in the CXR. Consequently, it is possible to make more accurate diagnoses by taking into account the relationship between labels. However, many existing CXR classification methods do not consider the dependence between labels and instead treat each label independently. These algorithms are known as “flat classification” methods [36]. Furthermore, some labels at the lower levels of the hierarchy, specifically leaf nodes, have very few positive examples, making the flat learning model susceptible to negative class bias. To address these issues, we must create a model that considers the hierarchical nature of the CXR.

Hierarchical multi-label classification methods have been successfully implemented in a variety of domains, including text processing [37], visual recognition [38], and genomic analysis [28]. A common technique [39] for exploiting such a hierarchy is to train a classifier on conditional data while ignoring all samples with negative parent-level labels and then reintroducing these samples to fine-tune the network across the entire dataset [39]. These approaches help the classifier focus on the relevant data during initial training, thus improving the prediction accuracy. However, these techniques are computationally expensive, as they require training a classifier on conditional data and then fine-tuning it on a full dataset. This makes them difficult to apply to real-world problems, where the amount of data is often very large. Another common strategy is a cascading architecture where different classifiers are trained at each level of the hierarchy. Although these techniques enable more granular data analysis (each classifier can focus on a specific level of the hierarchy), they require a substantial amount of computational resources. Other existing deep learning-based approaches often use complex combinations of CNNs and recurrent neural networks (RNNs) [31, 32].

1.3 Methods

In this study, we introduce a unique method that improves the accuracy and interpretability of multi-label classification, with potential applications in areas such as chest radiography. We propose two distinct strategies. The first strategy termed as “loss-based”, requiring the availability of ground truth labels, incorporates the hierarchical relationships among different classes directly into the loss function. In contrast, the second strategy termed as “logit-based” utilizes these hierarchical relationships to modify the logit values before calculating the predicted probabilities for each class. These two strategies, which utilize a transfer learning approach, foster the use and fine-tuning of pre-existing models, thereby expanding their adaptability to new tasks. By improving the accuracy of classifying different pathologies, these techniques could potentially enhance disease diagnosis and treatment. The proposed technique is adaptable to the available computational resources. When ample computational resources are available, the “loss-based” strategy can be utilized. Alternatively, in scenarios with limited computational resources, to avoid the need for optimization of the network from scratch, the “logit-based” strategy can be utilized. One notable advantage of our proposed techniques lies in enhancing interpretability. By categorizing classes into a hierarchical structure and capitalizing on their relationships, the model not only improves classification performance but also provides insights into the relationships among predicted classes. This additional layer of interpretability can help radiologists in understanding the reasoning behind the model’s predictions, fostering trust in the model’s output and facilitate its integration into clinical workflows. Furthermore, the hierarchical nature of the taxonomy allows radiologists to explore predictions at various levels of granularity, depending on the level of detail required for a specific case.

1.3.1 Problem Formulation

Mathematical Formulation of Sigmoid Function

In the context of neural networks, a logit refers to the raw, unscaled output of a neuron. This output is obtained at the last layer of a neural network model prior to the application of the sigmoid layer [40]. Logit values can range from negative to positive infinity. The term “logit” originally comes from logistic regression, and it is the inverse of the logistic sigmoid function. In machine learning, it’s often desirable for our model to produce real numbers ranging from 0 to 1. Applying the sigmoid function to

the logit ensures this, as the sigmoid function maps any real number to the interval $[0, 1]$. The equation representing the sigmoid function is:

$$p = \text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (1.3.1)$$

When we apply this sigmoid function to the logit values produced by the neural network, the result is a predicted probability ranging from 0 to 1. This property is particularly useful in binary classification tasks, where the aim is to model the probability of a given input pertaining to a certain class. In a binary classification scenarios, if we apply the sigmoid function to the logit value and obtain output p , we interpret this as the model's estimated probability that the input belongs to the class. Finally, the equation for the logit (also known as the log-odds) can be given as

$$x = \text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (1.3.2)$$

where p is the probability of a positive event. This function maps a probability p from the interval $(0, 1)$ to any real number.

Glossary of Symbols

Let us define the following parameters:

- $C = \{c_k\}_{k=1}^K$: the set of classes (categories) in the multi-label dataset, where c_k is the name of the k -th class.
- \mathcal{E} : set of edges representing parent-child relationships between classes.
- $\mathcal{G} = \{C, \mathcal{E}\}$: Graph representing the taxonomy of thoracic diseases.
- $c_j = \Lambda(c_k) \in C$: parent class of class c_k in graph \mathcal{G} .
- $\mathcal{J}(c_j) \subset C$: set of child classes of class c_j in graph \mathcal{G}
- $y_k^{(i)} \in \{0, 1\}$: true label for the k -th class of instance i .

- $q_k^{(i)} \in (-\infty, 0)$: logits obtained in the last layer of the neural network model before the sigmoid layer.
- $p_k^{(i)} = \text{sigmoid}(q_k^{(i)}) = \frac{1}{1+\exp(-q_k^{(i)})}$: predicted probability for the k -th class (c_k) of instance i with a value between 0 and 1. $p_k^{(i)}$ represents the likelihood that class k is present in instance i and is obtained by passing logits $q_k^{(i)}$ through a sigmoid function.
- θ_k : Binarization threshold for class k . To obtain this, we can utilize any existing thresholding technique (for example, in one technique, we analyze the ROC curve and find the corresponding threshold where the difference between the true positive rate (sensitivity) and false positive rate (1-specificity) is maximum; alternatively, we could simply use 0.5).
- $t_k^{(i)} = \begin{cases} 1 & \text{if } p_k^{(i)} \geq \theta_k \\ 0 & \text{otherwise.} \end{cases}$: predicted label obtained by binarizing the $p_k^{(i)}$
- $\hat{p}_k^{(i)} \in (0, 1)$: updated predicted probability for the k -th class of instance i with a value between 0 and 1.
- $\hat{t}_k^{(i)} = \begin{cases} 1 & \text{if } \hat{p}_k^{(i)} \geq \theta_k \\ 0 & \text{otherwise.} \end{cases}$: updated predicted label for the k -th class of instance i .
- K : number of categories (aka classes) in a multi-class, multi-label problem. For example, suppose that we have a dataset that is labeled for the presence of cats, dogs, and rabbits in any given image. If a given image $X^{(i)}$ has cats and dogs but not rabbits, then $Y^{(i)} = \{1, 1, 0\}$.
- N : Number of instances.
- $X^{(i)}$: Data for instance i .
- $Y^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \dots, y_K^{(i)}\}$: True label set for instance i . For example, consider a dataset that is labeled for the presence of cats, dogs, and rabbits in any given instance. If a given instance $X^{(i)}$ has cats and dogs but not rabbits, then $Y^{(i)} = \{1, 1, 0\}$.

- $P^{(i)} = \{p_k^{(i)}\}_{k=1}^K$: Predicted probability set obtained in the output of the classifier $F(\cdot)$ representing the probability that each class k is present in the sample.
- $T^{(i)} = \{t_k^{(i)}\}_{k=1}^K$: predicted label set for instance i .
- $\mathbb{X} = \{X^{(i)}\}_{i=1}^N$: Set of all instances.
- $\mathbb{Y} = \{Y^{(i)}\}_{i=1}^N$: Set of all true labels.
- $\mathbb{D} = \{\mathbb{X}, \mathbb{Y}\}$: Dataset containing all instances and all true labels.
- $l_k^{(i)} = \mathcal{L}(y_k^{(i)}, p_k^{(i)})$: $\mathcal{L}(\cdot)$ is an arbitrary loss function (e.g., binary cross entropy) that takes the true label $y_k^{(i)}$ and predicted probability $p_k^{(i)}$ for class k and instance i and outputs the loss value $l_k^{(i)}$. We refer to this as the “base loss function” throughout this paper.
- $\text{Loss}(\theta)$: Measured loss for all classes and instances. This value is obtained using a modified version of the base loss function $\mathcal{L}(\cdot)$ (e.g., with added regularization, etc.).
- $\omega_k^{(i)}$: Estimated weight for k -th class c_k of instance i with respect to its parent class Γ_k .
- $\hat{l}_k^{(i)} = \omega_k^{(i)} l_k^{(i)}$: updated loss for class k and instance i .

Let us define the multi-label classification problem as follows. Let $\mathbb{X} = \{X^{(i)}\}_{i=1}^N$ be a set of N chest radiograph images and $\mathbb{Y} = \{Y^{(i)}\}_{i=1}^N$ be their corresponding ground truth labels. The ground-truth labels for the dataset were provided by experienced radiologists who annotated each image with the corresponding abnormalities. Given the set of disease classes $C = \{c_1, c_2, \dots, c_K\}$, let us define a graph $\mathcal{G} = \{C, \mathcal{E}\}$ representing the taxonomy of thoracic diseases, where \mathcal{E} is the set of edges representing parent-child relationships between these classes. For each node $c_k \in C$, let Λ_k be the parent node of class c_k and let $\mathcal{J}_k \subset C$ be the set of child classes of class c_k in graph \mathcal{G} .

In the context of multi-label classification problems, each sample may have multiple labels assigned to it simultaneously. To this end, we use a deep neural network, with multiple hidden layers and the sigmoid activation function in the final layer. Let's denote the input to this neural network by $x^{(i)}$, which represents the instance i 's data (data type can be a 1D feature vector, 2D image, or 3D volume). This network is trained to predict the probabilities for each class being present in a given sample.

Hence, the output of the final layer of the neural network for instance i is passed through a sigmoid function to generate a set of values, each ranging from 0 to 1, corresponding to the label set \mathcal{C} . The outcome of this operation is a set of K predicted probabilities $P^{(i)} = \{p_k^{(i)}\}_{k=1}^K$. Each of these predicted probabilities, derived from the sigmoid activation function, can be interpreted as the likelihood that the input sample belongs to each class. Furthermore, let $\omega_k^{(i)}$ be a scalar weight assigned to the class c_k of instance i with respect to its parent class Λ_k . Each of these predicted probabilities, derived from the sigmoid activation function, can be interpreted as the likelihood that the input sample belongs to each class. A loss function is utilized to quantify the similarity between predicted probabilities and true labels. This function guides the learning process of the neural network by providing a measure of the prediction error, which is minimized during the training phase. Let us denote the loss value as $l_k = \mathcal{L}(p_k^{(i)}, y_k^{(i)})$, $k \in \{1, 2, \dots, K\}$ where $\mathcal{L}(\cdot)$ is an appropriate single-class loss function for the task (e.g., binary cross-entropy, Dice, etc.) that is used to calculate the difference between the predicted probability $p_k^{(i)}$ and the true class label $y_k^{(i)}$ for instance i and class k .

1.3.2 Label Taxonomy Structure

To exploit the hierarchical relationships between thoracic abnormalities, the first step is to define a disease taxonomy that demonstrates different abnormalities' interrelationships. In this taxonomy, diseases are structured hierarchically in a graph, with higher levels representing broader disease categories and lower levels representing more nuanced distinctions between related diseases. The taxonomy is structured such that if a disease is present then its parent disease is also present. Furthermore, in the presence of multiple parent classes for a given child class, the taxonomy structure only utilizes the more dominant parent (e.g., if class c_1 has two parent classes c_3 and c_5 , while the c_5 is also the parent of c_3 class (hence it's both the parent and grandparent of c_1 class), in this scenario, we assume c_5 as the parent class of both c_1 and c_3 . For example, pleural effusion and pneumothorax can be classified as subcategories of pleural abnormalities, whereas atelectasis and consolidation can be classified under pulmonary opacity [7]. This hierarchical structure enables the model to take advantage of the relationships between diseases to improve its classification performance. In medical imaging, classes are frequently organized as graphs to represent the hierarchical relationships between different classes. For example, a graph can be used to represent the human body's organs, with each node representing a different organ and the edges representing the relationships between organs (e.g., the liver is part of the

abdominal cavity). Using a graph structure for labels in medical imaging has a number of advantages, including improved accuracy and interpretability of classification algorithms, which are essential for making sense of the vast amounts of data generated by medical imaging technologies. In medical imaging, hierarchies of labels are typically constructed by subject matter experts with a comprehensive understanding of human anatomy and physiology, such as radiologists. To create the label taxonomy shown in Figure 1.1, we combined the taxonomies provided by Irvin [7] for the CheXpert dataset, Chen [41] for the PADCHEST [6] and the CXR portion of the prostate, lung, colorectal and ovarian (PLCO) dataset [42]. In order to maintain uniformity, we adopted the renaming scheme introduced by Cohen [43] for the pathology names. Subsequently, the key pathologies were identified and extracted to build the hierarchical taxonomy structure illustrated in Figure 1.1.

1.3.3 Approach 1: Conditional Predicted Probability

When computational resources are limited, this technique can be applied to test samples without the need to fine-tune the pre-trained, multi-label classification model. This adaptability ensures that the benefits of considering hierarchical relationships between labels can be realized in a wide range of practical scenarios, without imposing excessive computational requirements. Directly updating the predicted probabilities presents potential benefits, including the following:

- **Simplicity:** Direct modification of predicted probabilities eliminates the need for substantial changes to the loss function, thus facilitating implementation.
- **Faster convergence:** In some cases, direct updates can accelerate convergence due to a more accurate representation of hierarchical relationships, thus reducing the overall training time.
- **Improved performance in specific scenarios:** Depending on the problem and dataset, direct updates may provide superior performance in certain circumstances, especially when incorporating class relationships into the loss function is challenging.
- **Easier calibration:** Direct modification of predicted probabilities can facilitate calibration of the model output to more closely match the true label distribution.

The proposed technique provides an easy way to improve the performance of existing pre-trained models during inference time by updating the value of the predicted logit for each class that was obtained at the last layer of the neural network based on the predicted logit of its corresponding parent class. The aim is to calculate the conditional predicted probability for each class k and instance i , taking into account the predicted probability of the parent class. We can formalize this by defining a new predicted probability for the k -th class (c_k) and instance i as follows.

$$\widehat{p}_k^{(i)} = \frac{1}{1 + \exp\left(-\left(q_k^{(i)} + \alpha_{k,j}q_j^{(i)}\right)\right)} \quad (1.3.3)$$

where $j = \Lambda_k$ is the index of the parent class of the k -th class, and $\alpha_{k,j}$ is the hyperparameter that controls the influence of different parent class logits on child class logits. When $\alpha_{k,j} = 0$, there is no influence from the parent class c_j on the child class c_k . By carefully selecting appropriate hyperparameter values, this transfer learning technique can be employed to effectively adjust the predicted probabilities of each class, considering the hierarchical relationship between classes, and potentially improving classification accuracy.

Parameter Selection and Tuning

The selection of appropriate hyperparameters is crucial for the effectiveness of the proposed transfer learning technique. In this study, we employ a systematic approach to tune the hyperparameters $\alpha_{k,j}$, which controls the dependency between the predicted probabilities of the child and parent classes. We utilize a grid search method along with cross-validation to determine the optimal values for these hyperparameters. The search space for both hyperparameters is defined based on preliminary experiments and domain knowledge, ensuring a balance between model complexity and predictive performance.

1.3.4 Approach 2: Conditional Loss

In a second approach, we propose a similar concept to the approach discussed in Section 1.3.3; however, rather than directly updating the predicted probability of each class, we instead update the loss value of each class based on the loss values of its parent classes. In the previous approach, we directly updated the predicted probability so that it could be applied unsupervised to existing pre-trained

models. Although this method is highly useful during inference time, it presents some challenges if we use it during the optimization phase of our classifier model. Among these disadvantages are the following.

- **Inconsistency with the optimization process:** Direct updating of predicted probabilities can misalign with the optimization procedure, which typically minimizes the loss function, potentially resulting in learning inconsistencies.
- **Difficulty in fine-tuning:** Direct updates can complicate fine-tuning the method's impact on the model, whereas adjusting the influence of various components is often simpler when updating the loss value through weighting factors or hyperparameters.
- **Potential overfitting:** Direct modification of predicted probabilities could inadvertently overfit the model to particular hierarchical relationships in the training data, thus hindering generalization to unseen data.

The utilization of the loss function approach can prove advantageous in certain scenarios, particularly in the context of multi-label classification tasks that involve hierarchical relationships, as it offers numerous benefits:

- **Emphasis on error minimization:** The loss values quantify the divergence between the predictions made by the model and the actual labels provided as ground truth. Integrating parent class loss values into child class loss calculations aims to minimize errors throughout the hierarchy, with the goal of improving prediction accuracy for both parent and child classes.
- **Enhanced gradient propagation:** During the training process of deep learning models, the model parameters are updated by backpropagating gradients through layers. Incorporating the loss values of the parent class through the calculation of the loss for the child classes can improve the gradient propagation between the parent and child classes. This may lead to more effective learning of hierarchical associations and speed up the training convergence.
- **Robustness to label noise:** Real-world datasets may exhibit inconsistencies or noise in their ground truth labels. The inclusion of loss values from parent classes in the computation of loss

values for child classes can enhance the consistency of the hierarchy by penalizing deviations from expected parent-child associations. This approach can result in improvement in the model's resilience to potential label inaccuracies in the dataset.

- **Improved interpretability:** The use of loss values rather than predicted probabilities enables a more straightforward understanding of the model's ability to capture hierarchical relationships among classes. When parent classes have high loss values, these losses influence their corresponding child classes' losses, underscoring the importance of improving the underlying model architecture and parameters to better present these hierarchical associations.

Formulation of the Proposed Technique

In multi-label classification problems, where each sample may belong to multiple classes, it is often necessary to combine the loss values for all classes to effectively train the model. Various methods can be employed to achieve this, depending on the specific problem. A common approach is to calculate the average loss across all classes for each sample by summing the losses for each class of a given sample and dividing the sum by the total number of classes to which the sample belongs. This method is effective when all classes are independent, of equal importance, and warrant equal weight in the total loss calculation. For example, in the case of cross-entropy loss, we have

$$l_k = \mathcal{L}(y_k^{(i)}, p_k^{(i)}) = -\left(y_k^{(i)} \log(p_k^{(i)}) + (1 - y_k^{(i)}) \log(1 - p_k^{(i)})\right) \quad (1.3.4)$$

$$\text{Loss}(\theta) = \sum_{i=1}^N \sum_{k=1}^K l_k \quad (1.3.5)$$

In this formulation, the objective is to minimize the loss function with respect to the model parameters θ , resulting in an optimal set of parameters that produce accurate predictions for multi-label classification tasks. However, class independence and equal importance between different classes cannot always be assumed. Inclusion of a hierarchical penalty or regularization term in the loss function is one way to push the loss function to take the taxonomy into account when optimizing the model hyperparameters (weights and biases). We use a regularization term β_k to penalize the loss for class c_k for each instance i in which there is a low probability that it also belongs to parent class c_j . This can be represented mathematically by adding a hierarchical penalty term $H(c_k|c_j)$ for the class c_k with respect to its

corresponding parent class c_j as follows:

$$\tilde{l}_k^{(i)} = l_k^{(i)} + \beta_k H(c_k | c_j) \quad (1.3.6)$$

where $c_j = \Lambda(c_k)$, and β_k is the hyperparameter that balances the contributions of class k 's own loss value and its parent class's loss values. There are multiple ways to define the hierarchical penalty. For example, we can define it as the loss value of the parent class $l_j = \mathcal{L}(y_j^{(i)}, p_j^{(i)})$ as follows:

$$H(k|j) = \mathcal{L}(y_j^{(i)}, p_j^{(i)}) \quad (1.3.7)$$

Another approach to incorporating the interdependence between different classes into the loss function is to apply the loss function \mathcal{L} to the true label of the parent class and the predicted probability of the child class as follows.

$$H(k|j) = \mathcal{L}(y_j^{(i)}, p_k^{(i)}) \quad (1.3.8)$$

The penalization term in Equations (1.3.7) and (1.3.8) encourages the model to correctly predict the corresponding parent class when predicting the child class, hence ensuring that the predicted labels align well with the hierarchical structure. The aforementioned approach, assumes a linear relationship between the child and parent losses. However, this may not always accurately capture the relationship between the parent-child classes, as the relationship may not necessarily be linear. The approach of multiplying losses introduces a greater adaptability in the representation of relationships between parent and child classes, as it can encapsulate both linear and potentially complex interrelations. Under the constraints of our problem - where the absence of a parent class guarantees the absence of its child class - both parent and child loss values would simultaneously increase or decrease (if the parent class is absent). In such a scenario, their summation or product would correspondingly escalate or diminish, thus demonstrating a linear relationship. However, the complexity arises when we consider the scenario where the parent's loss value is significantly low in comparison to the child's loss. Here, a simple additive model might undervalue the parent's loss impact, as adding a small parent loss value to a considerably larger child loss value might not significantly alter the new updated loss for that child class. On the contrary, a multiplicative model amplifies the influence of each parent loss on the total, even if the parent's loss is relatively small. By defining the new loss for child classes in such way that

their updated loss values are proportional to their corresponding parent's losses, we may enhance the hierarchical relationships' portrayal. To define such a loss value measurement scheme, we can modify the loss measurements presented in Equations (1.3.7) and (1.3.8) to be based on the multiplication of losses rather than their addition.

$$\widehat{l}_k^{(i)} = l_k^{(i)} H(k|j) \quad (1.3.9)$$

where the hierarchical penalty term is

$$H(k|j) = \begin{cases} 1 & \text{otherwise.} \\ \alpha_k l_j^{(i)} + \beta_k & c_j \text{ is parent of } c_k \end{cases} \quad (1.3.10)$$

where c_j is the parent class of the child class c_k , and l_j is the parent loss value for instance i .

In Equation (1.3.9), $\widehat{l}_k^{(i)}$ represents the new loss value that we calculate by multiplying the original loss value $l_k^{(i)}$ for child class k and instance i with the hierarchical penalty term $H(k|j)$ which is calculated based on the parent class j . The hierarchical penalty term $H(k|j)$, defined in Equation (1.3.10), adjusts based on the hierarchical relationships between classes. The terms α_k and β_k are parameters that can be adjusted to control the degree of influence the hierarchical relationships have on the learning process.

The parameter α_k directly scales the parent's loss $l_j^{(i)}$. If α_k is increased, the penalty term becomes larger, and thus the total loss $\widehat{l}_k^{(i)}$ becomes more sensitive to the parent's loss. This, in effect, increases the degree of influence that hierarchical information has on the learning process. The parameter β_k serves as a baseline or offset. If β_k is increased, the penalty term increases irrespective of the parent's loss value. This means that even if the parent's loss is low, the total loss $\widehat{l}_k^{(i)}$ can still be high, thus maintaining the influence of hierarchical information in the learning process. However, if β_k is set too high, it may lead to an overemphasis on hierarchy, possibly at the expense of other important learning elements. The regulation of parameters α_k and β_k allow us to balance the degree to which hierarchical information influences the learning process, thus improving the reflection of the hierarchical structure in the model outputs, while remaining flexible to diverse learning scenarios.

1.3.5 Updating Loss Values and Predicted Probabilities

In the previous section, we introduced a taxonomy-based loss function with the goal of improving the classification accuracy of multi-class problems. However, one of the main advantages of our proposed technique is that it enables efficient utilization of pre-trained models and leverages the existing knowledge, thus reducing the computational cost and training time associated with re-optimization. In this section, we illustrate how both of our proposed approaches can be seamlessly integrated into an existing classification framework without the necessity to re-run the optimization phase of the classifier (e.g., DenseNet121). This can be achieved by focusing on updating the loss values (approach 2 shown in Section 1.3.4) and predicted probabilities (approach 1 shown in Section 1.3.3) to incorporate the hierarchical relationships present in the taxonomy structure. During a training phase of a classifier (e.g., DenseNet121), an optimization algorithm such as gradient descent is used to determine the predicted probabilities that minimize the loss across the entire dataset. However, this approach is only valid during the training phase and only shows the predicted probability with respect to the original loss values measured by the classifier. In the following, we show how to calculate the updated predicted probabilities from their updated loss values obtained from Equation (1.3.9) without re-doing the optimization process. Let us assume that binary cross entropy is used for the choice of the loss function $\mathcal{L}(\cdot)$. Let us denote $\hat{q}_k^{(i)}, \hat{p}_k^{(i)}$ as the updated values for logit and predicted probability of class k and instance i after applying the proposed technique. As previously discussed, to calculate the predicted probabilities, we need to pass the logits $\hat{q}_k^{(i)}$ into a sigmoid function:

$$\hat{p}_k^{(i)} = \text{sigmoid}(\hat{q}_k^{(i)}) = \frac{1}{1 + \exp(-\hat{q}_k^{(i)})} \quad (1.3.11)$$

The sigmoid activation function maps any value to a number ranging from zero to one. The gradient of the sigmoid function (shown below) provides the direction in which the predicted probability must be updated.

$$\frac{\partial \text{sigmoid}}{\partial \hat{q}_k^{(i)}} = \text{sigmoid}(\hat{q}_k^{(i)}) (1 - \text{sigmoid}(\hat{q}_k^{(i)})) \quad (1.3.12)$$

$$= \hat{p}_k^{(i)} (1 - \hat{p}_k^{(i)}) \quad (1.3.13)$$

The loss gradient gives us the direction in which the predicted probability needs to be updated to minimize the loss. The gradient of the binary cross-entropy loss is calculated as follows:

$$\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} = \frac{y_k^{(i)}}{\hat{p}_k^{(i)}} - \frac{1 - y_k^{(i)}}{1 - \hat{p}_k^{(i)}} \quad (1.3.14)$$

where $y_k^{(i)}$ and $\hat{p}_k^{(i)}$ are the true label and predicted probability, respectively, for instance i and class k . We now show how we can use the predicted probability, the gradient loss shown in Equation (1.3.14) and the derivative of the sigmoid function shown in Equation (1.3.12) to calculate the updated predicted probability as follows:

$$\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \hat{q}_k^{(i)}} = \left(\frac{y_k^{(i)}}{\hat{p}_k^{(i)}} - \frac{1 - y_k^{(i)}}{1 - \hat{p}_k^{(i)}} \right) \hat{p}_k^{(i)} (1 - \hat{p}_k^{(i)}) \quad (1.3.15)$$

$$= y_k^{(i)} - \hat{p}_k^{(i)} \quad (1.3.16)$$

Hence, we can conclude that

$$\hat{p}_k^{(i)} = \begin{cases} -\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \hat{q}_k^{(i)}} & y_k^{(i)} = 1 \\ -\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \hat{q}_k^{(i)}} & \text{otherwise.} \end{cases} \quad (1.3.17)$$

We would like to modify this equation so that it does not directly depend on the true value and instead rely on the gradient loss. If we simplify the loss gradient shown in Equation (1.3.14) we obtain the following:

$$\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} = \frac{y_k^{(i)}}{\hat{p}_k^{(i)}} - \frac{1 - y_k^{(i)}}{1 - \hat{p}_k^{(i)}} \quad (1.3.18)$$

$$= \frac{y_k^{(i)} - \hat{p}_k^{(i)}}{\hat{p}_k^{(i)} (1 - \hat{p}_k^{(i)})} \quad (1.3.19)$$

In this equation, we see that when the true label is positive ($y_k^{(i)} = 1$), the loss gradient can only be 0 or a positive number. Similarly, when zero ($y_k^{(i)} = 0$), the loss gradient can only take the value 0 or a

negative number. Thus, we can modify Equation (1.3.17) as follows:

$$\hat{p}_k^{(i)} = \begin{cases} -\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \hat{q}_k^{(i)}} + 1 & \text{if } \frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \geq 0 \\ -\frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \hat{q}_k^{(i)}} & \text{otherwise.} \end{cases} \quad (1.3.20)$$

Finally, Equation (1.3.20) can be simplified as follows:

$$\hat{p}_k^{(i)} = \begin{cases} \exp(-\tilde{l}_k^{(i)}) & \text{if } \frac{\partial \mathcal{L}(\hat{p}_k^{(i)}, y_k^{(i)})}{\partial \hat{p}_k^{(i)}} \geq 0 \\ 1 - \exp(-\tilde{l}_k^{(i)}) & \text{otherwise} \end{cases} \quad (1.3.21)$$

where, $\tilde{l}_k^{(i)}$ is the updated loss for class k and instance i .

The following demonstrates Equation (1.3.21) based on predicted probability to demonstrate its similarity to Equation (1.3.3) in Approach 1 (Section 1.3.3). From Equation (1.3.10) we have $\hat{l}_k^{(i)} = l_k^{(i)} (\alpha_k l_j^{(i)} + \beta_k)$. By substituting that into $\exp(-\tilde{l}_k^{(i)})$, for $y_k^{(i)} = 1$ we obtain:

$$\exp(-\tilde{l}_k^{(i)}) = \exp(-l_k^{(i)} (\alpha_k l_j^{(i)} + \beta_k)) \quad (1.3.22)$$

$$= (p_k^{(i)})^{-\alpha_k \log(p_j^{(i)}) + \beta_k} \quad (1.3.23)$$

Furthermore, $1 - \exp(-\tilde{l}_k^{(i)})$, for $y_k^{(i)} = 0$ is as follows:

$$\begin{aligned} 1 - \exp(-\tilde{l}_k^{(i)}) &= 1 - \exp(-l_k^{(i)} (\alpha_k l_j^{(i)} + \beta_k)) \\ &= 1 - (1 - p_k^{(i)})^{-\alpha_k \log(1 - p_j^{(i)}) + \beta_k} \end{aligned} \quad (1.3.24)$$

By substituting Equations (1.3.22) and (1.3.24) into Equation (1.3.21) we obtain

$$\hat{p}_k^{(i)} = \begin{cases} (p_k^{(i)})^{-\alpha_k \log(p_j^{(i)}) + \beta_k} & \text{if } y_k^{(i)} = 1 \\ 1 - (1 - p_k^{(i)})^{-\alpha_k \log(1 - p_j^{(i)}) + \beta_k} & \text{otherwise.} \end{cases} \quad (1.3.25)$$

1.3.6 Experimental Setup

Datasets

Three diverse and publicly available datasets are used to evaluate the proposed hierarchical multi-label classification techniques: CheXpert [7], PADCHEST [6], and VinDr-CXR [44]. These datasets contain a diverse range of chest radiographic images covering various thoracic diseases, providing a comprehensive evaluation of the effectiveness of our method. The description of the three datasets are as follows.

- **CheXpert** [7] is a large-scale dataset containing 224,316 chest radiographs of 65,240 patients, labeled with 14 radiographic findings.
- **PADCHEST** [6] consists of 160,000 chest radiographs of 67,000 patients, annotated with 174 radiographic findings. This dataset is highly diverse and includes a wide variety of thoracic diseases.
- **NIH** [5] includes 112,120 chest radiographs of 30,805 patients labeled with 14 categories of thoracic diseases.

Preprocessing: The chest radiographs were pre-processed to ensure consistency across the datasets. The images were resized to a resolution of 224×224 pixels, with the pixel intensities normalized to a range of 0 and 1. Data augmentation techniques, such as rotation, translation, and horizontal flipping, were applied to increase the dataset's size and diversity, consequently enhancing the model's generalization capability.

Model Optimization

The DenseNet121 [16] architecture and the pre-trained weights provided by Cohen [43] was used as the baseline model. The model was fine-tuned on a subset of CheXpert [7], NIH [5], PADCHEST [6] for 18 thoracic diseases. A series of transformations were applied to all train images, including rotation of up to 45 degrees, translation of up to 15%, and scaling up to 10%. Binary cross entropy losses and Adam optimizer were used.

Parallelization for multiple CPU cores: To effectively optimize the hyperparameters of our proposed taxonomy-based transfer learning methods, we utilize parallelization techniques that distribute the computational load across multiple CPU cores. By leveraging the power of parallel processing, we can drastically reduce the overall computation time and accelerate the optimization procedure, making the method more applicable to large-scale and high-dimensional datasets. Different parallelization libraries, such as joblib and Python multiprocessing, were employed to facilitate the implementation of parallelism, ensuring seamless integration with existing frameworks and offering a scalable and hardware-adaptable solution.

Optimum Threshold Determination: Determining the optimal threshold is a crucial aspect of evaluating the performance of the proposed method, as it determines the point at which the predictions for multi-label classification tasks are translated into binary class labels. To determine the optimal threshold value, we used receiver operating characteristic (ROC) analysis, a common method for evaluating the performance of classification models. ROC analysis provides a comprehensive view of the model's performance at various threshold values, allowing us to determine the optimal point for balancing the true positive rate (sensitivity) and the false positive rate (specificity) (1-specificity). By plotting the ROC curve and calculating the area under the curve (AUC), we can quantitatively evaluate the discriminatory ability of the model and compare its performance at various threshold values. The optimal threshold is determined by locating the point on the ROC curve closest to the upper left corner, which represents the highest true positive rate and the lowest false positive rate. By incorporating ROC analysis and optimal threshold determination into our experimental design, we ensure that our results not only accurately reflect the performance of the model but also provide valuable insight into the practical applicability of our approach in real-world settings.

Evaluation: To assess the performance of the proposed techniques in accurately classifying samples compared to a baseline model, several evaluation metrics were used. The metrics were selected based on their ability to provide a comprehensive assessment of the model's performance in terms of accuracy, precision, recall, and the ability to differentiate between true and false positives. The evaluated metrics are as follows.

- **Accuracy** measures the proportion of correctly classified samples to the total number of samples.

- **F1-score** is the harmonic mean of precision and recall, providing a balanced assessment of the method's performance.
- **Area Under the Receiver Operating Characteristic Curve (AUROC):** The ROC curve is a graphical representation of the diagnostic performance of a binary classifier system as its discrimination threshold is varied. The ROC curve is derived by plotting the true positive rate (TPR) versus the false positive rate (FPR) at different thresholds. The AUC provides a single scalar value representing the expected performance of the classifier. An AUC of 1 indicates that the classifier can distinguish perfectly between the two classes (e.g., “positive” and “negative”), whereas an AUC of 0.5 indicates that the classifier is no better than random chance.
- **t-stat (t-statistic)** is a measurement of the magnitude of the difference relative to the variance in sample data. The t-value quantifies the statistical significance of the difference. It is used to test hypotheses regarding the mean or the difference between two means when the standard deviation of the population is unknown.
- **p-value:** In hypothesis testing, the p-value is a function used to determine the significance of the results. It represents the probability that test results were generated at random. If the p-value is small (typically 0.05), there is strong evidence that the null hypothesis should be rejected.
- **Cohen's Kappa** measures the concordance between two raters who classify items into mutually exclusive categories. Primarily, it is used to determine the degree of agreement between two raters. The Kappa score takes into account the possibility that the agreement occurred by chance. A Kappa score of 1 indicates perfect concordance between two raters. A Kappa score of less than 1 indicates less than perfect agreement, and a Kappa score of less than 0 indicates either no agreement or agreement that is worse than random.
- **BF10 (Bayes Factor)** rates the strength of the evidence in favor of one statistical model over another, given the available data. BF10 specifically contrasts the evidence supporting a null hypothesis (H_0) with an alternative hypothesis (H_1). The data are equally likely to be true under the null and alternative hypotheses, according to a BF10 value of 1. A BF10 value greater than 1 denotes support for H_1 , while a value lower than 1 denotes support for H_0 . In general, values

between $1/3$ and 3 are regarded as inconclusive, values above 3 as some evidence for H_1 , and values below $1/3$ as evidence for H_0 .

- **Cohen's d** is a measure of effect size in the context of a t-test for the difference between two means. It can be calculated as the difference between two means divided by the data's standard deviation. Typically, small, medium, and large effect sizes are referred to as Cohen's d values of 0.2, 0.5, and 0.8, respectively. It is a common method of estimating the difference between two groups after adjusting for variance and sample size variations.
- **Power (Statistical Power)** is the likelihood that a test will correctly reject the null hypothesis when the alternative hypothesis is true (i.e., the test will not make a Type II error). Power is typically desired to be 0.8 or higher, meaning there is an 80% or greater chance of discovering a true effect if it is present. Many variables, such as the effect size, sample size, significance level, and data variability, can have an impact on power. Calculating power can be used to determine the sample size required to detect an effect of a given size when designing a study.

Some limitations of these metrics are as follows. While accuracy is a useful metric for evaluating overall performance, it may not be the most appropriate metric for unbalanced datasets in which the number of samples in each class is significantly different. Similarly, F1-score may be biased towards the class with a larger sample size, and AUROC may not be appropriate for datasets with a high degree of class overlap. In addition, outliers or non-normal distributions may influence the t-statistic and p-value, whereas Cohen's Kappa may not be applicable to non-categorical data. BF10 may be affected by the selection of prior probabilities, and Cohen's d may not apply to non-parametric data. The choice of significance level and the data variability may have an effect on the power.

1.4 Results

Figure 1.1 shows the created taxonomy structure. This comprehensive classification system accumulated using taxonomy graphs in Irvin [7], and Chen [41] helps categorize various disease manifestations observed in public datasets, such as CheXpert, PADCHEST and NIH and serves as a framework for understanding and analyzing chest radiograph abnormalities.

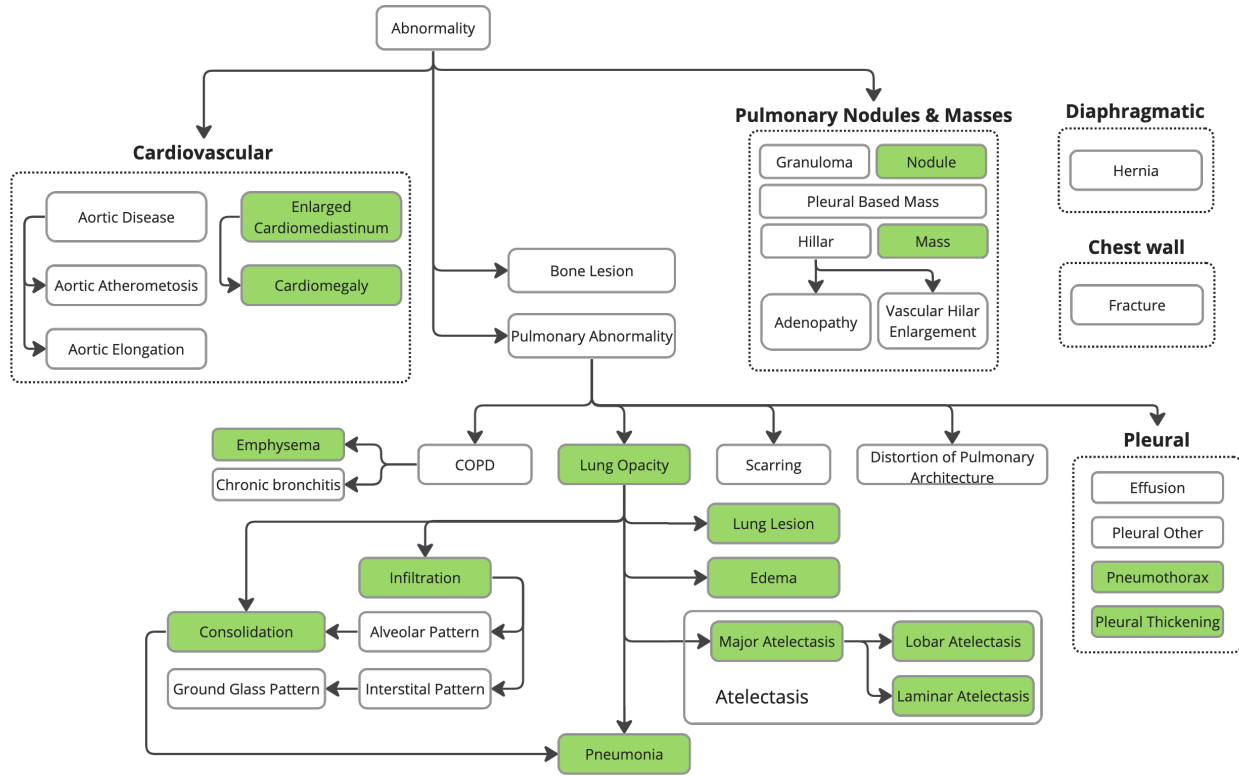


Figure 1.1: Taxonomy structure of lung pathologies in chest radiographs.

In this study, we investigated the frequency of various pathological labels in three distinct medical imaging datasets: CheX [7], PADCHEST [6], NIH [5]. Table 1.1 depicts the presence of each pathology label across these datasets. To conform to Cohen [43] work, the same 18 pathologies selected by Cohen [43] are used for model optimization. Special consideration is given to the pathologies that appear in at least two of the three datasets and are included in our taxonomy. These pathologies are marked with a green color in the table and include **Atelectasis, Consolidation, Infiltration, Edema, Pneumonia, Cardiomegaly, Lung Lesion, Lung Opacity, Enlarged Cardiomeastinum**. This selection is significant because it reflects the pathologies that not only manifest more frequently across multiple data sources, but are also reflected in the hierarchical taxonomy used in our study. The study's analysis and conclusions are focused on these specific pathologies due to their consistent presence and relevance within the taxonomy structure. Further, the cross-dataset presence of these pathologies enhances the generalizability of our study, as the developed models are validated on multiple independent datasets. The pathologies that are not highlighted, i.e., those occurring in just one or none of the datasets or

not included in our taxonomy, were not included in the final evaluation of this study. Their exclusion is mainly due to the lack of sufficient data for a robust comparison or their non-alignment with the studied taxonomy structure. Table 1.2 shows the number of instances that has a specific pathology in

Table 1.1: Representation of pathologies across datasets

Pathologies	NIH	PADCHEST	CheX	Pathologies	NIH	PADCHEST	CheX
Air Trapping		X		Hemidiaphragm Elevation		X	
Aortic Atheromatosis		X		Hernia	X	X	
Aortic Elongation		X		Hilar Enlargement		X	
Aortic Enlargement				ILD			
Atelectasis	X	X	X	Infiltration	X	X	
Bronchiectasis		X		Lung Lesion			X
Calcification				Lung Opacity			X
Calcified Granuloma				Mass	X	X	
Cardiomegaly	X	X	X	Nodule/Mass			
Consolidation		X	X	Nodule	X	X	
Costophrenic Angle Blunting		X		Pleural Other			X
Edema	X	X	X	Pleural Thickening	X	X	
Effusion	X	X	X	Pneumonia	X	X	X
Emphysema	X	X		Pneumothorax	X	X	X
Enlarged Cardiomediastinum			X	Pulmonary Fibrosis			
Fibrosis	X	X		Scoliosis		X	
Flattened Diaphragm		X		Tuberculosis		X	
Fracture		X	X	Tube		X	
Granuloma		X					

each of the three studied datasets (CheX [7], PADCHEST [6], NIH [5]). Prior to applying the proposed technique a set of preprocessing steps are applied to ground truth label set. In medical images with multiple classes, it is common for the labeler to only label the pathologies that their study requires. This sometimes result in situations where some instances of data are labeled for the presence of some of the child pathologies but not their corresponding parent pathologies. To compensate for this lack of labeling for some parent classes which is necessary for the effectiveness of the proposed techniques, we updated the label value indicating the presence of classes with at least one child class to **TRUE** (indicating the class exist in that instance). This preprocessing is applied to all pathologies which are

not labeled in the original ground truth label set. As can be seen in Table 1.2 (highlighted cells), while the Lung Opacity and Enlarged Cardiomeastinum classes were not present in the original ground truth label sets of NIH and PADCHEST datasets (Table 1.1), by updating the ground truth label set we end up with multiple instances where based on the presence of their child classes' presence we have determined the presence of the respective parent class.

Table 1.2: Sample distribution per pathology in evaluated datasets (CheX, NIH, and PC)

Pathologies\Dataset	CheXpert		NIH		PADCHEST	
	PA	AP	PA	AP	PA	AP
Atelectasis	2460	11643	1557	1016	2419	232
Consolidation	1125	4956	384	253	475	77
Infiltration	0	0	3273	1131	4309	587
Pneumothorax	1060	4239	243	253	97	15
Edema	1330	15117	39	237	108	130
Emphysema	0	0	264	193	546	30
Fibrosis	0	0	556	61	341	8
Effusion	5206	19349	1269	654	1625	311
Pneumonia	992	2064	175	89	1910	211
Pleural_Thickening	0	0	745	145	2075	34
Cardiomegaly	2117	8284	729	203	5387	261
Nodule	0	0	1609	460	2190	95
Mass	0	0	1213	493	506	17
Hernia	0	0	81	13	988	38
Lung Lesion	1655	3110	0	0	0	0
Fracture	1115	3463	0	0	1662	69
Lung Opacity	7006	28183	4917	2216	6947	861
Enlarged Cardiomeastinum	1100	4577	729	203	5387	261
Total	20543	53359	28868	9060	61692	2445

Figure 1.2 presents the comparison of the performance of our proposed techniques “logit” and “loss”

against the “baseline” technique for a series of nine medical conditions related to lung and heart diseases on three datasets (CheXpert, PADCHEST, NIH). These nine pathologies include the two parent classes (**Lung Opacity**, and **Enlarged Cardiomeastinum**) and their corresponding child classes, as shown in Figure 1.1. The individual subplots exhibit overlaid receiver operating characteristic (ROC) curves and their corresponding AUC scores.

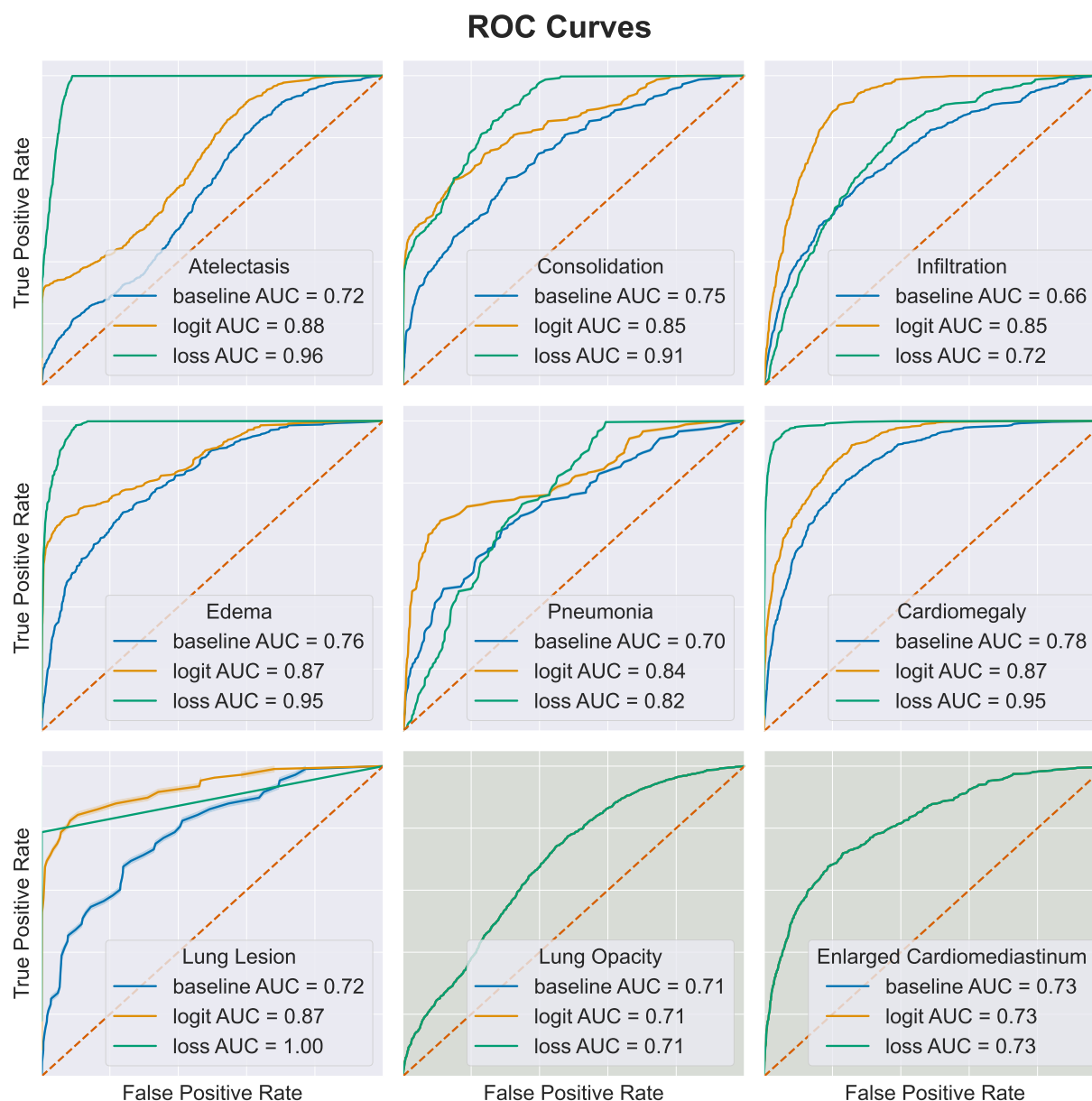


Figure 1.2: Comparative analysis of the ROC curves for nine thoracic pathologies using the “logit” and “loss” techniques as well as the baseline. The subplots highlighted with a darker background, represent parent class diseases.

Table 1.3 presents the comparison of the performance of our proposed techniques “logit” and “loss” against the “baseline” technique for various statistical metrics. The “logit” method (upper table), shows a considerable improvement over the “baseline” across all conditions. We observe kappa values between

0.495 and 1. The kappa statistic is a measure of agreement between two methods, with a value of 1 indicating perfect agreement. The p-value for all child classes show a value less than 0.05 (ranging from $2.1\text{E-}89$ to $2.9\text{E-}16$), which indicates the statistically significant superiority of the “logit” method over the “Baseline”. High t-statistics and power values of 1 further confirm the robustness of our method. In terms of Bayes factor, results for the “logit” method are extremely strong for all conditions suggesting a high amount of evidence in favor of the “logit” method for these conditions. The second proposed method, “loss”, also presents promising results when compared to the “baseline”, but with more variation. Kappa values ranged from a low of 0.059 for Lung Lesion to a high of 0.836 for Infiltration. The p-values indicate statistically significant differences for most conditions, although Infiltration and Pneumonia present p-values greater than 0.05 (0.053 and 0.207 respectively), suggesting that the performance improvement over the “baseline” for these conditions may not be statistically significant. T-statistics are high, and power values are 1 for all conditions except Infiltration and Pneumonia. The cohen-d values for this method are generally larger than those in the “logit” method, indicating a larger effect size. In terms of Bayes factor, results for the “loss” method are extremely strong for conditions such as Atelectasis and Edema, while being much lower for conditions like Infiltration and Pneumonia, suggesting a reduced amount of evidence in favor of the “loss” method for these conditions. Both the “logit” and “loss” methods show significant improvements over the “baseline” method in the diagnosis of several heart and lung conditions, albeit with some variance in the degree of improvement. While the “logit” method demonstrates a more consistent level of improvement across all conditions, the “loss” method shows potential for even greater improvement in certain conditions but with less consistency across the conditions studied.

Table 1.3: Statistical performance comparison between the proposed techniques “logit” and “loss” and the “baseline” technique across various pathologies. The upper table displays the findings of the “logit” technique, while the lower table displays the findings of the “loss” technique. The reported metrics for each pathology are the Kappa statistic, p-value, t-statistic, statistical power, Cohen’s d, and Bayes Factor (BF10). A kappa value of 1 indicates perfect agreement between techniques, whereas a larger Bayes factor indicates greater support for the “logit” or “loss” technique over the baseline.

		kappa	p_value	t_stat	power	cohen-d	BF10
L	Atelectasis	0.495	2.1E-89	20.2	1	0.346	3.0E+85
	Consolidation	0.508	2.0E-18	8.8	1	0.150	8.3E+14
O	Infiltration	0.620	2.7E-28	11.1	1	0.190	4.9E+24
	Edema	0.614	1.2E-52	15.3	1	0.263	7.2E+48
G	Pneumonia	0.573	2.9E-16	8.2	1	0.140	6.3E+12
	Cardiomegaly	0.615	1.9E-72	18.1	1	0.310	3.9E+68
I	Lung Lesion	0.580	7.0E-23	9.9	1	0.169	2.1E+19
	Lung Opacity	1	1	0	0.05	0	0.019
T	Enlarged Cardiomedastinum	1	1	0	0.05	0	0.019

		kappa	p_value	t_stat	power	cohen-d	BF10
	Atelectasis	0.222	4.9E-183	29.3	1	0.502	7.7E+178
L	Consolidation	0.310	4.3E-116	23.1	1	0.396	1.2E+112
	Infiltration	0.836	0.053	1.9	0.49	0.033	0.125
O	Edema	0.343	4.4E-190	29.9	1	0.512	8.2E+185
	Pneumonia	0.394	0.207	1.3	0.24	0.022	0.043
S	Cardiomegaly	0.501	1.2E-101	21.6	1	0.370	4.7E+97
	Lung Lesion	0.059	1.2E-207	31.3	1	0.537	2.9E+203
S	Lung Opacity	1	1	0	0.05	0	0.019
	Enlarged Cardiomedastinum	1	1	0	0.05	0	0.019

Figure 1.3 compares the performance of the proposed “loss” and “logit” techniques to the “baseline”

across three key metrics: accuracy (ACC), area under the receiver operating characteristic curve (AUC), and F1 score for various pathologies. The accuracy metric presents a clear advantage for the “loss” and “logit” methods over the “baseline” for the child classes of pathologies, a pattern that is consistent with the kappa statistics presented earlier. For example, in Atelectasis, the “loss” method has an accuracy of 0.922 compared to 0.686 for the “baseline”, while the “logit” method stands at 0.874. As expected, the parent classes, lung opacity and enlarged cardiomeastinum, show no difference between the techniques, with an accuracy of 0.663 and 0.696, respectively. The AUC, a model performance metric that accounts for both sensitivity and specificity, demonstrates once more that the “loss” and “logit” methods for the child classes are superior. For example, in the case of cardiomegaly, the AUC is improved by 21% and 11% using the loss and logit methods, respectively. The AUC values for lung opacity and an enlarged cardiomeastinum, the parent classes, are identical for all three methods. The F1 score, which is the harmonic mean of precision and recall, sheds additional light on the enhanced performance of our proposed techniques. Notably, lung lesion increases from 0.094 in the “baseline” method to 0.982 in the “loss” method and 0.263 in the “logit” method. These results further validate our earlier findings that the “logit” and “loss” methods provide significant performance improvements over the “baseline” method for the majority of the child classes. Across all metrics and conditions, the “loss” method appears to perform marginally better than the “logit” method.

	ACC			AUC			F1		
Atelectasis	0.686	0.922	0.874	0.721	0.957	0.876	0.291	0.685	0.541
Consolidation	0.687	0.844	0.767	0.747	0.910	0.851	0.137	0.278	0.199
Infiltration	0.604	0.650	0.801	0.664	0.715	0.845	0.359	0.406	0.575
Edema	0.685	0.911	0.819	0.755	0.946	0.865	0.235	0.563	0.370
Pneumonia	0.749	0.740	0.842	0.701	0.823	0.838	0.097	0.126	0.179
Cardiomegaly	0.741	0.915	0.878	0.784	0.946	0.874	0.312	0.620	0.501
Lung Lesion	0.760	0.999	0.915	0.723	1.000	0.875	0.094	0.982	0.263
Lung Opacity	0.663	0.663	0.663	0.705	0.705	0.705	0.475	0.475	0.475
Enlarged Cardiomediastinum	0.696	0.696	0.696	0.731	0.731	0.731	0.253	0.253	0.253
	baseline	loss	logit	baseline	loss	logit	baseline	loss	logit

Figure 1.3: Heatmap visualization of model performance metrics across all three datasets. The subplots from left to right correspond to the Accuracy (ACC), Area Under the ROC Curve (AUC), and F1 Score for the baseline, “loss”, and “logit” techniques respectively. The pathologies are shared on the y-axis. Darker colors signify higher values, indicating better model performance. Each cell represents the value of the corresponding metric for the given technique on a specific pathology

1.5 Discussion and Conclusion

In this work, we presented two hierarchical multi-label classification methods aimed at enhancing thoracic disease diagnosis in chest radiography. The first approach, denoted as the “loss” method, refines the loss value for each pathology, factoring in the influence of parent pathologies in the hierarchical structure. The “logit” method is a powerful tool for improving the performance of pre-trained models. This approach updates the logit values of each pathology based on the corresponding logit values of their parent pathologies, leveraging the inherent taxonomical relationships between pathologies. This strategy is particularly useful when ground truth labels are not readily available, as it allows for efficient enhancement of model performance without requiring access to such labels. By avoiding the need to build new models from scratch, the “logit” method offers a streamlined and effective approach

to optimizing model performance in a variety of contexts. Whether working with complex medical data or other types of information, this approach can help researchers and practitioners achieve more accurate and reliable results with greater ease and efficiency. The results of the study demonstrate the effectiveness of proposed hierarchical multi-label classification methodologies in improving the precision of thoracic disease diagnosis. Various performance metrics, including accuracy, AUC, and F1 scores, as well as Cohen's d, Cohen's kappa, t-statistics, p-value, and Bayes factor, are used to evaluate the performance of the proposed techniques against the baseline on three public datasets (CheXpert [7], PADCHEST [6], and NIH [5]), showing substantial improvements in the proposed techniques against the baseline. These findings suggest that these methods can be used as reliable tools for accurate and efficient diagnosis of thoracic diseases. Further research is needed to explore the potential benefits of these methods in clinical practice. The utilization of logit adjustments represents a simple yet powerful approach for incorporating label hierarchy into a model without requiring significant modifications to the existing framework. However, this approach has the potential to obscure the optimization and learning process and results. On the other hand, modifying loss values is more closely aligned with the optimization process of the model, enabling a more refined adjustment of the hierarchical influence through weighting factors. This methodology promotes resilience to inaccuracies in labeling and improves conformity with established hierarchical relationships. In essence, both the "loss" and "logit" techniques effectively leverage disease taxonomy to bolster classification performance, reinforcing the significance of exploiting label relationships in classification tasks. Moreover, these hierarchical techniques can potentially aid clinicians by improving the interpretability of the models' predictions. Exploring predictions at varying levels of granularity based on taxonomy could facilitate personalized diagnoses tailored to individual clinical needs. Additionally, the techniques could be integrated into computer-aided diagnosis systems to provide more accurate and efficient diagnoses, potentially reducing the workload of clinicians and improving patient outcomes. However, these methodologies exhibit certain constraints. Extending these methodologies to other applications would necessitate the development of a taxonomical structure for the labels of the corresponding dataset. The construction of such a structure may present difficulties for complex applications and typically necessitates the consensus of multiple domain experts. Additionally, the performance of the proposed techniques may be influenced by the quality and consistency of the labeling in the datasets, which may vary across different sources. Future studies should aim to evaluate the techniques on a broader range

of datasets and consider the impact of labeling quality on performance.

Appendices

Acknowledgements

References

- [1] N. Bellaviti, F. Bini, L. Pennacchi, G. Pepe, B. Bodini, R. Ceriani, C. D’Urbano, and A. Vaghi, “Increased Incidence of Spontaneous Pneumothorax in Very Young People: Observations and Treatment,” *CHEST*, vol. 150, p. 560A, Oct. 2016.
- [2] L. Delrue, R. Gosselin, B. Ilse, A. Van Landeghem, J. de Mey, and P. Duyck, “Difficulties in the Interpretation of Chest Radiography,” in *Comparative Interpretation of CT and Standard Radiography of the Chest* (E. E. Coche, B. Ghaye, J. de Mey, and P. Duyck, eds.), Medical Radiology, pp. 27–49, Berlin, Heidelberg: Springer, 2011.
- [3] N. Crisp and L. Chen, “Global Supply of Health Professionals,” *N Engl J Med*, vol. 370, pp. 950–957, Mar. 2014.
- [4] J. Silverstein, “Most of the World Doesn’t Have Access to X-Rays,” *The Atlantic*, Sept. 2016.
- [5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, (Honolulu, HI), pp. 3462–3471, IEEE, July 2017.
- [6] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, “Padchest: A Large Chest X-Ray Image Dataset With Multi-Label Annotated Reports,” *Medical Image Analysis*, vol. 66, p. 101797, Dec. 2020.
- [7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, “CheXpert: A Large Chest Radiograph Dataset With Uncertainty Labels and Expert Comparison,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 590–597, July 2019.

- [8] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial Transformer Networks," in *Adv. Neural Inf. Process. Syst.* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.
- [9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.
- [10] M. Eshghali, D. Kannan, N. Salmanzadeh-Meydani, and A. M. Esmayeeli Sikaroudi, "Machine Learning Based Integrated Scheduling and Rescheduling for Elective and Emergency Patients in the Operating Theatre," *Ann Oper Res*, Jan. 2023.
- [11] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues, "Identifying Pneumonia in Chest X-Rays: A Deep Learning Approach," *Measurement*, vol. 145, pp. 511–518, Oct. 2019.
- [12] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," *Radiology*, vol. 284, pp. 574–582, Aug. 2017.
- [13] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization," *Sci Rep*, vol. 9, p. 6268, Dec. 2019.
- [14] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic Lung Cancer Prediction From Chest X-Ray Images Using the Deep Learning Approach," in *11th Biomed. Eng. Int. Conf. BMEiCON*, (Chiang Mai), pp. 1–5, IEEE, Nov. 2018.
- [15] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *Int. J. Data Warehous. Min.*, vol. 3, pp. 1–13, July 2007.
- [16] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017*, 2017.
- [17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Med. Image Comput. Comput.-Assist. Interv. – MICCAI 2015* (N. Navab, J. Hornegger,

- W. M. Wells, and A. F. Frangi, eds.), vol. 9351, (Cham), pp. 234–241, Springer International Publishing, 2015.
- [18] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, “Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks,” 2017.
 - [19] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, “Diagnose Like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification,” 2018.
 - [20] H. Liu, L. Wang, Y. Nan, F. Jin, Q. Wang, and J. Pu, “SDFN: Segmentation-Based Deep Fusion Network for Thoracic Disease Classification in Chest X-Ray Images,” *Computerized Medical Imaging and Graphics*, vol. 75, pp. 66–73, July 2019.
 - [21] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, “Iterative Attention Mining for Weakly Supervised Thoracic Disease Pattern Localization in Chest X-Rays,” in *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2018* (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), Lecture Notes in Computer Science, (Cham), pp. 589–598, Springer International Publishing, 2018.
 - [22] S. Guendel, F. C. Ghesu, S. Grbic, E. Gibson, B. Georgescu, A. Maier, and D. Comaniciu, “Multi-Task Learning for Chest X-Ray Abnormality Classification on Noisy Labels,” May 2019.
 - [23] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, “Thoracic Disease Identification and Localization With Limited Supervision,” in *IEEECVF Conf. Comput. Vis. Pattern Recognit.*, (Salt Lake City, UT), pp. 8290–8299, IEEE, June 2018.
 - [24] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, “Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-Rays,” in *Int. Conf. Bioinforma. Comput. Biol. Health Inform.*, (Washington DC USA), pp. 103–110, ACM, Aug. 2018.
 - [25] M. L. Zhang and Z. H. Zhou, “A Review on Multi-Label Learning Algorithms,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, pp. 1819–1837, Aug. 2014.
 - [26] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, “On Label Dependence and Loss

- Minimization in Multi-Label Classification,” *Mach Learn*, vol. 88, pp. 5–45, July 2012.
- [27] H. Harvey and B. Glocker, “A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology,” in *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks* (E. R. Ranschaert, S. Morozov, and P. R. Algra, eds.), pp. 61–72, Cham: Springer International Publishing, 2019.
- [28] W. Bi and J. T. Kwok, “Bayes-Optimal Hierarchical Multilabel Classification,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, pp. 2907–2918, Nov. 2015.
- [29] H. Pourghassem and H. Ghassemian, “Content-Based Medical Image Classification Using a New Hierarchical Merging Scheme,” *Computerized Medical Imaging and Graphics*, vol. 32, pp. 651–661, Dec. 2008.
- [30] I. Dimitrovski, D. Koccev, S. Loskovska, and S. Džeroski, “Hierarchical Annotation of Medical Images,” *Pattern Recognition*, vol. 44, pp. 2436–2449, Oct. 2011.
- [31] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew, “CNN-RNN: A Large-Scale Hierarchical Image Classification Framework,” *Multimed Tools Appl*, vol. 77, pp. 10251–10271, Apr. 2018.
- [32] K. Kowsari, D. E. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber, and L. E. Barnes, “HDLTex: Hierarchical Deep Learning for Text Classification,” in *16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA*, (Cancun, Mexico), pp. 364–371, IEEE, Dec. 2017.
- [33] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in *IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, (Honolulu, HI), pp. 6517–6525, IEEE, July 2017.
- [34] D. Roy, P. Panda, and K. Roy, “Tree-Cnn: A Hierarchical Deep Convolutional Neural Network for Incremental Learning,” *Neural Networks*, vol. 121, pp. 148–160, Jan. 2020.
- [35] S. Van Eeden, J. Leipsic, S. F. Paul Man, and D. D. Sin, “The Relationship Between Lung Inflammation and Cardiovascular Disease,” *Am J Respir Crit Care Med*, vol. 186, pp. 11–16, July 2012.
- [36] N. Alaydie, C. K. Reddy, and F. Fotouhi, “Exploiting Label Dependency for Hierarchical Multi-

- Label Classification,” in *Advances in Knowledge Discovery and Data Mining* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P.-N. Tan, S. Chawla, C. K. Ho, and J. Bailey, eds.), vol. 7301, pp. 294–305, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
- [37] R. Aly, S. Remus, and C. Biemann, “Hierarchical Multi-Label Classification of Text With Capsule Networks,” in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist. Stud. Res. Workshop*, (Florence, Italy), pp. 323–330, Association for Computational Linguistics, 2019.
- [38] W. Bi and J. T. Kwok, “Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 25, pp. 2275–2287, Dec. 2014.
- [39] H. Chen, S. Miao, D. Xu, G. D. Hager, and A. P. Harrison, “Deep Hierarchical Multi-Label Classification of Chest X-Ray Images,” in *Proc. 2nd Int. Conf. Med. Imaging Deep Learn.*, pp. 109–120, PMLR, May 2019.
- [40] G. Furnieles, “Sigmoid and SoftMax Functions in 5 minutes,” 2022.
- [41] H. Chen, S. Miao, D. Xu, G. D. Hager, and A. P. Harrison, “Deep Hierarchical Multi-Label Classification Applied to Chest X-Ray Abnormality Taxonomies,” *Medical Image Analysis*, vol. 66, p. 101811, Dec. 2020.
- [42] J. K. Gohagan, P. C. Prorok, R. B. Hayes, and B.-S. Kramer, “The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status,” *Controlled Clinical Trials*, vol. 21, pp. 251S–272S, Dec. 2000.
- [43] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, and H. Bertrand, “TorchXRyVision: A Library of Chest X-Ray Datasets and Models,” in *Proc. 5th Int. Conf. Med. Imaging Deep Learn.*, pp. 231–249, PMLR, Dec. 2022.
- [44] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu, “VinDr-CXR: An

Open Dataset of Chest X-Rays With Radiologist's Annotations," *Sci Data*, vol. 9, p. 429, July 2022.