# Graphical Abstract

**A Hierarchical Multilabel Classification Method for Enhanced Thoracic Disease Diagnosis in Chest Radiography**

Mohammad S. Majdi, Jeffrey J. Rodriguez

# Highlights

**A Hierarchical Multilabel Classification Method for Enhanced Thoracic Disease Diagnosis in Chest Radiography**

Mohammad S. Majdi, Jeffrey J. Rodriguez

- Research highlight 1
- Research highlight 2

# A Hierarchical Multilabel Classification Method for Enhanced Thoracic Disease Diagnosis in Chest Radiography

Mohammad S. Majdi[a], Jeffrey J. Rodriguez[a]

[a]*Dept. of Electrical and Computer Engineering, University of Arizona, Tucson 85721, AZ, USA*

**Abstract**

Accurate diagnosis of thoracic diseases from chest radiographs is a challenging task that can lead to diagnostic errors and negative patient outcomes. In this study, we propose a novel hierarchical multilabel classification technique that utilizes the taxonomical relationship between different pathologies to improve classification accuracy. Two methods are proposed to encompass both scenarios where the ground truth is available (referred to as "loss" in this paper) and when it is not (referred to as "logit"). The proposed methods leverage a predefined disease taxonomy to account for interrelationships among diseases, thereby augmenting their generalizability to novel tasks. The "logit" approach can be seamlessly integrated into existing pre-trained models without the need for re-optimization, ensuring efficiency and broad applicability. The "loss" approach can be incorporated into the existing technique during the training phase by modifying the loss function. To evaluate the effectiveness of the proposed technique, experiments were conducted on three diverse and publicly available chest radiograph datasets (CheXpert, PadChest, and NIH Chest-Xray14). The results demonstrate that the proposed technique significantly improves the accuracy and interpretability of machine learning models for thoracic disease on chest radiography. This approach has the potential to promote an accurate and efficient diagnosis by providing radiologists with an additional layer of decision support, ultimately leading to better patient outcomes.

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

*June 2, 2023*

## 1. Introduction

The timely diagnosis and effective treatment of diseases depend on the precise and efficient detection of anomalies in medical imaging. Deep learning techniques have made substantial progress in the medical imaging domain, exhibiting impressive success across various applications Litjens et al. (2017). Nonetheless, conventional classification methods primarily designed for single-label predictions struggle with multi-label classification, which requires predicting multiple labels for each input sample.

Chest radiography (CXR) is a prevalent radiological examination for diagnosing lung and heart disorders, constituting a significant share of ordered imaging studies. Swift and accurate detection of different conditions, such as pneumothorax, is crucial for optimal patient care Bellaviti et al. (2016). However, interpreting CXRs can be challenging due to similarities between multiple diseases, which may result in misinterpretations even by experienced radiologists Delrue et al. (2011). Consequently, devising an accurate system to identify and localize common thoracic diseases can aid radiologists in minimizing diagnostic errors Crisp and Chen (2014); Silverstein (2016).

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 Jaderberg et al. (2015). Progress in natural language processing (NLP) has enabled the collection of extensive annotated datasets such as ChestX-ray8 Wang et al. (2017), MIMIC-CXR Johnson et al. (2019), and CheXpert (Irvin et al., 2019b), allowing researchers to develop more efficient and robust supervised learning algorithms.

Regarding 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 Tsoumakas and Katakis (2007). 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 Jaiswal et al. (2019); Lakhani and Sundaram

(2017); Pasa et al. (2019); Ausawalaithong et al. (2018), leaving ample opportunities to investigate a unified deep learning framework that can efficiently detect a broad spectrum of common thoracic diseases.

This paper aims to tackle the challenges of multi-label classification within the realm of medical imaging by introducing a hierarchical framework that can be employed in a transfer learning approach without necessitating costly computational resources. The rest of this paper is structured as follows. Section 2 discusses related work on multi-label classification and hierarchical loss functions; Section 3 describes our proposed method for integrating label hierarchy into multi-label loss functions; Section 4 presents experimental results using the chest radiograph dataset; and Section 5 concludes the paper and outlines future research directions.

## 2. Related Work

The introduction of the ChestX-ray8 dataset and its associated model Wang et al. (2017) 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 Islam et al. (2017), attention mechanisms Guan et al. (2018); Liu et al. (2019), and localization techniques Cai et al. (2018); Guendel et al. (2019); Li et al. (2018); Yan et al. (2018). 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 label, assuming independence between labels Zhang and Zhou (2014). However, BR-based techniques do not account for label dependence, either conditional (instance-specific label dependence. In a given instance, the presence or absence of one label may impact another's.) or marginal (dataset-specific label dependence: certain labels may co-occur more frequently.) Dembczyński et al. (2012).

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 Harvey and Glocker (2019); Tsoumakas and Katakis (2007). 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 Bi and Kwok (2015). Early methods used hierarchical and multi-label gener-

alizations of traditional algorithms, such as nearest-neighbor or multi-layer perceptrons Pourghassem and Ghassemian (2008) and decision trees Dimitrovski et al. (2011). With the rise of deep learning, the adaptation of convolutional neural networks (CNN) for hierarchical classification has gained increasing attention Guo et al. (2018); Kowsari et al. (2017); Redmon and Farhadi (2017); Roy et al. (2020).

**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 labels in the hierarchy Van Eeden et al. (2012). 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, followed by specific patterns of lung consolidation suggesting pneumonia (child label). Consequently, it is possible to make more accurate diagnoses based on the data considering the relationships 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 Alaydie et al. (2012). 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 Aly et al. (2019), visual recognition Bi and Kwok (2014), and genomic analysis Bi and Kwok (2015). A common technique Chen et al. (2019) 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 Chen et al. (2019). These approaches help the classifier focus on the relevant data during initial training, improving prediction accuracy. It also allows the classifier to consider label hierarchies. 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. Additionally, they may not always perform satisfactorily, as it may not be possible to find a good set of parent-level labels that accurately capture the hierarchical relationships in the data. Another common strategy is cascading architecture where dif-

ferent 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) Guo et al. (2018); Kowsari et al. (2017).

We propose a method that takes advantage of hierarchical relationships between labels without imposing computational requirements. Our proposed method is adaptable to the computational capacity of the user. If sufficient computational resources are available, it can be used as a standalone loss function during the optimization process, or it can be applied to test samples without the need to fine-tune the pre-trained ML model. The proposed loss function is based on the following hypothesis.

- The highest level of taxonomy contained the most general labels, whereas the lowest level contained the least.
- Each node contains a collection of granular child labels.

## 3. Methods

In this section, we present a comprehensive methodology to enhance multi-label classification performance using chest radiograph (CXR) data, which is applicable not only during the training phase, but also as a transfer learning approach during the testing phase. Our proposed strategy encompasses the formulation of a multi-label classification problem for CXR, the establishment of an evaluation protocol, and the incorporation of a loss measurement technique that leverages hierarchical label relationships. As a transfer learning approach, our method facilitates the adaptation and fine-tuning of pre-trained models, thereby augmenting their generalizability to novel tasks. This ultimately contributes to the improvement of disease diagnosis and treatment through increased accuracy in detecting abnormalities within CXR images.

### *3.1. Notations*

Let us denote the following parameters:

- $\mathcal{C} = \{c_k\}_{k=1}^{K}, c_k \in \{0, 1\}$: 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 directed edges representing parent-child relationships between classes.
- $y_k^{(i)} \in \{0, 1\}$: true label for the $k$-th class($c_k$) of instance $i$.
- $q_k^{(i)} \in (-\infty, 0)$: logits obtained in the last layer of the neural network model before the sigmoid layer.
- $p_k^{(i)} = \text{sigmoid}\left(q_k^{(i)}\right) = \frac{1}{1+\exp\left(-q_k^{(i)}\right)}$: 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.
- $\theta_k$: Binarization threshold for class $k$. To obtain this, we can utilize
- $t_k^{(i)} = \begin{cases} 1 & \text{if } p_k^{(i)} \geq \theta_k \\ 0 & \text{otherwise.} \end{cases}$ : predicted label obtained by binarizing the $p_k^{(i)}$
- $\widehat{p}_k^{(i)} \in (0, 1)$: updated predicted probability for the $k$-th class of instance $i$ with a value between 0 and 1.
- $t_k^{(i)} = \begin{cases} 1 & \text{if } p_k^{(i)} \geq \theta_k \\ 0 & \text{otherwise.} \end{cases}$ : 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 the $i$-th instance.
- $Y^{(i)} = \left\{y_k^{(i)}\right\}_{k=1}^{K}$: true label set, for instance $i$.
- $P^{(i)} = \left\{p_k^{(i)}\right\}_{k=1}^{K}$: predicted probability set, for instance $i$.
- $T^{(i)} = \left\{t_k^{(i)}\right\}_{k=1}^{K}$: predicted label set, for instance $i$.

- $\mathbb{X} = \left\{X^{(i)}\right\}_{i=1}^{N}$: Set of all instances.
- $\mathbb{Y} = \left\{Y^{(i)}\right\}_{i=1}^{N}$: Set of all true labels.
- $\mathbb{D} = \{\mathbb{X}, \mathbb{Y}\}$: Dataset containing true labels.
- $\mathcal{L}\left(y_k^{(i)}, p_k^{(i)}\right)$: $\mathcal{L}$ is an arbitrary loss function (e.g., binary cross entropy) that takes the true label and predicted probability for class k and instance I and outputs the loss value $l_k^{(i)}$. We will refer to this as the "base loss function" throughout this paper.
- Loss($\theta$): Measured loss in all cases and instances. This value will be obtained using a modified version of the base loss function $\mathcal{L}(\cdot)$ (e.g., with added regularization, etc.).
- $\mathcal{G} = \{\mathcal{C}, \mathcal{E}\}$: directed acyclic graph (DAG) $\mathcal{G}$ represents the taxonomy of thoracic diseases, where $\mathcal{C}$ is the set of disease classes and $E$ is the set of directed edges representing parent-child relationships between these classes.
- $\Lambda(c_k) \subset \mathcal{C}$: set of parent classes of class $c_k$ in DAG $\mathcal{G}$.
- $\mathcal{J}(c_k) \subset \mathcal{C}$: set of child classes of class $c_k$ in DAG $\mathcal{G}$
- $\omega_k^{(i)}$: Estimated weight for $k$-th class $c_k$ of instance $i$ with respect to its parent class $\Gamma_k$.
- $\widehat{l}_k^{(i)} = \omega_k^{(i)}\ l_k^{(i)}$: updated loss for class $k$ and instance $i$.
- $\widehat{p}_k^{(i)} = \omega_k^{(i)}\ p_k^{(i)}$: updated predicted probability for the $k$ -th class.

### *3.2. Problem Formulation*

Let us define the multi-label classification problem as follows. Let $\mathbb{X} = \left\{X^{(i)}\right\}_{i=1}^{N}$ be the set of $N$ chest radiograph images and $\mathbb{Y} = \left\{Y^{(i)}\right\}_{i=1}^{N}$ be their corresponding ground truth labels. In the context of chest radiograph interpretation, the label set $\mathcal{C}$ typically includes various thoracic abnormalities such as pneumothorax, consolidation, atelectasis, and cardiomegaly. 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 $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, let us define a directed acyclic graph (DAG) $\mathcal{G} = \{\mathcal{C}, \mathcal{E}\}$ representing the taxonomy of thoracic diseases, where $E$ is the set of directed edges representing parent-child relationships between these classes. For each node $c_k \in \mathcal{C}$, let $\Lambda(c_k) \in \mathcal{C} : c_k \neq \Lambda(c_k)$ the parent node of class $c_k$ and denote $\mathcal{J}(c_k) \subset \mathcal{C}$ the set of child classes of class $c_k$ in DAG $\mathcal{G}$. The root node does not represent an abnormality (ie, normal chest radiograph).

Let $\omega_k^{(i)}$ be a scalar weight assigned to the class $c_k$ of instance $i$ with respect to its parent class $\Lambda(c_k)$. In multi-label classification problems, each sample can have multiple labels simultaneously assigned to it; thus, the sigmoid function is utilized to predict the probabilities for each class being present in a given sample. 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 between 0 and 1 corresponding to the label set $\mathcal{C}$ to obtain a set of $K$ predicted probabilities $P^{(i)} = \left\{p_k^{(i)}\right\}_{k=1}^{K}$. These predicted probabilities, derived from the sigmoid activation function, can be interpreted as the probability that the input sample belongs to each class. Consequently, the loss function quantifies the similarity between predicted and true labels.

Let us denote $l_k = \mathcal{L}\left(p_k^{(i)}, y_k^{(i)}\right),\ k \in \{1, 2, \ldots, K\}$ where $\mathcal{L}(\cdot)$ is an arbitrary and 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 sample $X^{(i)}$ and class $k$.

### *3.3. Label Taxonomy and Hierarchy*

To exploit the inherent hierarchical relationships between thoracic abnormalities, the first step is to define a disease taxonomy that demonstrates different abnormalities interrelationships. In this taxonomy, diseases will be structured hierarchically, with higher levels representing broader disease categories and lower levels representing more nuanced distinctions between related diseases. For example, pleural effusion and pneumothorax can be categorized as subcategories of pleural abnormalities, whereas atelectasis and consolidation can be classified under pulmonary opacity. This hierarchical structure enables the model to take advantage of the relationships between diseases to improve its classification performance.

In medical imaging, labels are frequently organized as trees or directed acyclic graphs (DAGs) to represent the hierarchical relationships between

different classes of labels. For example, a DAG 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 tree or DAG 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. Construction of these hierarchies can be challenging and time-consuming because it requires in-depth knowledge of the subject matter and the ability to organize complex data into clean and intuitive structures.

To develop a comprehensive label taxonomy for lung diseases, we integrated the taxonomies presented by Irvin Irvin et al. (2019) for the CheXpert dataset and Chen Chen et al. (2020) for the PadChest and PLCO datasets. This unified taxonomical structure can be applied to various chest radiography datasets. In the following two sections, we propose two methods for incorporating taxonomy information to improve accuracy.

- In the first approach, we use taxonomy information to update the predicted probability of each class based on the predicted probability of its respective parent classes. This method can be easily applied unsupervised to existing, pre-trained models without the need for true labels.

- In our second approach, we propose a similar concept. 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.

### *3.4. Approach 1: Conditional Predicted Probability*

A transfer learning-based approach that uses hyperparameters to update the predicted probability of a class based on the predicted probability of its parent class can be devised to further enhance the accuracy of classification by considering the interrelationship between different classes. In this approach, our aim is to calculate the conditional predicted probability for each class $k$ and instance $i$, taking into account the predicted probabilities 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.

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

where $j$ is defined so that $c_j = \Lambda(c_k)$, and $\alpha_{k,j}$ is the hyperparameter controlling the influence of different parent class logits on child class logits. When $\alpha_{k,j} = 0$, there is no influence from the parent classes, and when $\alpha_{k,j} > 0$, it introduces a degree of dependency between the child and parent classes in terms of their predicted probabilities.

By carefully selecting appropriate hyperparameter values, this transfer learning-based technique can be employed to effectively adjust the predicted probabilities of each class, considering the hierarchical relationship between classes, and potentially improving classification accuracy.

*3.4.1. Parameter Selection and Tuning*

The selection of appropriate hyperparameters is crucial for the effectiveness of the proposed transfer learning-based technique. In this study, we employ a systematic approach to tune the hyperparameters $\alpha_{k,j}$, which control 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.

*3.4.2. Adaptive Computation for Real-World Applications*

The proposed transfer learning-based technique is adaptable to the user's computational capacity, making it suitable for real-world applications with varying computational resources. When sufficient computational resources are available, the method can be employed as a standalone loss function during the optimization process. On the other hand, when computational resources are limited, the 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 the calibration of the model output to more closely match the true label distribution.

### *3.5. Approach 2: Conditional Loss*

#### *3.5.1. Disadvantages of conditional predicted probability*

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 the testing phase, it presents some challenges if we use it during the training phase of our classifier model. Among these disadvantages are the following.

- **Loss of interpretability:** Direct updates can obscure the effects of the optimization procedure, as the relationship between loss and predictions may become obscured.
- **Inconsistency with the optimization process:** Directly updating 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.

*3.5.2. Advantages of conditional loss function*

This approach offers several benefits in the context of multi-label classification tasks with hierarchical relationships.

- **Emphasis on error minimization:** Loss values represent the discrepancy between model predictions and ground-truth labels. Incorporating parent class loss values into child class loss calculations focuses on minimizing errors across the hierarchy, thereby ensuring accurate predictions for both parent and child classes.

- **Enhanced gradient propagation:** Gradients are backpropagated through layers to update the model parameters during deep learning model training. Using parent class loss values in child class loss calculations strengthens the connection between parent and child classes in terms of gradient propagation, which could result in more efficient learning of hierarchical relationships and faster convergence during training.

- **Robustness to label noise:** Ground truth labels may contain inconsistencies or noise in real-world datasets. Incorporating parent-class loss values into child-class loss calculations promotes hierarchy consistency by penalizing deviations from expected parent-child relationships, thereby increasing the model's robustness to potential label noise within the dataset.

- **Improved interpretability:** Using loss values rather than predicted probabilities enables a more direct interpretation of the model's ability to capture hierarchical relationships between classes. High loss values for parent classes have a greater effect on the losses of their respective child classes, highlighting the need for improvement in certain areas to more accurately represent these relationships.

*3.5.3. 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 instance, in the case of cross-entropy loss, we have:

$$l_k = -\left(y_k^{(i)} \log(p_k^{(i)}) + (1 - y_k^{(i)}) \log(1 - p_k^{(i)})\right) \tag{2}$$

$$\text{Loss}(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} l_k \tag{3}$$

In this formulation, the objective is to minimize the loss function with respect to the model parameters $\theta$, 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 (e.g., classification of thoracic diseases). In this study, we demonstrate how to incorporate the interdependence between different classes (i.e., thoracic diseases) into our loss measurement, thus improving the overall classification accuracy. We introduce multiple approaches in which this can be achieved using either the parent class loss value or its predicted probability value.

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 hyperparameters. This term penalizes the loss for class $k$ for each instance $i$ in which the likelihood that its parent class exists in that instance is low. This can be represented mathematically by adding a hierarchical penalty term equal to the sum of the loss values of all parent classes of class $k$ as follows.

$$\widehat{l}_k^{(i)} = l_k^{(i)} + \beta_k H(k|j) \tag{4}$$

where $j$ is defined such that $c_j = \Lambda(c_k)$, and $\beta_k$ is the hyperparameter that balances the contributions of the class k's own loss value and its parent classes' loss values. $\Lambda(c_k)$ denote the set of parent classes for class $k$.

There are multiple ways to define the hierarchical penalty. For example, in one approach, we can define it as the loss value of the parent class $l_j = L\left(y_j^{(i)}, p_j^{(i)}\right)$ as follows.

$$H(k|j) = \mathcal{L}\left(y_j^{(i)}, p_j^{(i)}\right) \tag{5}$$

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.

$$H\left(k|j\right) = \mathcal{L}\left(y_j^{(i)}, p_k^{(i)}\right) \tag{6}$$

In both Equations (5) and (6) the penalization term encourages the model to correctly predict the parent labels when predicting the child labels, ensuring that the predicted label set adheres to the hierarchical structure. In the aforementioned approaches, we assume a linear relationship between child and parent losses, which can simplify the optimization process. However, this may not always accurately capture the relationship between the parent-child classes, as the relationship may not always be linear. Furthermore, the impact of the parent's loss on the total loss could be less significant, particularly if the child's loss is considerably greater than the parent's loss.

To address this problem, we can modify the loss measurements presented in Equations (5) and (6) to be based on the multiplication of losses rather than their addition. Multiplying losses allows for a more flexible relationship between the child and parent losses, as it can model both linear and nonlinear relationships. Furthermore, the parent's loss can have a more significant impact on the total loss, since it is multiplied by the child's loss, ensuring that the hierarchical relationships are better captured. To achieve this, we can define the new loss as follows.

$$\widehat{l}_k^{(i)} = l_k^{(i)} H\left(k|j\right) \tag{7}$$

where the hierarchical penalty term is defined as follows.

$$H(k|j) = \begin{cases} 1 & \text{if } \Lambda(c_k) = \varnothing \\ a_k l_j^{(i)} + \beta_k & \text{otherwise.} \end{cases} \tag{8}$$

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

The modified loss function in Equation (7) aims to ensure that predictions adhere to hierarchical relationships between labels by penalizing deviations from these established relationships. Adjusting the weighting parameters $\alpha_k$ and $\beta_k$, we can regulate the extent to which hierarchical information influences the learning process.

### *3.6. Updating Loss Values and Predicted Probabilities*

In the previous section, we showed how a taxonomy-based loss function can be used to improve 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 our proposed taxonomy-based transfer learning approach can be seamlessly integrated into the existing classification framework without the necessity to re-run the optimization phase of our classifier (e.g., DenseNet121). This can be achieved by focusing on updating the loss values and predicted probabilities to incorporate the hierarchical relationships present in the taxonomy. We demonstrate how the interdependence between different classes, as represented by the hierarchical taxonomy, can be effectively captured through the adjustment of loss values and predicted probabilities. This ensures that the classifier's performance is enhanced while respecting the inherent structure of the disease taxonomy, ultimately leading to improved diagnosis and better patient outcomes.

To calculate the predicted probability based on the loss, we must first define a loss function that quantifies the difference between the predicted probability and the true label. Once the loss function has been defined, during a training phase of a classifier (e.g., DenseNet121), an optimization algorithm such as gradient descent can be 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 this section, we demonstrate how we can directly calculate the new predicted probabilities from their new loss values obtained in Equation (7). Let us assume that binary cross entropy is used for the choice of the loss function $\mathcal{L}(\cdot)$. Furthermore, let us denote $\widehat{q}_k^{(i)}, \widehat{p}_k^{(i)}$ as the updated values we are looking for showing the logit and predicted probability of class $k$ and instance $i$ after applying the proposed technique. As discussed above, to

calculate the predicted probabilities, we need to pass the logits $\widehat{q}_k^{(i)}$ into a sigmoid function as shown below:

$$\widehat{p}_k^{(i)} = \text{sigmoid}\left(\widehat{q}_k^{(i)}\right) = \frac{1}{1+\exp\left(-\widehat{q}_k^{(i)}\right)} \tag{9}$$

The sigmoid activation function maps any value to a number between zero and one. The sigmoid function is defined as follows. The gradient of the sigmoid function provides the direction in which the predicted probability must be updated.

$$\text{sigmoid}'\left(\widehat{q}_k^{(i)}\right) = \frac{\partial \text{sigmoid}}{\partial q} = \text{sigmoid}\left(\widehat{q}_k^{(i)}\right)\left(1-\text{sigmoid}\left(\widehat{q}_k^{(i)}\right)\right) = \widehat{p}_k^{(i)}\left(1-\widehat{p}_k^{(i)}\right) \tag{10}$$

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 will be as follows.

$$\frac{\partial \mathcal{L}\left(\widehat{p}_k^{(i)},\ y_k^{(i)}\right)}{\partial \widehat{p}} = \frac{y_k^{(i)}}{\widehat{p}_k^{(i)}} - \frac{1-y_k^{(i)}}{1-\widehat{p}_k^{(i)}} \tag{11}$$

where $y_k^{(i)}$ and $\widehat{p}_k^{(i)}$ are the true label and predicted probability, respectively, for instance $i$ and class $k$.

In the following equations, we show how we can use the predicted probability, the gradient loss shown in Equation (11) and the derivative of the sigmoid function shown in Equation (10) to calculate the updated predicted probability.

$$\frac{\partial \mathcal{L}\left(p_k^{(i)},\ y_k^{(i)}\right)}{\partial p}\ \text{sigmoid}'\left(\widehat{q}_k^{(i)}\right) = \left(\frac{y_k^{(i)}}{\widehat{p}_k^{(i)}} - \frac{1-y_k^{(i)}}{1-\widehat{p}_k^{(i)}}\right)\widehat{p}_k^{(i)}\left(1-\widehat{p}_k^{(i)}\right) = y_k^{(i)} - \widehat{p}_k^{(i)} \tag{12}$$

Hence, we can conclude the following.

$$\hat{p}_k^{(i)} = \begin{cases} -\frac{\partial \mathcal{L}\left(p_k^{(i)},\, y_k^{(i)}\right)}{\partial p}\ \text{sigmoid}'\left(\widehat{q}_k^{(i)}\right) + 1 & y = 1 \\ -\frac{\partial \mathcal{L}\left(p_k^{(i)},\, y_k^{(i)}\right)}{\partial p}\ \text{sigmoid}'\left(\widehat{q}_k^{(i)}\right) & \text{otherwise.} \end{cases} \tag{13}$$

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 (11) we will have the following:

$$\frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}} = \frac{y_k^{(i)}}{\widehat{p}_k^{(i)}} - \frac{1 - y_k^{(i)}}{1 - \widehat{p}_k^{(i)}} = \frac{y_k^{(i)} - \widehat{p}_k^{(i)}}{\widehat{p}_k^{(i)}(1 - \widehat{p}_k^{(i)})} \tag{14}$$

In this equation, we can see that when the true label is positive $\left(y_k^{(i)} = 1\right)$, the loss gradient can only be 0 or a positive number. Similarly, when $\left(y_k^{(i)} = 0\right)$, the loss gradient can only take the value 0 or a negative number. Thus, we can modify the Equation (13) to look as follows.

$$\widehat{p}_k^{(i)} = \begin{cases} -\frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}} \operatorname{sigmoid}'(q_k^{(i)}) + 1 & \text{if} \quad \frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}} \geq 0 \\ -\frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}} \operatorname{sigmoid}'(q_k^{(i)}) & \text{otherwise.} \end{cases} \tag{15}$$

Finally, the Equation (15) can be simplified as follows.

$$\widehat{p}_k^{(i)} = \begin{cases} \exp(-\widehat{l}_k^{(i)}) & \text{if} \quad \frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}} \geq 0 \\ 1 - \exp(-\widehat{l}_k^{(i)}) & \text{otherwise} \end{cases} \tag{16}$$

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

The following demonstrates the Equation (16) based on predicted probability syntax to demonstrate its similarity to Equation (1) in Approach 1. From Equation (8) we have $l_k^{(i)} = l_k^{(i)}\left(\alpha_k\, l_j^{(i)} + \beta_k\right)$. By substituting that into $\exp\left(-\widehat{l}_k^{(i)}\right), y_k^{(i)} = 1$ we would have the following equation.

$$\exp\left(-\widehat{l}_k^{(i)}\right) = \exp\left(-l_k^{(i)}\left(\alpha_k\, l_j^{(i)} + \beta_k\right)\right) = \left(p_k^{(i)}\right)^{-\alpha_k \log\left(p_j^{(i)}\right) + \beta_k} \tag{17}$$

Furthermore, $1 - \exp\left(-\widehat{l}_k^{(i)}\right),\ y_k^{(i)} = 0$ will be as follows.

$$1 - \exp\left(-\widehat{l}_k^{(i)}\right) = 1 - \exp\left(-l_k^{(i)}\left(\alpha_k\, l_j^{(i)} + \beta_k\right)\right) = 1 - \left(1 - p_k^{(i)}\right)^{-\alpha_k \log\left(1 - p_j^{(i)}\right) + \beta_k} \tag{18}$$

By substituting the Equations (17) and (18) into Equation (16) we will have the following.

$$\widehat{p}_k^{(i)} = \begin{cases} \left(p_k^{(i)}\right)^{-\alpha_k \log(p_j^{(i)})+\beta_k} & \text{if} \quad y_k^{(i)} = 1 \\ 1 - \left(1 - p_k^{(i)}\right)^{-\alpha_k \log\left(1-p_j^{(i)}\right)+\beta_k} & \text{otherwise} \end{cases} \tag{19}$$

### *3.7. Interpretability Enhancement*

One of the key benefits of our proposed method is the enhancement of interpretability. By organizing diseases into a hierarchical structure and leveraging their relationships, the model not only improves classification performance, but also provides insights into the relationships among predicted diseases. This additional layer of interpretability can help radiologists understand the rationale behind the model's predictions, building trust in the model output, and facilitating 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.

### *3.8. Experimental Setup*

#### *3.8.1. Datasets*

We utilized three diverse and publicly available datasets for the evaluation of our proposed hierarchical multi-label classification technique: CheXpert Irvin et al. (2019), PadChest Bustos et al. (2020), and VinDr-CXR Nguyen et al. (2022). These datasets contain a diverse range of chest radiographic images covering various thoracic diseases, providing a comprehensive evaluation of the effectiveness of our method.

**Dataset Description**

- **CheXpert** Irvin et al. (2019) is a large-scale dataset containing 224,316 chest radiographs of 65,240 patients, labeled with 14 radiographic findings.
- **PadChest** Bustos et al. (2020) 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** Wang et al. (2017) includes 112,120 chest radiographs of 30,805 patients labeled with 14 categories of thoracic diseases.

**Preprocessing**

The input chest radiographs were pre-processed to ensure consistency across the datasets. The images were resized to a resolution of $224 \times 224$ pixels, with the pixel intensities normalized to a range of [0, 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.

*3.8.2. Model Architecture and Training Details*

The pre-trained model provided by Cohen Cohen et al. (2022) was used as the base model. The model was fine-tuned using DenseNet121 Huang et al. (2017) architecture on a subset of CheXpert Irvin et al. (2019), NIH Wang et al. (2017), PadChest Bustos et al. (2020) for 18 toracic 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 method, we utilized 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. In this investigation, we employed parallelization libraries, such as joblib and Python multiprocessing, which enable the concurrent execution of multiple tasks while sharing available resources. These libraries facilitate the implementation of parallelism in our optimization process, ensuring seamless integration with the existing framework and offering a scalable and hardware-adaptable solution.

**Optimum Threshold Determination**

Determining the optimal threshold is a crucial aspect of evaluating the performance of our 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 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.

*3.8.3. Evaluation*

**Model Evaluation and Comparison**

The performance of proposed methods were tested by training a Desenet121 architecture on all three public datasets (CheXpert, NIH, PADCHEST). To assess the performance of the proposed techniques, we employ several evaluation metrics that capture different aspects of the classification problem. These metrics include precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). By analyzing the performance of the proposed methods across these metrics, we can gain insights into the effectiveness of the methods in capturing inter-class relationships and improving classification accuracy. Additionally, the comparison of the proposed method with baseline approaches provides a comprehensive understanding of the method's improvements and potential for real-world implementation. The following metrics were used to evaluate the effectiveness of the proposed method.

- **Accuracy**: Proportion of correctly classified samples to the total number of samples.
- **F1-score**: Harmonic mean of precision and recall, providing a balanced assessment of the method's performance.
- **Area Under the Receiver Operating Characteristic Curve (AUROC)**: a summary measure of the true positive rate versus the false positive rate at different classification thresholds.

## 4. Results

### *4.1. Distribution of Pathologies in Public Chest Radiograph Datasets*

In this section, we present the distribution of different pathologies across three major public chest radiograph datasets: CheX, NIH, and PC. Table 1 compares the original and updated label sets for each dataset. The original counts represent the raw number of samples in the datasets, while the updated counts account for the absence of parent pathology labels in some datasets by considering a parent pathology as true if at least one of its child pathologies is present for that sample.

In the original label sets, the total number of samples across the datasets were 20,543 for CheX, 28,868 for NIH and 61,692 for PC. After updating the label sets, the total number of samples remained unchanged for CheX and NIH, while for PC, it increased to 61,692. The distribution of pathologies in the datasets varies, with some pathologies such as Atelectasis, Effusion and Lung Opacity having a higher prevalence, while others such as Hernia and Fibrosis showing a lower prevalence across the datasets.

[Table 1 about here.]

[Figure 1 about here.]

Table 1: Details of the datasets included in this library. The number of images shows total images / usable frontal images. Usable frontal means images that are readable, have all the necessary metadata, and are in AP, PA, AP Supine, or AP Erect view.

Table 2: Labels available for each dataset, the total number of positive samples for each class across all datasets, and the total number of examples in each dataset, and the sum over each row in the right column. The COVID-19 datasets are excluded from this table because they have many unique pathologies.

### *4.2. Model Performance on Public Chest Radiograph Datasets*

*AUC*

In this section, we present the performance of the three methods—baseline, "logit", and "loss"; on the three public chest radiograph datasets (CheX, NIH, and PC) in terms of AUC metrics for various pathologies. The baseline represents the original model's performance, while "logit"

and "loss" refer to the proposed modified logits and modified loss approaches, respectively. A single model was trained on all three datasets and evaluated on the test cases from each dataset 2.

The results demonstrate varying performance across the pathologies and datasets. In general, the proposed "logit" and "loss" approaches show improved performance compared to the baseline, with several pathologies, such as Atelectasis, Consolidation, Edema, and Pneumonia, exhibiting significant improvements in AUC metrics. For instance, in the CheX dataset, the AUC for Atelectasis increased from 0.811 in the baseline to 0.960 and 1.000 in the "logit" and "loss" methods, respectively.

However, for some pathologies like Pneumothorax, Emphysema, and Pleural_Thickening, the performance remained consistent across all three approaches. In some cases, such as Mass and Nodule in the PC dataset, the performance was notably lower than in the other datasets.

[Table 2 about here.]

*F1-score*

In this section, we discuss the F1-score performance of the three methods—baseline, "logit", and "loss"—on the three public chest radiograph datasets (CheX, NIH, and PC) for various pathologies 3. The baseline represents the original model's performance, while "logit" and "loss" refer to the proposed modified logits and loss approaches, respectively. As before, a single model was trained on all three datasets and evaluated on the test cases from each dataset.

The F1-scores reveal that the proposed "logit" and "loss" approaches generally show improved performance compared to the baseline method. For example, in the CheX dataset, the F1-score for Atelectasis increased from 0.201 in the baseline to 0.353 and 0.927 in the "logit" and "loss" methods, respectively. Similarly, the F1-score for Edema increased from 0.352 in the baseline to 0.733 and 0.829 in the "logit" and "loss" methods, respectively.

However, for some pathologies such as Pneumothorax, Emphysema, and Pleural_Thickening, the F1-scores remained consistently low across all three approaches. This result indicates that there is room for further improvement in these areas.

[Table 3 about here.]

The accuracy of the three methods (baseline, "logit", and "loss") was evaluated for the CheX, NIH, and PC datasets for different pathologies, as presented in Table 4. Comparing the baseline method to the "logit" and "loss" methods, we observed improvements in accuracy across most pathologies in all three datasets. For instance, in the CheX dataset, the accuracy for Atelectasis increased from 0.593 in the baseline method to 0.796 and 0.992 in the "logit" and "loss" methods, respectively. Similarly, in the NIH dataset, the accuracy for Edema improved from 0.972 in the baseline method to 0.971 and 0.972 for the "logit" and "loss" methods, respectively.

In some cases, the "logit" and "loss" methods achieved comparable performance, while in others, one method outperformed the other. For example, in the PC dataset, the accuracy for Pneumonia improved from 0.862 in the baseline to 0.806 and 0.887 in the "logit" and "loss" methods, respectively, with the "loss" method yielding higher accuracy. These results demonstrate the effectiveness of the proposed modifications in improving the performance of the models across various chest radiograph pathologies.

[Table 4 about here.]

[Figure 2 about here.]

Figure 2 illustrates a comparison between the performance of the baseline technique and the proposed logit-based method (Approach 1 discussed in the Methods section) in detecting eight thoracic pathologies. These eight pathologies include the pathologies with child classes and their respective child classes, as shown in Figure 1. The individual subplots exhibit overlaid receiver operating characteristic (ROC) curves, each of which corresponds to a specific pathology. The present analysis employed a model that was derived through the application of the test dataset from the NIH dataset to a pre-trained model. The latter had undergone training on various publicly available thoracic datasets, with the aim of allowing the identification of 18 different pathologies related to the thorax. The last two subplots, showcased with a darker background, showcase the roc curves for parent classes diseases. As these parent class diseases were not influenced by the proposed technique, their ROC curves and corresponding Area Under the Curve (AUC) values remain consistent with the baseline technique. The ROC curve and its corresponding AUC for the six child classes demonstrate a significant improvement for the proposed technique in comparison to the baseline technique.

[Figure 3 about here.]

Figure 3 illustrates a comparison between the performance of the baseline technique and the proposed loss-based method (Approach 2 discussed in the Methods section) in detecting eight thoracic pathologies. These eight pathologies include the pathologies with child classes and their respective child classes, as shown in Figure 1. The individual subplots exhibit overlaid ROC curves, each of which corresponds to a specific pathology. The present analysis employed a model that was derived through the application of the test dataset from the NIH dataset to a pre-trained model. The latter had undergone training on various publicly available thoracic datasets, with the aim of allowing the identification of 18 different pathologies related to the thorax. The last two subplots, showcased with a darker background, showcase the roc curves for parent classes diseases. As these parent class diseases were not influenced by the proposed technique, their ROC curves and corresponding AUC values remain consistent with the baseline technique. The ROC curve and its corresponding AUC for the six child classes demonstrate a significant improvement for the proposed technique in comparison to the baseline technique.

## 5. Discussion and Conclusion

The study presented two Hierarchical Multilabel Classification Methods for Enhanced Thoracic Disease Diagnosis in Chest Radiography. One method refered to as "loss" udpates the value of loss for pathologies according to their parent pathologies. This method is purticularly useful for both fine-tuning the existing pre-trained models and training from scratch. A second method refered to as "logit" is also proposed where the logit values of each pathology is updated based on the logit value of their parent pathologies. This technique is particularly useful when the ground truth labels are not available. This method improves the performance of existing pre-trained models solely based on the taxonomical relationship of pathologies in the model without any need for availablility of ground truth labels.

The results, as indicated in Tables **??** and 3, show the F1 score and AUC performance of two methods ("logit", and "loss") with respect to baseline on the CheX, NIH, and PC chest radiograph datasets for various pathologies. Hierarchical multi-label classification approaches demonstrated a significant improvement in the accuracy and efficiency of thoracic disease diagnosis. This

was particularly evident in the AUC performance, where the "loss" and "logit" methods consistently outperformed the "baseline" across most pathologies in the CheX, NIH, and PC datasets.

Modifying logits provides a simple yet effective means of incorporating the label hierarchy without substantially changing existing model architectures. However, this approach can obscure the effects of optimization and learning. On the contrary, modifying loss values more directly aligns with the model optimization process and allows fine-tuning of the hierarchical influence through weighting factors. This approach also promotes consistency with established hierarchical relationships and robustness to label noise.

In general, both techniques were effective in using the disease taxonomy to enhance classification performance, indicating the value of leveraging label relationships in medical image classification. The loss-based technique generally showed higher performance gains, suggesting that it may more accurately capture hierarchical dependencies during model training.

The improvement in interpretability offered by these hierarchical techniques can potentially aid clinicians by providing insight into the models' predictions. The ability to explore predictions at different levels of granularity based on taxonomy may facilitate personalized diagnoses based on specific clinical needs.

However, limitations remain. Techniques require predefined label hierarchies, which can be challenging to construct for complex diseases. Further refinement of the hyperparameter tuning procedures may yield even higher performance. Future work could explore the integration of label hierarchies directly into model architectures to achieve end-to-end learning of hierarchical relationships.

In summary, incorporating hierarchical label relationships through modifying either logits or loss functions presents an effective strategy for improving the multi-label classification tasks.

## Appendices

## Acknowledgements

## References

Alaydie, N., Reddy, C. K., and Fotouhi, F. (2012). Exploiting Label Dependency for Hierarchical Multi-Label Classification. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C.,

Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Tan, P.-N., Chawla, S., Ho, C. K., and Bailey, J., editors, *Advances in Knowledge Discovery and Data Mining*, volume 7301, pages 294–305. Springer Berlin Heidelberg, Berlin, Heidelberg.

Aly, R., Remus, S., and Biemann, C. (2019). Hierarchical Multi-Label Classification of Text With Capsule Networks. In *Proc. 57th Annu. Meet. Assoc. Comput. Linguist. Stud. Res. Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Ausawalaithong, W., Thirach, A., Marukatat, S., and Wilaiprasitporn, T. (2018). Automatic Lung Cancer Prediction From Chest X-Ray Images Using the Deep Learning Approach. In *11th Biomed. Eng. Int. Conf. BMEiCON*, pages 1–5, Chiang Mai. IEEE.

Bellaviti, N., Bini, F., Pennacchi, L., Pepe, G., Bodini, B., Ceriani, R., D'Urbano, C., and Vaghi, A. (2016). Increased Incidence of Spontaneous Pneumothorax in Very Young People: Observations and Treatment. *CHEST*, 150(4):560A.

Bi, W. and Kwok, J. T. (2014). Mandatory Leaf Node Prediction in Hierarchical Multilabel Classification. *IEEE Trans. Neural Netw. Learning Syst.*, 25(12):2275–2287.

Bi, W. and Kwok, J. T. (2015). Bayes-Optimal Hierarchical Multilabel Classification. *IEEE Trans. Knowl. Data Eng.*, 27(11):2907–2918.

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2020). Padchest: A Large Chest X-Ray Image Dataset With Multi-Label Annotated Reports. *Medical Image Analysis*, 66:101797.

Cai, J., Lu, L., Harrison, A. P., Shi, X., Chen, P., and Yang, L. (2018). Iterative Attention Mining for Weakly Supervised Thoracic Disease Pattern Localization in Chest X-Rays. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C., and Fichtinger, G., editors, *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2018*, Lecture Notes in Computer Science, pages 589–598, Cham. Springer International Publishing.

Chen, H., Miao, S., Xu, D., Hager, G. D., and Harrison, A. P. (2019). Deep Hierarchical Multi-Label Classification of Chest X-Ray Images. In *Proc. 2nd Int. Conf. Med. Imaging Deep Learn.*, pages 109–120. PMLR.

Chen, H., Miao, S., Xu, D., Hager, G. D., and Harrison, A. P. (2020). Deep Hiearchical Multi-Label Classification Applied to Chest X-Ray Abnormality Taxonomies. *Medical Image Analysis*, 66:101811.

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., and Bertrand, H. (2022). TorchXRayVision: A Library of Chest X-Ray Datasets and Models. In *Proc. 5th Int. Conf. Med. Imaging Deep Learn.*, pages 231–249. PMLR.

Crisp, N. and Chen, L. (2014). Global Supply of Health Professionals. *N Engl J Med*, 370(10):950–957.

Delrue, L., Gosselin, R., Ilsen, B., Van Landeghem, A., de Mey, J., and Duyck, P. (2011). Difficulties in the Interpretation of Chest Radiography. In Coche, E. E., Ghaye, B., de Mey, J., and Duyck, P., editors, *Comparative Interpretation of CT and Standard Radiography of the Chest*, Medical Radiology, pages 27–49. Springer, Berlin, Heidelberg.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. (2012). On Label Dependence and Loss Minimization in Multi-Label Classification. *Mach Learn*, 88(1-2):5–45.

Dimitrovski, I., Kocev, D., Loskovska, S., and Džeroski, S. (2011). Hierarchical Annotation of Medical Images. *Pattern Recognition*, 44(10-11):2436–2449.

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., and Yang, Y. (2018). Diagnose Like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification.

Guendel, S., Ghesu, F. C., Grbic, S., Gibson, E., Georgescu, B., Maier, A., and Comaniciu, D. (2019). Multi-Task Learning for Chest X-Ray Abnormality Classification on Noisy Labels.

Guo, Y., Liu, Y., Bakker, E. M., Guo, Y., and Lew, M. S. (2018). CNN-RNN: A Large-Scale Hierarchical Image Classification Framework. *Multimed Tools Appl*, 77(8):10251–10271.

Harvey, H. and Glocker, B. (2019). A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology. In Ranschaert, E. R., Morozov, S., and Algra, P. R., editors, *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*, pages 61–72. Springer International Publishing, Cham.

Huang, G., Liu, Z., Maaten, L. V. D., and Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. In *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017*.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset With Uncertainty Labels and Expert Comparison. In *Proc. AAAI Conf. Artif. Intell.*, volume 33, pages 590–597.

Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K. (2017). Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks.

Jaderberg, M., Simonyan, K., Zisserman, A., and kavukcuoglu, k. (2015). Spatial Transformer Networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Adv. Neural Inf. Process. Syst.*, volume 28. Curran Associates, Inc.

Jaiswal, A. K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., and Rodrigues, J. J. P. C. (2019). Identifying Pneumonia in Chest X-Rays: A Deep Learning Approach. *Measurement*, 145:511–518.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). MIMIC-CXR, a De-Identified Publicly Available Database of Chest Radiographs With Free-Text Reports. *Sci Data*, 6(1):317.

Kowsari, K., Brown, D. E., Heidarysafa, M., Jafari Meimandi, K., Gerber, M. S., and Barnes, L. E. (2017). HDLTex: Hierarchical Deep Learning for

Text Classification. In *16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA*, pages 364–371, Cancun, Mexico. IEEE.

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

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., and Fei-Fei, L. (2018). Thoracic Disease Identification and Localization With Limited Supervision. In *IEEECVF Conf. Comput. Vis. Pattern Recognit.*, pages 8290–8299, Salt Lake City, UT. IEEE.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88.

Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., and Pu, J. (2019). SDFN: Segmentation-Based Deep Fusion Network for Thoracic Disease Classification in Chest X-Ray Images. *Computerized Medical Imaging and Graphics*, 75:66–73.

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., Ngo, D. T., Nguyen, N. T., Nguyen, N. T., Dao, M., and Vu, V. (2022). VinDr-CXR: An Open Dataset of Chest X-Rays With Radiologist's Annotations. *Sci Data*, 9(1):429.

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

Pourghassem, H. and Ghassemian, H. (2008). Content-Based Medical Image Classification Using a New Hierarchical Merging Scheme. *Computerized Medical Imaging and Graphics*, 32(8):651–661.

Redmon, J. and Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In *IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, pages 6517–6525, Honolulu, HI. IEEE.

Roy, D., Panda, P., and Roy, K. (2020). Tree-Cnn: A Hierarchical Deep Convolutional Neural Network for Incremental Learning. *Neural Networks*, 121:148–160.

Silverstein, J. (2016). Most of the World Doesn't Have Access to X-Rays. *The Atlantic*.

Tsoumakas, G. and Katakis, I. (2007). Multi-Label Classification: An Overview. *Int. J. Data Warehous. Min.*, 3(3):1–13.

Van Eeden, S., Leipsic, J., Paul Man, S. F., and Sin, D. D. (2012). The Relationship Between Lung Inflammation and Cardiovascular Disease. *Am J Respir Crit Care Med*, 186(1):11–16.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). 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*, pages 3462–3471, Honolulu, HI. IEEE.

Yan, C., Yao, J., Li, R., Xu, Z., and Huang, J. (2018). Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-Rays. In *Int. Conf. Bioinforma. Comput. Biol. Health Inform.*, pages 103–110, Washington DC USA. ACM.

Zhang, M. L. and Zhou, Z. H. (2014). A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837.

## List of Figures

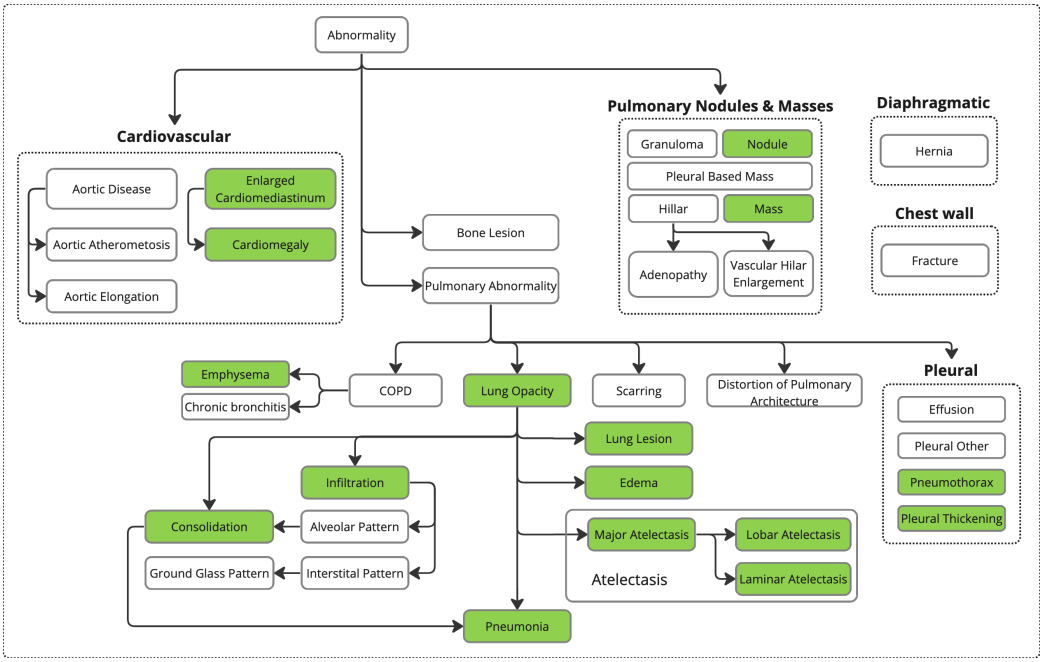
Abnormality
Cardiovascular
Aortic Disease
Enlarged Cardiomediastinum
Aortic Atherometosis
Cardiomegaly
Aortic Elongation
Bone Lesion
Pulmonary Abnormality
Pulmonary Nodules & Masses
Granuloma
Nodule
Pleural Based Mass
Hillar
Mass
Adenopathy
Vascular Hilar Enlargement
Diaphragmatic
Hernia
Chest wall
Fracture
Emphysema
Chronic bronchitis
COPD
Lung Opacity
Scarring
Distortion of Pulmonary Architecture
Pleural
Effusion
Pleural Other
Pneumothorax
Pleural Thickening
Lung Lesion
Edema
Infiltration
Consolidation
Alveolar Pattern
Ground Glass Pattern
Interstital Pattern
Major Atelectasis
Lobar Atelectasis
Laminar Atelectasis
Atelectasis
Pneumonia

Figure 1

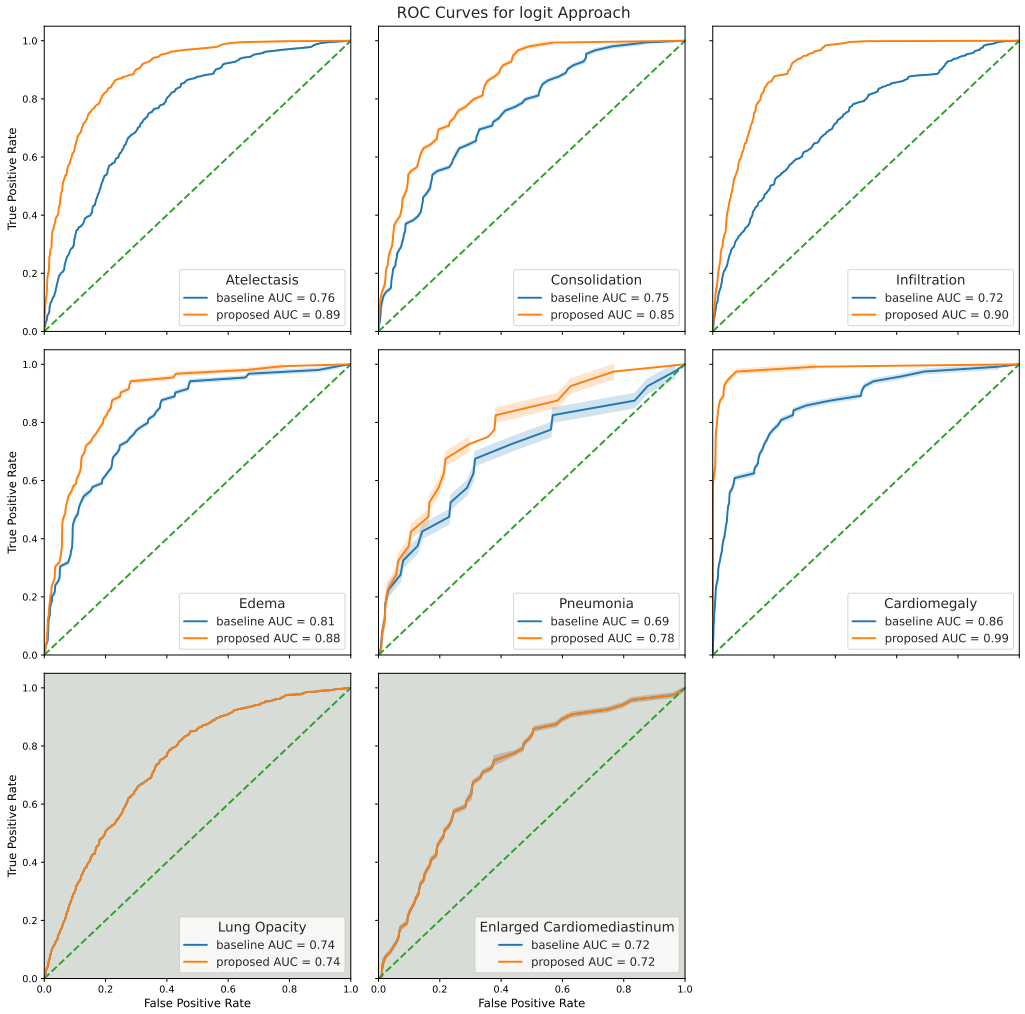
ROC Curves for logit Approach
True Positive Rate
False Positive Rate
Atelectasis
baseline AUC = 0.76
proposed AUC = 0.89
Consolidation
baseline AUC = 0.75
proposed AUC = 0.85
Infiltration
baseline AUC = 0.72
proposed AUC = 0.90
Edema
baseline AUC = 0.81
proposed AUC = 0.88
Pneumonia
baseline AUC = 0.69
proposed AUC = 0.78
Cardiomegaly
baseline AUC = 0.86
proposed AUC = 0.99
Lung Opacity
baseline AUC = 0.74
proposed AUC = 0.74
Enlarged Cardiomediastinum
baseline AUC = 0.72
proposed AUC = 0.72

Figure 2

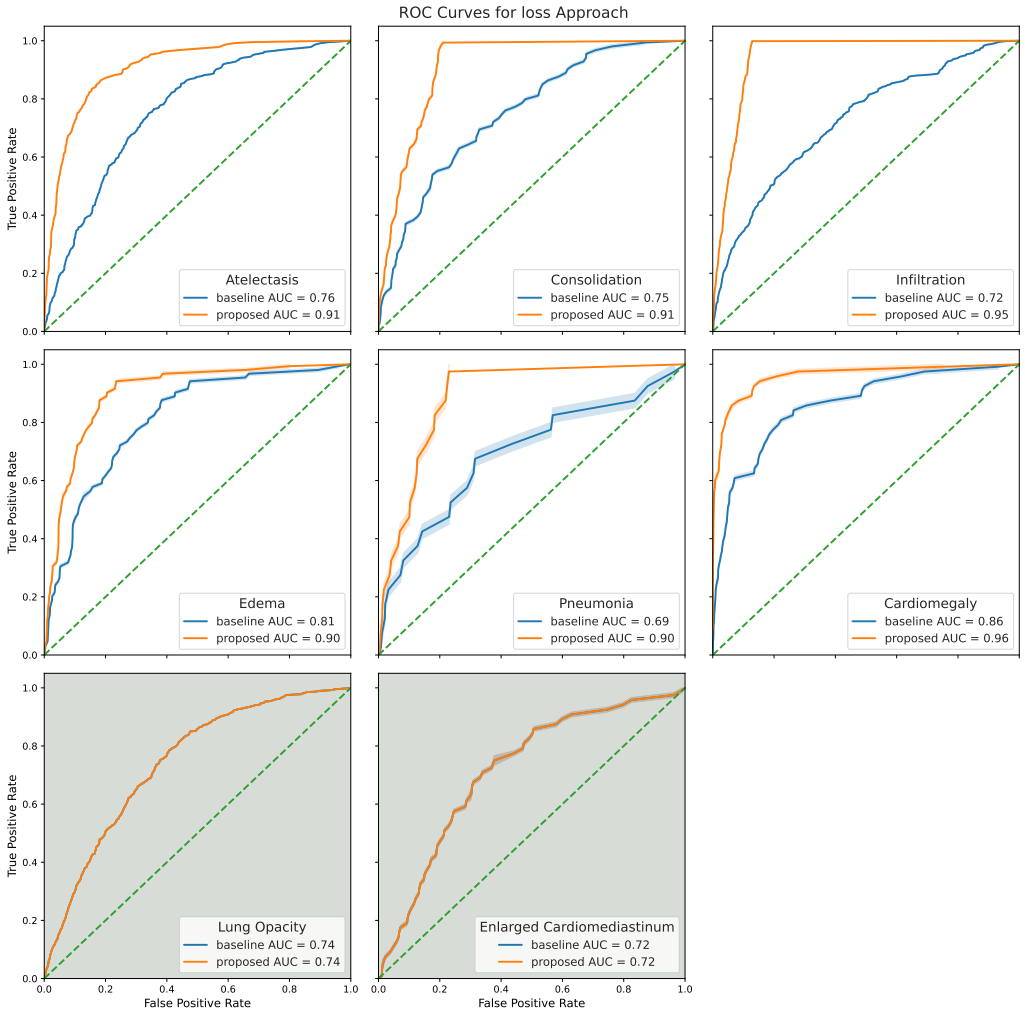
ROC Curves for loss Approach
True Positive Rate
False Positive Rate
Atelectasis
baseline AUC = 0.76
proposed AUC = 0.91
Consolidation
baseline AUC = 0.75
proposed AUC = 0.91
Infiltration
baseline AUC = 0.72
proposed AUC = 0.95
Edema
baseline AUC = 0.81
proposed AUC = 0.90
Pneumonia
baseline AUC = 0.69
proposed AUC = 0.90
Cardiomegaly
baseline AUC = 0.86
proposed AUC = 0.96
Lung Opacity
baseline AUC = 0.74
proposed AUC = 0.74
Enlarged Cardiomediastinum
baseline AUC = 0.72
proposed AUC = 0.72

Figure 3

## List of Tables

Table 1: Comparison of the number of samples for different chest radiograph public datasets (CheX, NIH, and PC) per pathology, considering both original and updated label sets. The original counts represent the raw number of samples in the datasets, while the updated counts show the number of samples after updating the label set for parent pathologies when a dataset does not contain labels for that parent class. In the updated label set, a parent pathology is considered true if at least one of its child pathologies is present for that sample; otherwise, it is set to false.

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

Table 2: AUC performance of the three methods (baseline, "logit", and "loss") on the CheX, NIH, and PC chest radiograph datasets for various pathologies.

| | CheX | | | NIH | | | PC | | |
|---|---|---|---|---|---|---|---|---|---|
| pathologies\approach | **baseline** | **on_logit** | **on_loss** | **baseline** | **on_logit** | **on_loss** | **baseline** | **on_logit** | **on_loss** |
| **Atelectasis** | 0.811 | 0.96 | 1 | 0.759 | 0.89 | 0.908 | 0.747 | 0.867 | 0.905 |
| **Consolidation** | 0.895 | 0.982 | 0.86 | 0.75 | 0.846 | 0.913 | 0.597 | 0.721 | 0.803 |
| **Infiltration** | | | | 0.723 | 0.903 | 0.946 | 0.758 | 0.897 | 0.945 |
| **Pneumothorax** | 0.774 | 0.774 | 0.774 | 0.739 | 0.739 | 0.739 | 0.4 | 0.4 | 0.4 |
| **Edema** | 0.853 | 0.969 | 0.995 | 0.81 | 0.883 | 0.901 | 0.81 | 0.861 | 0.906 |
| **Emphysema** | | | | 0.749 | 0.749 | 0.749 | 0.853 | 0.853 | 0.853 |
| **Fibrosis** | | | | 0.775 | 0.775 | 0.775 | | | |
| **Effusion** | 0.872 | 0.872 | 0.872 | 0.866 | 0.866 | 0.866 | 0.847 | 0.847 | 0.847 |
| **Pneumonia** | 0.854 | 0.947 | 0.999 | 0.693 | 0.779 | 0.898 | 0.62 | 0.74 | 0.846 |
| **Pleural_Thickening** | | | | 0.718 | 0.718 | 0.718 | 0.841 | 0.841 | 0.841 |
| **Cardiomegaly** | 0.86 | 0.911 | 0.998 | 0.859 | 0.986 | 0.96 | 0.776 | 0.97 | 0.911 |
| **Nodule** | | | | 0.751 | 0.751 | 0.751 | 0.383 | 0.383 | 0.383 |
| **Mass** | | | | 0.797 | 0.797 | 0.797 | 0.913 | 0.913 | 0.913 |
| **Hernia** | | | | 0.999 | 0.999 | 0.999 | 0.806 | 0.806 | 0.806 |
| **Lung Lesion** | 0.788 | 0.93 | 1 | | | | | | |
| **Fracture** | 0.736 | 0.736 | 0.736 | | | | 0.742 | 0.742 | 0.742 |
| **Lung Opacity** | 0.804 | 0.804 | 0.804 | 0.742 | 0.742 | 0.742 | 0.782 | 0.782 | 0.782 |
| **Enlarged Cardiomediastinum** | 0.852 | 0.852 | 0.852 | 0.717 | 0.717 | 0.717 | 0.665 | 0.665 | 0.665 |

Table 3: F1-score performance of the three methods (baseline, "logit", and "loss") on the CheX, NIH, and PC chest radiograph datasets for various pathologies.

| | CheX | | | NIH | | | PC | | |
|---|---|---|---|---|---|---|---|---|---|
| pathologies\approach | **baseline** | **on_logit** | **on_loss** | **baseline** | **on_logit** | **on_loss** | **baseline** | **on_logit** | **on_loss** |
| **Atelectasis** | 0.201 | 0.353 | 0.927 | 0.007 | 0.123 | 0.007 | 0.079 | 0.4 | 0.081 |
| **Consolidation** | 0.207 | 0.784 | 0.188 | 0 | 0.048 | 0 | 0 | 0.069 | 0 |
| **Infiltration** | | | | 0.058 | 0.325 | 0.059 | 0.204 | 0.575 | 0.214 |
| **Pneumothorax** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Edema** | 0.352 | 0.733 | 0.829 | 0 | 0 | 0 | 0.107 | 0.241 | 0.115 |
| **Emphysema** | | | | 0 | 0 | 0 | 0 | 0 | 0 |
| **Fibrosis** | | | | 0 | 0 | 0 | | | |
| **Effusion** | 0.328 | 0.328 | 0.328 | 0.27 | 0.27 | 0.27 | 0.35 | 0.35 | 0.35 |
| **Pneumonia** | 0.156 | 0.759 | 0.623 | 0.056 | 0.078 | 0.078 | 0.192 | 0.26 | 0.224 |
| **Pleural_Thickening** | | | | 0 | 0 | 0 | 0 | 0 | 0 |
| **Cardiomegaly** | 0.432 | 0.46 | 0.889 | 0.116 | 0.333 | 0.125 | 0.305 | 0.519 | 0.396 |
| **Nodule** | | | | 0.014 | 0.014 | 0.014 | 0 | 0 | 0 |
| **Mass** | | | | 0.278 | 0.278 | 0.278 | 0 | 0 | 0 |
| **Hernia** | | | | 0 | 0 | 0 | 0.267 | 0.267 | 0.267 |
| **Lung Lesion** | 0.126 | 0.602 | 0.69 | | | | | | |
| **Fracture** | 0.124 | 0.124 | 0.124 | | | | 0 | 0 | 0 |
| **Lung Opacity** | 0.345 | 0.345 | 0.345 | 0.446 | 0.446 | 0.446 | 0.537 | 0.537 | 0.537 |
| **Enlarged Cardiomediastinum** | 0.425 | 0.425 | 0.425 | 0 | 0 | 0 | 0.025 | 0.025 | 0.025 |

Table 4: Accuracy performance of the three methods (baseline, "logit", and "loss") on the CheX, NIH, and PC chest radiograph datasets for various pathologies.

| | CheX | | | NIH | | | PC | | |
|---|---|---|---|---|---|---|---|---|---|
| pathologies\approach | **baseline** | **on_logit** | **on_loss** | **baseline** | **on_logit** | **on_loss** | **baseline** | **on_logit** | **on_loss** |
| **Atelectasis** | 0.593 | 0.796 | 0.992 | 0.89 | 0.895 | 0.89 | 0.905 | 0.885 | 0.907 |
| **Consolidation** | 0.81 | 0.984 | 0.789 | 0.972 | 0.971 | 0.972 | 0.952 | 0.926 | 0.955 |
| **Infiltration** | | | | 0.88 | 0.887 | 0.882 | 0.765 | 0.819 | 0.779 |
| **Pneumothorax** | 0.983 | 0.983 | 0.983 | 0.973 | 0.973 | 0.973 | 0.995 | 0.995 | 0.995 |
| **Edema** | 0.768 | 0.948 | 0.973 | 0.972 | 0.971 | 0.972 | 0.932 | 0.914 | 0.937 |
| **Emphysema** | | | | 0.98 | 0.98 | 0.98 | 0.988 | 0.988 | 0.988 |
| **Fibrosis** | | | | 0.994 | 0.994 | 0.994 | | | |
| **Effusion** | 0.678 | 0.678 | 0.678 | 0.925 | 0.925 | 0.925 | 0.858 | 0.858 | 0.858 |
| **Pneumonia** | 0.938 | 0.994 | 0.992 | 0.975 | 0.957 | 0.983 | 0.862 | 0.806 | 0.887 |
| **Pleural_Thickening** | | | | 0.982 | 0.982 | 0.982 | 0.989 | 0.989 | 0.989 |
| **Cardiomegaly** | 0.796 | 0.788 | 0.978 | 0.978 | 0.982 | 0.979 | 0.888 | 0.929 | 0.925 |
| **Nodule** | | | | 0.948 | 0.948 | 0.948 | 0.959 | 0.959 | 0.959 |
| **Mass** | | | | 0.933 | 0.933 | 0.933 | 0.985 | 0.985 | 0.985 |
| **Hernia** | | | | 1 | 1 | 1 | 0.985 | 0.985 | 0.985 |
| **Lung Lesion** | 0.874 | 0.983 | 0.991 | | | | | | |
| **Fracture** | 0.943 | 0.943 | 0.943 | | | | 0.975 | 0.975 | 0.975 |
| **Lung Opacity** | 0.564 | 0.564 | 0.564 | 0.74 | 0.74 | 0.74 | 0.72 | 0.72 | 0.72 |
| **Enlarged Cardiomediastinum** | 0.838 | 0.838 | 0.838 | 0.975 | 0.975 | 0.975 | 0.895 | 0.895 | 0.895 |