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

Mohammad S. Majdi, Jeffrey J. Rodriguez

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

Abstract

This paper introduces two innovative multi-label classification methods that utilize hierarchical taxonomy of labels to improve the diagnostic accuracy of lung diseases from chest X-rays. These diseases are often challenging to distinguish due to their similar characteristics, even for seasoned radiologists. The first method, termed as the "logit" technique, adjusts the neural network logit outputs based on the hierarchy of class relationships. The second method, termed as "loss", integrates these hierarchical relationships directly into the loss function. We apply these methods to categorize lung abnormalities in chest X-rays, using three publicly available datasets - CheXpert, PADCHEST, and NIH for evaluation. The "logit" and "loss" techniques consistently surpass the standard approach in terms of performance metrics such as accuracy, AUC, and F1 scores. Additional statistical measures, including Cohen's d, Cohen's kappa, t-statistics, p-value, and Bayes factor further validate these performance enhancements.

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

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 Bellaviti et al. (2016). However, interpreting CXRs can be challenging due to similarities between different thoracic diseases, which may result in misinterpretation 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). Progress in natural language processing (NLP) has enabled the collection of extensive annotated datasets such as ChestX-ray8 Wang et al. (2017), PADCHEST Bustos et al. (2020), and CheXpert Irvin et al. (2019), 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 Jaderberg et al. (2015). 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 Litjens et al. (2017); Eshghali et al. (2023). 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. 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 Tsoumakas and Katakis (2007).

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 Huang et al. (2017) or U-Net Ronneberger et al. (2015)). 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 2 discusses related work on multi-label classification and hierarchical loss functions; Section 3 describes the proposed techniques for integrating label hierarchy into multi-label classification techniques; 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 class Zhang and Zhou (2014). 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 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 generalizations of traditional algorithms, such as nearest-neighbor or multi-layer perceptron 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 class 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. 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 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, 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) Guo et al.

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

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.

3.1. Problem Formulation

3.1.1. 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 Furnieles (2022). 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:

$$p = \mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

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

$$x = \text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{2}$$

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

3.1.2. Glossary of Symbols

Let us define the following parameters:

- $\mathcal{C} = \{c_k\}_{k=1}^{K}$: 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.
- $\mathcal{G} = \{\mathcal{C}, \mathcal{E}\}$: Graph representing the taxonomy of thoracic diseases.
- $c_j = \Lambda(c_k) \in \mathcal{C}$: parent class of class c_k in graph $\mathcal{G}$.
- $\mathcal{J}(c_j) \subset \mathcal{C}$: 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.
- $p_k^{(i)} = \text{sigmoid}\left(q_k^{(i)}\right) = \frac{1}{1+\exp\left(-q_k^{(i)}\right)}$: 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.

- 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 $p_k^{(i)}$

- $\widehat{p}_k^{(i)} \in (0,1)$: updated predicted probability for the k-th class of instance i with a value between 0 and 1.

- $\widehat{t}_k^{(i)} = \begin{cases} 1 & \text{if } \widehat{p}_k^{(i)} \geq \theta_k \\ 0 & \text{otherwise.} \end{cases}$: 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.

- $Y^{(i)} = \left\{ y_1^{(i)}, y_2^{(i)}, \ldots, y_K^{(i)} \right\}$: 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 $Y^{(i)} = \{1, 1, 0\}$.

- $P^{(i)} = \left\{p_k^{(i)}\right\}_{k=1}^{K}$: Predicted probability set obtained in the output of the classifier $F(\cdot)$ representing the probability that each class k is present in the sample.
- $T^{(i)} = \left\{t_k^{(i)}\right\}_{k=1}^{K}$: 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 all instances and all true labels.
- $l_k^{(i)} = \mathcal{L}\left(y_k^{(i)}, p_k^{(i)}\right)$: $\mathcal{L}(\cdot)$ 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.
- Loss(θ): 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.
- $\widehat{l}_k^{(i)} = \omega_k^{(i)}\, l_k^{(i)}$: updated loss for class k and instance i.

Let us define the multi-label classification problem as follows. Let $\mathbb{X} = \left\{X^{(i)}\right\}_{i=1}^{N}$ be a set of N chest radiograph images and $\mathbb{Y} = \left\{Y^{(i)}\right\}_{i=1}^{N}$ 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

$\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, let us define a graph $\mathcal{G} = \{\mathcal{C}, \mathcal{E}\}$ representing the taxonomy of thoracic diseases, where $\mathcal{E}$ is the set of edges representing parent-child relationships between these classes. For each node $c_k \in \mathcal{C}$, let Λ_k be the parent node of class c_k and let $\mathcal{J}_k \subset \mathcal{C}$ 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 data for instance i (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 $P^{(i)} = \left\{p_k^{(i)}\right\}_{k=1}^{K}$. 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

$l_k = \mathcal{L}\left(p_k^{(i)}, y_k^{(i)}\right), \ k \in \{1, 2, \ldots, K\}$ 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.

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 (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 Irvin et al. (2019). 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, we combined the taxonomies provided by Irvin Irvin et al. (2019) for the CheXpert dataset, Chen Chen et al. (2020) for the PADCHEST Bustos et al. (2020) and the CXR portion of the prostate, lung, colorectal and ovarian (PLCO) dataset Gohagan et al. (2000). In order to maintain uniformity, we adopted the renaming scheme introduced by Cohen Cohen et al. (2022) for the pathology names. Subsequently, the key pathologies were identified and extracted to build the hierarchical taxonomy structure illustrated in Figure 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.

$$\widehat{p}_k^{(i)} = \frac{1}{1 + \exp\left(-\left(q_k^{(i)} + \alpha_{k,j} q_j^{(i)}\right)\right)} \tag{3}$$

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.

3.3.1. 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.

3.4. Approach 2: Conditional Loss

In a second approach, we propose a similar concept to the approach discussed in Section 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.

3.4.1. 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

$$l_k = \mathcal{L}\left(y_k^{(i)}, p_k^{(i)}\right) = -\left(y_k^{(i)} \log(p_k^{(i)}) + (1 - y_k^{(i)}) \log(1 - p_k^{(i)})\right) \tag{4}$$

$$\text{Loss}(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} l_k \tag{5}$$

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:

$$\widehat{l}_k^{(i)} = l_k^{(i)} + \beta_k H\left(c_k|c_j\right) \tag{6}$$

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 $l_j = \mathcal{L}\left(y_j^{(i)}, p_j^{(i)}\right)$ as follows:

$$H(k|j) = \mathcal{L}\left(y_j^{(i)}, p_j^{(i)}\right) \tag{7}$$

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.

$$H(k|j) = \mathcal{L}\left(y_j^{(i)}, p_k^{(i)}\right) \tag{8}$$

The penalization term in Equations (7) and (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 (7) and (8) to be based on the multiplication of losses rather than their addition.

$$\widehat{l}_k^{(i)} = l_k^{(i)} H(k|j) \tag{9}$$

where the hierarchical penalty term is

$$H(k|j) = \begin{cases} 1 & \text{otherwise.} \\ \alpha_k l_j^{(i)} + \beta_k & c_j \text{ is parent of } c_k \end{cases} \tag{10}$$

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 (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 (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.

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 3.4) and predicted probabilities (approach 1 shown in Section 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 (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 $\widehat{q}_k^{(i)}, \widehat{p}_k^{(i)}$ 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 $\widehat{q}_k^{(i)}$ into a sigmoid function:

$$\widehat{p}_k^{(i)} = \text{sigmoid}\left(\widehat{q}_k^{(i)}\right) = \frac{1}{1 + \exp\left(-\widehat{q}_k^{(i)}\right)} \tag{11}$$

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.

$$\begin{aligned} \frac{\partial \text{sigmoid}}{\partial \widehat{q}_k^{(i)}} &= \text{sigmoid}\left(\widehat{q}_k^{(i)}\right)\left(1 - \text{sigmoid}\left(\widehat{q}_k^{(i)}\right)\right) && (12) \\ &= \widehat{p}_k^{(i)}\left(1 - \widehat{p}_k^{(i)}\right) && (13) \end{aligned}$$

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:

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

where $y_k^{(i)}$ and $\widehat{p}_k^{(i)}$ 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 (14) and the derivative of the sigmoid function shown in Equation (12) to calculate the updated predicted probability as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(p_k^{(i)},\ y_k^{(i)}\right)}{\partial \widehat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \widehat{q}_k^{(i)}} &= \left(\frac{y_k^{(i)}}{\widehat{p}_k^{(i)}} - \frac{1-y_k^{(i)}}{1-\widehat{p}_k^{(i)}}\right) \widehat{p}_k^{(i)} \left(1-\widehat{p}_k^{(i)}\right) && (15) \\
&= y_k^{(i)} - \widehat{p}_k^{(i)} && (16)
\end{aligned}$$

Hence, we can conclude that

$$\hat{p}_k^{(i)} = \begin{cases} -\dfrac{\partial \mathcal{L}\left(p_k^{(i)},\, y_k^{(i)}\right)}{\partial \widehat{p}_k^{(i)}} \dfrac{\partial \text{sigmoid}}{\partial \widehat{q}_k^{(i)}} & y_k^{(i)} = 1 \\ -\dfrac{\partial \mathcal{L}\left(p_k^{(i)},\, y_k^{(i)}\right)}{\partial \widehat{p}_k^{(i)}} \dfrac{\partial \text{sigmoid}}{\partial \widehat{q}_k^{(i)}} & \text{otherwise.} \end{cases} \tag{17}$$

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 (14) we obtain the following:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}_k^{(i)}} &= \frac{y_k^{(i)}}{\widehat{p}_k^{(i)}} - \frac{1-y_k^{(i)}}{1-\widehat{p}_k^{(i)}} && (18) \\
&= \frac{y_k^{(i)} - \widehat{p}_k^{(i)}}{\widehat{p}_k^{(i)}\left(1-\widehat{p}_k^{(i)}\right)} && (19)
\end{aligned}$$

In this equation, we 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 zero $\left(y_k^{(i)} = 0\right)$, the loss gradient can only take the value 0 or a negative number. Thus, we can modify Equation (17) as follows:

$$\widehat{p}_k^{(i)} = \begin{cases} -\frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \widehat{q}_k^{(i)}} + 1 & \text{if} \quad \frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}_k^{(i)}} \geq 0 \\ -\frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}_k^{(i)}} \frac{\partial \text{sigmoid}}{\partial \widehat{q}_k^{(i)}} & \text{otherwise.} \end{cases} \tag{20}$$

Finally, Equation (20) can be simplified as follows:

$$\widehat{p}_k^{(i)} = \begin{cases} \exp(-\widehat{l}_k^{(i)}) & \text{if} \quad \frac{\partial \mathcal{L}(\widehat{p}_k^{(i)}, y_k^{(i)})}{\partial \widehat{p}_k^{(i)}} \geq 0 \\ 1 - \exp(-\widehat{l}_k^{(i)}) & \text{otherwise} \end{cases} \tag{21}$$

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

The following demonstrates Equation (21) based on predicted probability to demonstrate its similarity to Equation (3) in Approach 1 (Section 3.3). From Equation (10) we have $\hat{l}_k^{(i)} = l_k^{(i)} \left(\alpha_k \, l_j^{(i)} + \beta_k \right)$. By substituting that into $\exp\left(-\widehat{l}_k^{(i)}\right)$, for $y_k^{(i)} = 1$ we obtain:

$$\begin{aligned} \exp\left(-\widehat{l}_k^{(i)}\right) &= \exp\left(-l_k^{(i)} \left(\alpha_k \, l_j^{(i)} + \beta_k\right)\right) \qquad (22) \\ &= \left(p_k^{(i)}\right)^{-\alpha_k \log\left(p_j^{(i)}\right) + \beta_k} \qquad (23) \end{aligned}$$

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

$$\begin{aligned} 1 - \exp\left(-\widehat{l}_k^{(i)}\right) &= 1 - \exp\left(-l_k^{(i)} \left(\alpha_k \, l_j^{(i)} + \beta_k\right)\right) \\ &= 1 - \left(1 - p_k^{(i)}\right)^{-\alpha_k \log\left(1 - p_j^{(i)}\right) + \beta_k} \qquad (24) \end{aligned}$$

By substituting Equations (22) and (24) into Equation (21) we obtain

$$\widehat{p}_k^{(i)} = \begin{cases} \left(p_k^{(i)}\right)^{-\alpha_k \log(p_j^{(i)}) + \beta_k} & \text{if} \quad y_k^{(i)} = 1 \\ 1 - \left(1 - p_k^{(i)}\right)^{-\alpha_k \log\left(1 - p_j^{(i)}\right) + \beta_k} & \text{otherwise.} \end{cases} \tag{25}$$

3.6. Experimental Setup

3.6.1. Datasets

Three diverse and publicly available datasets are used to evaluate the proposed hierarchical multi-label classification techniques: CheXpert Irvin et al.

(2019), PADCHEST Bustos et al. (2020), and NIH Wang et al. (2017). 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** 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 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.

3.6.2. Model Optimization

The DenseNet121 Huang et al. (2017) architecture and the pre-trained weights provided by Cohen Cohen et al. (2022) was used as the baseline model. The model was fine-tuned on a subset of CheXpert Irvin et al. (2019), NIH Wang et al. (2017), PADCHEST Bustos et al. (2020) 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.

3.6.3. 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.

3.6.4. 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.

3.6.5. 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 (AU-ROC)**: 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 (H0) with an alternative hypothesis (H1). 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 H1, while a value lower than 1 denotes support for H0. In general, values between 1/3 and 3 are regarded as inconclusive, values above 3 as some evidence for H1, and values below 1/3 as evidence for H0.

- **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.

4. Results

4.1. Taxonomy Structure

In this study, we devised a detailed taxonomy, depicted in Figure 1, to classify various lung pathologies observable in chest radiographs. This classification system, inspired by the works of Irvin Irvin et al. (2019), Chen Chen et al. (2020), and Gohagan Gohagan et al. (2000), integrates prevalent disease manifestations evident in widely used datasets like CheXpert Irvin et al. (2019), PADCHEST Bustos et al. (2020), and NIH Wang et al. (2017). The taxonomy provides a structured approach to categorize these disease manifestations and offers a framework to facilitate the comprehension and analysis of abnormalities in chest radiographs.

[Figure 1 about here.]

4.2. Datasets

The prevalence of different pathology labels across three distinct medical imaging datasets: CheXpert Irvin et al. (2019), PADCHEST Bustos et al. (2020), and NIH Wang et al. (2017) are examined. Table 1 provides an overview of each pathology label's prevalence across these datasets. To utilize the TorchXRayVision software Cohen et al. (2022), the same 18 pathologies as their work, were chosen for model fine-tuning. For the purpose of assessing the proposed methodologies, particular emphasis is placed on pathologies that are recurrent across different datasets (appear in at least two of the three datasets) and are included in our formulated taxonomy. Furthermore, the cross-dataset presence of these pathologies enhances the generalizability of our study, as the developed models are validated on multiple independent datasets. These pathologies, highlighted in green in the table, comprise **Atelectasis**, **Consolidation**, **Infiltration**, **Edema**, **Pneumonia**, **Cardiomegaly**, **Lung Lesion**, **Lung Opacity**, and **Enlarged Cardiomediastinum**. 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 taxonomy structure studied.

[Table 1 about here.]

The distribution of samples per pathology in each dataset is presented in Table 2. Before applying the proposed technique, a series of preprocessing steps are performed on the ground truth label set. In the context of medical images containing multiple classes, it is a prevailing practice for the individual responsible for labeling to solely annotate the pathologies that are pertinent to their specific study requirements. Occasionally, there are situations wherein certain instances of data are classified as having specific child pathologies, but not their corresponding parent pathologies. In order to address the absence of labels for certain parent classes, which is crucial for the efficacy of the proposed techniques, we have modified the label value to signify the presence of classes with at least one child class as TRUE, indicating the existence of the class in that particular instance. A preprocessing step is applied to classes that do not have corresponding labels in the original ground truth label set. In the context of this study, the Lung Opacity and Enlarged Cardiomediastinum classes are absent from the original ground truth label sets of the NIH and PADCHEST datasets (Table 1). By revising the ground truth label set, we have identified several instances where the presence of the respective parent class can be inferred based on the presence of their respective child classes as shown in Table 2 (cells highlighted in green).

[Table 2 about here.]

4.3. Techniques Evaluation

The performance comparison of our proposed methods, namely “logit” and “loss”, with the “baseline” technique is illustrated in Figure 3. This comparative analysis centers on nine distinct medical conditions associated with pulmonary and cardiovascular diseases within three datasets. These nine pathologies encompass two parent classes (**Lung Opacity**, and **Enlarged Cardiomediastinum**) and their respective child classes, as illustrated in

Figure 1. Each subplot exhibits the receiver operating characteristic (ROC) curves for each methodology superimposed on one another, accompanied by their respective AUC (Area Under Curve) scores annotated. AUC (Area Under the Curve) scores are computed for each pathology class across all test samples in all studies datasets. We can see a notable improvement in AUC scores for all pathologies possessing parent classes. The aforementioned findings serve as compelling evidence for the effectiveness of the proposed methodologies, as they showcase their ability to improve the accuracy of classification in scenarios involving hierarchical class structures. AUC scores for two parent classes, "Lung Opacity" and "Enlarged Cardiomediastinum", remain unchanged as expected. The techniques proposed in this study are designed to exploit the hierarchical structure of classes, and therefore only bring about improvements where a class possesses a parent class.

[Table 3 about here.]

The comparative analysis presented in Figure 2 examines the performance of the proposed "loss" and "logit" methods in comparison to the "baseline" method across three important metrics: Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUC), and F1 score for different pathologies.

The "loss" and "logit" methods exhibit a distinct advantage over the "baseline" method in terms of accuracy. In the case of Atelectasis, the "loss" method demonstrates a notably higher accuracy of 0.922 compared to the "baseline" method's accuracy of 0.686. Additionally, the "logit" method achieves an accuracy of 0.874. As predicted, there is no noticeable disparity in accuracy

between the methods for the parent classes, Lung Opacity and Enlarged Cardiomediastinum, as indicated by scores of 0.663 and 0.696, respectively.
The AUC, a performance measure that takes into account both sensitivity and specificity, provides further evidence of the superior performance of the "loss" and "logit" techniques. In the case of Cardiomegaly, the area under the curve (AUC) demonstrates improvements of 21% and 11% when employing the loss and logit techniques, respectively. The AUC values for the parent classes, Lung Opacity and Enlarged Cardiomediastinum, are consistent across all three methods.

The F1 score, which is calculated as the harmonic means of precision and recall, serves to emphasize the improved performance of our proposed methods. Significantly, in the case of Lung Lesion, the F1 score exhibits a notable increase from 0.094 in the "baseline" approach to 0.982 in the "loss" approach, and 0.263 in the "logit" approach.

The obtained results provides further support for our previous findings, which indicate that the utilization of the "logit" and "loss" methods leads to substantial improvements in performance compared to the "baseline" method across most child classes. In all measured aspects and scenarios, the "loss" method exhibits slightly superior performance compared to the "logit" method.

[Figure 2 about here.]

[Figure 3 about here.]

Table 3 provides a comparative analysis of the performance of our proposed "logit" and "loss" techniques with the "baseline" method, using various statistical metrics. The "logit" technique, as indicated in the upper table, suggest a significant performance enhancement compared to the “baseline” across all evaluation tests, with kappa values ranging between 0.495 and 1. The kappa statistic is used to measure the level of agreement between two techniques, where a value of 1 signifies perfect alignment. The p-value for all child classes is below 0.05, ranging from 2.1E-89 to 2.9E-16, thereby implying a statistically significant improvement of the “logit” method over the “baseline”. High t-statistics and power values of 1 further underscore the robustness of our technique. The Bayes factor results for the “logit” technique are exceptionally strong across all conditions, suggesting substantial evidence favoring the “logit” method for these scenarios.

The proposed “loss” technique demonstrates encouraging results when benchmarked against the “baseline”, albeit with more variability. Kappa values spanned from a minimum of 0.059 for Lung Lesion to a maximum of 0.836 for Infiltration. While the p-values indicate statistically significant improvement for most conditions, Infiltration and Pneumonia had p-values exceeding 0.05 (0.053 and 0.207, respectively), hinting that the performance improvement over the “baseline” for these conditions may not be statistically significant. High t-statistics and power values of 1 were observed for all conditions except Infiltration and Pneumonia. The cohen-d values for the “loss” technique were generally larger than those for the “logit” technique, signifying a larger effect size. The Bayes factor results for the “loss” technique were exceedingly strong for conditions such as Atelectasis and Edema, but considerably lower for conditions like Infiltration and Pneumonia, indicating less evidence supporting the “loss” technique for these conditions.

Both the “logit” and “loss” techniques shows considerable improvements over the “baseline” technique, though the degree of improvement varied. The “logit” technique exhibited a more consistent level of improvement across all conditions, whereas the “loss” technique showed potential for even larger improvements in certain conditions, albeit with less consistency across the conditions studied.

5. Discussion and Conclusion

In this research, we introduce two innovative hierarchical multi-label classification techniques designed to increase both the accuracy and understandability of results in applications where a hierarchical structure exists among classes. The techniques are intended to improve classification accuracy, resistance to labeling inaccuracies and enhance alignment with hierarchical class structures. The proposed “loss” technique is devised to be integrated into any loss function (e.g., binary cross entropy loss) that is used for optimizing the model parameters, enabling a more refined adjustment of the hierarchical influence through introducing a regularization term in the loss function of the model. The proposed “logit” technique offers a straightforward yet potent method to integrate label hierarchy into a model without necessitating considerable changes to the existing structure. Our results affirm the effectiveness of the introduced hierarchical multi-label classification techniques in increasing the classification accuracy of thoracic diseases. Several performance indicators, including accuracy, AUC, and F1 scores, along with Cohen’s d, Cohen’s kappa, t-statistics, p-value, and Bayes factor, attest to the substantial performance improvements of these methods over the baseline across three major public datasets (CheXpert, PADCHEST, and

NIH). These findings suggest that the proposed techniques can be more reliable tools for improving classification accuracy as well as a higher level of interpretability in the findings. The “loss” and “logit” techniques harness the disease taxonomy to enhance classification performance, underscoring the value of using label relationships in classification tasks. These hierarchical techniques could potentially aid healthcare professionals by enhancing the comprehensibility of the model’s predictions. Providing predictions with varying detail levels based on taxonomy could enable personalized diagnoses that better meet individual clinical needs. Moreover, the techniques could be integrated into computer-aided diagnosis systems to deliver more precise and efficient diagnoses, possibly reducing clinicians’ workload and improving patient outcomes. However, further research is needed to explore their potential benefits in a clinical setting. There are also some limitations to these methods. For example, applying these techniques to other applications would necessitate the creation of a taxonomical structure for the dataset labels, which could be challenging for complex applications and usually requires consensus among several domain experts. Also, the effectiveness of the introduced techniques could be influenced by the quality and consistency of dataset labeling, which may vary across different sources. Future research should aim to evaluate these techniques across a broader array of datasets and investigate the impact of labeling quality on performance.

Acknowledgements

Appendices

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.

Eshghali, M., Kannan, D., Salmanzadeh-Meydani, N., and Esmaieeli Sikaroudi, A. M. (2023). Machine Learning Based Integrated Scheduling and Rescheduling for Elective and Emergency Patients in the Operating Theatre. *Ann Oper Res.*

Furnieles, G. (2022). Sigmoid and SoftMax Functions in 5 minutes.

Gohagan, J. K., Prorok, P. C., Hayes, R. B., and Kramer, B.-S. (2000). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening

Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials*, 21(6):251S–272S.

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.

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.

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.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Med. Image Comput. Comput.-Assist. Interv. – MICCAI 2015*, volume 9351, pages 234–241, Cham. Springer International Publishing.

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

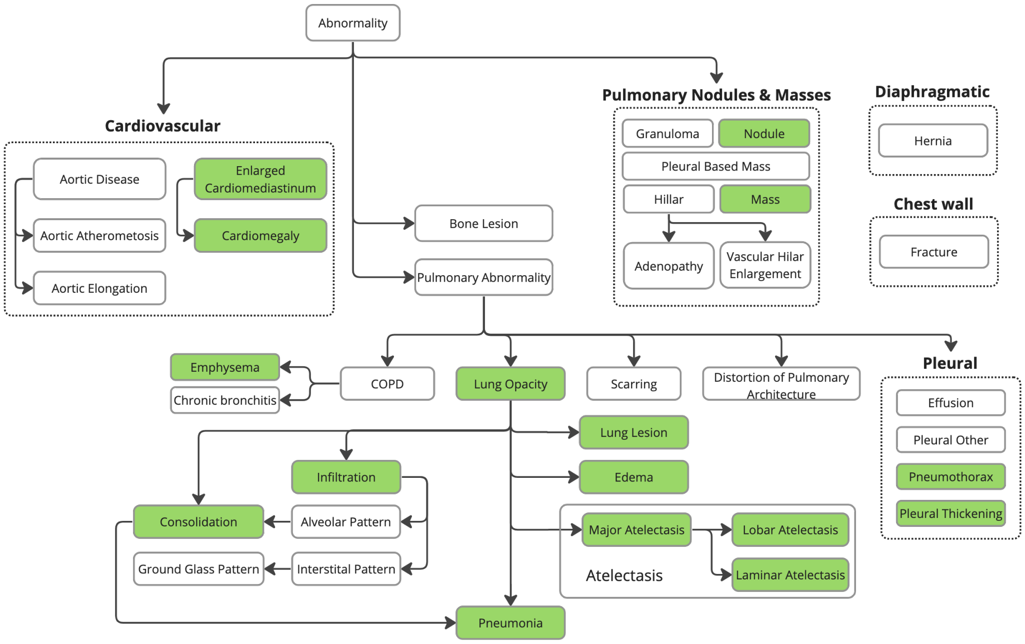

Figure 1: Taxonomy structure of lung pathologies in chest radiographs.

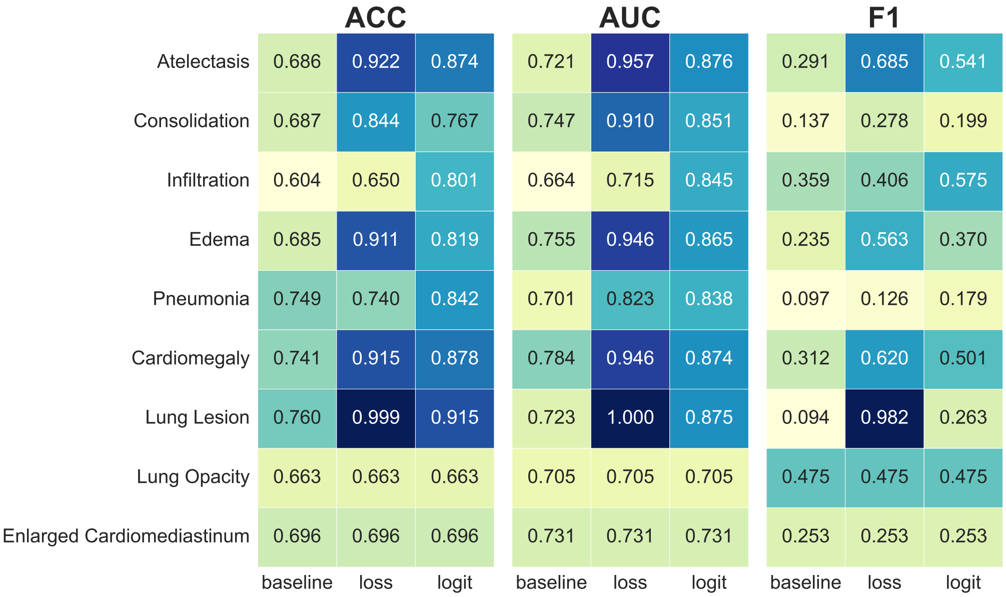

Figure 2: Heatmap visualization of model performance metrics across all three datasets. The subplots from left to right correspond to the Accuracy (ACC), Area Under the ROC Curve (AUC), and F1 Score for the baseline, “loss”, and “logit” techniques respectively. The pathologies are shared on the y-axis. 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

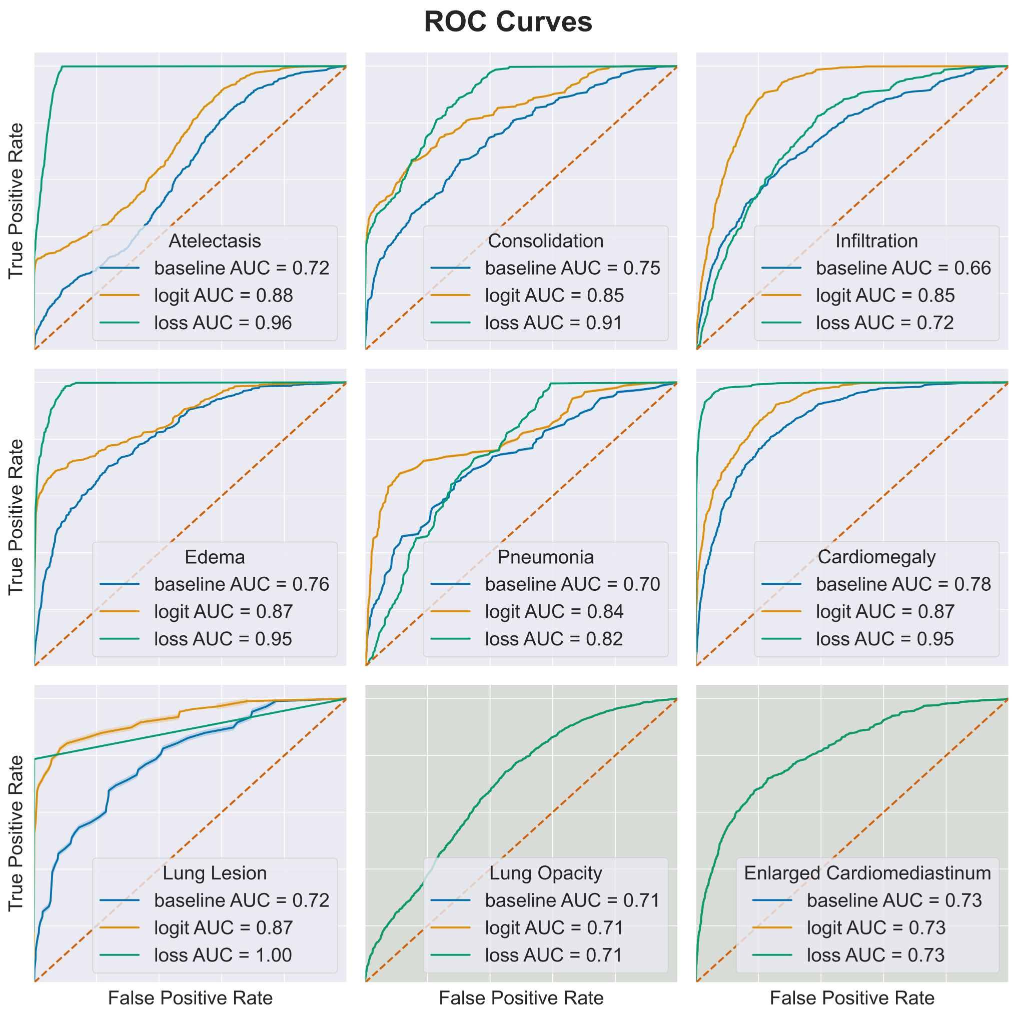

Figure 3: Comparative analysis of the ROC curves for nine thoracic pathologies using the "logit" and "loss" techniques as well as the baseline. The subplots highlighted with a darker background, represent parent class diseases.

List of Tables

Table 1: Representation of pathologies across datasets

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

Table 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 Cardiomediastinum	1100	4577	729	203	5387	261
Total	20543	53359	28868	9060	61692	2445

Table 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 Cardiomediastinum	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 Cardiomediastinum	1	1	0	0.05	0	0.019