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[1], Jeffrey J. Rodriguez[1]

[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 15, 2023

1. Introduction

Chest radiography (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 **?**. However, interpreting CXRs can be challenging due to similarities between different thoracic diseases, which may result in misinterpretation even by experienced radiologists **?**. Consequently, devising an accurate system to identify and localize common thoracic diseases can aid radiologists in minimizing diagnostic errors **??**. Progress in natural language processing (NLP) has enabled the collection of extensive annotated datasets such as ChestX-ray8 **?**, PADCHEST **?**, and CheXpert (Irvin et al., 2019b), 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 **?**.

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 **??**. 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 **????**, 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 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 **?**.

This paper aims to tackle the challenges of multi-label classification by

introducing a hierarchical framework that incorporates the relationship between different classes to provide a more accurate classification framework. Two different approaches are proposed for scenarios where ground truth are available, in which the proposed technique is employed into the baseline loss function, and scenario where the ground truth are not available in which its applied to the logit values. The latter provides a transfer learning approach that improves the classification accuracy 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 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 **?** 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 **?**, attention mechanisms **??**, and localization techniques **????**. 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 **?**. 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's or marginal (dataset-specific label dependence) where certain labels may co-occur more frequently. **?**.

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 **??**. 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 **?**. Early methods used hierarchical and multi-label generalizations of traditional algorithms, such as nearest-neighbor or multi-layer perceptrons **?** and decision trees **?**. With the rise of deep learning, the adaptation of convolutional neural networks (CNN) for hierarchical classification has gained

increasing attention ????.

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 ?. 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 ?. 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 ?, visual recognition ?, and genomic analysis ?. A common technique ? 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 ?. 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 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) ??.

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.

3. Methods

We propose a novel method that improves the accuracy and interpretability of multi-label classification with applications such as chest radiograph (CXR). Two different approaches are proposed. In the first approach which requires access to ground truth labels, the hierarchical relationships between different classes are embedded into the loss function. In a second approach, the hierarchical relationships are used to update the value of logits prior to the calculation of predicted probabilities for each class. As a transfer learning approach, these two techniques facilitate the adoption and/or 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 within applications where there is hierarchical relationship between abnormalities.

One of the key benefits of the proposed techniques 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 predictions, build trust in the model 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. Glossary of Symbols

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.
- $\mathcal{G} = \{\mathcal{C}, \mathcal{E}\}$: Directed acyclic graph (DAG) $\mathcal{G}$ representing the taxonomy of thoracic diseases.
- $c_j = \Lambda(c_k) \in \mathcal{C}$: parent class of class c_k in DAG $\mathcal{G}$.
- $\mathcal{J}(c_j) \subset \mathcal{C}$: set of child classes of class c_j in DAG $\mathcal{G}$

- $y_k^{(i)} \in \{0, 1\}$: 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 will refer to this as the "base loss function" throughout this paper.
- Loss(θ): Measured loss for all classes 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.).
- $\omega_k^{(i)}$: 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.

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 $\mathcal{E}$ is the set of directed 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 denote $\mathcal{J}_k \subset \mathcal{C}$ the set of child classes of class c_k in DAG $\mathcal{G}$.

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. 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 instance i and class k.

3.3. Label Taxonomy Structure

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 classified 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.

A comprehensive label taxonomy for lung diseases was developed based on the taxonomies presented by Irvin **?** for the CheXpert dataset and Chen **?** for the PadChest and PLCO datasets. This unified taxonomical structure is designed to be applied to various chest radiography datasets. The developed taxonomy structure is depicted in Figure **??**.

3.4. 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.

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

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-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.

3.5. Approach 2: Conditional Loss

In a second approach, we propose a similar concept to the approach discussed in section **??**, 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 based 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 represent the difference between the predictions made by the model and the actual labels provided as ground-truth. Incorporating parent class loss values into child class loss calculations aims to minimize errors throughout the hierarchy, thereby assuring accurate predictions for both parent and child classes.
- **Enhanced gradient propagation:** During the training 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 class improves the connections between the parent and child classes with respect to the propagation of gradients. This may lead to more effective acquisition of hierarchical associations and expedited convergence in the course of training.

- **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 enhances the consistency of the hierarchy by penalizing deviations from anticipated parent-child associations. This approach improves the model's resilience to possible label inaccuracies in the dataset.

- **Improved interpretability:** Employing loss values instead of predicted probabilities facilitates a more straightforward comprehension of the model's ability to capture hierarchical interrelationships among classes. The impact of high loss values on parent classes is more pronounced on the losses of their corresponding child classes, indicating the necessity to improve specific areas to better reflect these associations.

3.5.1. 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 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 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 the probability that its parent class c_j exists in that instance is low. 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.

$$\widehat{l}_k^{(i)} = l_k^{(i)} + \beta_k H\left(c_k|c_j\right) \tag{4}$$

where $c_j = \Lambda(c_k)$, and β_k is the hyperparameter that balances the contributions of the class k's own loss value and its parent classes' 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 $l_j = L\left(y_j^{(i)}, p_j^{(i)}\right)$ 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 (**??**) and (**??**) the penalization term encourages the model to correctly predict the parent classe when predicting the child class, ensuring that the predicted label set adheres to the hierarchical structure. In the aforementioned approach, 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 (**??**) and (**??**) 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 classes, 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.

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

where the hierarchical penalty term is defined as follows.

$$H(k|j) = \begin{cases} 1 & \text{otherwise.} \\ a_k l_j^{(i)} + \beta_k & c_j \text{ has a parent} \end{cases} \tag{8}$$

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

The modified loss function in Equation (**??**) aims to ensure that predictions adhere to hierarchical relationships between classes by penalizing deviations from these established relationships. By adjusting the parameters α_k and β_k, we can regulate the degree to which hierarchical information influences the learning process.

3.6. 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 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 (approach 2 shown in section **??**) and predicted probabilities (approach 1 shown in section **??**) 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 (**??**) 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 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 gradient of the sigmoid function (shown below) 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 (**??**) and the derivative of the sigmoid function shown in Equation (**??**) 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 (**??**) 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 (**??**) 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 (**??**) 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 (**??**) based on predicted probability to demonstrate its similarity to Equation (**??**) in Approach 1 (section **??**). From Equation (**??**) 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 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, $1 - \exp\left(-\widehat{l}_k^{(i)}\right)$, for $y_k^{(i)} = 0$ 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 (**??**) and (**??**) into Equation (**??**) 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. Experimental Setup

3.7.1. Datasets

Three diverse and publicly available datasets are used to evaluate the proposed hierarchical multi-label classification techniques: CheXpert **?**, PadChest **?**, and VinDr-CXR **?**. 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 ?** is a large-scale dataset containing 224,316 chest radiographs of 65,240 patients, labeled with 14 radiographic findings.
- **PadChest ?** 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 ?** 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.

3.7.2. Model Optimization

The DenseNet121 **?** architecture and the pre-trained weights provided by Cohen **?** was used as the baseline model. The model was fine-tuned on a subset of CheXpert **?**, NIH **?**, PadChest **?** 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 methods, we utilize parallelization techniques that distribute the computational load across multiple CPU cores. By leveraging the power of parallel processing, we can drastically reduce the overall computation time and accelerate the optimization procedure, making the method more applicable to large-scale and

high-dimensional datasets. Different parallelization libraries, such as joblib and Python multiprocessing, were employed to facilitate the implementation of parallelism, ensuring seamless integration with existing frameworks and offering a scalable and hardware-adaptable solution.

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

Evaluation - To assess the performance of the proposed techniques, several evaluation metrics were used to analyze the performance of the model compared to a baseline model. The metrics utilized are as follows.

- **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

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

[Figure 1 about here.]

Table **??** shows the labels present in each of the three studied datasets. Highlighted with green rows shows the pathologies selected for the final evaluation which are selected based on their presence in atleast two of the three datasets as well as in the taxonomy structure shown shown in Figure **??**.

Table **??** shows the number of instances that has a specific pathology in each of the three studied datasets (CheX **?**, PADCHEST **?**, NIH **?**). Prior to applying the proposed technique a set of preprocessing steps are applied to ground truth label set. In medical images with multiple classes, it is common for the labeler to only label the pathologies that their study requires. This sometimes result in sitautions where some instances of data are labeled for the presence of some of the child pathologies but not their corresponding parent pathologies. To compensate for this lack of labeling for some parent classes which is necessary for the effectiveness of the proporsed technques, we updated the label value indicating the presence of classes with at least one child class to 1 (indicating the class exist in that instance). This preprocessing is applied to all pathologies which are not labeled in the original ground truth label set. As can be seen in Table **??** (highlighted cells), while the Lung Opacity and Enlarged Cardiomediastinum classes were not present in the original ground truth label sets of NIH and PADCHEST datasets (As showcased in Table **??**); by updating the ground truth label set we have ended up with multiple instances where based on the presence of their child class presence we can be certain that the parent calss should have existed as well.

[Table 1 about here.]

[Table 2 about here.]

< Reviewed Until Here >

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

List of Figures

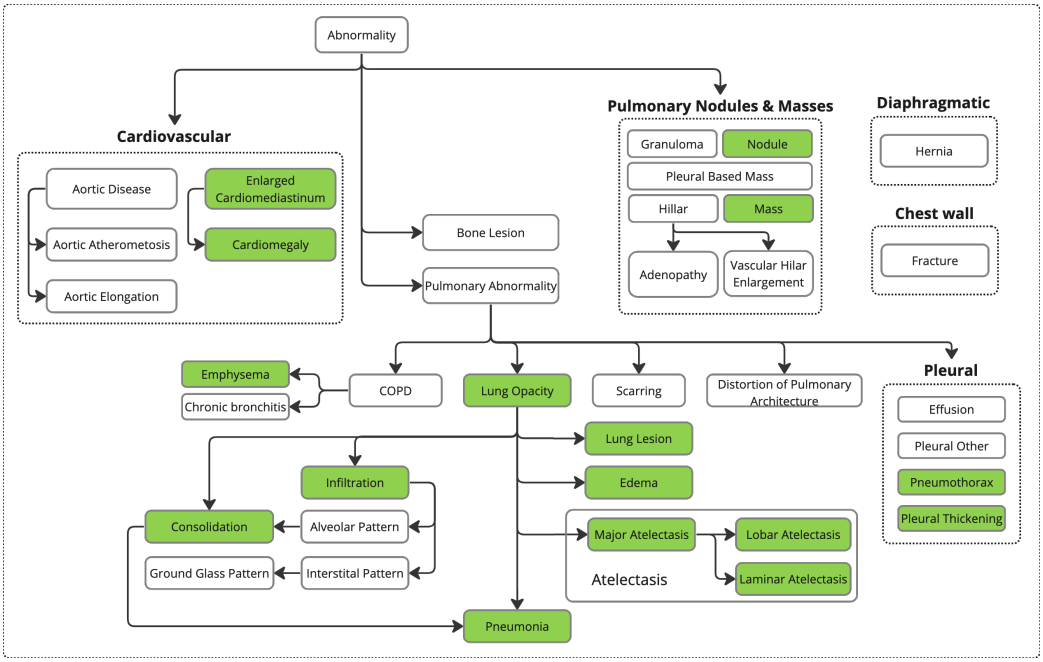
Abnormality
Cardiovascular
Aortic Disease
Enlarged Cardiomediastinum
Aortic Atherometosis
Cardiomegaly
Aortic Elongation
Bone Lesion
Pulmonary Abnormality
Pulmonary Nodules & Masses
Granuloma
Nodule
Pleural Based Mass
Hillar
Mass
Adenopathy
Vascular Hilar Enlargement
Diaphragmatic
Hernia
Chest wall
Fracture
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

List of Tables

Table 1: Pathologies present in each dataset

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	

Pathologies
Hemidiaphragm Elevatio
Hernia
Hilar Enlargement
ILD
Infiltration
Lung Lesion
Lung Opacity
Mass
Nodule/Mass
Nodule
Pleural Other
Pleural Thickening
Pneumonia
Pneumothorax
Pulmonary Fibrosis
Scoliosis
Tuberculosis
Tube

Table 2: Number of samples present in the evaluated datasets (CheX, NIH, and PC) per pathology.

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