Multi-Label Classification of Thoracic Diseases using Dense Convolutional Network on Chest Radiographs

Dipkamal Bhusal[a] and Sanjeeb Prasad Panday[a]

[a]Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, Lalitpur, Nepal

ARTICLE HISTORY
Compiled October 17, 2023

ABSTRACT
Traditional methods of identifying pathologies in X-ray images rely heavily on skilled human interpretation and are often time-consuming. The advent of deep learning techniques has enabled the development of automated disease diagnosis systems, but the performance of such systems is opaque to end-users and limited to the detection of single pathology. In this paper, we propose a multi-label disease diagnosis model that allows the detection of more than one pathology at a given test time. We use a dense convolutional neural network (DenseNet) for disease diagnosis and gradcam for model interpretability. Our proposed model achieved the highest AUC score of 0.896 for the condition Cardiomegaly with an accuracy of 0.826, while the lowest AUC score was obtained for Nodule, at 0.655 with an accuracy of 0.66. To build trust in decision-making, we generated heatmaps on X-rays to visualize the regions where the model paid attention to make certain predictions. Our proposed automated disease diagnosis model obtained highly confident high-performance metrics in multi-label disease diagnosis tasks. We believe this work will contribute towards the development of reliable and trustworthy automated diagnosis systems for disease diagnosis.

KEYWORDS
deep learning; disease diagnosis; model explainability; chest x-rays; convolutional neural network; dense network; thoracic diseases

1. Introduction

Deep learning has revolutionized image classification by achieving remarkable improvements in performance and accuracy (Krizhevsky et al. (2012); Ren et al. (2015); Simonyan and Zisserman (2015); Szegedy et al. (2017); He et al. (2016a)). The availability of large labeled datasets has enabled researchers to classify and identify images accurately. Deep learning has also shown immense potential in health analytics, particularly in automating the diagnosis process. This is especially important given that approximately 3 billion people lack access to medical imaging expertise, as reported by the World Health Organization (Litjens et al. (2017)). An automated diagnosis system can be particularly beneficial in areas where medical expertise is limited.

Given a medical image of a patient as input, a disease diagnosis system provides the probability of the occurrence of a disease. This approach represents a single-label classification problem. Examples of such diagnoses include diabetic retinopathy in

CONTACT Sanjeeb Prasad Panday Email: sanjeeb@ioe.edu.np

eye fundus images, skin cancer in skin lesion images, and pneumonia in chest X-rays (Figure 1, Figure 2, and Figure 3). However, in certain cases, multi-label diagnosis becomes crucial as it provides the probabilities of multiple pathologies occurring within the same medical image. This is particularly important when there are possibilities of more than one disease being present.

Thoracic diseases pose a significant threat to the global population, with recent events such as the COVID-19 pandemic causing respiratory illnesses worldwide (He et al. (2020)). In addition to COVID-19, chest radiography plays a crucial role in the screening and diagnosis of common thoracic diseases, including pneumonia, cardiomegaly, and pneumothorax. It is estimated that more than 2 billion chest radiography procedures are performed annually (Raoof et al. (2012)). However, the increasing workload for radiologists has led to inefficiencies in disease identification due to fatigue, potentially resulting in cognitive or perceptual errors in diagnosis. Consequently, there is a growing interest in first-world countries to develop computer-aided diagnosis systems that can assist medical professionals in evaluating X-rays. These systems have the potential to assist medical professionals and reduce diagnostic errors.

Figure 1. A sample of fundus photo (Tymchenko et al. (2020))

Figure 2. A sample of skin image (Li and Shen (2018))

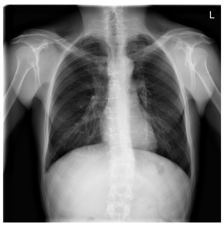


Figure 3. A sample of chest x-ray (Wang et al. (2017a))

Most existing studies on disease diagnosis using chest X-rays primarily focus on detecting a single pathology, such as pneumonia or COVID-19 (Bar et al. (2015); Cicero et al. (2017); Rajpurkar et al. (2017); Dasanayaka and Dissanayake (2021); Hussain et al. (2023)). However, it is important to note that an X-ray image can exhibit multiple pathological conditions simultaneously. Detecting multiple pathologies can provide a more comprehensive view of the patient's health from a single image, which single-label classification can overlook. Single-label classifications may produce false negatives when a patient has multiple diseases, as they focus solely on the primary condition. Multi-label classification can help reduce false negatives by identifying secondary or co-occurring diseases. Multi-label classification can also be valuable in epidemiological studies and public health research. It can provide insights into the prevalence and co-occurrence of diseases in specific populations, aiding in resource allocation and healthcare planning. In this research, we employ a 121-layer DenseNet architecture to perform diagnostic predictions for 14 distinct pathological conditions found in chest X-rays. Additionally, we utilize the GRADCAM explanation method to localize specific areas within the chest radiograph to visualize the regions to which the model paid attention in order to make disease predictions, which enhances our understanding of the model's predictions. The detection of these 14 different pathology conditions, including 'Atelectasis', 'Cardiomegaly', 'Consolidation', 'Edema', 'Emphysema', 'Effusion', 'Fibrosis', 'Hernia', 'Infiltration', 'Mass', 'Nodule', 'Pneumothorax', 'Pleural Thickening', and 'Pneumonia', presents a multi-label classification problem. The input to the DenseNet architecture is a chest X-ray image, and the output is a label that provides the probability of each pathology being present in the X-ray. The code for our approach is available on Github[1].

Our paper is structured as follows: We begin with providing the necessary background information in Section 2. This section covers a comprehensive overview of deep learning and highlights the existing works in the field of disease diagnosis and model interpretability.

Section 3 focuses on explaining our methodology, including the dataset used, the architecture of our model, and the various steps involved in developing the model. We also introduce the evaluation criteria utilized in assessing the performance of our proposed diagnostic model.

In Section 4, we present the evaluation results of our diagnostic model. We analyze and interpret the findings and present various quantitative results of disease diagnosis, showcasing the accuracy of our model in detecting different pathologies. We also display the qualitative results of model interpretation, demonstrating the effectiveness of our model in generating meaningful visualizations.

Finally, we conclude our paper with limitations in Section 5 and conclusions in Section 6, summarizing our key findings and contributions. We also discuss the implications of our work and highlight future directions for research in the field of thoracic disease diagnosis using deep learning techniques.

2. Background

2.1. Convolutional Neural Network

Convolutional Neural Networks (CNNs) Fukushima and Wake (1991) are a widely used type of neural network architecture in image classification tasks. They are efficient in extracting and learning image features through convolution and pooling layers. The pioneering CNN architecture, AlexNet Krizhevsky et al. (2012), employed multiple convolutional and fully connected layers, achieving state-of-the-art performance on the ImageNet dataset. VGG Net Simonyan and Zisserman (2015) further improved upon AlexNet by introducing deeper models with 16 or 19 weight layers, known as VGG16 and VGG19, respectively. However, increasing the depth of CNNs can lead to overfitting Goodfellow et al. (2016). To address this, Inception Net Szegedy et al. (2017) proposed the use of filters of different sizes within the same level to widen the network rather than making it deeper. Residual Networks (ResNets) He et al. (2016a) introduced skip connections to enable training of even deeper models. By doing so, ResNets achieved state-of-the-art results in various image classification benchmarks. DenseNet Huang et al. (2017) parallelized this approach by connecting each layer to all preceding and succeeding layers, addressing the vanishing gradient problem in deep neural networks.

Although these architectures differ in their topology for transmitting features across layers, they share the fundamental CNN principle, consisting of convolution, sub-sampling, dense, and softmax layers. In the convolution layer, filters are applied to the input image to extract features. The sub-sampling layer reduces the spatial size of the output from the convolution layer. The dense layer is a fully connected layer that processes the output from the sub-sampling layer. Finally, the softmax layer computes the probability distribution over the output classes.

[1] `https://github.com/dipkamal/chestxrayclassifier`

2.2. Deep learning for thoracic disease diagnosis

The impressive advancements achieved by deep learning architectures in computer vision have garnered significant interest in their application to medical imaging. Accurate diagnosis of diseases or malignancies in patient images, such as X-rays or MRIs, is crucial in medical imaging. Deep learning models have shown notable improvements in diagnosis accuracy across various medical imaging applications, including the detection of diabetic retinopathy Tymchenko et al. (2020) and skin cancer Li and Shen (2018). Given the substantial impact of thoracic diseases on public health and the widespread use of chest X-rays in medical diagnosis, numerous research projects have explored the performance of deep learning in detecting these conditions.

In 2015, Bar et al. (2015) proposed a convolutional neural network (CNN) classifier for the classification of pathologies in chest radiographs. They achieved an area under the curve (AUC) of 0.87-0.94 on a test set of 433 images, demonstrating the feasibility of detecting X-ray pathology using a pre-trained model Donahue et al. (2014). In 2017, Cicero et al. (2017) presented a similar CNN classifier that achieved an AUC of 0.964 using a medium-sized dataset of 35,000 X-rays annotated by 2443 radiologists. The authors achieved an overall sensitivity and specificity of 91% using GoogleNet Szegedy et al. (2015). These positive results indicate that deep neural network architectures can successfully detect common pathologies, even with modest-sized medical datasets. Maduskar et al. (2013) evaluated the performance of CNNs in tuberculosis detection using a small dataset of 1007 chest X-rays. They experimented with pre-trained and untrained versions of two architectures, AlexNet Krizhevsky et al. (2012) and GoogleNet Szegedy et al. (2015), and obtained the best performance with an ensemble of both architectures in the pretrained condition (AUC = 0.99). The pretrained models consistently outperformed the untrained models. Similarly, Lakhani and Sundaram (2017) compared the performance of a computer-aided tuberculosis diagnosis system (CAD4TB) with that of health professionals and found that the tuberculosis assessment of CAD4TB was comparable to that of health officers. The tests were performed on 161 subjects. In 2016, Wang et al. (2017a) proposed weakly controlled multi-label classification and localization of thoracic diseases using deep learning on a dataset of 108,948 images annotated for eight different diseases. They considered multi-label losses and aimed to detect common thoracic diseases using deep learning. While they achieved promising results, they expressed skepticism about utilizing deep neural architectures in fully automated high-precision computer-aided diagnosis systems. In 2017, Rajpurkar et al. (2017) designed a deep learning model called CheXNet, which utilized a 121-layer CNN with dense connections and batch normalization for the detection of pneumonia. The model was trained on a publicly available dataset of 100,000 chest X-ray images and outperformed the average radiologist performance. These disease diagnosis models mostly differ in employed network architecture and size of the dataset.

In Bar et al. (2018), feature selection was performed to improve disease diagnosis. In addition to a set of classical pathology features, various features were extracted from layers of a CNN to identify thoracic diseases. The authors used a pre-trained model on a non-medical dataset and fine-tuned it with feature selection techniques. Dasanayaka and Dissanayake (2021) presented segmentation techniques to detect pulmonary tuberculosis. The pipeline utilized deep learning techniques to improve the accuracy of tuberculosis diagnosis. In Patel and Kashyap (2023), the authors utilized the Littlewood-Paley Empirical Wavelet Transform (LPEWT) to decompose lung images into sub-bands and extract robust features for lung disease detection. These features were then used to train a support vector machine (SVM) network using different wavelet functions.

Deep learning has also been extensively applied in the detection and diagnosis of COVID-19. In Bhuyan et al. (2022), regional deep-learning approaches were used to detect infected areas by the coronavirus. They classified the infected region into COVID-19 or Non-COVID-19 regions using a full-resolution convolutional network (FrCN). Similarly, Farooq et al. Farooq and Hafeez (2020) designed COVID-ResNet, a deep neural network architecture using ResNet-50 He et al. (2016b) to differentiate three types of COVID-19 infections from common pneumonia disease. Yang et al. Yang et al. (2020) developed a system for diagnosing the x-ray images for COVID-19 by providing a severity score. Li et al. Li et al. (2020) developed a system to identify COVID-19 using the lung CT severity index (CTSI), which quantifies the severity of lung lesions. Zhou et al. Zhou et al. (2020) implemented the COVID examination using the non-contrast CTSI of 62 COVID-19 patients evaluated by CT scan. Several other studies (Pushparaj et al. (2022); Irene D and Beulah (2022); Kesav and MG (2022); Hussain et al. (2023); s and s (2023); Rajarajeswari et al. (2022); Noshad et al. (2021); More and Saini (2022); Dhruv et al. (2023)) have demonstrated the potential of deep learning in assisting medical professionals in diagnosing COVID-19 and monitoring disease severity.

Most disease diagnosis models focus on single-label classification, where the model only detects the presence of a single pathology. However, multi-label disease classification can offer several advantages over single-label classification in disease diagnosis. Multi-label diagnosis is akin to realistic representation since, in clinical practice, it's common for patients to have multiple medical conditions. Multi-label classification allows a single instance (e.g., an x-ray image) to be associated with multiple disease labels. This provides a more comprehensive view of the patient's health, as many patients may suffer from more than one medical condition simultaneously. Single-label classification may force a medical professional to make a decision about which disease is the "primary" one when a patient has multiple conditions. This can lead to information loss, as secondary conditions may be overlooked. A multi-label classification doesn't require this decision and captures all relevant conditions. So, multi-label models can serve as valuable decision-support tools for healthcare professionals, helping them consider multiple disease possibilities when diagnosing patients.

However, it should be noted that deep learning models should not be considered as a replacement for clinical diagnosis by medical professionals. These models should be used as complementary tools to aid medical professionals in making more accurate diagnoses. It is also crucial to validate the accuracy and reliability of these models on diverse and representative datasets. Interpretability and explainability of deep learning models in medical imaging tasks remain challenging areas of research.

2.3. Deep learning interpretability

Interpretable machine learning models have the ability to provide explanations for their decisions or predictions in a way that is understandable to humans. Unlike interpretable models like decision trees, black box models lack transparency, and their internal workings are not easily comprehensible to humans. Explanation methods aim to address this by either simplifying the model's complexity or generating post hoc explanations for individual test samples after the model has been trained Murdoch et al. (2019).

Deep learning models are examples of black box predictors that often lack interpretability. End-users typically struggle to understand the complex inner workings of deep neural networks and the reasoning behind their output predictions. However, understanding the underlying process of a model's prediction is crucial for building trust with users Ribeiro et al. (2016). When users need to make decisions based on a model's predictions, having insights into the reasons behind those predictions becomes essential. Therefore, explanations from deep learning models are vital for establishing trust and confidence in their outputs Pieters (2011).

Explanations derived from black box models can also be utilized by model designers to verify if the model is functioning as intended. Moreover, model users can employ explanation methods to feel more comfortable and confident when using black box models by gaining information about the model's predictions on specific test samples Bhusal et al. (2023). Interpreting black box models also plays a significant role in evaluating factors such as fairness, privacy, reliability, causality, and trust Doshi-Velez and Kim (2017). Consequently, the development of explainable AI models is essential for ensuring transparency and accountability in decision-making processes.

Post-hoc explanation is a well-studied approach that focuses on explaining individual predictions made by a model, rather than providing a comprehensive understanding of the entire decision-making process Bodria et al. (2021). In the context of a black box model $F(x)$, where $x = x_1, x_2, ..., x_N$ represents the input and $F(x) = y$ is the model's prediction, a post hoc explanation strategy denoted as $\gamma(x)$ is employed to generate an explanation vector $E_k(x)$ that indicates the relevance or importance of the k given features.

In the case of image data, pixel attribution methods are commonly used for post-hoc explanations. These methods aim to highlight the pixels in an image that significantly contribute to a specific classification made by the model. By visualizing these highlighted pixels, it becomes possible to understand which regions of the image played a crucial role in the model's prediction. Figure 4 provides an overview of post-hoc explanations in image classifiers. We provide a brief overview of some major post-hoc explanation methods underneath:

(1) **Perturbation-based methods:** Perturbation-based methods involve perturbing or occluding specific regions of the input and observing the impact on the model's output to assess the importance of different regions. Two popular perturbation-based post-hoc explanation methods are Local Interpretable Model-Agnostic Explanations (LIME) Ribeiro et al. (2016) and Shapley Additive Explanation (SHAP) Lundberg and Lee (2017). LIME creates an interpretable surrogate model by perturbing a test instance and generating new data samples, using linear regression to explain the predictions. The weights of the surrogate model indicate the importance of features. SHAP uses a game-theoretic approach to compute Shapley values for each feature, generating new samples around a given instance and fitting an interpretable linear model. However, unlike LIME, SHAP weights the new instances according to the weight a coalition would receive in the Shapley value estimation rather than their closeness to the original sample. SHAP provides reliable and consistent explanations with a theoretical foundation, while LIME is faster and easier to implement. Both methods have been successful in providing insights into black box models, allowing users to understand specific predictions. However, they are primarily suited for tabular data and not commonly used for unstructured data like images Slack et al. (2020).

(2) **Gradient-based methods:** Gradient-based methods calculate the gradients of

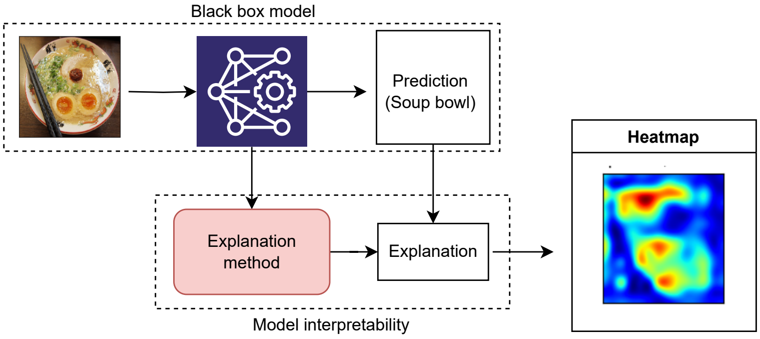


Figure 4. Pixel attributions or saliency maps for an image-classifier test case using Grad-CAM Molnar (2022)

the model's output with respect to the input features to measure their importance. The Gradient method Simonyan et al. (2013) computes the gradient of the model output with respect to the input features, providing pixel attributions. Integrated Gradient (IG) Sundararajan et al. (2017) accumulates gradients along a linear path from a baseline to the test sample, addressing some limitations of the Gradient method. SmoothGrad Smilkov et al. (2017) and NoiseGrad Bykov et al. (2022) are improvements over gradient-based methods. SmoothGrad adds Gaussian noise to generate multiple samples, averaging their pixel attributions. NoiseGrad introduces noise to model parameters, generating multiple models and averaging feature attributions. GradientSHAP Erion et al. (2021) combines SHAP and SmoothGrad with integrated gradients, selecting a baseline randomly and averaging the resulting attributions. Grad-CAM Selvaraju et al. (2017) and Grad-CAM++ Chattopadhay et al. (2018) are class activation map methods that compute feature-importance maps of convolutional layers with respect to specific classes. These methods backpropagate gradients from the model prediction to the convolutional layers, obtaining coarse localization maps of feature attribution. Gradient-based methods are popular due to their simplicity, speed, and effectiveness in providing feature attributions, particularly in image classification tasks.

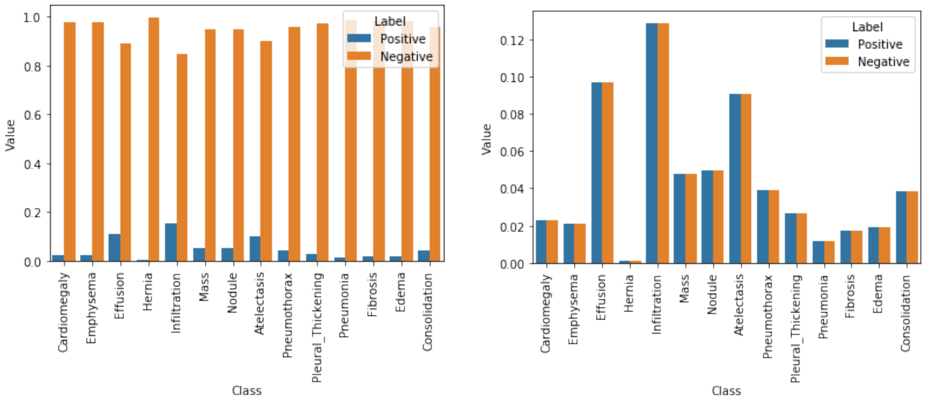
3. Methodology

3.1. Data preparation

3.1.1. Addressing Data Leakage

To prevent data leakage in medical image analysis, we took precautions to ensure that images from the same patient were not present in both the training and test sets. Data leakage can occur when the model sees images of the same patient during both training and testing, leading to biased results Rathore et al. (2017).

We utilized the ChestX-ray8 dataset Wang et al. (2017a) as our primary dataset and randomly selected 99,000 images. However, we implemented a patient-level division of the dataset for obtaining the train and test set to avoid data leakage. This means that all images belonging to a particular patient were assigned exclusively to either the training or test set, but not both. By separating the images at the patient level, we



(a) Highly skewed data distribution before adjusting class imbalance

(b) After adjusting class imbalance problem in the dataset

Figure 5. Solving class-imbalance problem

minimized the risk of data leakage and avoided introducing any biases into our deep learning model.

This approach helps maintain the integrity of the evaluation process, ensuring that the model generalizes well to unseen patients and accurately reflects its performance in real-world scenarios.

3.1.2. Addressing Class Imbalance

The class imbalance problem in the dataset refers to the significant difference in the number of positive cases (images with diseases) compared to negative cases (images without diseases). This imbalance can pose challenges during model training, as the algorithm may prioritize the majority class and overlook the minority class, leading to biased predictions.

Figure 5 illustrates the class imbalance problem in our dataset, highlighting the low proportion of positive cases compared to negative cases for certain pathological conditions. For example, the Hernia class has a ratio of positive to negative cases of approximately 0.02, indicating a highly imbalanced distribution. Similarly, the Infiltration class, which has the highest number of positive labels, exhibits a ratio of around 0.18.

Various techniques can be employed to mitigate class imbalance. One common approach is to assign different weights to the positive and negative classes during the loss calculation. By giving more importance to the minority class (positive cases) in the loss calculation, the model is encouraged to focus on correctly predicting these cases. We employ this technique in our study and modify the cross-entropy loss function. Other methods for addressing class imbalance include oversampling the minority class, undersampling the majority class, or using a combination of both. These techniques create a balanced training set by either replicating instances of the minority class or reducing instances of the majority class.

The normal cross-entropy loss, which is commonly used in classification models, for the i^{th} example is given by:

$$L_{cross-entropy}(x_i) = -(y_i \log(g(x_i)) + (1 - y_i) \log(1 - g(x_i))) \tag{1}$$

Here, x_i and y_i denote the features and label of the given image, respectively, and $g(x_i)$ represents the model prediction. Either y_i or $(1 - y_i)$ will contribute to the loss at any given time since, when y_i equals one, $(1-y_i)$ is zero and vice versa. This means that in an imbalanced dataset, one label will dominate the loss. For an entire training set of size N, the cross-entropy loss is given by:

$$L_{cross-entropy}(D) = -(1/N)(\sum_{\text{positive examples}} \log(g(x_i)) + \sum_{\text{negative examples}} \log(1 - g(x_i))) \tag{2}$$

Here, the first summation term represents the loss for all positive examples, while the second summation term represents the loss for all negative examples. This loss function leads to a bias towards the majority class in an imbalanced dataset.

To address the problem of highly skewed data distribution, it is crucial to ensure that each class's labels make an equal contribution. This can be achieved by multiplying each example from each class by a class-specific weight factor, denoted as w_{pos} and w_{neg}. To obtain equal contribution from both positive and negative classes, we aim to satisfy the following condition:

$$w_{pos} \times freq_p = w_{neg} \times freq_n \tag{3}$$

For this condition to hold, we compute weight factors that are determined based on the frequency of positive and negative examples in the dataset:

$$w_{pos} = freq_{neg} \tag{4}$$

and

$$w_{neg} = freq_{pos} \tag{5}$$

Here, $freq_p$ and $freq_n$ represent the frequency of positive and negative examples, respectively, defined as:

$$freq_p = (\text{number of positive examples})/N \tag{6}$$

and

$$freq_n = (\text{number of negative examples})/N \tag{7}$$

By adjusting the class imbalance, we achieve a balanced contribution of positive and negative labels to the loss function as shown in Figure 5. Therefore, the final weighted loss after computing the positive and negative weights is given by:

$$Lcross - entropy^{w}(x) = -(w_{pos} y \log(f(x)) + w_{neg}(1-y)\ \log(1-f(x))) \quad (8)$$

Here, y represents the true label, $f(x)$ represents the predicted label, and w_{pos} and w_{neg} represent the class-specific weight factors for positive and negative examples, respectively.

3.1.3. Pre-processing

To prepare the images for training the deep convolutional network, several preprocessing steps were performed on the training dataset. These steps aimed to standardize the data distribution and make it compatible with the chosen architecture and pre-trained model Abdou (2022).

The first step in preprocessing involved normalization of the mean and standard deviation of the input data. This normalization process ensures that the pixel values across the images have a standardized distribution, which can help improve the training process and model convergence.

Next, the x-ray images in the dataset have different sizes and hence, were resized to a uniform dimension of 320 by 320 pixels. This size is a suitable dimension for a deep convolutional network architecture used in the study. Resizing the images to a specific dimension ensures consistency in the input size for the model.

To facilitate transfer learning, we utilized a pre-trained model from ImageNet. However, the pre-trained model was trained on RGB images, while the Chest X-ray images in the dataset are single-channel grayscale images. To overcome this, we converted the 1-channel X-ray images to a 3-channel format. This conversion involved duplicating the grayscale image to create three identical channels, mimicking the RGB format required by the pre-trained model. This step enables the utilization of the pre-trained model's learned features and weights.

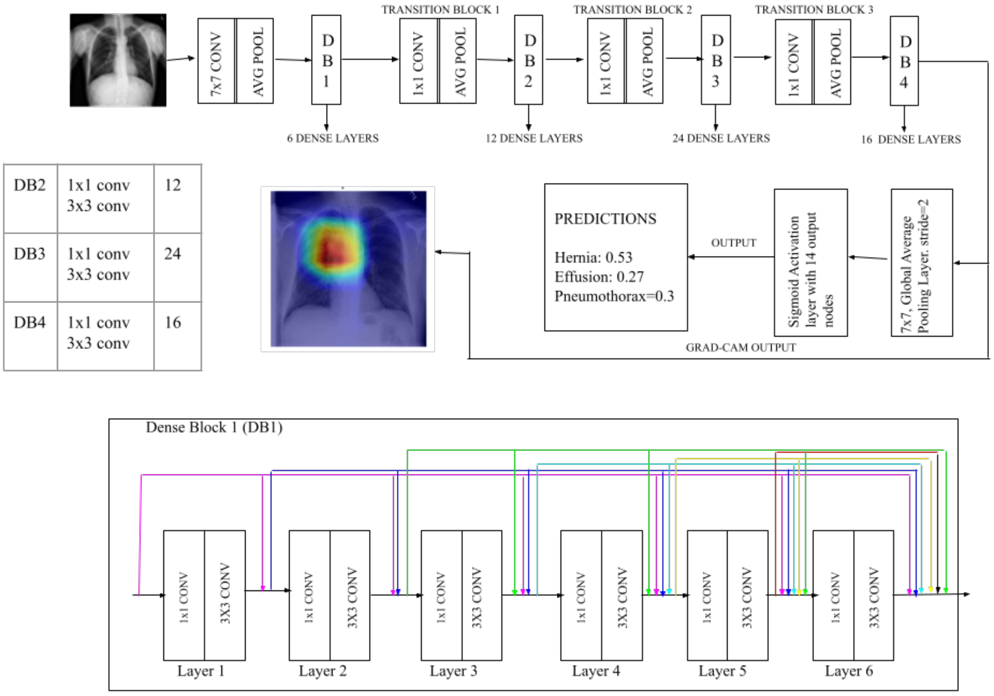
In addition to preprocessing the training data, the test data also underwent normalization. The normalization process involved using the statistics (mean and standard deviation) calculated from the training set. By normalizing the test data using the training set's statistics, the overall distribution of data during training and testing remained consistent. This ensures that the model's performance on the test set reflects its ability to generalize to new, unseen data.

By performing these preprocessing steps, the images were appropriately prepared for training the deep convolutional network, enabling effective transfer learning and maintaining consistency between the training and test datasets.

3.2. Network Architecture

Neural networks are composed of multiple layers that process input data to produce output predictions. Each layer takes a previous state vector, represented as $h(n)$, and applies a function F to generate a new state vector, $h(n+1) = F(h(n))$. The function F can involve operations like activation functions, summation or convolutional blocks, or LSTM cells. However, simply adding more layers to a network does not always lead to improved function approximation or accuracy. In fact, it can result in a decrease in accuracy for both the training and test sets due to the vanishing gradient problem.

The vanishing gradient problem occurs when the gradients of the loss function decrease rapidly as they propagate backward through multiple layers via the chain



DB2	1x1 conv 3x3 conv	12
DB3	1x1 conv 3x3 conv	24
DB4	1x1 conv 3x3 conv	16

Figure 6. Network architecture for the proposed diagnostic model using DenseNet.

rule. Eventually, these gradients become very small or vanish, preventing the weights from updating and impeding learning in the network. Consequently, the accuracy of the network may stagnate at a saturation point or degrade over time, a phenomenon known as the degradation problem.

The DenseNet architecture Huang et al. (2017) addresses the vanishing gradient problem by incorporating dense connections between layers. In DenseNet, each layer is directly connected to all the preceding layers in the network. This connectivity pattern enables each layer to have access to the feature maps produced by all the earlier layers. By facilitating the efficient flow of gradient signals and information throughout the network, DenseNet promotes more effective learning, leading to improved accuracy. This connectivity scheme mitigates the vanishing gradient problem and promotes better information flow within the network, enabling deeper networks to be trained successfully. As a result, DenseNet has been shown to achieve state-of-the-art performance in various tasks, such as image classification and object detection.

In our project, we employed the DenseNet architecture, which comprises two main blocks: the DenseBlock and the TransitionBlock. The DenseBlock keeps the feature size dimension constant while varying the number of filters. Within a DenseBlock, each layer performs a 1x1 convolution for feature extraction and a 3x3 convolution to reduce the feature depth. On the other hand, the TransitionBlock is positioned between DenseBlocks and serves to downsample the feature size by applying a 1x1 convolution and a 2x2 average pooling function. This downsampling operation reduces spatial resolution, which helps control computational cost and prevent overfitting. The output of the TransitionBlock feeds into the next DenseBlock, repeating the process. This unique design of DenseNet enables efficient feature reuse and propagation, leading to improved accuracy and reduced overfitting. For our specific implementation, we used the DenseNet-121 architecture, which consists of four DenseBlocks. Dense Block 1 has six dense layers, Dense Block 2 has twelve dense layers, Dense Block 3 has twenty-four dense layers, and Dense Block 4 has sixteen dense layers. There are three TransitionBlocks between the DenseBlocks. The name ”DenseNet-121” indicates the total number of layers with trainable weights, which in this case is 121.

3.3. Training

To initialize the network, we employed transfer learning by utilizing pre-trained weights from the ImageNet dataset. The early layers of the DenseNet, which capture general image features like edges, were left unchanged. We skipped the top layers, which contain more specific image features, and added two additional layers: a Global Average Pooling layer and a Dense layer with sigmoid activation. The Global Average Pooling layer computes the average of the last convolutional layer's output, while the Dense layer with sigmoid activation provides predictions for all target classes. It's important to note that our project deals with a multi-label classification problem, where each X-ray image can predict the probabilities for multiple pathologies independently. To accommodate this, we applied a sigmoid function to each raw output independently. The loss function used in this project is the weighted loss function described by Equation 8.

Once the neural network architecture was defined, we trained the model using backpropagation with mini-batch stochastic gradient descent. We used mini-batches of 8 images and the Adam optimizer with a default learning rate of 0.001. Backpropagation is a supervised learning algorithm that allows the model to adjust its parameters to minimize the difference between the predicted output and the actual output. Stochastic gradient descent is an optimization algorithm that updates the model parameters using a random subset of the training data rather than the entire dataset. Mini-batch stochastic gradient descent is a variant of stochastic gradient descent where the data is divided into small batches, and the model parameters are updated after each batch is processed. The Adam optimizer Kingma and Ba (2014) is an adaptive learning rate optimization algorithm that computes adaptive learning rates for each parameter based on the first and second moments of the gradients. This results in faster convergence and better performance compared to other optimization algorithms.

4. Experiment and Results

4.1. Data

We utilized the ChestX-ray8 dataset as our primary dataset. The dataset consists of Chest X-ray images and is commonly used for research in the thoracic disease field. From this dataset, we randomly selected 99,000 images as our training set. Each image in the dataset was annotated with labels that identify 14 distinct pathological conditions. These labels provide information about the presence or absence of specific conditions in the X-ray images. Table 1 in the study provides a snapshot of the dataset, showing a summary of the dataset's characteristics, such as the pathological condition and patient ID. This table serves as a reference to understand the dataset's composition and the distribution of labels across the different pathological conditions.

We addressed the data leakage as explained in Section 3.1.1 and created a train and

Table 1. Detail annotation in the dataset[a]

Image	Atelectasis	Consolidation	Edema	Effusion	PatientId
015.png	0	0	0	0	8270
001.png	1	0	0	0	29855
000.png	0	0	0	0	1297
002.png	0	0	0	0	12359
001.png	0	0	0	0	17951

[a] All 14 pathological conditions are not displayed due to width constraint.

test set. We use the training dataset to train our supervised model. We employed deep learning architecture as explained in Section 3.2 to develop a model that can predict the presence or absence of the 14 pathological conditions based on the input Chest X-ray images. The training process involved feeding the model with the labeled images and adjusting its parameters to learn the patterns and features associated with each condition.

To evaluate the model's performance and assess its generalization capabilities, we created a separate test set. This test set comprised 500 images randomly selected from the test dataset. We applied the trained model to the test set and measured various evaluation criteria as explained in Section 4.2.

4.2. Evaluation Metrics

We computed several metrics to assess the generalization of our diagnostic models. These metrics include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), the receiver operating characteristic (ROC) curve, and the F1 score.

Sensitivity, also known as the **true positive rate** or **recall**, measures the probability that the model correctly predicts the presence of a disease given that the patient actually has the disease. It is computed as TP/(TP+FN), where TP represents the number of correctly predicted positive samples and FN represents the number of incorrectly predicted negative samples.

Specificity, also known as the **true negative ratio**, measures the probability that the model correctly predicts a patient as disease-free given that the patient is actually normal. It is computed as TN/(TN+FP), where TN represents the number of correctly predicted negative samples and FP represents the number of incorrectly predicted positive samples.

Diagnostically, sensitivity and specificity are not helpful alone. While sensitivity tells the probability that the test results positive given that the person already has the condition, the information of probability that the person has the disease given that the test gives positive is important. **Positive predictive value (PPV)**, also called **precision**, provides the probability that a patient has the disease, given that the model predicts a positive result. It is computed as TP/(TP+FP).

Negative predictive value (NPV) provides the probability that a patient does not have the disease when the model predicts a negative result. It is computed as TN/(TN+FN).

The **ROC curve** is a graphical representation of the true positive rate (sensitivity) versus the false positive rate (1-specificity) at different threshold settings for classification. The area under the ROC curve (AUC) is a threshold-independent measure of the model's goodness of fit and its ability to distinguish between positive and negative samples. In medical literature, this number also gives the probability that a randomly

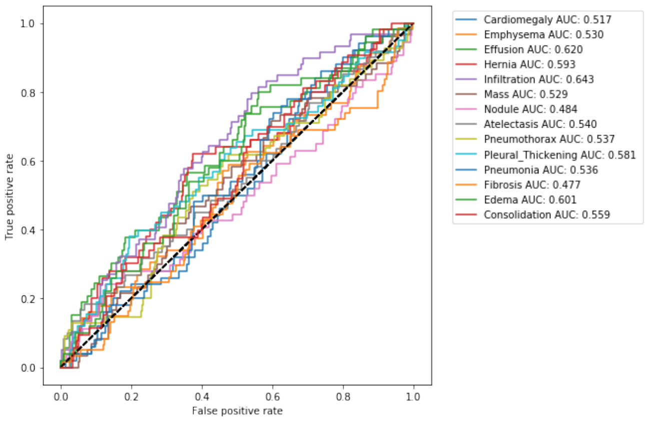


Figure 7. ROC curve on the classification of thoracic pathologies by a model trained on a small dataset of 1000 images.

selected patient who experienced a condition had a higher risk score than a patient who had not experienced the event. A higher AUC indicates better performance of the model in distinguishing between positive and negative samples.

The **F1 score** is computed by taking the harmonic mean of precision and recall. It is a single metric that balances both precision and recall. The best possible value of the F1 score is 1 (perfect precision and recall), while the worst value is 0. If a single F1 score is required for multiclass classification, a micro-average (weighted by class frequency) or macro-average (same weights for all classes) approach can be used.

By analyzing these metrics, we gain insights into the model's performance in terms of sensitivity, specificity, predictive values, discrimination power (ROC curve), and overall classification accuracy (F1 score). These evaluations help us understand the strengths and limitations of the model in accurately predicting different pathologies.

4.3. Evaluation of diagnostic model

We trained the model on different sizes of the training dataset with different training epochs to observe the generalization accuracy of the model against an unseen test set. Specifically, we wanted to analyze the following from our experiments:

(1) *Effect of Dataset Size*: Data quantity can have a significant impact on a data-hungry deep learning model. By training the model with different dataset sizes, we can understand the relationship between data quantity and model performance. Typically, larger datasets lead to better generalization. However, datasets can over or underfit the model. Smaller datasets with smaller training epochs can underfit, whereas smaller datasets with higher training epochs are more prone to overfitting.

(2) *Impact of Training Epochs*: The number of training epochs affects the convergence of a model by affecting how well the model converges to an optimal solution. Too few epochs may result in underfitting, while too many epochs may lead to overfitting.

The network was initially trained on a random dataset of 1000 images from the training set for five epochs. The model's performance was then assessed by plotting the receiver operating characteristic (ROC) curve, as shown in Figure 7. The ROC curve illustrates the model's ability to distinguish between true positive and false positive rates for different classification thresholds.

From the results obtained, it can be observed that the model performed poorly, as the ROC curve for each pathology closely approximated the diagonal line. This suggests that the model acted as a poor classifier, exhibiting limited discriminative ability in predicting diseases accurately. Furthermore, the corresponding area under the ROC curve (AUC) scores further support this observation, indicating inaccurate disease prediction across various thresholds.

Based on these findings, we hypothesize that the underperformance of the model can be attributed to underfitting and the limited size of the training dataset, as well as the relatively small number of training epochs. With a small dataset, the model may not have had enough examples to learn the complex patterns and variations present in the medical images, leading to inadequate generalization to unseen data.

To address this issue and improve the model's performance, it is anticipated that training on a larger dataset and increasing the number of training epochs will yield better results. This would allow the model to capture a more comprehensive range of patterns and variations, leading to improved accuracy in disease prediction.

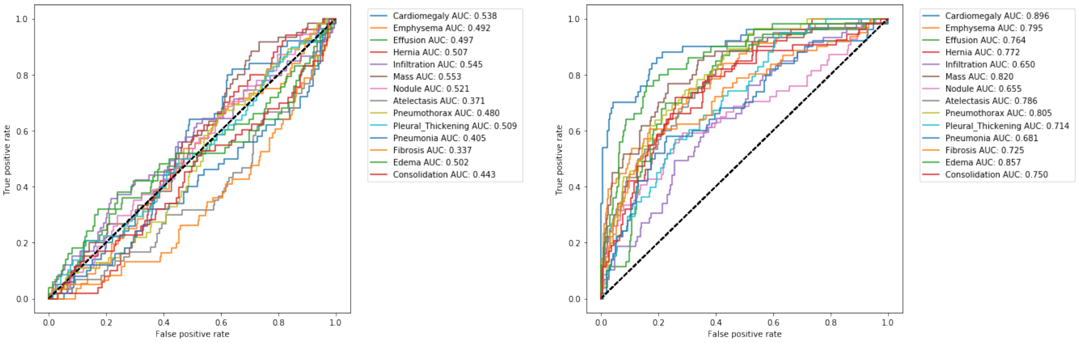
Table 2 shows the evaluation metrics for the model trained on 1000 X-ray images and tested on a test dataset. The results indicate poor generalization of the model, as reflected in its performance across all metrics. Although accuracy can be deceptive, particularly for conditions such as Pneumothorax, Hernia, and Pleural Thickening, the F1 scores provide a more reliable estimate of the model's ability to generalize, which is deemed unacceptable. Additionally, some conditions lack precision and recall values due to the absence of true positives and false positives. This highlights the limitations of the model and the need for further improvement.

Table 2. Table of evaluation metrics after training on 1000 images

	Accuracy	**Prevalence**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**F1**	**Threshold**
Cardiomegaly	0.445	0.119	0.54	0.432	0.114	0.874	0.517	0.188	0.5
Emphysema	0.864	0.133	0	0.997	0	0.866	0.53	0	0.5
Effusion	0.524	0.126	0.66	0.504	0.161	0.911	0.62	0.259	0.5
Hernia	0.881	0.119	0	1	NaN	0.881	0.593	0	0.5
Infiltration	0.54	0.14	0.712	0.512	0.193	0.916	0.643	0.303	0.5
Mass	0.167	0.143	0.9	0.044	0.136	0.727	0.529	0.236	0.5
Nodule	0.595	0.129	0.352	0.631	0.123	0.868	0.484	0.183	0.5
Atelectasis	0.41	0.143	0.65	0.369	0.147	0.864	0.54	0.239	0.5
Pneumothorax	0.869	0.131	0	1	NaN	0.869	0.537	0	0.5
Pleural_Thickening	0.862	0.138	0	1	NaN	0.862	0.581	0	0.5
Pneumonia	0.514	0.119	0.54	0.511	0.13	0.892	0.536	0.209	0.5
Fibrosis	0.855	0.145	0	1	NaN	0.855	0.477	0	0.5
Edema	0.624	0.119	0.54	0.635	0.167	0.911	0.601	0.255	0.5
Consolidation	0.65	0.126	0.377	0.689	0.149	0.885	0.559	0.214	0.5

To address the limitations, we attempted two strategies. Firstly, we increased the size of the training dataset to 99,000 images while keeping the number of epochs constant. Unfortunately, this did not lead to an improvement in the generalization performance of the model. The ROC plot in Figure 8(a) shows the ROC curve below the diagonal, indicating that the model was still unable to capture the underlying patterns in the data. Our second strategy included increasing the number of training epochs to 100 and introducing regularization techniques for adaptive learning rate and dropout Srivastava et al. (2014) of 10%. These techniques were implemented to prevent overfitting on the training data and improve the model's ability to capture more complex patterns.

The ROC curve shown in Figure 8(b) illustrates the improved performance of the classifier for different pathology conditions across various thresholds. The correspond-



(a) Under-fit model.

(b) Final model.

Figure 8. ROC curve on classification of thoracic pathologies by a model trained on complete dataset.

ing evaluation metric table in Table 3 provides the final measures of evaluation criteria for each condition. The results demonstrate a significant improvement in the discriminative performance of the model compared to our previous experiments. Specifically, the substantial improvement in model performance when moving from 1,000 images to 99,000 images with large training epochs indicates that data quantity plays a critical role in training deep CNN models. The model also did not overfit the data during the extended training period of 100 epochs.

In particular, the model achieved high sensitivity and accuracy on all conditions, indicating its ability to correctly identify positive cases. However, it is important to note that the positive predictive value (PPV) of the predictions can still be low. For example, for the Pneumonia condition, the sensitivity is 0.6, but given that the model predicted a positive result, the probability that the person actually has Pneumonia is only 0.18. These results highlight the importance of considering both sensitivity and PPV in evaluating the performance of the model.

Table 3. Table of evaluation metrics after training on 99000 images with regularizer

	Accuracy	**Prevalence**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**F1**	**Threshold**
Cardiomegaly	0.826	0.119	0.78	0.832	0.386	0.966	0.896	0.517	0.5
Emphysema	0.762	0.133	0.589	0.788	0.3	0.926	0.795	0.398	0.5
Effusion	0.681	0.126	0.736	0.673	0.245	0.946	0.764	0.368	0.5
Hernia	0.705	0.119	0.66	0.711	0.236	0.939	0.772	0.347	0.5
Infiltration	0.598	0.14	0.644	0.59	0.204	0.91	0.65	0.31	0.5
Mass	0.764	0.143	0.75	0.767	0.349	0.948	0.82	0.476	0.5
Nodule	0.66	0.129	0.593	0.669	0.209	0.918	0.655	0.309	0.5
Atelectasis	0.674	0.143	0.75	0.661	0.269	0.941	0.786	0.396	0.5
Pneumothorax	0.7	0.131	0.745	0.693	0.268	0.948	0.805	0.394	0.5
Pleural_Thickening	0.6	0.138	0.672	0.588	0.207	0.918	0.714	0.317	0.5
Pneumonia	0.633	0.119	0.6	0.638	0.183	0.922	0.681	0.28	0.5
Fibrosis	0.65	0.145	0.623	0.655	0.235	0.911	0.725	0.341	0.5
Edema	0.783	0.119	0.8	0.781	0.331	0.967	0.857	0.468	0.5
Consolidation	0.607	0.126	0.792	0.58	0.214	0.951	0.75	0.337	0.5

In Table 4 and Table 5, we compare the performance of our proposed model against single and multi-label diagnosis models for selected pathologies. Table 4 shows that our proposed multi-label approach was able to outperform single-label models. In Table 5, the results indicate that our proposed architecture outperforms Wang et al. Wang et al. (2017b) and Irvin et al. Irvin et al. (2019) in multiple detections whereas betters performance of CheXNext Rajpurkar et al. (2018), which is the state-of-the-art chest x-ray diagnosis model, for cardiomegaly condition only.

Table 4. Comparing performance of the proposed architecture against single-label diagnosis models. (-) indicates the unavailability of the model prediction on that pathology.

Pathology	Crosby Crosby et al. (2020)	Taylor Taylor et al. (2018)	Cha Cha et al. (2019)	VGG16-Islam Islam et al. (2017)	Ours
Cardiomegaly	-	-	-	0.87	0.896
Nodule	-	-	0.732	-	0.655
Pneumothorax	0.801	0.75	-	-	0.805

Table 5. Comparing performance of the proposed architecture against multi-label diagnosis models. (-) indicates the unavailability of the model prediction on that pathology.

Pathology	Wang et al. Wang et al. (2017b)	Irvin et al. Irvin et al. (2019)	CheXNext Rajpurkar et al. (2018)	Ours
Cardiomegaly	0.807	0.854	0.831	0.896
Edema	-	0.928	0.924	0.857
Mass	0.706	-	0.909	0.82
Consolidation	0.708	0.937	0.893	0.75
Pneumothorax	0.806	-	0.944	0.805

4.4. Model interpretation

Interpreting deep learning models is a challenging task due to their complex architecture. Class Activation Maps (CAMs) Kwaśniewska et al. (2017) have emerged as a popular method for generating visual explanations of model predictions, particularly for convolutional neural networks (CNNs). CAMs provide insights into where the model is focusing its attention when making a classification, aiding in model interpretation.

In our study, we utilized GRADCAM, a technique based on CAMs, to generate heatmaps for highlighting important regions in X-ray images during prediction. While GRADCAM does not provide detailed explanations for the model's reasoning, it can be valuable for expert validation, allowing experts to assess whether the model is attending to relevant regions in the image for making associated predictions.

Figure 9 presents the visualization of pathology prediction using GRADCAM. We generated heatmaps for two classes with the highest-performing AUC measures and one class with the lowest AUC measure. The activation map generates a heatmap that highlights the region of interest identified by the model when making predictions for Mass and Pneumothorax. The heatmap indicates the areas where the model is focusing its attention, providing insights into the features it considers important for the classification task. On the other hand, since the prediction for Cardiomegaly was very low, no heatmap is produced on the X-ray image, suggesting that the model did not identify any specific region as crucial for making a correct decision.

These visualizations using GRADCAM can be valuable in helping experts validate

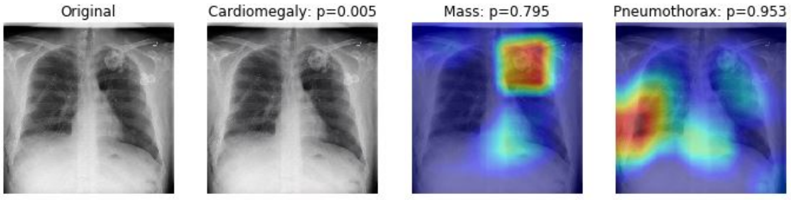


Figure 9. Visualization of pathology prediction using GRADCAM

the model's performance and gain a better understanding of the model's decision-making process. By visualizing the areas of focus, experts can assess whether the model's attention aligns with their own knowledge and intuition, aiding in building trust and confidence in the model's predictions.

4.5. Uncertainty estimation

Computing the confidence interval of a measurement is a valuable technique that allows us to understand the impact of sampling and provides an estimation of uncertainty in our predictions based on the training dataset. It quantifies our confidence in the reliability of the model's predictions by considering that the training dataset is only a sample of the real-world data.

In our experiment, we computed the confidence interval of the AUC score for each pathology and obtained the results, as presented in Table 6. These findings provide valuable insights into the reliability and consistency of our model predictions. The confidence interval represents a range of values within which we can be confident that the true AUC score falls. The narrower the interval, the more precise and confident we are in our predictions.

Table 6. Table of confidence interval for AUC score.

Pathology	Mean AUC (CI 5%-95%)
Cardiomegaly	0.89 (0.86-0.92)
Emphysema	0.79 (0.76-0.82)
Effusion	0.76 (0.73-0.80)
Hernia	0.77 (0.73-0.80)
Infiltration	0.65 (0.62-0.68)
Mass	0.82 (0.79-0.85)
Nodule	0.65 (0.60-0.70)
Atelectasis	0.79 (0.76-0.81)
Pneumothorax	0.81 (0.77-0.83)
Pleural_Thickening	0.71 (0.68-0.74)
Pneumonia	0.68 (0.62-0.73)
Fibrosis	0.72 (0.68-0.76)
Edema	0.86 (0.82-0.88)
Consolidation	0.75 (0.72-0.78)

Observing the confidence intervals, we can see that they are relatively narrow for almost all classes, indicating a high level of confidence in the results. This suggests that our proposed approach is robust and produces consistent results regardless of the specific sample of the training dataset. Narrow confidence intervals imply that the model's performance is consistent and reliable, reinforcing our confidence in the generalization ability of the model.

5. Limitations

The deep learning model presented in this study has several limitations that should be acknowledged. These limitations include:

(1) *Limited Image Types:* The model was trained using only frontal X-ray images. The unavailability of other types of images, such as lateral radiographs, restricts the model's exposure to different views and perspectives of the thoracic region.

Incorporating multiple image types could enhance the model's performance and its ability to capture a comprehensive understanding of various pathologies.

(2) *Lack of Medical History:* The model solely relies on the information extracted from the X-ray images and does not consider the patient's medical history. Medical history, including patient symptoms, previous diagnoses, and other relevant clinical information, plays a crucial role in disease diagnosis. By incorporating such contextual information, the model could potentially improve its accuracy and diagnostic capabilities.

(3) *Evaluation by Medical Professionals:* While the model demonstrates high sensitivity and specificity, it is essential to emphasize that it has not been evaluated by medical professionals. Comparing the model's performance to human performance is challenging, as it requires expert evaluation and validation. Therefore, further assessment by medical professionals is necessary to establish the model's clinical utility and compare its performance to that of human experts.

6. Conclusion and future works

In conclusion, this study demonstrates the potential of deep neural network models in the diagnosis of thoracic diseases using chest X-ray images. The evaluation of the model's performance indicates its effectiveness in detecting various pathologies. Additionally, the use of interpretable techniques such as Class Activation Maps (CAMs) provides insights into the model's decision-making process and aids in expert validation.

However, there are still several areas for improvement and further research. Incorporating different types of X-ray images, such as lateral radiographs, could enhance the model's performance and expand its applicability. Integration of patient medical history and contextual information may improve the accuracy and diagnostic capabilities of the model. Furthermore, collaboration with healthcare professionals is necessary to validate the model's predictions and assess its performance compared to human experts.

The development of an automated diagnosis system that combines the power of deep learning models with interpretability and expert collaboration holds great potential in improving access to reliable and efficient disease detection. By leveraging these technologies, we can work towards the goal of providing accurate and accessible diagnostic tools for thoracic diseases, ultimately benefiting patients worldwide and improving healthcare outcomes.

Data Availability

The dataset is publicly available at NIHCC website[2].

Funding

No funding was received to assist with the research work or preparation of this manuscript.

[2] `https://nihcc.app.box.com/v/ChestXray-NIHCC`

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors. The article uses open-source datasets.

References

Abdou MA. 2022. Literature review: efficient deep neural networks techniques for medical image analysis. Neural Computing and Applications. 34(8):5791–5812.

Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. 2015. Chest pathology detection using deep learning with non-medical training. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI). IEEE. p. 294–297.

Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. 2018. Chest pathology identification using deep feature selection with non-medical training. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 6(3):259–263.

Bhusal D, Shin R, Shewale AA, Veerabhadran MKM, Clifford M, Rampazzi S, Rastogi N. 2023. Sok: Modeling explainability in security analytics for interpretability, trustworthiness, and usability. In: The 18th International Conference on Availability, Reliability and Security (ARES 2023).

Bhuyan HK, Chakraborty C, Shelke Y, Pani SK. 2022. Covid-19 diagnosis system by deep learning approaches. Expert Systems. 39(3):e12776.

Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. 2021. Benchmarking and survey of explanation methods for black box models. arXiv preprint arXiv:210213076.

Bykov K, Hedström A, Nakajima S, Höhne MMC. 2022. Noisegrad—enhancing explanations by introducing stochasticity to model weights. In: Proceedings of the AAAI Conference on Artificial Intelligence; vol. 36. p. 6132–6140.

Cha MJ, Chung MJ, Lee JH, Lee KS. 2019. Performance of deep learning model in detecting operable lung cancer with chest radiographs. Journal of thoracic imaging. 34(2):86–91.

Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). p. 839–847.

Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, Barfett J. 2017. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Investigative radiology. 52(5):281–287.

Crosby J, Rhines T, Li F, MacMahon H, Giger M. 2020. Deep learning for pneumothorax detection and localization using networks fine-tuned with multiple institutional datasets. In: Medical Imaging 2020: Computer-Aided Diagnosis; vol. 11314. SPIE. p. 70–74.

Dasanayaka C, Dissanayake MB. 2021. Deep learning methods for screening pulmonary tuberculosis using chest x-rays. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 9(1):39–49.

Dhruv B, Mittal N, Modi M. 2023. Hybrid particle swarm optimized and fuzzy c means clustering based segmentation technique for investigation of covid-19 infected chest ct. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 11(2):197–204.

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. PMLR. p. 647–655.

Doshi-Velez F, Kim B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608.

Erion G, Janizek JD, Sturmfels P, Lundberg SM, Lee SI. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature machine intelligence. 3(7):620–631.

Farooq M, Hafeez A. 2020. Covid-resnet: A deep learning framework for screening of covid19 from radiographs. arXiv preprint arXiv:200314395.

Fukushima K, Wake N. 1991. Handwritten alphanumeric character recognition by the neocognitron. IEEE transactions on Neural Networks. 2(3):355–365.

Goodfellow I, Bengio Y, Courville A. 2016. Deep learning. MIT press.

He F, Deng Y, Li W. 2020. Coronavirus disease 2019: What we know? Journal of medical virology. 92(7):719–725.

He K, Zhang X, Ren S, Sun J. 2016a. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770–778.

He K, Zhang X, Ren S, Sun J. 2016b. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770–778.

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 4700–4708.

Hussain K, Khan SA, Muzaffar M, Khan ZR, Farooq SU. 2023. Performance evaluation of cnn architectures for covid-19 detection from x-ray images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 11(1):80–93.

Irene D S, Beulah JR. 2022. An efficient covid-19 detection from ct images using ensemble support vector machine with ludo game-based swarm optimisation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 10(6):675–686.

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence; vol. 33. p. 590–597.

Islam MT, Aowal MA, Minhaz AT, Ashraf K. 2017. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. arXiv preprint arXiv:170509850.

Kesav N, MG J. 2022. A deep learning approach with bayesian optimized kernel support vector machine for covid-19 diagnosis. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization:1–15.

Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980.

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems; vol. 25. Curran Associates, Inc.

Kwaśniewska A, Rumiński J, Rad P. 2017. Deep features class activation map for thermal face detection and tracking. In: 2017 10Th international conference on human system interactions (HSI). IEEE. p. 41–47.

Lakhani P, Sundaram B. 2017. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 284(2):574–582.

Li K, Fang Y, Li W, Pan C, Qin P, Zhong Y, Liu X, Huang M, Liao Y, Li S. 2020. Ct image visual quantitative evaluation and clinical classification of coronavirus disease (covid-19). European radiology. 30:4407–4416.

Li Y, Shen L. 2018. Skin lesion analysis towards melanoma detection using deep learning network. Sensors. 18(2):556.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. 2017. A survey on deep learning in medical image analysis. Medical image analysis. 42:60–88.

Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems. 30.

Maduskar P, Muyoyeta M, Ayles H, Hogeweg L, Peters-Bax L, Van Ginneken B. 2013. Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers. The International journal of tuberculosis and lung disease. 17(12):1613–1620.

Molnar C. 2022. Interpretable machine learning. 2nd ed. Available from: `https://christophm.github.io/interpretable-ml-book`.

More PS, Saini BS. 2022. Competitive verse water wave optimisation enabled covid-net for covid-19 detection from chest x-ray images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization:1–13.

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences. 116(44):22071–22080.

Noshad A, Arjomand P, Khonaksar A, Iranpour P. 2021. A novel method for detection of covid-19 cases using deep residual neural network. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 9(5):555–564.

Patel RK, Kashyap M. 2023. Machine learning-based lung disease diagnosis from ct images using gabor features in littlewood paley empirical wavelet transform (lpewt) and lle. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization:1–15.

Pieters W. 2011. Explanation and trust: what to tell the user in security and ai? Ethics and information technology. 13(1):53–64.

Pushparaj TL, Irudaya Raj EF, Irudaya Rani EF. 2022. A detailed review of contrast-enhanced fluorescence magnetic resonance imaging techniques for earlier prediction and easy detection of covid-19. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization:1–13.

Rajarajeswari P, Chattopadhyay P, O AB. 2022. A deep learning computational approach for the classification of covid-19 virus. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization:1–10.

Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. PLoS medicine. 15(11):e1002686.

Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:171105225.

Raoof S, Feigin D, Sung A, Raoof S, Irugulpati L, Rosenow III EC. 2012. Interpretation of plain chest roentgenogram. Chest. 141(2):545–558.

Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C. 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages. NeuroImage. 155:530–548.

Ren S, He K, Girshick R, Sun J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 28:91–99.

Ribeiro MT, Singh S, Guestrin C. 2016. " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. p. 1135–1144.

s D, s S. 2023. Diagnosis and detection of covid-19 infection on x-ray and ct scans using deep learning based generative adversarial network. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization:1–11.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. p. 618–626.

Simonyan K, Vedaldi A, Zisserman A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:13126034.

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. ICLR.

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. p. 180–186.

Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. 2017. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:170603825.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 15(1):1929–1958.

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. In: International conference on machine learning. PMLR. p. 3319–3328.

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 1–9.

Taylor AG, Mielke C, Mongan J. 2018. Automated detection of moderate and large pneumothorax on frontal chest x-rays using deep convolutional neural networks: A retrospective study. PLoS medicine. 15(11):e1002697.

Tymchenko B, Marchenko P, Spodarets D. 2020. Deep learning approach to diabetic retinopathy detection. arXiv preprint arXiv:200302261.

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 2017a. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 2097–2106.

Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 2017b. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 2097–2106.

Yang R, Li X, Liu H, Zhen Y, Zhang X, Xiong Q, Luo Y, Gao C, Zeng W. 2020. Chest ct severity score: an imaging tool for assessing severe covid-19. radiol cardiothorac imaging 2 (2): e200047.

Zhou Z, Guo D, Li C, Fang Z, Chen L, Yang R, Li X, Zeng W. 2020. Coronavirus disease 2019: initial chest ct findings. European radiology. 30:4398–4406.