Project Number 6

Title

Transfer Learning Using ResNet-50 for Lung Disease Detection

Abstract

This project focuses on the use of the convolutional neural network ResNet-50 for detecting lung diseases. Methods such as transfer learning, data augmentation, and other techniques to overcome overfitting were applied. The dataset used was the publicly available ”NIH-Chest-X-ray-dataset” from Hugging Face. After training, the model was evaluated based on the AUC-ROC curve and sensitivity/specificity for each class individually.

Keywords

ResNet-50, Transfer learning, Lung Disease Detection, Data augmentation, Chest-X-Ray dataset, AUC-ROC curve

Introduction

The diagnosis of lung diseases using imaging techniques, such as chest X-rays, plays a key role in medical practice. However, manual interpretation of these images is time-consuming and prone to subjective errors. In recent years, deep neural networks, especially convolutional neural networks (CNNs), have proven to be an effective tool for the automated analysis of medical images.

This work focuses on the detection of lung diseases using the ResNet-50 model, which is known for its depth and ability to extract relevant features from complex image data. To accelerate and improve the learning process, the transfer learning approach was applied, allowing a pre-trained model to be adapted to a new, specific task. In addition, techniques such as data augmentation and methods for reducing overfitting were used to improve the model's generalization.

Description of the method

During training, several methods were used to overcome overfitting, such as transfer learning, data augmentation, class weighting, and dropout regularization. In addition, the model was trained using different hyperparameters to determine which are the most suitable for the given model and dataset.

Hyperparameter	Value
learning rate	0.0001
batch size	32
optimizer	Adam
loss function	BCE with Logits Loss
number of epochs	5

Table 6.1: The final hyperparameters

The dataset was split in the ratio of 70% for the training set, and 15% each for the validation and test sets.

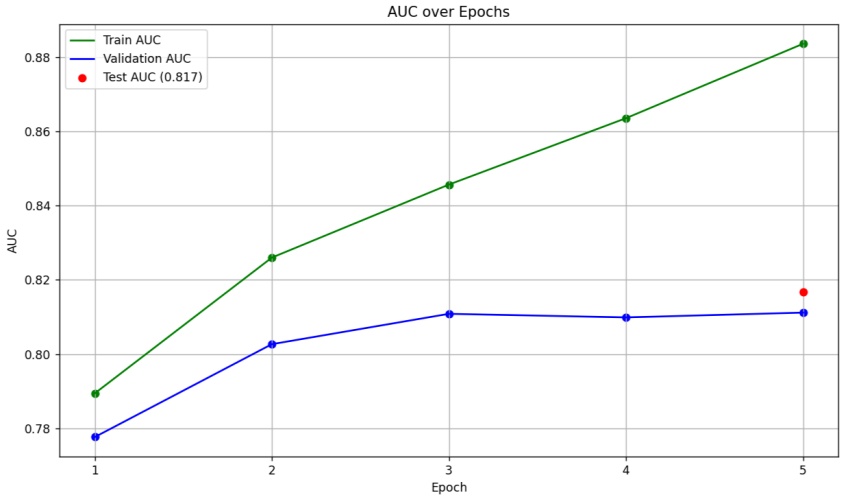


Figure 6.1: Training of the model

The model was initialized with weights pre-trained on the large-scale ImageNet dataset. These weights capture fundamental image features such as edges, shapes, and textures, which are also valuable for analyzing medical images.

In transfer learning, these pre-trained layers—especially the lower convolutional layers—are either kept frozen or fine-tuned slightly, while the final layers of the network are adapted to the specific task, in this case, the detection of lung diseases.

To address the issue of class imbalance in the data, positive class weighting was applied for each class. The weights were calculated as the ratio of the number of negative samples to the number of positive samples in the training set. Incorporating these weights into the loss function increases the penalty for errors on the rarer classes, which improves the model's ability to recognize less represented diseases and enhances its overall accuracy.

To reduce the risk of overfitting, dropout regularization with a rate of 0.5 was applied in the model. Dropout is a regularization technique that randomly "turns off" (sets to zero) 50% of the neurons in a given layer during each training step. This prevents the model from becoming too tailored to specific patterns in the training data and forces it to learn more robust and generalizable representations.

One of the main methods used to improve the model's generalization and balance classes with very few images was data augmentation. Since the dataset consists of X-ray images, the available augmentation options were limited. Therefore, aggressive augmentations like strong rotations, cropping, or other drastic transformations could not be applied. So, we applied rather cautious (conservative) augmentations.

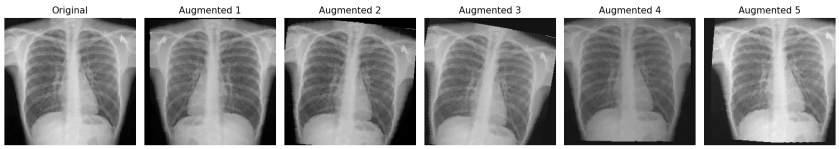


Figure 6.2: An example of the augmentations used.

We applied augmentations only to the training set.

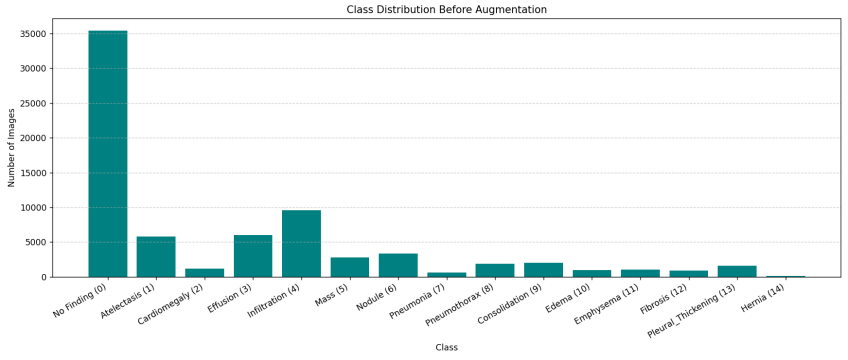


Figure 6.3: Class distribution before augmentation

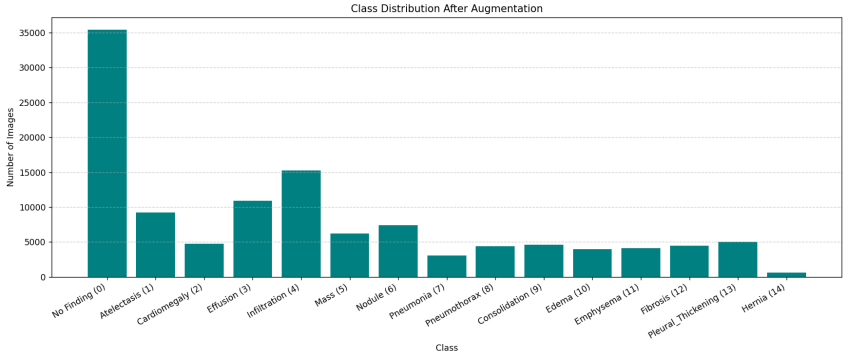


Figure 6.4: Class distribution after augmentation

Topology of the model

The ResNet-50 model was used as the base architecture. It is a deep convolutional neural network (CNN) known for its ability to efficiently extract visual features from complex image data. ResNet-50 consists of 50 layers and is characterized by the use of residual blocks, which enable layer skipping through identity mapping. This approach helps prevent degradation issues that can occur in very deep networks.

The network topology includes an initial convolutional layer followed by normalization and a max pooling layer, after which several residual blocks of varying depths are stacked. These blocks combine convolutional, normalization, and activation layers. At the end of the network, there is a global average pooling layer and a fully connected output layer, which was adapted in this project to match the number of target classes (15 lung diseases). Additionally, a dropout layer with a rate of 0.5 was added before the final linear layer to improve generalization.

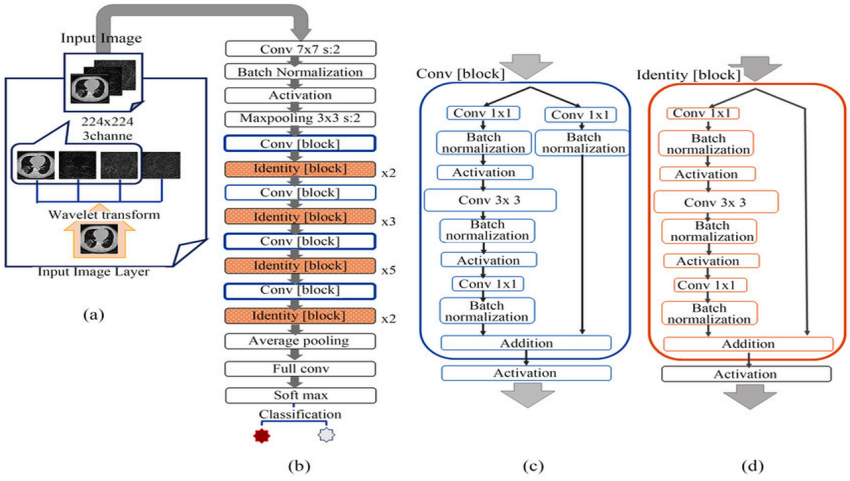


Figure 6.5: Topology of ResNet-50

Datasets description

In this work, the publicly available NIH Chest X-ray dataset was used, which contains 112,120 frontal chest X-ray images from 30,805 unique patients. Each image is annotated with one or more of the 14 most common lung diseases. These diseases include: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia.

The dataset is multi-label, meaning that a single image can be associated with multiple diagnoses at the same time. The images are in PNG format and have been anonymized to ensure the protection of patient privacy.

An important characteristic of this dataset is the significant class imbalance—some diagnoses appear in only a small number of images compared to others. This class imbalance poses a challenge during model training and requires the use of special techniques, such as class weighting during learning, to prevent the model from favoring more frequent diseases.

Experimental results

Since this is a multi-label classification task, the main evaluation metric is the AUC-ROC curve (Area Under the Receiver Operating Characteristic Curve), which is considered a standard for measuring the performance of binary classifiers and is well-suited for extension to multi-label cases.

The ROC curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different threshold values. In other words, the ROC shows how well the model distinguishes between positive and negative examples.

The AUC (Area Under Curve) value represents the area under this curve. The closer this value is to 1, the better the model classifies. An AUC of 0.5 corresponds to random guessing, while an AUC of 1.0 indicates perfect class separation.

In the case of multi-label classification, the AUC-ROC is computed separately for each class and then averaged. The most commonly used approach is the macro-average, which treats each class as equally important, regardless of its frequency. This is especially important in the context of class imbalance, where the classifier needs to correctly detect even the rarest diagnoses.

This metric is suitable because it is independent of the choice of decision threshold and provides a comprehensive view of the model's ability to distinguish between classes across all possible probability thresholds.

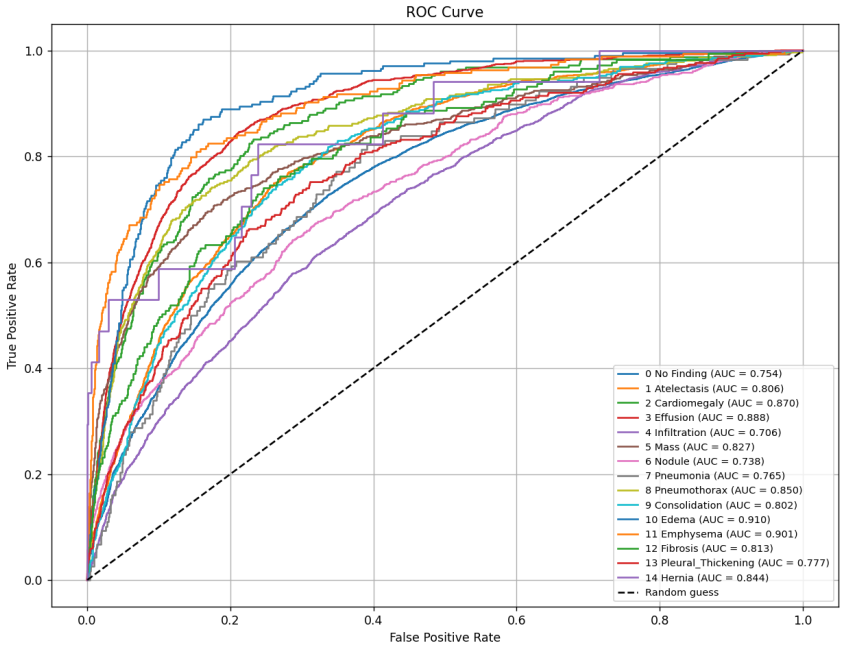


Figure 6.6: AUC-ROC curves

The image above shows the ROC curves for each of the 15 classes (diagnoses) in the multi-label classification task. Each curve represents the model's performance in distinguishing the given pathology, accompanied by the AUC value quantifying its accuracy. The graph reveals that the model achieved varying degrees of success in classifying individual diseases.

The highest AUC values were reached for the classes Edema (0.910), Emphysema (0.901), Effusion (0.888), and Cardiomegaly (0.870), indicating that the model very effectively identifies these diagnoses. These high AUC scores mean the model has excellent ability to distinguish positive cases from negative ones in these categories.

Conversely, lower AUC values were observed for classes such as Infiltration (0.706) and Nodule (0.738). These diagnoses tend to be visually less distinct or more easily confused with other conditions, complicating accurate classification. This may also be related to their lower representation in the training dataset.

The class No Finding achieved an AUC of 0.754, reflecting the challenge of distinguishing between healthy findings and latent or less obvious pathologies that may be present to a lesser extent.

All curves lie above the diagonal line of a random classifier (AUC = 0.5), confirming that the model performs better than random chance. Moreover, the steep initial segments of the curves suggest the model achieves high sensitivity at low false positive rates.

These results also highlight a significant class imbalance in the dataset — some classes are substantially underrepresented, which negatively impacts classifier performance for these cases. Therefore, the use of the AUC-ROC metric is particularly important, as it is independent of any specific decision threshold and provides a robust evaluation across all classes even with uneven distribution.

Subsequently, threshold values for each class were determined on the validation set. These thresholds represent the boundaries at which a sample is classified as belonging to a given class or not. The model generates probabilities or scores for individual classes during prediction, and these thresholds decide from which value the prediction is considered positive for that class.

These threshold values were determined based on Youden's index, which is calculated as the difference between the true positive rate (TPR or sensitivity) and the false positive rate (FPR), i.e.:

$$J = \text{TPR} - \text{FPR} \tag{6.1}$$

Selecting the threshold using Youden's index means finding the value at which this difference (and thus the trade-off between sensitivity and specificity) is maximized. This approach ensures that the model best balances correctly identifying positive cases while minimizing the number of false positive classifications.

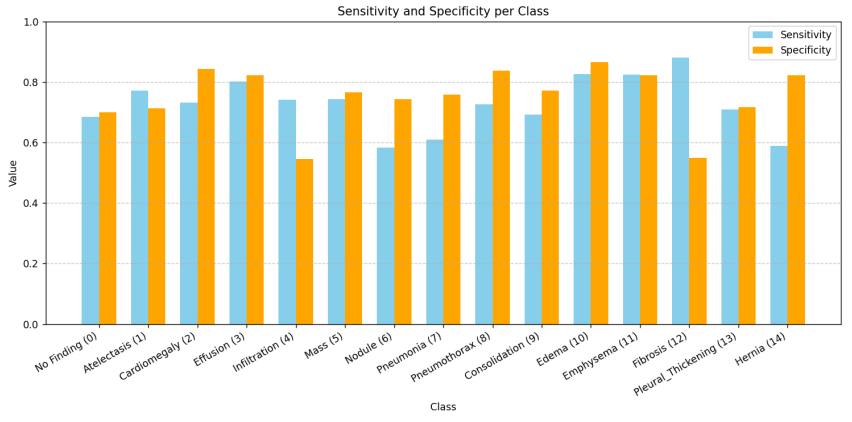


Figure 6.7: Sensitivity and specificity per class

After determining these optimal threshold values, the model was tested on the training set and evaluated using sensitivity and specificity metrics for each class separately. This evaluation

allowed assessing the model's performance in classifying individual classes based on the chosen thresholds. The formulas for these metrics are:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \text{Specificity} = \frac{TN}{TN+FP} \tag{6.2}$$

Conclusion

This project demonstrated the effectiveness of transfer learning using the convolutional neural network ResNet-50 for detecting lung diseases from chest X-ray images. By leveraging pre-trained weights from the ImageNet dataset and applying techniques such as data augmentation, regularization via dropout layers, and class weighting, the model achieved solid performance in classifying various pulmonary conditions, despite significant class imbalance.

The AUC-ROC metric showed that the model performed best on diagnoses such as edema, emphysema, and effusion, where the AUC exceeded 0.88. In contrast, the model achieved lower performance on less represented or visually ambiguous classes such as infiltration and nodule, highlighting the need for further improvements in data balance and the model's sensitivity to subtle image features.

Despite these limitations, the results confirm that deep neural networks, when properly adapted through transfer learning, can effectively support automated analysis of medical imaging data and have the potential to serve as decision-support tools in clinical practice. Future work could include training on larger and more balanced datasets, incorporating clinical metadata, and exploring ensemble approaches to further enhance diagnostic accuracy.

Citations

This project utilizes several key resources. For an in-depth understanding of the ResNet-50 architecture, refer to [1]. The foundational work on deep residual learning is detailed in [2]. The dataset employed in this study is the NIH Chest X-ray Dataset [3].

[1] A. Rastogi, "Understanding resnet-50 in depth: Architecture, skip connections, and advantages over other networks," *Wisdom ML*, 2023, Accessed: 2025-06-03. [Online]. Available: https://wisdomml.in/understanding-resnet-50-in-depth-architecture-skip-connections-and-advantages-over-other-networks/.

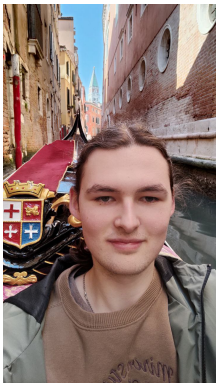
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385.

[3] Alkzar90, *Nih chest x-ray dataset*, Accessed: 2025-06-03, 2023. [Online]. Available: https://huggingface.co/datasets/alkzar90/NIH-Chest-X-ray-dataset.

Authors

Dmytro Skrypchenko

Pavlo Yarovyi



Nikita Dakhno