AI Project in Healthcare: Utilizing Computer Vision for Medical Images

MASTER THESIS

Mit Ashokbhai Desai

Submitted on 14 August 2025

Friedrich-Alexander-Universität Erlangen-Nürnberg
Faculty of Engineering, Department Computer Science
Professorship for Open Source Software

Supervisor:
Prof. Dr. Dirk Riehle, M.B.A.
Prof. Dr. Frauke Liers

Friedrich-Alexander-Universität
Faculty of Engineering

Declaration of Originality

I confirm that the submitted thesis is original work and was written by me without further assistance. Appropriate credit has been given where reference has been made to the work of others. The thesis was not examined before, nor has it been published. The submitted electronic version of the thesis matches the printed version.

Erlangen, 14 August 2025

License

Erlangen, 14 August 2025

Abstract

*T*he rising rates of respiratory diseases like pneumonia and COVID-19 have made the need for fast, reliable, and accessible tools to analyze medical images important and necessary. Chest X-rays (CXRs) are still one of the most prevalent imaging modalities to be used in the diagnosis of these conditions, but they normally require expert radiological evaluation, which in certain situations may not be conveniently accessible in underprivileged conditions. This thesis addresses the use of deep learning methods in the automatic interpretation of CXR images in aiding the process of lung disease detection over healthy images. Several convolutional neural network (CNN) models were tested against each other, such as VGG, ResNet. ResNet-50 was used as a transfer learning model, and several ensemble mechanisms were tried. ResNet-50 and VGG-16 combined with equal weights as an ensemble, had produced the best results. Interestingly, in all models, training was performed on image-level labels only (excluding annotated lesion coordinates), which shows that high performance can be achieved even without detailed annotations. The accuracy of the final ensemble model is 94.98%, precision is 95.26%, recall is 94.98%, and the F1-score is 94.96%. It has, as well, an AUROC value of 0.9812 and a 97.37% specificity, which emphasize its high discriminative ability. A graphical user interface (GUI) has been designed to implement usability, requiring the user to enter the symptoms and also to upload a CXR image. Both image-based and symptom-based diagnostic predictions are returned by the system, and this aspect means that the system can be fully integrated into clinical workflows. Future directions will involve further enhancement in the interpretability of models, classification of other thoracic disorders, and evaluation of the system in practice in clinical practice.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

CXRs Chest X-rays

CNN Convolutional Neural Network

ResNet Residual Network

VGG Visual Geometry Group

Grad-CAM Gradient-weighted Class Activation Mapping

ViT Vision Transformer

LIME Local Interpretable Model-Agnostic Explanations

SHAP Shapley Additive Explanations

PLI Pixel-Level Interpretability

VQA Visual Question Answering

Lr Learning Rate

DenseNet Densely Connected Convolutional Network

ROI Region of Interest

XAI Explainable Artificial Intelligence

GPU Graphics Processing Unit

IoU Intersection over Union

CPU Central Processing Unit

AdamW Adaptive Moment Estimation with Weight Decay

PA view Posteroanterior View

ReLU Rectified Linear Unit

OCR Optical Character Recognition

FC Fully Connected

CSS style Cascading Style Sheets Style

Acc Accuracy

AUROC Area Under the Receiver Operating Characteristic Curve

OvR One-vs-Rest

FPR False Positive Rate

TPR True Positive Rate

1 Introduction

1.1 Motivation

The issue of respiratory diseases remains a significant health burden throughout the world and Pneumonia and COVID generate a significant morbidity/ mortality rate. In-hospital mortality of COVID patients in Europe, based on cohort and observational data, has ranged between about 16-20% with one German study reported 18.9% of mortality (Hedberg et al., 2024; Smith, 2024). In Germany, it is confirmed that Pneumonia is a significant cause of infectious deaths since the report projections, according to the report show that in 2024, there will be approximately 19,240 deaths caused by Pneumonia (ReportLinker, 2024). CXRs play a major role in diagnostics of these conditions because of their availability and comparatively low cost. However, their accurate interpretation demands specialized expertise, which is not accessible regularly in underserved and remote locations; also problematic in cities, as the interpretation relies on human experience and the most-experienced doctor is not always available.

The COVID pandemic placed an additional strain on the healthcare systems and showed a gap in the diagnostic capabilities and the necessity of scalable and efficient solutions. In that regard, AI-supported diagnostic instruments, especially those using deep learning and computer vision, present great potential. They could be used to automate CXR reading with great level of accuracy and eventual delay of effective diagnoses, and to reach rural areas with limited medical assets in a bid to provide quality healthcare. This thesis aims at the objective to contribute to the solution of these issues creating AI models that can aid the practical and correct diagnosis in the real-world healthcare setting.

1.2 Background

Artificial intelligence (AI) has been a game changer in the field of medical imaging, especially in automating the recognition of disease on radiology scans. CNNs, as well as architectures such as ResNet or VGG, have shown to be effective in

diagnosing chest ailments, specifically COVID and Pneumonia, when augmented with transfer learning. On the same note, transformer-based models have demonstrated potential in modeling complex medical imaging patterns because of their attention and global context representation capability.

The interpretability of such models has seen rising concern so as to facilitate their use in clinical decision-making. Visual explanations become possible due to such techniques as Grad-CAM, which highlights areas on the image that impacted the prediction of the model. This is particularly very crucial where transparency is important in clinical settings as to trust and adoption.

Nevertheless, most of these methods have the assumption that detailed annotations, e.g., bounding boxes or segmentation masks, indicating the pathological areas are availed. Creating such kind of annotations is cumbersome and needs skilled radiologists, and as such it becomes inapplicable in most real-life or resource-limited scenarios. As such, most of the literature remains abstract and assumptions made in literature may not translate well into clinical practices, thus reducing the general applicability and scalability of approaches.

This context reveals the necessity of diagnostic systems which can work with less supervision by utilizing solely image level labels, but also retain interpretability and clinical validity. It will discuss the reduction of the reliance on fully annotated data and strategies to alleviate such limitations without compromising the diagnostic performance in the following sections.

1.3 Problem Statement

Although current AI frameworks have demonstrated promising results in the classification of CXRs, they mainly assume having extensive supervision available throughout the training process. Nonetheless, the present task of the accurate classification and localization of conditions like Pneumonia and COVID mainly based on the coarse, image-level labels has a substantial gap. Many models do not produce clinically useful visual explanations when guidance was not provided as annotations because they can point to irrelevant parts of the image. Furthermore, the vast majority of methods are computationally expensive, which makes their implementation in the low-resource environment difficult. That demands a lightweight, explainable, and annotation-cost effective diagnosis tool, which can generalize to clinical practice without the assistance of localization information offered by the experts.

1.4 Research Questions

This research seeks to address the following questions:

RQ1: Would it be possible to develop a pretrained deep learning model able to correctly annotate chest X-ray images with either of the three categories, Normal, Pneumonia, or COVID, with image-level labels only and without the use of pixel-level annotations?

RQ2: What novel techniques can be used to identify clinically meaningful regions in CXRs in the absence of ground-truth bounding boxes or masks?

RQ3: What are the benefits of unsupervised or weakly supervised lung segmentation on the specificity and explanatory value of Grad-CAM visualizations, and how precisely does it inform local decisions on lung areas of scientific interest?

1.5 Objectives

The primary objectives of this thesis are:

1. To come up with a classification and localization pipeline not only with pretrained CNNs but also with Transformer-based models.
2. Extend explainable artificial intelligence (XAI) methods, namely Gradient-weighted Class Activation Mapping (Grad-CAM), to mark class-discriminative chest X-ray image patches.
3. To perform unsupervised or self-supervised methods of lung segmentation that will restrict visual explanations to anatomically reasonable regions of interest.
4. To illustrate and test the models using the publicly accessible chest X-ray datasets with a particular emphasis on the accuracy of classification, the quality of localization, and interpretability of a model.
5. To explore the use of hybrid ensemble approaches, e.g. combining features or outputs of other models (e.g., VGG and ResNet), and to provide results of at least one trained model that has been trained from scratch in order to compare.

1.6 Thesis Outline

This thesis is structured as follows:

- **Chapter 2: Literature Review** review investigates the history of AI-related practices in the sphere of medical imaging, with particular regard to CXRs classification and localization. It explains the substantial contributions and advancement of the gaps of prior research that are motivating this direction of the thesis.
- **Chapter 3: Theoretical Background** describes the basic technical groundworks used throught the thesis. Topics include CNNs, VGG and ResNet architectures, ensemble and transfer learning, vision transformers (ViT), weak supervision, Grad-CAM explainability, lung segmentation techniques, performance metrics, Frontend via Streamlit and Deployment using Docker.
- **Chapter 4: Requirements** introduces the needs of the proposed system in both functional and technical terms. It covers the problems of mismatch between classes of a dataset, the lack of sufficient annotations, and the preprocessing demands. It also tells which datasets was chosen and establishes the norm of designing and testing the model.
- **Chapter 5: Architecture** describes the architectural design of proposed models, including ResNet,VGG, transformer-based model,transfer learning model and ensemble model pipelines.
- **Chapter 6: Design and Implementation** presents the training pipeline, which includes data augmentation, transfer learning techniques, fine-tuning, and adding segmentation to the training setup.
- **Chapter 7: Evaluation** defines performance of the designed models by the procedure of the classification evaluation (accuracy, precision, recall, F1 score), as well as the localization assessment, via Grad-CAM visualization. It is a comparison of alternative architecture of models and approaches to segmentation in which their advantages and shortcomings are examined.
- **Chapter 8: Conclusion** describes the entire process of research; starting with the definition of the problem and the preparation of data to the training and evaluation of the processes. It remarks on the contribution that the proposed approach makes to annotation-free localization and provides the planned further directions of the work, particularly those regarding its clinical application and model generalization.

2 Literature Review

2.1 Grad-CAM as a Means of Increasing the Interpretability of Medical Imaging

Grad-CAM is an initial explainability methodology presented by Selvaraju (Selvaraju et al., 2017a) to identify the image parts that attribute to CNN decisions by producing class-discriminative heatmaps. It does the issue by finding the gradients of the classification score's output with respect to the feature maps of the network's last convolutional layer, scaling the feature maps with their gradients, and then using a ReLU activation to remove the features that are damaging the score.

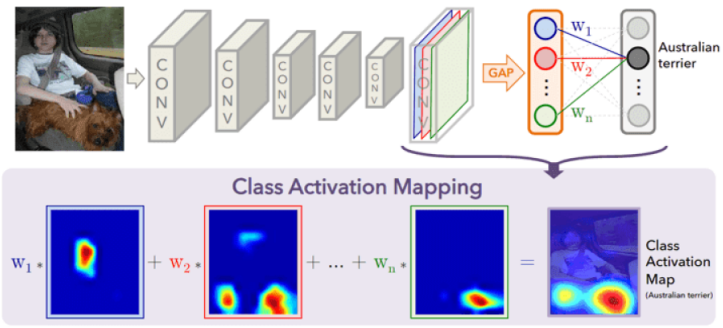


Figure 2.1: Grad-CAM pipeline showing how gradients weight convolutional feature maps to produce a class activation map. Adapted from Selvaraju(Selvaraju et al., 2017b).

Grad-CAM was extensively used in many clinical applications like lung cancer identification, breast cancer segmentation, brain tumor detection as well as lymph node metastasis segmentation in medical imaging (Reyes et al., 2020; Suara et al., 2023). Its main strength lies in providing visual explanations that align with established clinical knowledge, thereby increasing clinicians' confidence in using AI for disease diagnosis (Suara et al., 2023). Grad-CAM makes it easier to

validate and question automatized decisions since it permits clinicians to visually validate model predictions.

However, Grad-CAM has remained to have significant drawbacks in the application of medical imaging. It was shown to tend low robustness, where small input perturbations can cause problems in the stability of heatmaps(Reyes et al., 2020; Suara et al., 2023). In addition, unlike the CNNs, its spatial resolution is rather rough, because it uses the last convolutional layer, and localization of some minor lesions or ambiguous ones is also reduced(Suara et al., 2023). Grad-CAM only reflects the high-level activations and therefore might ignore clinically important mid-relevant or anatomical contexts hence its semantic interpretability is lower. These difficulties come with clinical risks, and more grounded and domain-adaptable explainability techniques must be discovered(Zoelden et al., 2020).

These problems were addressed in recent proposals towards more robust gradient computations(Suara et al., 2023), an injection of prior anatomical knowledge to keep explanations within relevant regions (Zoelden et al., 2020), or hierarchical or multi-layer schemes to achieve better resolution and variety of features covered(Hahn et al., 2020; Robb et al., 2021). As one illustrating example, utilize Grad-CAM in a model to detect cancer metastasis in histopathological lymph nodes and explain how such explainability can promote clinically efficient and reliable work processes(Suara et al., 2023).

It is important to get familiar with the leaps and pitfalls of Grad-CAM in order to build clinically-salient explainability pipelines. Although Grad-CAM is prevalently available, robustness, improved spatial localization, and expertise of the medical domain will have to be attacked in the future to improve transparent AI implementations in a high-stake context such as healthcare.

2.2 Comparative Understanding in Medical Imaging AI: Pixel-Level and Grad-CAM Methodologies

The use of the AI in medical imaging, namely in the methods of deep learning, has delivered a substantial rise in diagnostic accuracy and efficiency (Ennab and Mcheick, 2025). However, the black box nature of these models will also continue to be a major blockage towards their applications in the clinical setting. This lack of transparency causes a level of mistrust between AIs and healthcare practitioners, especially in high-stakes situations where the choice accordingly affects patient outcomes(Ennab and Mcheick, 2025).

To solve this problem, scientists have investigated methods of interpretability,

which are capable of showing how and why an AI model makes a certain prediction. CNNs Local Interpretable Model-agnostic Explanation-based methods such as Local Interpretable Model-Agnostic Explanations (LIME),Grad-CAM,and Shapley Additive Explanations(SHAP) provide some insights but are ultimately inferior to providing pixel-level granularity necessary in localizing pathologies in medical images effectively(Ennab and Mcheick, 2025).

Pixel-Level Interpretability (PLI)

To address the shortcomings of Grad-CAM, the Pixel-Level Interpretability (PLI) framework was proposed by Ennab and Mcheick (Ennab and Mcheick, 2025). PLI is based on fuzzy logic, combined with CNN feature maps, to generate high-resolution, pixel-level heatmaps, which are much more interpretable and more precise in diagnosis. In contrast Grad-CAM, PLI is able to come up with exposures conformable to clinical rationale by giving each pixel a localized level of significance(Ennab and Mcheick, 2025).

This model uses an extended VGG-19 architecture augmented with fuzzy logic process i.e. fuzzification, fuzzy inference and defuzzification operations to convert complicated neural activation into interpretable diagnostic values comprehensible to humans. Such colourful representations boost the confidence of clinicians in the AI-aided choices and render the model showings clinically practical.

Role of Fuzzy Logic in Interpretability

Fuzzy logic allows the models to deal with ambiguity and partly truth—any indispensable feature in medical fields where the choices are hardly ever binary. The fuzzy logic of PLI transforms CNN activations into fuzzy subsets, which makes it possible to interpret the results at a pixel level. This method identifies better with the reasoning process of clinicians that is based on uncertainty(Ennab and Mcheick, 2025).

Integration into this Thesis

The conceptual framework introduced by Ennab and Mcheick, specifically, the PLI framework, was of great inspiration and was used to inform both the design of models and the interpretability targets. Their workflow Figure 3 here is (Figure 2.2) has been modified and simplified to be applied in a multi-modal context by generalizing the concept of pixel-wise explainability. The architecture also internalizes a related concept of detailed visual elaboration, although the particular knockoff rules of fuzzy membership were not imitated.

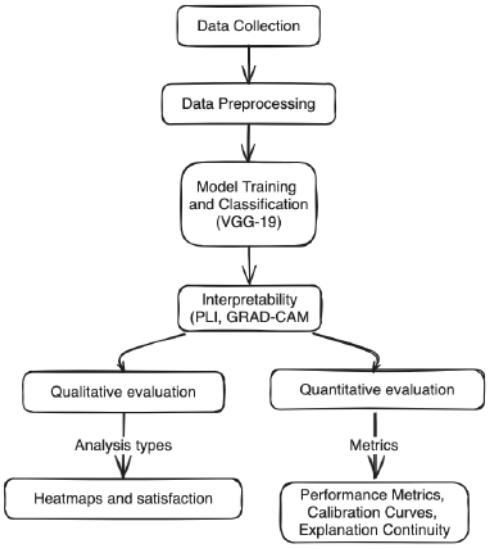


Figure 2.2: Adapted workflow for model training and interpretability evaluation using PLI and Grad-CAM. Inspired by the methodology presented in(Ennab and Mcheick, 2025).

2.3 Grad-CAM: The way to Apply Gradient-Based Localisation to Get Visual Explanations from Deep Networks

In high-stakes applications such as medical diagnostics, interpretability in deep neural networks, and specifically CNNs, is of great significance to understand the significance of transparency and model trustworthiness. An effective way to solve this requirement is described by Selvaraju, the so-called Grad-CAM. Grad-CAM produces visual explanations of CNN-based model predictions by utilizing the activations of CNN filters that flow into the deepest convolutional layer to produce the class-discriminative localization heatmaps correction without any changes in the architecture and retraining(Selvaraju et al., 2017a; Suara et al., 2023).

The Grad-CAM approach figures out the gradient of a certain classification score in relation to the feature maps of a chosen convolutional layer. After that, these gradients are aggregated globally and averaged together to get the neurone significance weights, which are used to weigh activation maps.. The resulting Map of heats identifies the regions within the image that plays the greatest role in the decision of the model.

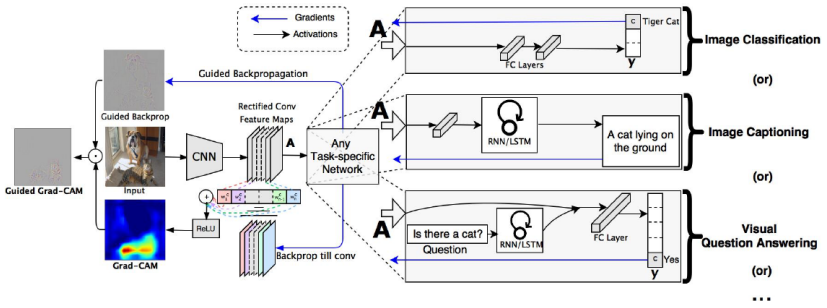


Figure 2.3: Overview of Grad-CAM: Gradients flowing into the final convolutional layer are combined to produce class-discriminative heatmaps.(Selvaraju et al., 2017a; Suara et al., 2023)

The authors proposed the Guided Grad-CAM to enhance visual resolution in Grad-CAM heatmaps, using the Guided Backpropagation in a combination with Grad-CAM. The resulting visualizations are both class-specific and high-resolution, which qualifies this hybrid strategy as being well-suited to a medical imaging application where localization of anatomical anomalies requires fine-grained localization.(Selvaraju et al., 2017a; Suara et al., 2023)

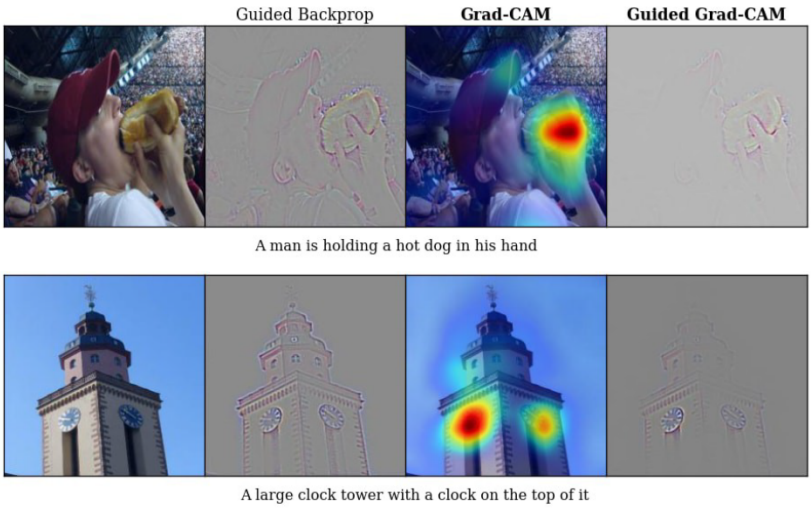


Figure 2.4: Comparison of Grad-CAM with alternative visualization methods like Guided Backpropagation and Deconvolution. Guided Grad-CAM offers the best combination of resolution and class-specific focus.Figure adapted from (Selvaraju et al., 2017a; Suara et al., 2023)

Grad-CAM has achieved high utility in various tasks, and these include:(a) classifying an image, which includes emphasizing areas related to the

disease in CXRs.(b) Image captioning by visualizing the relevant areas of the image.(c) Visual Question Answering (VQA), which can be used to display model attention when answering queries(Suara et al., 2023).

In the medical imaging context, these visualization tools are invaluable for: explaining errors or equivocal forecasts, Building clinician confidence in AI-based decision systems,Identifying and addressing biases in datasets that may affect generalization (Stodt et al., 2023; Suara et al., 2023).

Grad-CAM has been extensively validated through the Common benchmarks in weakly-supervised localization including ImageNet(Selvaraju et al., 2017a; Suara et al., 2023), Human-subject studies indicating Grad-CAM-based explanations are more believable and clearer(Suara et al., 2023), Comparative analysis where it outperforms CAM, Guided Backpropagation, and Deconvolution in faithfulness and explainability(Stodt et al., 2023).

3 Theoretical Background

3.1 Convolutional Neural Networks (CNNs)

CNN are a category of deep neural networks specially adapted to work on data whose topology can be expressed as a grid, most commonly a raster graphics image. CNNs were initially suggested by LeCun (LeCun et al., 1998) as applicable in digit recognition tasks, though these have since formed the basis of most recent computer vision systems, including analysis of medical images. A typical CNN refers to a system of numerous loss functions that allows the automatic classification of hierarchical features of raw image information.

The key elements of CNN are:

Convolutional layers: To get local spatial information from the input picture, such as edges, textures, and forms, it uses a series of learnable filters, often known as kernels. The filters go over the image to make a feature representation, which is a map of where a certain pattern may be found in each space (Krizhevsky et al., 2012).

Activation Functions: After any convolution operation, a nonlinear activation function (usually ReLU—Rectified Linear Unit) is introduced, making the model learn complex mappings(He et al., 2015b; Krizhevsky et al., 2012).

Pooling Layers: They are downsampling operations, which shrink the spatial size of the feature maps without removing the most valuable information within them. The most well-known one is called max pooling and takes the largest value in a region to keep major features of it(Krizhevsky et al., 2012; Wikipedia contributors, -).

Fully Connected Layers: After all the activated features become extracted, a final fully connected layer can combine all these extracted features together to make predictions at the end of the network. These layers act in the same manner as conventional neural networks and make the final classification(LeCun et al., 1998).

The use of raw image inputs allows CNNs to gain exceptional performance in image classification, segmentation, and localization tasks(Chen et al., 2025; He et al., 2016; Sarvamangala and Kulkarni, 2021; Simonyan and Zisserman, 2015; Tajbakhsh et al., 2017).

CNNs are especially beneficial in the setting of CXRs classification such as:(a) as they allow models to learn features that are useful (e.g., lung opacity, consolidation, texture changes) without using hand-crafted preprocessing.(b) learn to generalize across a wide variety of imaging conditions (e.g., noise, brightness, anatomical variation).(c) provide end-to-end learning at a great-to-medium scale on big data(Chen et al., 2025; Jia et al., 2024; Kourounis et al., 2023; Mienye et al., 2025)

With such strengths in mind, CNNs like VGG and ResNet have become centrepieces of most state-of-the-art medical image pipelines, including those dedicated to diagnosing Pneumonia and COVID through CXRs(Chen et al., 2025; Kourounis et al., 2023; Sarvamangala and Kulkarni, 2021; Tajbakhsh et al., 2017).

3.2 VGG and ResNet Architectures

VGG (Visual Geometry Group) and ResNet (Residual Network) are two of the most common CNN architectures in medical imaging. These two have been shown to be very useful in the classification of images and are frequently employed as transfer learning baselines in medical imaging applications, such as in the interpretation of CXRs (He et al., 2016; Ikechukwu et al., 2021; Sarvamangala and Kulkarni, 2021; Simonyan and Zisserman, 2015; Vellandurai et al., 2023).

VGG Architecture

The architecture proposed by Simonyan and Zisserman (Simonyan and Zisserman, 2015), called VGG, is simple and has a consistent structure.It takes smaller 3x3 convolutional filters and makes stacks that are deep (up to 19 weight layers) to give more representational power to the network but holds complexity manageable. Those convolutional layers are punctuated with the ReLU activation function and have the intervening max-pooling layers to reduce spatial dimensions.

VGG and its major advantages in relation to medical image analysis are uniform structure:(1) Flexible to extend or add variations to complete other tasks.(2) Good feature detail: Saves small-scale spatial details that can be used to identify slight radiographic results.(3) Pretrained availability: There are pretrained versions readily available to be pretrained on ImageNet and, thus, very appropriate to apply them to transfer learning in low-data medical domains.

ResNet Architecture

The concept of residual learning to deal with the issue of the vanishing gradient in very deep neural networks developed by He(He et al., 2016). They also came up with the ResNet network. It allows you to train networks that are much deeper (like ResNet-50 and ResNet-101) by adding so-called "skip connections" that bring back layers that have been left out. This feature enables gradients to go through the network without stopping while the network is being recomputed. Instead of learning a mapping $H(x)$, ResNet is trained to learn a residual mapping $F(x) = H(x) - x$. This technique changes the difficulty of learning a mapping to $H(x) = F(x) + x$. This little idea speeds up convergence and makes deep networks better at generalizing. (for More info 5.5)(He et al., 2016)

The benefits of ResNet used in medical imaging for better deep architecture convergence, more efficient reuse of features, making transfer learning possible, robustness to overfitting, especially when trained on limited datasets such as those found in healthcare (He et al., 2016; Ikechukwu et al., 2021; Sarvamangala and Kulkarni, 2021; Vellandurai et al., 2023).

Use in This Thesis

The two base networks, VGG and ResNet, were implemented in the same research as a base model to classify Pneumonia and COVID based on CXRs. They were all lumped together into an ensemble model (A machine learning technique to collectively aggregate multiple models with a view to predicting a more accurate and dependable value than any particular model may do alone (Alotaibi, 2025; Dietterich, 2000; Z.-H. Zhou, 2012)) to utilize the capability of VGG to detect fine-grain texture and the ability of ResNet to leverage deep hierarchical abstraction. The ultimate label of classification was calculated as the mean of the results of both heads, and only the final several layers were adapted to the domain-specific training. This combination of the architectures enhanced performance and stability as compared to single architecture use.

3.3 Model Ensemble (VGG + ResNet)

Ensemble methods/models are very applicable in the case of deep learning in medical imaging, as they help to minimize variance and generalize better. It is particularly true in the case where the available amount of data is not large or data is noisy, as is the case with analysis of CXRs (Abad et al., 2024; Mohanty et al., 2024; Müller et al., 2022).

Why Ensemble?

The results of predictions given by one model could depend on the way it is configured, its training, or the balancing of data. In contrast, ensembles provide some advantages like,(a) Lead to improved classification accuracy,(b) as mistakes individual models are averaged out,(c) get characteristics that are easy to operate jointly in order that various models can learn various things off the same input data,(d) give more robust predictions where the variation is high, e.g., weakly labelled medical datasets(Alotaibi, 2025; Dietterich, 2000; Z.-H. Zhou, 2012). Marginal improvements in the precision or easy comprehension of the medical diagnostics will result in a better patient outcome(Nguyen et al., 2021).

VGG and the ResNet Ensemble

In this thesis, the ensemble model was designed by mixing the outputs of two pre-trained architectures: VGG and ResNet. These two models were developed with the chest X-ray data and were equally used to contribute to the final classification. Their outputs were logits of the final fully connected layers, averaged to yield final class probabilities.

This design help to Fine-grained feature detection of VGG that can detect local abnormalities in CXRs. The deep residual learning ResNet with the high level of abstract representation and global patterns of the lungs.

Benefits and Trade-offs

Ensemble learning can be used to attain better model performance but it has its own trade-offs. The limitation is the reduced speed and the requirement of more memory, as multiple models are applied at once (Khan et al., 2024; Müller et al., 2022). Also ensembles take more work to train and tune, especially when there is multiple models heads being matched or synced (Alotaibi, 2025; Nguyen et al., 2021). Regardless of these challenges, this additional complexity is often defensible, particularly in high-stakes areas such as healthcare, where better performance and interpretability can present a meaningful difference in clinical decision-making (Dietterich, 2000; Z.-H. Zhou, 2012). However, it is well worth the added complexity to improve the performance and to make interpretations in clinical contexts where the stakes are high, such as in healthcare(Dietterich, 2000; Z.-H. Zhou, 2012).

3.4 Transfer Learning

Overview

In deep learning, transfer learning is a common method in which a model that has been trained on one task is used to do an additional task that is similar enough to the first one to be effective. Transfer learning means that a model that has been trained on a big dataset, like ImageNet, is then fine-tuned on a smaller dataset that is particular to that domain, being able to thus keep general knowledge about images but adapt it to the new task (Raghu et al., 2019; Shin et al., 2016; Yosinski et al., 2014; Yu et al., 2022).

Why Medical Imaging Transfer Learning?

Medical data tends to have the problem of Scarce annotated data (e.g., x-rays with COVID or Pneumonia), expensive annotation process. Learning fundamental features (edges, shapes, textures) that are useful in different domains can be done by pretraining on large, diverse datasets, such as ImageNet. Medical images can be quickly converged and achieve superior performance using less labelled data using fine-tuning of such models (Salehi, Salehi et al., 2023; Shin et al., 2016; Yu et al., 2022).

Within the present thesis, it can be argued that transfer learning is essential in the following contexts such as:(a) decrease in training and computation cost.(b) retaining generalizable visual patterns during the acquisition of information about disease-specific appearances.

Strategy of Fine-Tuning

Using these models, pre-trained VGG-16 and ResNet-50 adopte as the basics in this work. To use transfer learning, the following strategy implies; (a) Frozen low-level layers: A majority of convolutional layers were kept frozen to maintain low-level feature extraction.(b) Models finer layers: The last several convolutional blocks were unscheduled to adapt to the characteristics of CXRs.(c) Swapped classification heads: The initial fully connected layers were swapped out and substituted by new layers to work with the input data of the three target classes: Normal, Pneumonia, and COVID.

Strengths and Weaknesses

The strength of Transfer learnings are (a) Data-efficient: Does not require large datasets.(b) computationally efficient: It does not take much training time.(c)

Avoids overfitting(Raghu et al., 2019; Salehi, Salehi et al., 2023; Yosinski et al., 2014; Yu et al., 2022).

The main **Challenge** of transfer learning is domain shift, where The pretrained models do not necessarily generalize to medical imagery optimally,also some time Architecture mismatch.

Despite this, transfer learning is likely to be among the most viable and sensible approaches to the real-world handling of deep learning in healthcare(Salehi, Salehi et al., 2023; Yu et al., 2022).

3.5 Vision Transformers (ViT)

ViT were first used in the applications of natural language processing (Vaswani et al., 2017), where they changed many operations such as machine translation or text summarization. The self-attention mechanism is the most important innovation of these researchers, as it enables the model to prioritize the significance of the various components of an input sequence on a parallel basis. Transformers can model global dependencies beginning at the first layer of the network, unlike CNNs, which aim at local patterns via convolution.

ViT developed by Dosovitskiy, applies the transformer architecture to the task of image classification, where, instead of performing the convolutional operations, the image is represented as a sequence of flattened image patches(Dosovitskiy et al., 2020). ViT use Specifically for the image is sliced up into segments of a set size, such as 16x16 pixels, Each patch is flattened and linearly embedded, position embeddings were introduced to keep the spatial information.

ViTs were excellent at both worldly and local relationships in the full image, so they are promising in cases concerning medical images with minimal spots extend out(Aburass et al., 2025; Azad et al., 2023).

ViT's Advantages and Disadvantages in Medical Imaging

Advantages:(a) ViT is Good at detecting intricate and obscure pathologies (such as interstitial lung disease),Unlike CNNs.(b) ViTs do not need preset spatial relationships and can learn them on the data, which can be applied to specific tasks more generally(Aburass et al., 2025; Azad et al., 2023).

Disadvantages:(a) ViTs need enormous sets of data to train successfully.(b) They do not perform well in a situation where data is insufficient and no pretraining is done when compared to CNNs. (Azad et al., 2023; Preprint, 2023).(c) Transformers generally require more memory and computational resources.

3.6 Weakly Supervised Learning

Definition and Motivation

Weakly supervised learning takes place when you train models using annotations that aren't accurate, comprehensive, or precise. Weak supervision, on the other hand, used labels that are easier to obtain, such as image-level class annotations, instead of complex ground-truth labels like bounding boxes or segmentation masks(Misera et al., 2024; B. Zhou et al., 2016).

Annotations that are more detailed are cost-intensive, highly time-consuming, and expert-requiring in the sphere of medical imaging, especially radiology. It would be impractical to identify the precise outline of lung infection on each X-ray in the case of large volumes. Hence, weakly supervised learning has become a viable and scalable solution.(Misera et al., 2024; Rajpurkar et al., 2017).

Use in Medical Imaging

When speaking of CXRs Fully supervised models need region of interest (ROI) or pixel-level annotations in the form of masks (B. Zhou et al., 2016),Weakly supervised models are trained to classify disease, as well as localize it, based on the diagnosis label provided to the image (i.e., Pneumonia or normal) (Kim et al., 2021; Rajpurkar et al., 2017). Its objective is to draw spatial localization (e.g., to indicate infected areas in the lung) without exposure to any spatial labels in the training process(Misera et al., 2024; B. Zhou et al., 2016).

This method is very much like the practice in real-life clinical settings, when only diagnostic labels may be noted in the patient databases.(Di Noto et al., 2022; Misera et al., 2024).

Weak Supervision in This Thesis

In the current thesis, the model is trained with only image-level labels (i.e., Normal, Pneumonia, COVID) but lacking coordinate labels or segmentation labels.

Localization is carried out with the following methods such as: Grad-CAM, lung segmentation masks used post-hoc as a method of constraining Grad-CAM outputs to areas of the lung whose segmentation is deemed to be anatomically sound (Misera et al., 2024).

This kind of setup is like a real-life clinical setting where no use of manually labelled bounding boxes (Rajpurkar et al., 2017).Visual interpretability is directed by model attention + anatomical prior (the lung mask) (Misera et al., 2024; Pan and Chen, 2023).The localization is indirectly trained in the form of classification loss using backpropagation (B. Zhou et al., 2016).

Benefits:(a) Very few annotations are required and scalable to large data.(b) Simulates actual hospital records (that would show only diagnoses)(Kim et al., 2021; Misera et al., 2024; Rajpurkar et al., 2017).

Limitations:(a) The localization is not exact as compared to the fully supervised method.(b) Heatmaps can also underscore areas of no interest without instructions (e.g., areas beyond lungs).(c) Needs some form of extra processing (such as lung segmentation) to make it interpretable.(Misera et al., 2024; B. Zhou et al., 2016)

3.7 Grad-CAM

Motivation for Interpretability in AI

In life-vital areas such as healthcare, one of the factors that contribute to gaining more trust in the AI-based decision-making system is interpretability (Reyes et al., 2020). Doctors should not only be aware of what a model will tell them to expect, but they must also know why it makes such a projection. To fulfil this requirement, there are now a lot of explainable AI (XAI) techniques out now. For example, Grad-CAM is one of the most common ways to show how deep convolutional networks pay attention to things (Selvaraju et al., 2017b).

How Grad-CAM Works

Grad-CAM (Selvaraju et al., 2017b) is a post-hoc interpretability technique, which provides a visual description of those parts of an input image that most contribute to the model decision.

The technique works by:

1. Doing a forward pass of the model to get the prediction of the class.
2. The calculation of the gradient of the class score predicted against feature maps of a selected convolutional layer.
3. Using the spatial averaging of gradients (global average pooling) to extract weights.
4. Carrying out a weighted summation of the feature maps with these gradients.
5. Using ReLU as a way of removing non-useful features that have a negative effect on the output.
6. Upsampling the resulting heatmap to create one that is the same size as the original image and overlapping it to visualize.

The produced class-discriminative heatmap shows areas that were most pertinent to the decision in the model selected—none of which involve alteration of the network architecture or training procedure(Panwar et al., 2020; Selvaraju et al., 2017b).

Application in This Thesis

Grad-CAM was used in this study on the head of the dual classification scheme (ResNet + VGG) and the ResNet head of the dual-head classification model (ResNet + VGG). Following the last convolutional block of ResNet, backpropagation of gradients was performed to come up with Normal, Pneumonia, and COVID-based class-specific heatmaps, Grad-CAM heatmaps were then masked based on lung segmentation output to enhance clinical relevance, that is, to make sure that the highlighted area was localized to the anatomical lung region, rather than leaking to other irrelevant areas (e.g., bones, edges, background).

This enabled the disease areas where localization was weak and no bounding boxes or segmentation masks were used in the training.

3.8 Unsupervised Lung Segmentation

Role of Segmentation in Medical Imaging

Image segmentation describes the problem of dividing an image into semantically relevant regions, usually defining anatomical structures or abnormalities (Candemir et al., 2013).
Lung segmentation would be of special importance in the interpretation of CXR during:(a) The process of segmenting areas of interest, specifically the left and right lung fields, is crucial.(b) Imaging out irrelevant visual artifacts (e.g., ribs, background, edges).(c) Improving the interpretability of the model by ensuring the nature of the concentration to be clinically valid (Mansoor et al., 2015).

In tandem with Grad-CAM, segmentation makes visual explanations anatomically restricted, which increases the level of reliability among radiologists (Candemir et al., 2013; Selvaraju et al., 2017b) .

Supervised vs. Unsupervised Segmentation

The conventional segmentation models are supervised, which involves mask drawing by human experts. Such masks are not always available in the real-world medical datasets, as they require expert effort to label, which is both expensive and time-consuming. Unsupervised segmentation avoids the costs of labelling with annotated masks and uses lung boundaries as approximations of processing

images and heuristics. That is why it is the perfect solution for weakly supervised learning pipelines, where no ground-truth masks are presented(Candemir et al., 2013; Mansoor et al., 2015).

Method Used in This Thesis

In this work, a lightweight unsupervised lung segmentation pipeline was implemented, inspired by classical computer vision techniques (Candemir et al., 2013):

Preprocessing: First,change the image into black/white level.Then, use contrast enhancement CLAHE (Contrast Limited Adaptive Histogram Equalisation) and apply Gaussian blur to decrease the noise.

Thresholding and mask: Use adaptive and Otsu thresholds to create binary masks and apply morphological operations (closing, opening) to clean the binary masks.

Contour Detection: Firstly ,Find the contours on the binary image,then filter out area, aspect ratio, and location to identify lung-like structures then Choose the two largest regions (purportedly the left and the right lungs).

Final Mask Construction: Draw outlines on a blank mask and small holes can be filled and spurious small objects discarded using connected components.

Unsupervised mask was applied to filter Grad-CAM heatmaps, and only the signals were kept within the lungs to be visualised (Mansoor et al., 2015; Selvaraju et al., 2017b).

Advantages and Limitations of Unsupervised Learning:

Advantages (Mansoor et al., 2015):(a) There is no requirement for labelled segmentation masks, which makes it usable with any public CXR dataset.(b) Fast and lightweight; does not involve the usage of neural networks or GPU inference.(c) Good for weak localization and Grad-CAM integration (Selvaraju et al., 2017b).

Limitations(Mansoor et al., 2015) :(a) Not as precise as the segmentation models, which are based in deep learning.(b) May not work on low-fidelity photographs, or lung edges are confused.(c) Sensitive to threshold settings and light and dark alteration of the images.

However, regarding poorly supervised learning, unsupervised segmentation is a convenient instrument when conducting interpretability and localizing the areas of interest in the case of disease(Candemir et al., 2013; Mansoor et al., 2015).

3.9 Evaluation Metrics

Importance of Evaluation in Medical AI:

Assessing the performance of the AI model in healthcare needs to be more than accuracy. Because the consequences of false positives or false negatives may be life-threatening, models should be evaluated on clinically meaningful grounds (Baldi et al., 2000; Chicco and Jurman, 2020; Esteva et al., 2017). This segment describes the conventional metrics of assessing the classification and localization performance adopted in this thesis.

Classification Metrics:

For the multi-class classification task (Normal,COVID, Pneumonia), the following metrics were used(Baldi et al., 2000; Chicco and Jurman, 2020; Powers, 2020):

- **Confusion Matrix:**A table showing counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). It provides detailed insight into per-class performance.

	Predicted	
	Positive	Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

- **Accuracy:** The proportion of correctly predicted samples among all samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** The proportion of true positives among all predicted positives. High precision implies fewer false alarms.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity / True Positive Rate):** The proportion of true positives identified out of all actual positives. It is crucial in healthcare to prevent missed diagnoses.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall. Useful when there is class imbalance, which is common in medical datasets.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Specificity (True Negative Rate):** The proportion of true negatives correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Localization Evaluation (Grad-CAM):

In spite of the fact that localization was done under weak supervision (no pixel-level ground truth), qualitative assessment was done with the following Visual inspection of Grad-CAM heatmaps, in order to confirm that highlighted areas corresponded to expected distributions of lung infections. Masked Grad-CAM overlays, where heatmaps were filtered using lung segmentation to verify anatomical relevance.(Reyes et al., 2020; Selvaraju et al., 2017b)

Although no standard measurement tool such as IoU (Intersection over Union) was used since no sign of annotated back-to-back bounding boxes was available, the localization performance was presented per clinical interpretability and anatomical stability (Reyes et al., 2020).

Summary

The combination of numerical measures (like accuracy, precision, and recall) and visual tools (like Grad-CAM overlays) used in this thesis shows that the AI model is not only successful in numbers but also understandable in a clinical setting and suitable for real-world diagnosis (Esteva et al., 2017; Reyes et al., 2020).

3.10 Frontend and Deployment

Motivation for Deployability

When it comes to medical AI, accuracy is not the only aspect of relevance. The first criterion is that the system has to be available, understandable, and easily embedded into the current workflow to enable clinicians to gain access to AI models. Such a thesis will thus feature not just backend model development but also frontend interface creation and a containerized deployment environment so that it is easy to interact with and repeatable.

Streamlit-based Interface

In order to provide interactive visualization and use the trained model, a light frontend was developed based on Streamlit Python library, which allows creating data-driven web applications with minimal code complexity.

The highlights of the interface are:(1) Upload image: The user can upload a chest X-ray image such that it can be analyzed in real-time.(2) Prediction display: The promised type (Normal, COVID or Pneumonia) is displayed as well as the confidence of a model.(3) Grad-CAM heatmap: Users will be able to check the areas highlighted in the overlay to the original X-ray to interpret how the model came to its conclusion.(4) Lung segmentation overlay: This makes focus anatomically acceptable and increases clinical interpretability.

Such an interface can assist in filling the gap between AI developers and the medical professionals since results are openly explained.

Containerized Deployment Using Docker

The system was designed to be containerized, i.e., carried out with Docker to guarantee that it can be transported and reproduced. Docker bundles the whole application, which includes:(a) Python environment and dependencies (PyTorch, OpenCV, Streamlit, and so on).(b) Model weights were trained.(c) Preprocessing logic and visualizing logic.

Advantages of Docker:(a) Cool to run the system on any machine without dependency differences.(b) Facilitates possible introduction in hospitals or cloud support.(c) Allows version control and future scalability of a model update.

Summary

The fact that the integration of Streamlit and Docker enables the system both to be technically reasonable and to ensure that the system can be used by non-tech stakeholders in practice, where a radiologist or a worker of any public healthcare institution will be able to use the solution. This will help to achieve one of the main objectives of the thesis of bringing closer to AI tools that are accurate but, more importantly, deployable, interpretable, and clinically applicable.

4 Requirements

This chapter presents system requirements that will be used to generate a deep learning-based diagnostic instrument that will be applied in the classification and localization of COVID and Pneumonia using CXRs and image-level annotations. Such requirements include functional capabilities, technical requirements, condition of the dataset, preprocessing, and system limitations. Due to the clinical importance of the issue and the insufficiency of the data available, it is necessary to develop an AI system that can adhere to a weakly supervised and interpretable strategy (Baltruschat et al., 2019; Çallı et al., 2021).

4.1 Problem Context

Deep learning has come forward with a chance of using medical images, mostly CXRs due to their easy access , and also because they are relevant(in the context of diagnosis (Baltruschat et al., 2019; Çallı et al., 2021; Esteva et al., 2017). The question arises in the field of how to achieve a high level of accuracy and explainability of AI systems and, at the same time, be able to work with a minimal amount of annotated data. When compared to the segmentation-based tasks, the system to be developed should process the classification/weak localization task only with image-level labels, as the location of the lesion delineation is not required here.

By doing this, the system would Clean up noisy and high-artifact process medical images, Perform effective classification into 3 classes: Normal, Pneumonia, and COVID; Generate outputs that can be declared valid since they can be understood by humans (Reyes et al., 2020; Selvaraju et al., 2017b),trained with the constraint of the absence of a ground-truth bounding box or separated element segmentation mask (Robb et al., 2021).

4.2 Functional Requirements

The following are the functional goals of the proposed system:

1. **Multi-class Classification:** Input as CXR image and Output as Type of classification to be made: Normal, Pneumonia, or COVID
2. **Weak Localization:** Localization without ground truth coordinates, Grad-CAM, or other attention maps will find the abnormal locations; However, attention limits the maps to stimulate regions within the lung areas exclusively, and any failure of the same would result in a false activation.
3. **Unsupervised Lung Segmentation:** Segment lungs using image-processing procedure (threshold, morphological filtering)(Candemir et al., 2013; Mansoor et al., 2015) and Use lung masks on Grad-CAM maps (Selvaraju et al., 2017b)
4. **Image Cleaning:** Text labels and clinical annotations, and other artifacts removal by means of OCR + inpainting (Channin et al., 2012; Sardar, 2023; Sharma et al., 2017)
5. **Explainability:** Use Grad-CAM heat maps to determine the area of decision-making (Selvaraju et al., 2017b), Lung regions are provided with Grad-CAM masks to perform localized interpretation (Candemir et al., 2013).
6. **Symptom-Aware Decision Fusion:** The system also accepts user-reported symptoms through a structured questionnaire (supporting both English and German) for important indicators such as fever, cough, shortness of breath, fatigue, and a loss of smell and some more. A symptom score is calculated using the premade clinical weights of each disease group (COVID, Pneumonia).The system combines this symptom-based probability with the image-based model response to provide a weighted equation:

$$\text{Final Score} = 0.6 \times \text{Image-Based Score} + 0.4 \times \text{Symptom Score}$$

   Once the prediction is completed, the system is developed to automatically create a PDF report containing results of the model, Grad-CAM heatmaps, and visualizations to segment lungs. The user will be able to download the report as a PDF file under the "Download Report (PDF)" button offered by the Streamlit interface that could serve as a clinical or personal reference.
7. **Streamlit UI Integration:** Provide outputs through an easy front-end view to be visually checked

4.3 Technical Requirements

1. **Model Architecture:** Take pretrained architectures (ResNet, VGG) & transfer learning, Combine multiple backbones (e.g., ensemble ResNet + VGG) for improved robustness.

2. **Hardware Compatibility:** Trained to fit in CPU configurations with limited memory. It is portable, allowing it to operate in a single-GPU environment
3. **Training Optimization:** Cross-entropy loss with label smoothing and class weights, Learning rate scheduler: ReduceLROnPlateau, Optimizer: AdamW, Early stopping based on validation loss.
4. **Evaluation Metrics:** Classification performance: Accuracy, Precision, Recall, Sensitivity, Specificity, F1-score (Baldi et al., 2000; Esteva et al., 2017), Localization qualitative assessment on Grad-CAM overlay maps.
5. **Tools and Libraries:** Python, PyTorch, OpenCV, Scikit-Image, Streamlit and Docker support for deployment.

4.4 Dataset Considerations

Source: COVID, Pneumonia, and Normal Chest X-ray PA Dataset (Asraf, Amanullah; Islam, Zabirul , "COVID, Pneumonia, and Normal Chest X-ray PA Dataset," Asraf and Islam, 2021)

Only image-level labels; no segmentation masks or bounding boxes(No Annotations),Balanced classes. Here for using this data Preproceses is require like: Resize CXR images in to 224x224 pixel, Normalization with ImageNet statistics.

4.5 System Integration Overview

The system is designed to function effectively in annotation-scarce medical environments, mirroring real-world clinical workflows. It integrates multiple components, including data preprocessing, dual CNN architecture (VGG + ResNet), Grad-CAM explainability, and unsupervised lung segmentation.

A novel addition is the symptom-aware decision fusion module, which enhances diagnostic robustness by incorporating user-reported symptoms alongside image-based predictions. This hybrid scoring approach increases clinical relevance, particularly in borderline imaging cases where symptoms offer additional context.

Together, these modules create a lightweight yet reliable diagnostic assistant for CXRs interpretation, balancing interpretability, deployability, and performance.

5 Architectures

This chapter outlines the architectural layout of the suggested deep learning pipeline for diagnosing CXR images in three categories: normal, COVID, and Pneumonia. The architecture is based on a combination of the backbones (using CNN) (ResNet-18/50 and VGG-16/19), an ensemble fusion, Transfer Learning and also an optional Vision Transformer (ViT). Besides this, components of preprocessing, including artifact removal, lung segmentation, and visual explainability with Grad-CAM, are also incorporated into the system to enhance its performance and interpolatability.

5.1 System Design Overview

The system is modular, comprising five main stages:

1. Artifact Removal to clean overprinted text and radiographic noise
2. Lung Region Segmentation to separate anatomical regions of interest.
3. Backbone Classification Models, CNNs, and Transformer-Based
4. Ensemble Fusion Layer that would collect predictions made by various models.
5. Grad-CAM-Based Explainability for visualizing significant features within the lung field

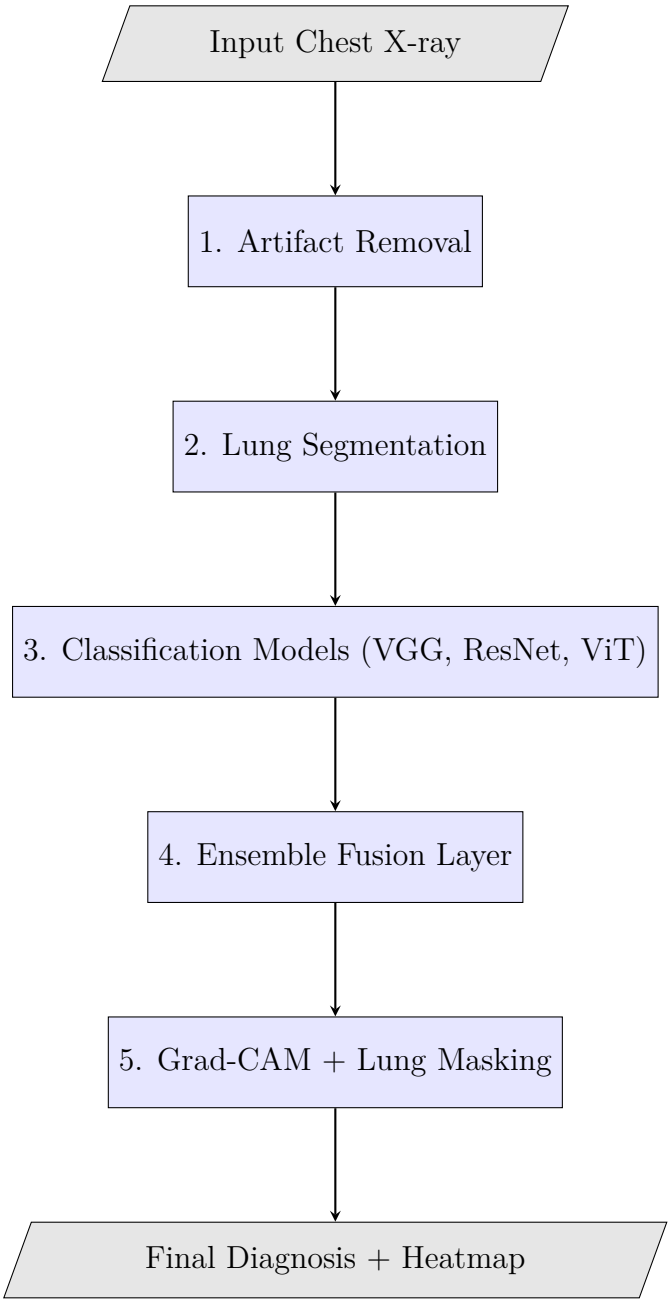


Figure 5.1: Full System Architecture – from preprocessing to prediction output, including segmentation and Grad-CAM visualization.

5.2 Preprocessing and Input Standardization

5.2.1 Artifact Removal

In order to clean raw CXRs, two complementary approaches are utilized:(1) Rule-based Filtering: Identifies and eliminates corner-based text and straight artifacts through morphological operations and intensity levels.(Aleksandr et al., 2019).(2) OCR-based Inpainting: Identifies the text overlays using the Tesseract OCR and cleans this text through inpainting. (Hogeweg et al., 2013)

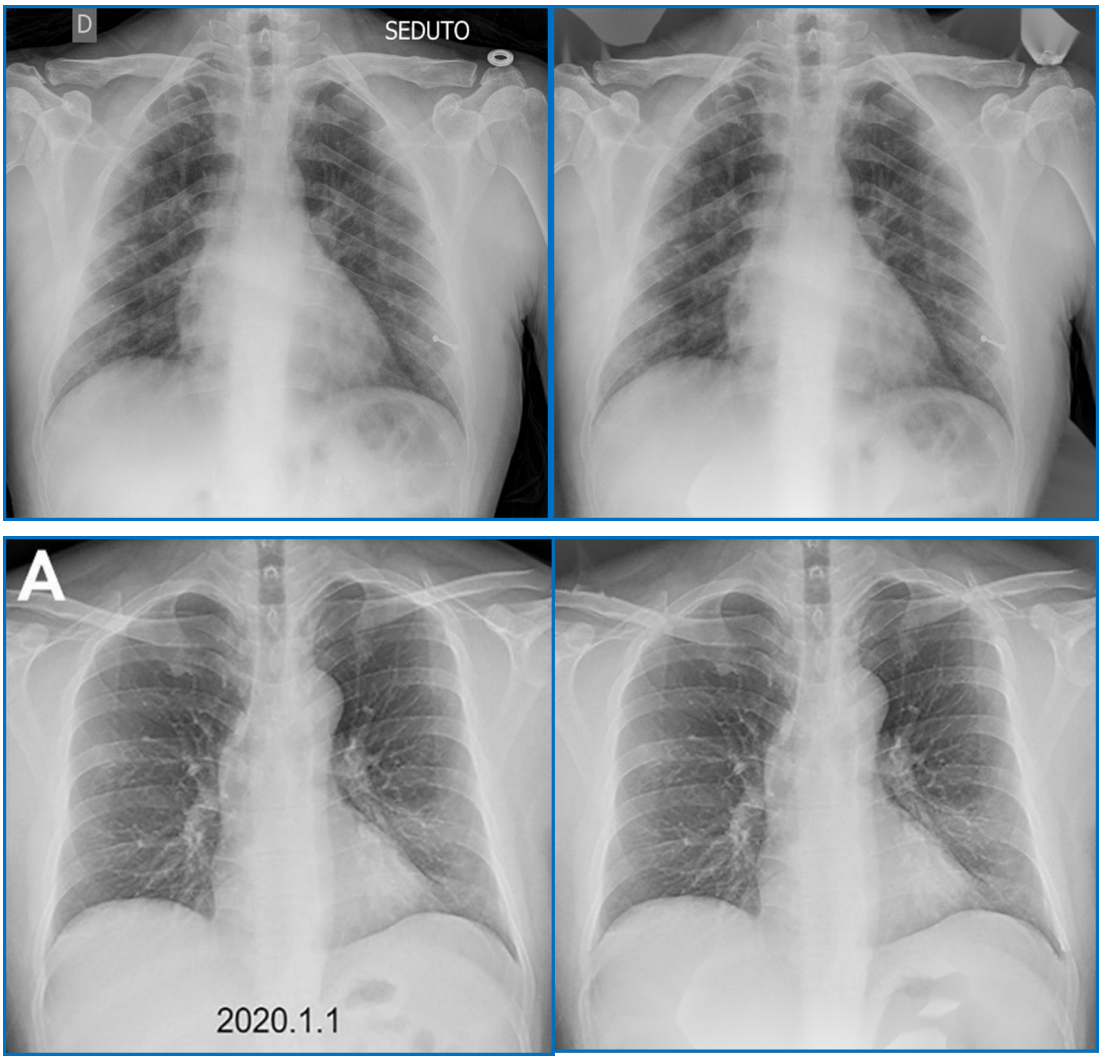


Figure 5.2: Original vs. Cleaned X-ray Image: Top and bottom rows show different examples illustrating the artifact removal process.

5.2.2 Lung Segmentation

The generation of lung masks is through the contrast enhancement using CLAHE, adaptive thresholding and Otsu thresholding,Morphological filtering and geometric constraints (aspect ratio, area, etc.). The final lung mask to apply with the Grad-CAM and optional overlays of visualizations.

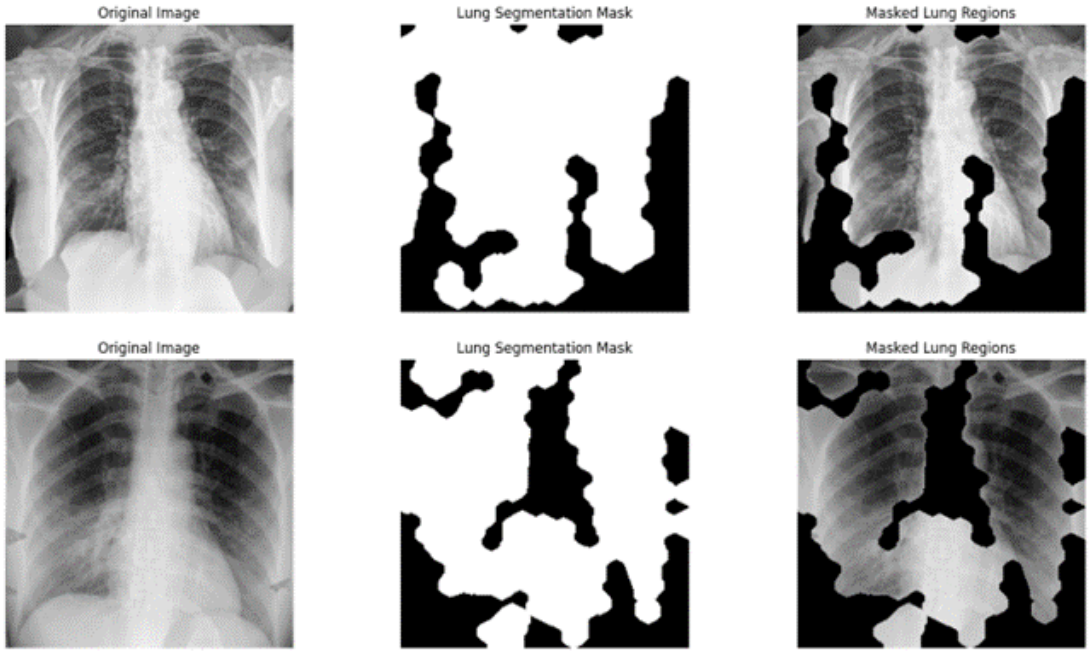


Figure 5.3: Lung Mask Extraction—Original, Mask, and Overlay: Two examples showing the stages of lung segmentation.

5.2.3 Image Normalization and Augmentation

The dimensions of the images are scaled to 224x224 pixels and rebuilt in RGB format after normalization to ImageNet statistics. Augmentation included random flipping, rotation, colour jittering, and centre cropping at training time.

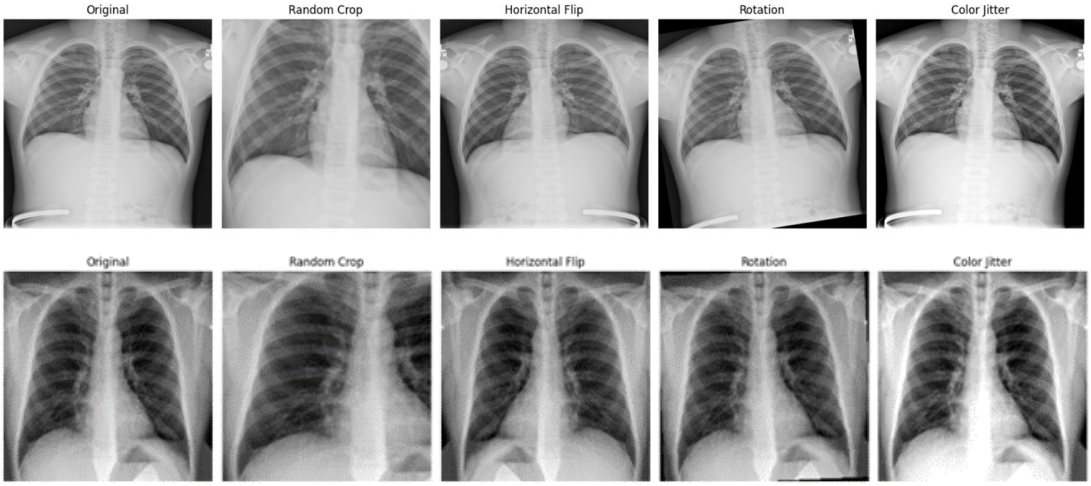


Figure 5.4: Image Augmentation for Model Training: Two examples demonstrating various augmentation techniques such as flipping, rotation, and cropping.

5.3 Backbone Model Architectures

5.3.1 ResNet-18 and ResNet-50 (He et al., 2015a)

ResNet is a kind of deep CNN consisting of deep network architecture, which allows training very deep models by using the residual learning core block. In a residual block, the network is trained to complete a residual mapping that is not a direct transformation. A skip connection (identity mapping) feeds forward the input to an intermediate layer to its output.

Diagram Elements: The diagram has an input tensor x,It consists of a series of rounds that are a convolutional block through which this information is transmitted. The layer in this block consists of a convolution, a batch normalization, and a ReLU activation function, followed by another convolution and batch normalization layer. The whole block is marked as $F(x)$. A skip connection is one that also adds the input directly. A skip connection is one that also adds the input directly as $F(x) + x$.

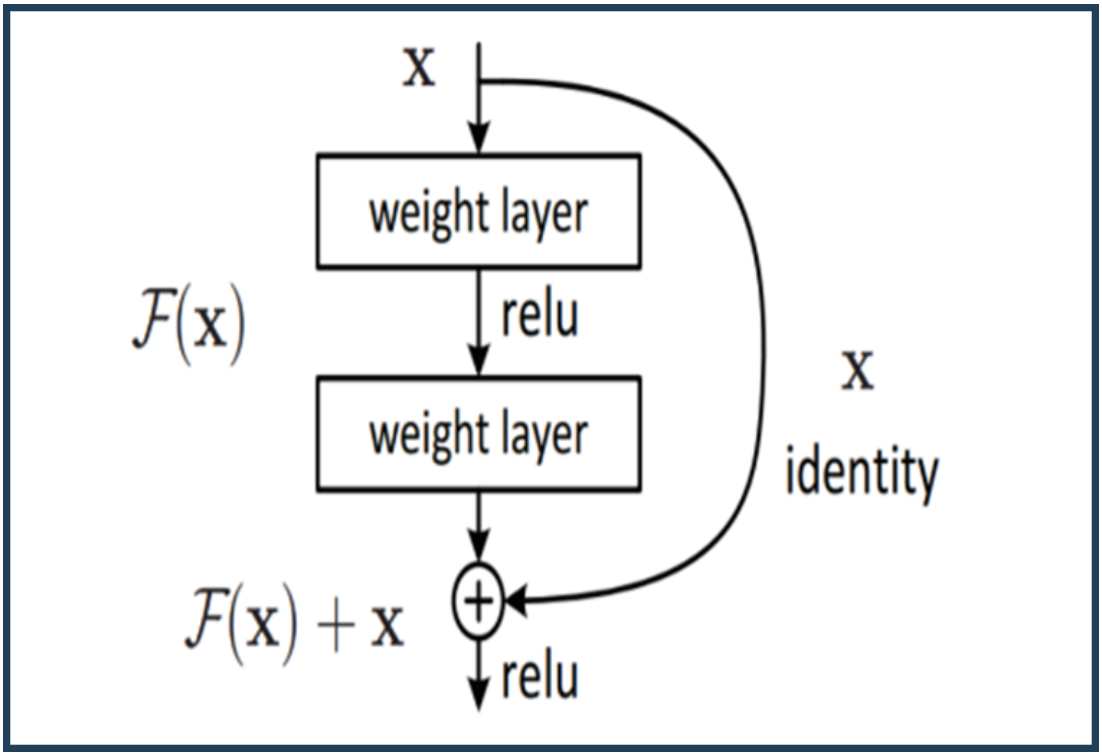


Figure 5.5: A residual learning block (Vligade, 2020)

Two variants are used.First, **ResNet-18** a Lightweight model with 18 layers and Second, **ResNet-50** Deeper model with bottleneck blocks, good for deep feature representations.

Differences between BasicBlock and BottleneckBlock that BasicBlock has Two 3x3 conv layers and Bottleneck has 1x1 – 3x3 – 1x1 convolutions (figure 5.7).

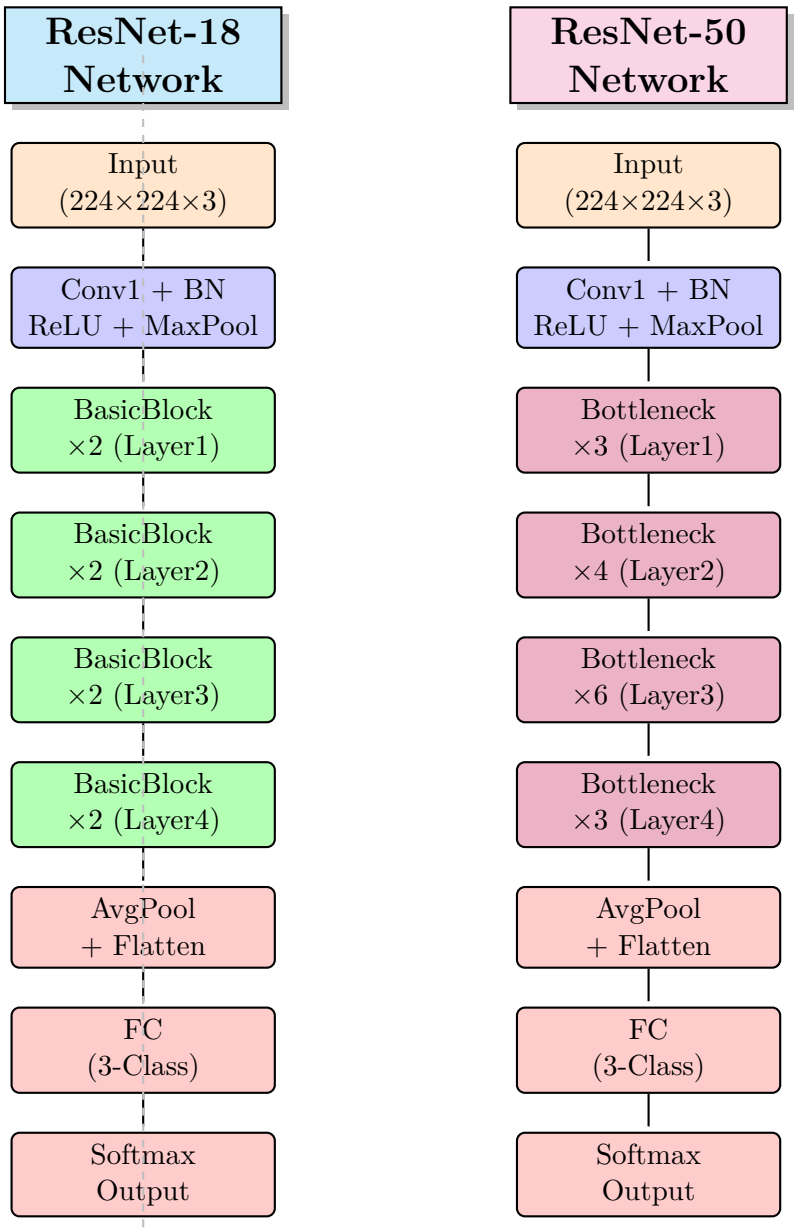


Figure 5.6: Architecture comparison between ResNet-18 and ResNet-50 for chest X-ray classification.
BasicBlock (2-layer) vs. Bottleneck (3-layer) structures highlighted.

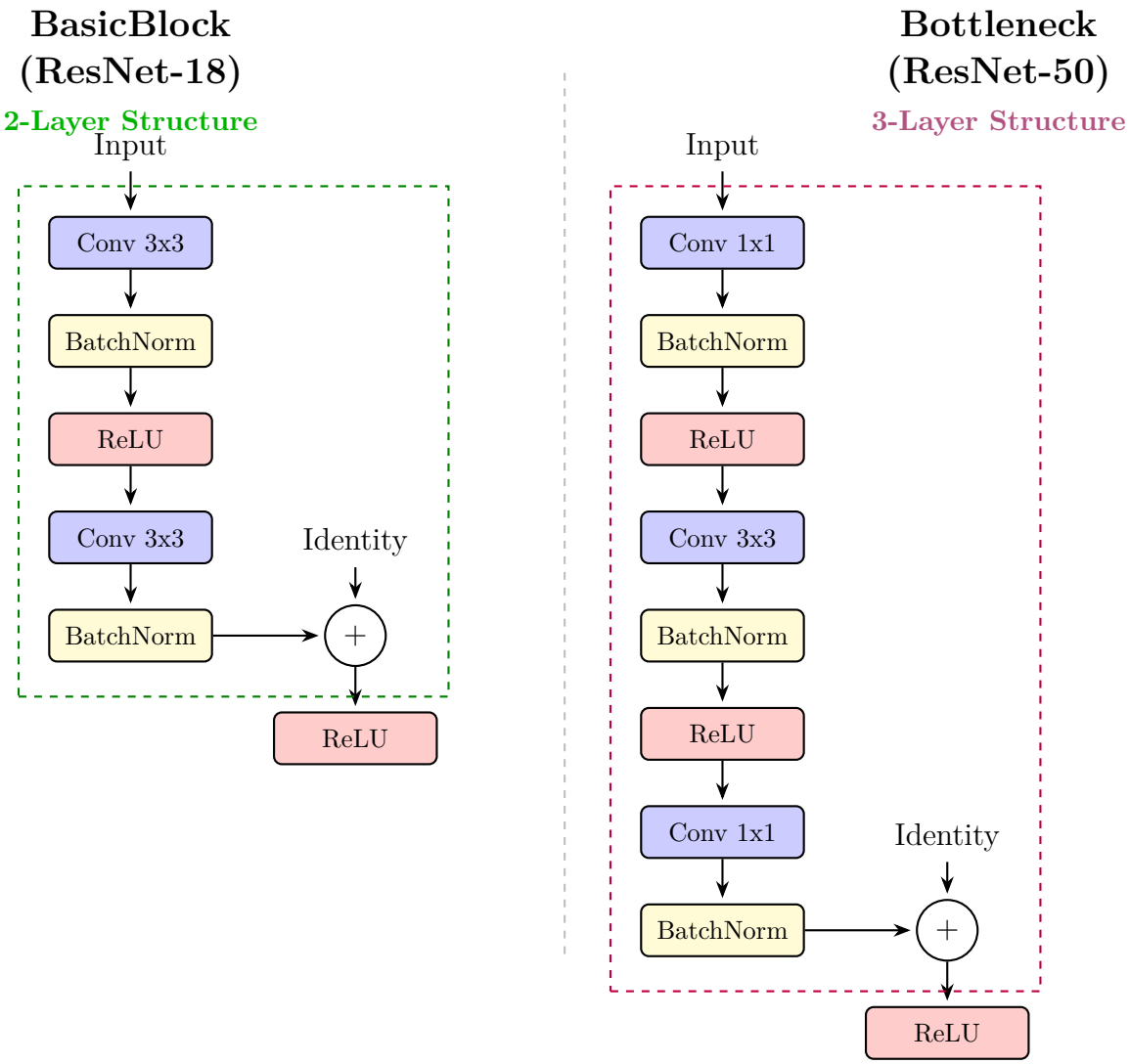


Figure 5.7: Detailed comparison of BasicBlock (ResNet-18) and Bottleneck (ResNet-50) structures.

5.3.2 VGG-16 and VGG-19 (Simonyan and Zisserman, 2014)

VGG models use stacked 3x3 convolution layers followed by fully connected classifiers.VGG-16 has 13 convlution + 3 FC layers and VGG-19 has deeper version which was Pretrained on ImageNet. Final classifier modified to output 3 classes. Last convolutional block and classifier head adapted; Earlier layers frozen of VGG-16. Grad-CAM visualizations derived from final convolutional block.

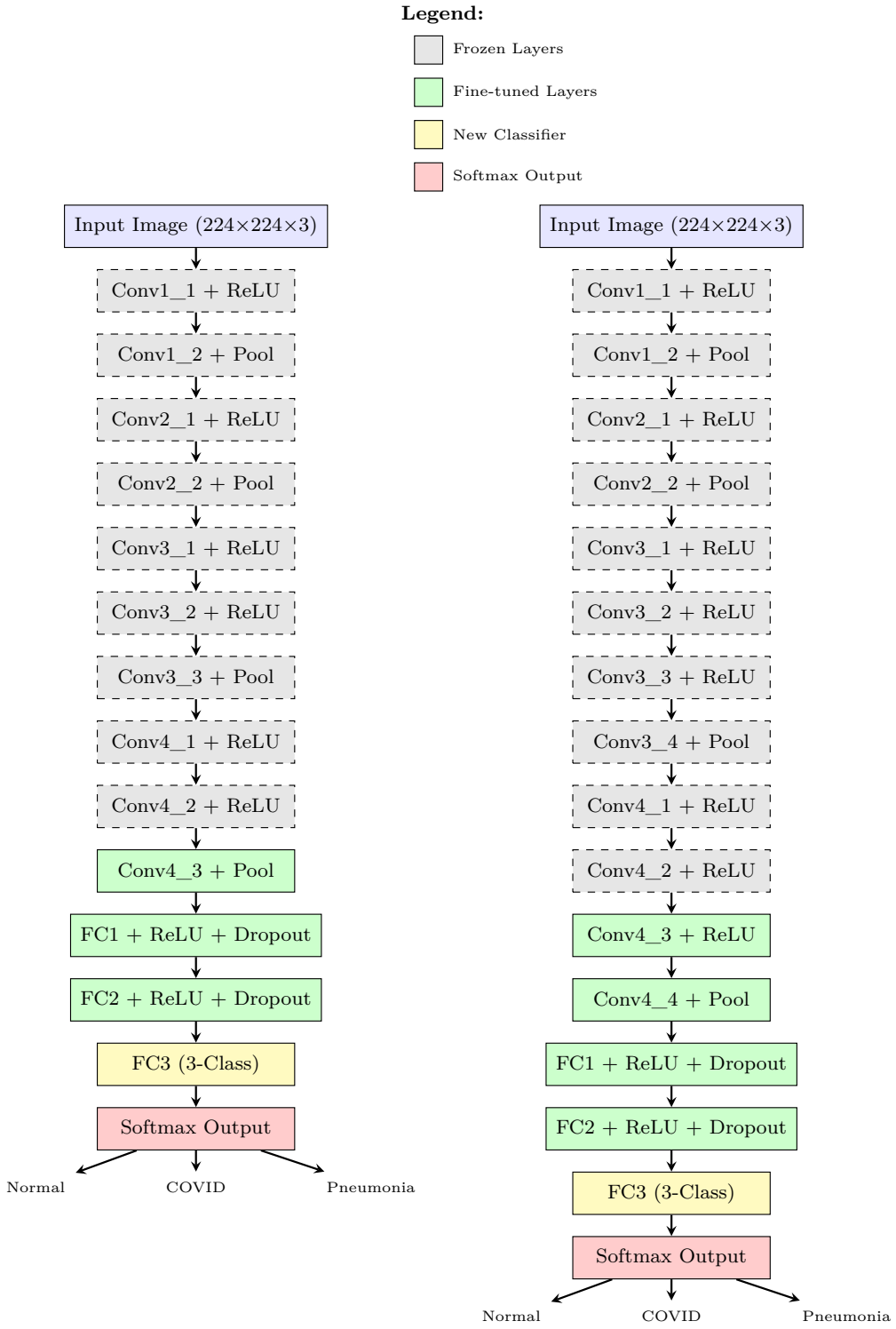


Figure 5.8: Side-by-side comparison of VGG-16 and VGG-19 architectures showing frozen layers, fine-tuned blocks, and classifier output for CXR classification. Output shows three-class classification (Normal, COVID, Pneumonia).

5.3.3 Vision Transformer (ViT) – Experimental

(Azad et al., 2023; Dosovitskiy et al., 2020; Henry et al., 2022; Preprint, 2023) ViT uses transformer-based self-attention, applied to flattened image patches. Adopted `vit_base_patch16_224` model from timm library. Accepts 224x224 RGB images, split into 196 patches of 16x16. Trained using ImageNet weights, classification head modified.

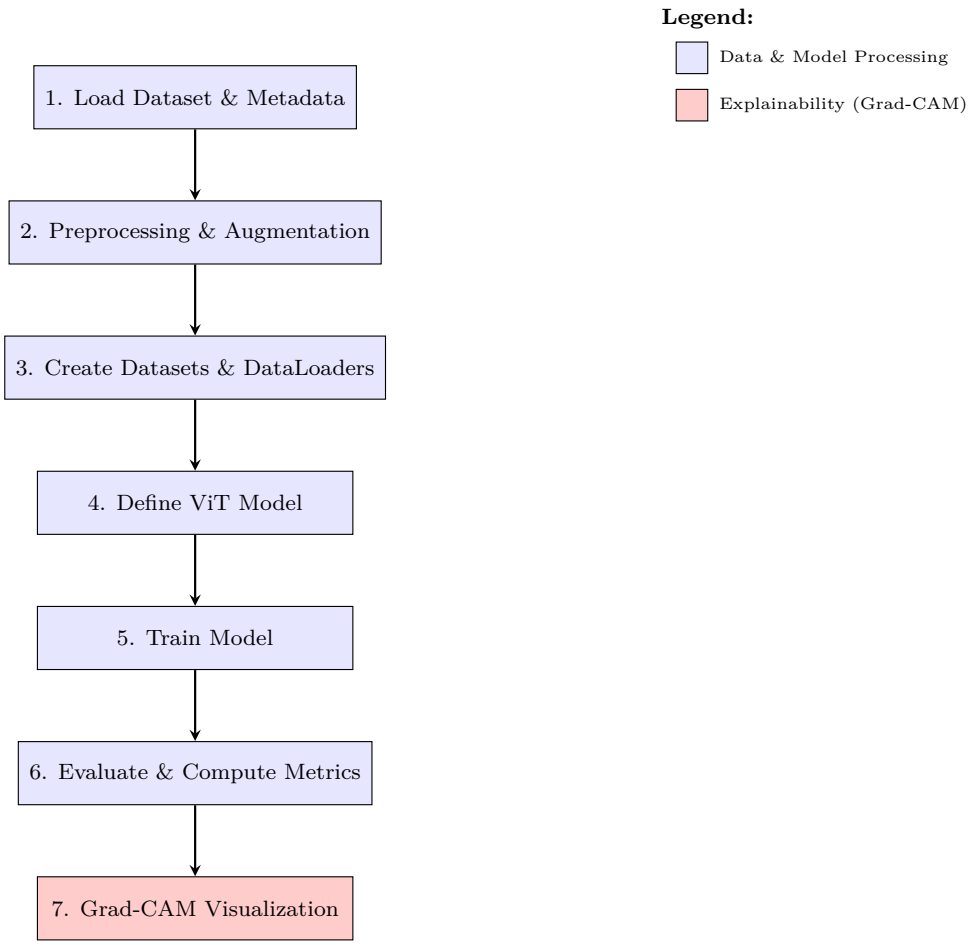


Figure 5.9: Step-by-step pipeline for chest X-ray classification using Vision Transformer (ViT) and Grad-CAM visualization. Blue boxes indicate data and model processing steps. Red box indicates explainability phase.

5.4 Transfer Learning with ResNet-50

The ResNet-50 architecture was used as a model backbone to support a transfer learning approach to support the development of a strong and effective process to classify CXRs. ResNet-50 has a deep residual architecture and drives more stable

optimization at greater depth. It has the advantage of being mostly successful in the natural and medical image analysis as well as analysis(He et al., 2015a; Suganyadevi et al., 2024).

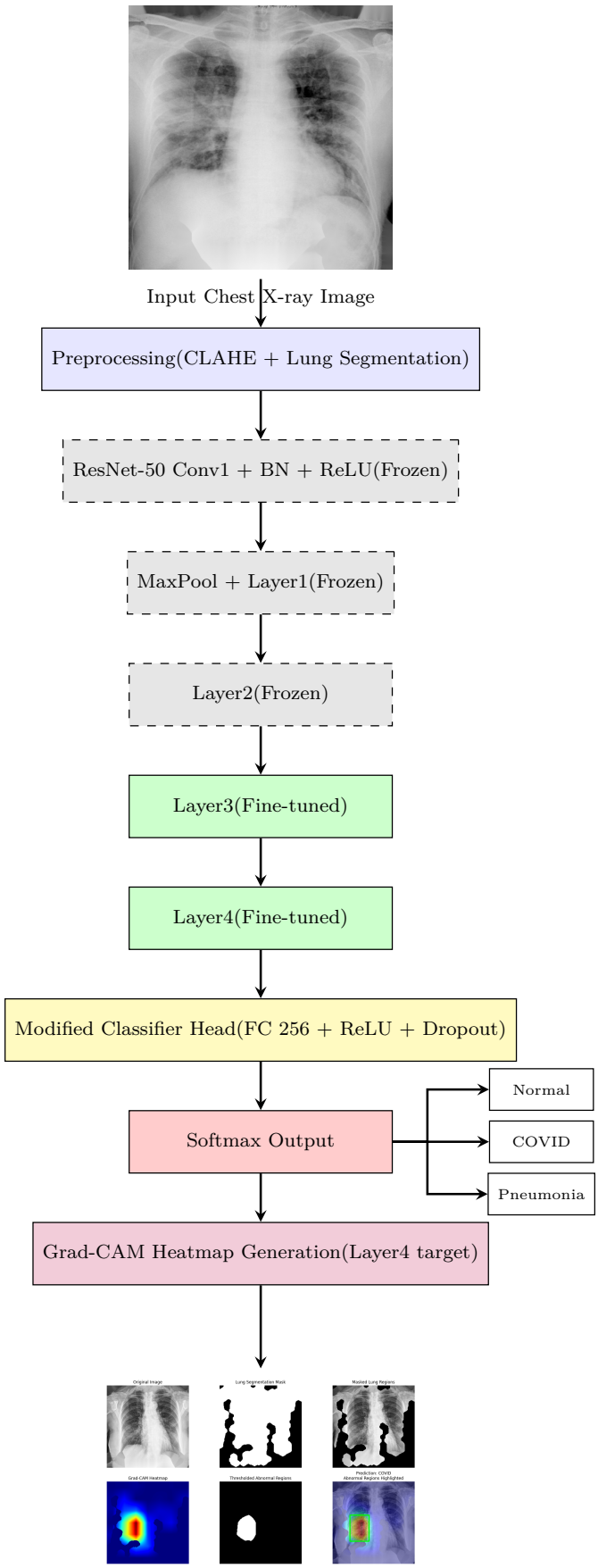
Architecture Adaptation

For feature extraction, a pretrained ResNet-50 (from ImageNet) was used. It was available as a fully connected classification head, replaced with a user-defined block:

$$\text{FC}(2048 \to 256) \to \text{ReLU} \to \text{Dropout}(0.4) \to \text{FC}(256 \to 3)$$

Training was only allowed on the deeper layers (layer2, layer3, layer4, and fc), but the earlier blocks of convolution remained frozen to maintain an overall feature representation. Through this type of fine-tuning, the model managed to learn domain-specific features (i.e., lung textures) intelligently without losing the bigger picture of visual knowledge it acquired on ImageNet.

Advantages of the Transfer Approach:(a) Less training time with the update on partial layers.(b) Better knowledge transfer in small medical data sets.(c) Good convergence with minimal epochs improved over full model training.(d) Explainable: Grad-CAM was conveniently plugged into the end convolutional block (see Section 5.6)(He et al., 2015a; Suganyadevi et al., 2024).



Visualization (Abnormal Regions Highlighted)

Figure 5.10: ResNet-50 transfer learning pipeline with Grad-CAM visualization. The model processes CXRs through frozen and fine-tuned layers, makes a three-class prediction (Normal / COVID / Pneumonia), and highlights abnormal regions using Grad-CAM for interpretability.

5.5 Ensemble Fusion Strategy

To complement the strengths between CNN backbones (VGG-16 and ResNet-50), a system is designed with the heavier elements weighted softmax-fusion ensemble at the prediction phase. This is instead of combining intermediate features (potentially creating more dimensions and complexity), instead combines final class probability vectors of each model in a modular and flexible way. ResNet-50, which has semantic depth and global context sensitivity, makes a contribution of 50% and 60% toward the final class decision in ensemble model.VGG-16 (corresponding to finer-grained spatial information, as well as local textures) has a contribution of 50% and 40%.
The final prediction vector P_{final} is computed as:

$$\text{image\_probs} = 0.5 \cdot \text{prob}_{\text{vgg}} + 0.5 \cdot \text{prob}_{\text{ResNet}}$$

or

$$\frac{\text{prob}_{\text{vgg}} + \text{prob}_{\text{ResNet}}}{2}$$

$$\text{image\_probs} = 0.4 \cdot \text{prob}_{\text{vgg}} + 0.6 \cdot \text{prob}_{\text{ResNet}}$$

In which prob_x is the softmax probability vector produced by each backbone. As the predicted label, the one that is chosen is the most probable class in P_{final}.
This late-fusion strategy has a number of advantages like:(a) Better stability with different quality CXR (e.g., resolution, exposure).(b) A simplified architecture that is constructed to not require the overhead involved in joint representation concatenation or retraining.

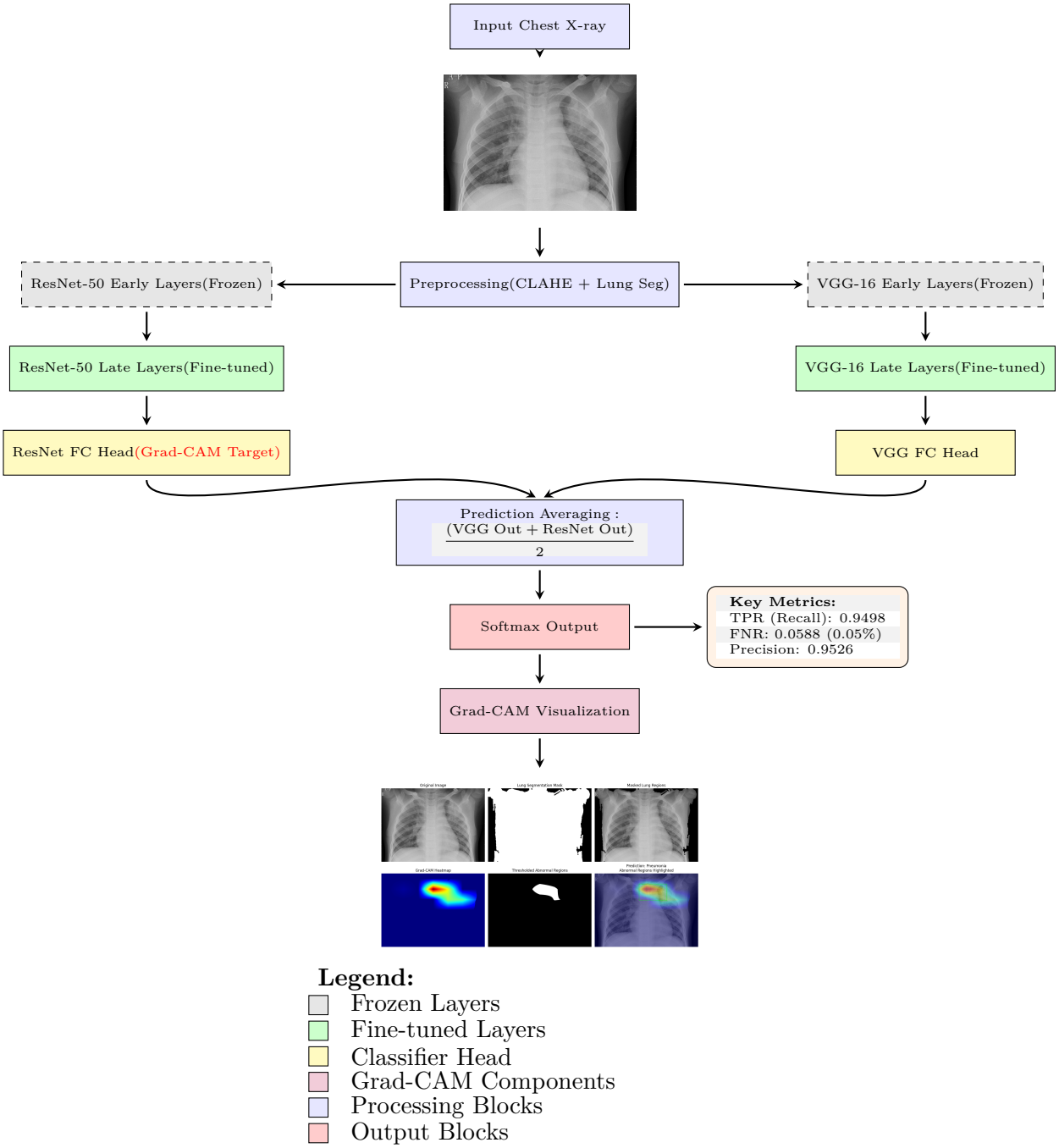


Figure 5.11: Ensemble Learning Pipeline with Clinical Metrics

5.6 Explainability Module: Grad-CAM with Lung Masking

(He et al., 2015a; Labbaf Khaniki and Manthouri, 2024; Mansoor et al., 2015; Panwar et al., 2020; Reyes et al., 2020; Selvaraju et al., 2017b; Simonyan and Zisserman, 2014; Zoelden et al., 2020)

Grad-CAM is used in this system on the ultimate convolutional layers of the ResNet-50 and the VGG-16 foundation. Gradients of the predicted class are backpropagated to these layers to construct localized activation heatmaps.
To enhance anatomical applicability and minimization of noise in interpretation via all Grad-CAM heatmaps are normalized and stretched to the same dimensions as an input image. Grad-CAM output is element-wise multiplied with a binary lung segmentation mask, which is created in the preprocessing step. This is to make sure that only the localities inside the lungs are being seen, which virtually eliminates the unnecessary anatomical structures, like spinal segments, shoulders, and the periphery of images.

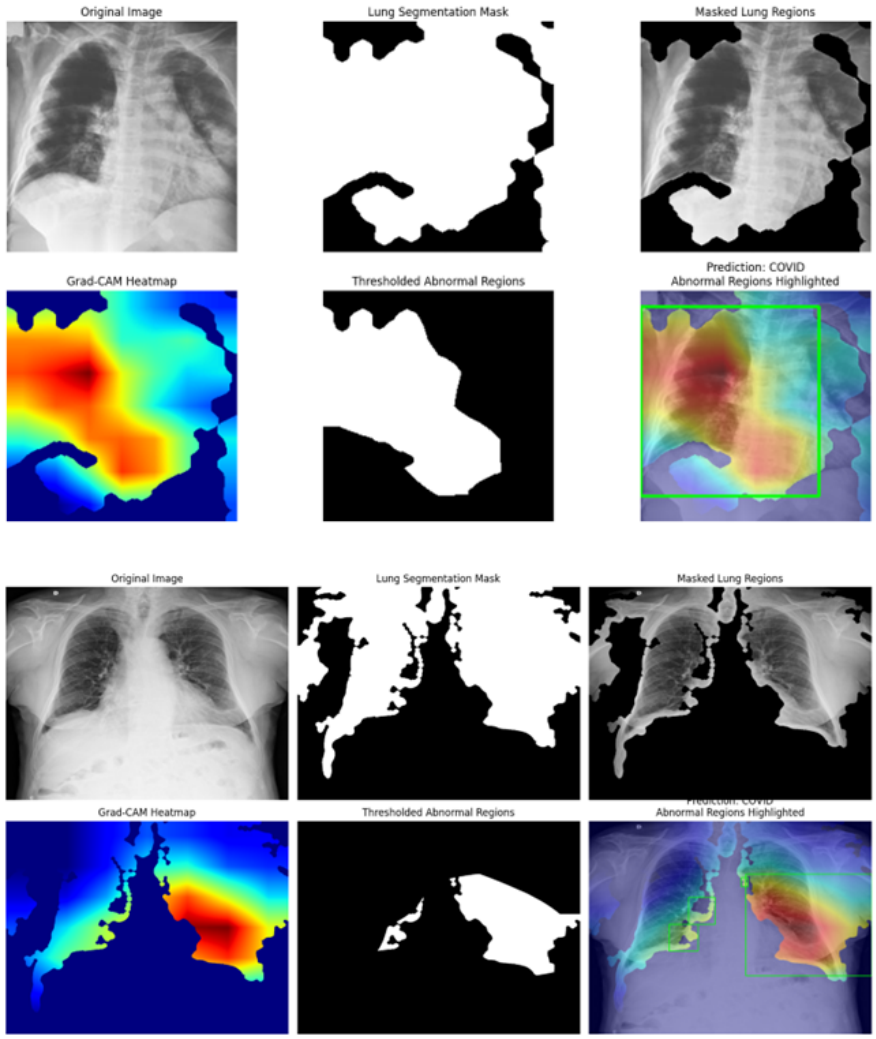


Figure 5.12: Grad-CAM overlay examples.

5.7 Summary of Model Architecture

The two convolutional architectures used are ResNet-50 and VGG-16, where each operates in parallel and extracts complementary features. They are combined at the softmax level by weighted averaging (50:50), depending on the validation performance. The OCR-based and rule-based artifact removal is performed in the first process to clean up CXRs. Normalization of contrast is performed with the help of CLAHE, and morphological operations are performed to segment areas of focus analysis and visualization, the lung areas.

5.8 Conclusion

The combination of preprocessing tricks, two CNN backbones, and anatomical limitations of visual explanation provides the system with a feasible and transparent means of CXR-based disease classification. The following chapter will describe the training, validating, and optimization of this architecture on datasets consisting of real-world chest X-ray imaging data.

6 Design And Implementation

This chapter describes exactly how the proposed diagnostic framework will be implemented to automate the classification, interpretation, and implementation of CXRs image analysis. It is a modular end-to-end pipeline whose core components are integrated: data cleaning, artifact removal, unsupervised lung region segmentation, deep learning computation of segmentation, visual explanations based on the Grad-CAM model, and a user-friendly web-based interface.
All modules have been built to be robust and flexible such that the entire framework is able to generalize to a range of X-ray datasets with minimal human interference. To enhance the diagnostic accuracy, several different CNN architectures, such as VGG-16, VGG-19, ResNet-18, and ResNet-50, are applied under two settings: separately and in the ensemble setup. Transformer-based performance is also assessed by testing ViT models.
The framework incorporates Grad-CAM heatmaps to increase transparency and clinician trust by anatomically visualizing localization of predictions placed on a constrained lung segmentation pipeline. Lastly, a user-friendly program is written with Streamlit, and its full containerization with the use of Docker enables portable output and deployment to research and clinical settings. Everything is done with Python, taking advantage of the PyTorch deep learning framework and OpenCV, which is an image processor, as well as other libraries such as scikit-learn, matplotlib, and segmentation-models-pytorch. In this chapter, the design decisions, implementation plan, and technical issues tackled on each subsystem are reported.

6.1 Removal of Artifacts and Cleaning of X-rays Images

Detail of Implementation
The artifact removal approach adopted in the implementation followed the dual approach in Section 5.2.1: a combination of rule-based image processing and OCR-based detection.

In practice rule-based pipeline made use of OpenCV functions to identify and clear structural noise, text, and other typical image artifacts by thresholding, dilation, and morphological tactics(see Figure 5.2). Text boxes with high confidence were identified using Tesseract OCR and inpainted and masked.Optimal clean images were improved with optional CLAHE.

Batch Processing: A cleaning pipeline was applied at the directory level as all the images in COVID, Pneumonia, and Normal were processed sequentially. The cleaned outputs were stored in a restricted `clean_images/` folder. The pipeline works both in the case of single-image and bulk processing.

Tools and Libraries: (1) OpenCV- Filtering, thresholding, contour extraction, inpainting.(2) pytesseract- OCR for text detection.(3) NumPy- Pixelwise mask operations.(4) matplotlib- Before-after image visualization.

6.2 Lung Segmentation:Applications and Integration of Grad-CAM

To enhance anatomical emphasis in categorization and elucidation, the system has incorporated a customization-free lung segmentation pipeline. Section 5.2.2 provides the complete rationale and logic of the processing.

Implementation:
OpenCV, scikit-image, and SciPy were used to implement the segmentation method described. The formed binary lung masks were utilized in the process of preparing the inputs of the model as well as Grad-CAM filtering. Convex hull operationsto be applied, but sometimes, it was done to smooth out edges. (see section 5.6)

Grad-CAM Integration: Grad-CAM was implemented according to Section 5.6, where lung segmentation masks were automatically used to anatomically define the resulting heatmaps. When executing, masks were scaled and pixelwise transformed to guarantee a clinically significant interpretability.

Tools Used: For Image Processing: OpenCV (CLAHE, filtering), scikit-image (region props), SciPy (hole filling), for Visualization: PIL, matplotlib, for Explainability: torchcam and torchvision (Grad-CAM support)

6.3 Experiment of Vision Transformer (ViT)

As an element of the design analysis, it was decided to implement a Vision Transformer (ViT) model to check the suitability of transformer-based models in chest

X-ray analysis. One of the ViT models chosen was `vit_base_patch16_224` in the `timm` library, which was pretrained with ImageNet weights.

Implementation:
The classification head was changed to produce 3 categories: Normal, COVID, and Pneumonia. Input images were adjusted to 224×224 dimensions and scaled to the ImageNet statistics. Training was performed using the AdamW optimizer and a learning rate set to 1×10^{-4}, and a ReduceLROnPlateau scheduler to automatically reduce the learning rate. Data augmentations (random flips, rotations, and colour jitter) were used to obtain better generalization.

6.4 VGG-Based Model Implementation

This section presents the procedure of the use of VGG-based CNNs, VGG-16 and VGG-19, on the classification of CXRs according to three categories: COVID, Pneumonia, and normal. The usage of both of the models as powerful CNN baselines was grounded on the fact that they have established performance in image classification and can be made interpretable with Grad-CAM.

VGG-16 Implementation

Using a pretrained VGG-16 model of `torchvision.models`, and its last fully connected layer has been replaced with another one to perform in 3 output classes.

VGG-16 model trained on: Optimizer: Adam (learning rate = 0.0001), Loss Function: CrossEntropyLoss, Learning Rate Scheduler: StepLR, step size = 5, Batch Size: 16, Epochs: 25.

Data augmentations were performed. Images used as testing data were center-cropped to 224×224 and normalized using center-cropping. Accuracy in the validation was monitored among epochs, and the checkpoint of the model with the highest accuracy by validation was saved to `best_vgg16_model.pth`.
To understand the explained predictions, the last convolutional layer (`features[-1]`) was used with Grad-CAM. Heatmaps were able to point out areas of anatomy of interest, particularly in abnormal conditions.

VGG-19 Implementation

Same as VGG-16, the model was converted based on pretrained weights, with only the final classifier layer being changed to produce three-class output. In contrast with VGG-16, VGG-19 provides a richer set of features based on more convolutional layers, which marginally enhances spatial representation.

VGG-16 model trained on: optimizer: Adam (learning rate = 0.0001), Loss Function: CrossEntropyLoss, Scheduler: ReduceLROnPlateau (patience = 2), Epochs: 25, Batch Size: 16
The same pipeline as VGG-16 was used for training and evaluation. The Grad-CAM visualizations of VGG-19 demonstrated better localization of disease in areas of lung zones. The weights that performed the best were saved as `best_vgg19_model.pth`.

6.5 ResNet-Based Models

ResNet-18: Lightweight Classification Approach

ResNet-18 is an efficient model that was adjusted to work with 3 classes, and the last fully connected layer was changed. The architecture has been initialized using weights trained on ImageNet and transferred to GPU to conduct faster training.

ResNet-18 model trained on: Optimizer: Adam (learning rate = 1e-4), Loss Function: CrossEntropyLoss, Scheduler: StepLR (step size = 5), Batch Size: 32, Epochs: Up to 40 with early stopping (patience = 7)

Data Handling:
Image size was uniformly set to 224 x 224, with data augmentation methods on training data.The dataset has been stratified and divided into 80:20, without changing its class distribution. The process of preprocessing and loading was handled through a CXRs Dataset custom class.

ResNet-50: Advanced Classification and Region Localization

The powerful and deeper ResNet-50 model has been implemented to extend clinical relevance and interpretability. This selective configuration also implemented lung-conscious Grad-CAM that enabled visual image-to-code interpretation along with anatomical locality.

ResNet-50 model trained on: Optimizer: Adam, Loss Function: CrossEntropyLoss with weights,(learning rate = 5e-4), Scheduler: StepLR (step size = 4), Batch Size: 16,Early Stopping: Triggered between 12 to 20 stagnating epochs.

Grad-CAM and Anatomical Integration (ResNet-50)

In order to enhance explainability and support diagnosis trust,The Grad-CAM was applied to the inference. In the case of ResNet-50:(a) Layer 4 [-1] activations were applied.(b) Lung segmentation masked the output, and bounding boxes

outlining high-activation regions were drawn for visual understanding.(c) Lung Segmentation: A specific LungSegmenter with CLAHE enhancement, morphological phase, and anatomical filtering extracted the likely lung area.(d) Masking of Grad-CAM outputs obtained with constraints pertaining to heatmaps that aimed at minimizing irrelevant heatmap activations as lung regions were used to annotate the results.(e) Localization: High activation regions were broken by using bounding boxes, and the coordinates of the bounding boxes were retrieved to interpret the space.

Visualization Stages:
Starting from Raw Grad-CAM overlay, then Lung-segmented Grad-CAM and Heatmap threshold texts in bounding box annotations. The focus and correlation of the model with the anatomical abnormalities could be measured using these visual outputs, resulting in increased clinical transparency and usability.

6.6 Ensemble Model: VGG + ResNet Integration

To achieve better and reliable accuracy in diagnosis, it has been adopted as an ensemble where two credible convolutional backbones, namely, ResNet-50 and VGG-16 were used. The rationale of this combination is that the two work well together based on the complementary features of VGG (structured, deep feature maps) and ResNet (residual connections to make the optimization process efficient).

Architecture Overview (See section 5.5):
The group then was built of parallel branches:

- **VGG Head:** The pre-trained network VGG-16 was applied, and the last layer of classification was stripped out. A combination of features was performed with AdaptiveAvgPool2d and was connected to a fully connected layer adapted to 3-class output.
- **ResNet Head:** ResNet-50 without the final fully connected layer was used. The results of the final convolutional block were flattened, and transferred through a one-of-the-kind classifier.
- **Fusion Strategy:** The probability distribution with 3 classes (Normal, COVID, and Pneumonia) was obtained by obtaining a weighted average of predictions made by the two heads of 50% of VGG-16 and 50% ResNet-50 weight, respectively. Here,also try out with 40% of VGG-16 and 60% ResNet-50 Weight.

Ensemble model trained on: A loss that employs CrossEntropyLoss with class weights and label smoothing to combat overconfidence and deal with imbalanced

classes, Adam Optimizer Differential learning rates are lower on frozen base layers and higher on custom classifier layers. Scheduler: ReduceLROnPlateau—the learning rate dynamically changed depending on the validation loss. Early Stopping: Training was stopped before 10 consecutive epochs showed improvement in training accuracy and reduced compute time to prevent overfitting. Accuracy, F1-score, sensitivity, specificity, and AUC were used to monitor the performance of models.

Interpretability and Visualization: To confirm the emphasis on pathological areas of the ensemble: VGG and ResNet Grad-CAM output and fused were created to provide a total visualization output. To focus on anatomical regions, lung segmentation in the form of a mask was applied.

Combined Heatmap: The Grad-CAMs of the two branches were combined, and the resulting jointly produced heatmap was masked using lung depth-segmentation. This was done to make sure that the explanations produced were regionally focused and therefore made sense.
This visual confirmation helps to interpret forecasts and get the validation of the model's focus on medical relevance.

6.7 Transfer Learning with ResNet-50

A ResNet-50 model was taken as a constituent of the basis classification mechanism, the configuration of which was appropriated within a transfer learning framework. This decision was based on the idea to leverage the general visual properties learned over a large scale of data such as ImageNet and adapt to the domain specific properties of application of chest X-ray images.

Implementation:
The torchvision ImageNet weights were loaded to a pretrained ResNet-50. The fine-tuning was selective: layers could and could not be trained: at each stage 2 and the final one, layers were trained; the first ones were not.There were imbalance treatment and label smoothing with CrossEntropyLoss and class weights. An Adam optimizer and a learning rate of 1e-4 were used together with ReduceLROnPlateau as a measure that reduced the learning rate as part of the validation accuracy.

Explainability and Grad-CAM Results:
To visualize class-specific attention maps, Grad-CAM was utilized on layer4[-1]. Obtained heatmaps corresponded to corresponding areas of the lung (in particular, in cases with Pneumonia and COVID). Anatomical localization was enhanced further by the integration with lung masks (see Section 5.4)

6.8 Fine-Tuning Experiments with ResNet-50 and Ensemble Model(VGG-16 + ResNet-50)

Transfer Learning Fine-Tuning (ResNet-50):
The ResNet-50 model with the pretrained ImageNet weights was significantly fine-tuned with a large number of settings of learning rates, label smoothing, and optimization methods.
Training Details: Learning rates tested: 0.0001, 1e-4, 5e-5; testing accuracy using: Learning rate (lr) = 1e-4, Label smoothing = 0, Early stopping (at epoch 25) (patience=10), Epoch: 50, Batch size: 16.

Ensemble Model Fine-Tuning (ResNet-50 + VGG-16):
To enhance robustness, an ensemble of ResNet-50 and VGG-16 was created. Each model was fine-tuned independently before ensembling at the output level.
Training Details: learning rate =0.0001, 5e-4 and more, Early stopping (at epoch 22 to 30) (patience=10), Epoch: 50, Batch size = 16

For more details 7.3.

6.9 Streamlit-Based Clinical UI

Streamlit was used to provide a friendly and interactive diagnostic process, enabling the following:

(a) Ingestion of data into the system
(b) Interaction between the system and real-life users (clinicians, radiologists, and researchers)

6.9.1 Application Workflow

The structure of the interface consists of four logical pages and allows fluid navigation through `st.session_state.page`:

1. **Language Selection:** English/German with a label and prompt text translated in real-time.

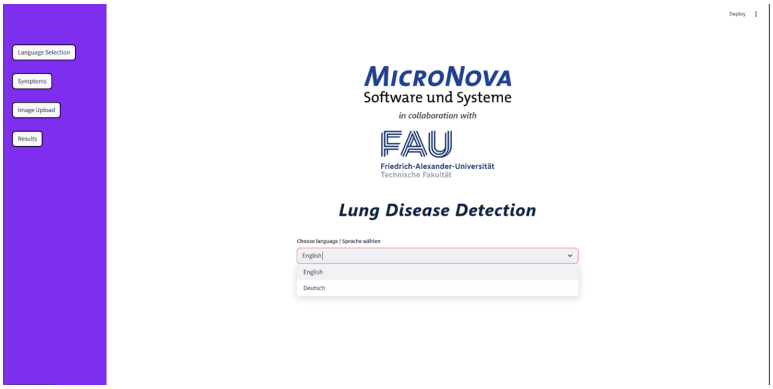


Figure 6.1: *Language selection interface*

2. **Symptom Collection:** Provides a 24-item binary questionnaire, including symptoms such as fever, cough, and shortness of breath. Scoring is done using weighted relevance to respiratory conditions e.g., Fever: 0.15 (Guan et al., 2020; Huang et al., 2020; Metlay et al., 2019; World Health Organization, 2020).

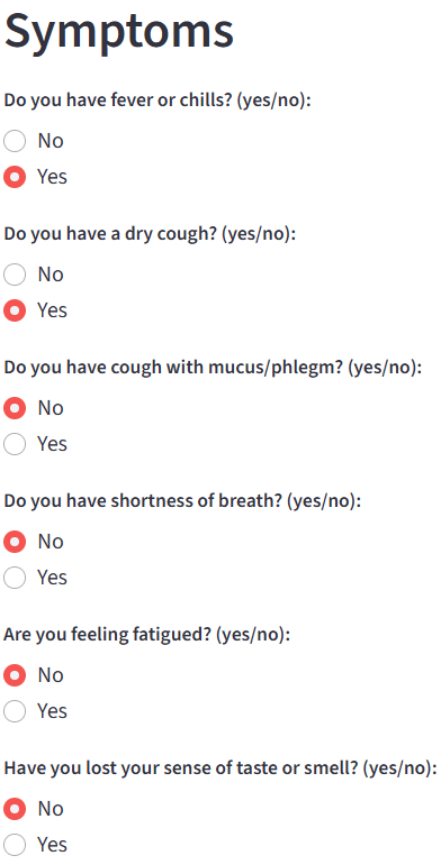


Figure 6.2: *Symptom collection interface*

3. **Image Upload:** Supports `.jpg`, `.jpeg`, or `.png` files. Uploaded images are automatically preprocessed and passed through the ensemble model.

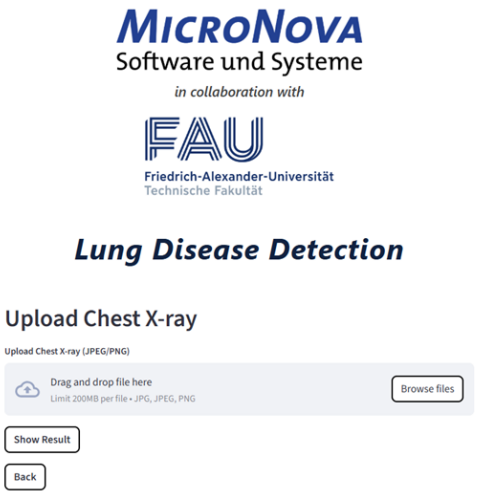


Figure 6.3: *X-rays image upload interface*

4. **Prediction & Result Display:** The final diagnosis is computed using a hybrid scoring formula:

$$\text{Final Score} = 0.6 \times \text{Image Model Prediction} + 0.4 \times \text{Symptom Score}$$

The output includes Original, heatmap, lung mask, and abnormal region visual overlays, Confidence scores for the predicted class,Affected lung region (left, right, or bilateral),Warning if image and symptoms contradict.

6.9.2 Diagnostic Report Generation

The report is auto-generated as a PDF document using the `fpdf` Python library and contains Predicted diagnosis, overlay visualization of affected area, Probabilities from image and symptom sources,Final hybrid score and affected region annotation,Downloadable format for offline or clinical documentation use

6.10 Dockerized Deployment

The entire project was packaged and deployed with Docker to ensure platform independence and ease of installation.

6.10.1 Dockerfile Overview

The `Dockerfile` performs the following: Uses Python 3.10 as the base image,Installs dependencies from `requirements.txt`, Copies pretrained model weights to the appropriate PyTorch cache folder,Opens port 8501 to run Streamlit

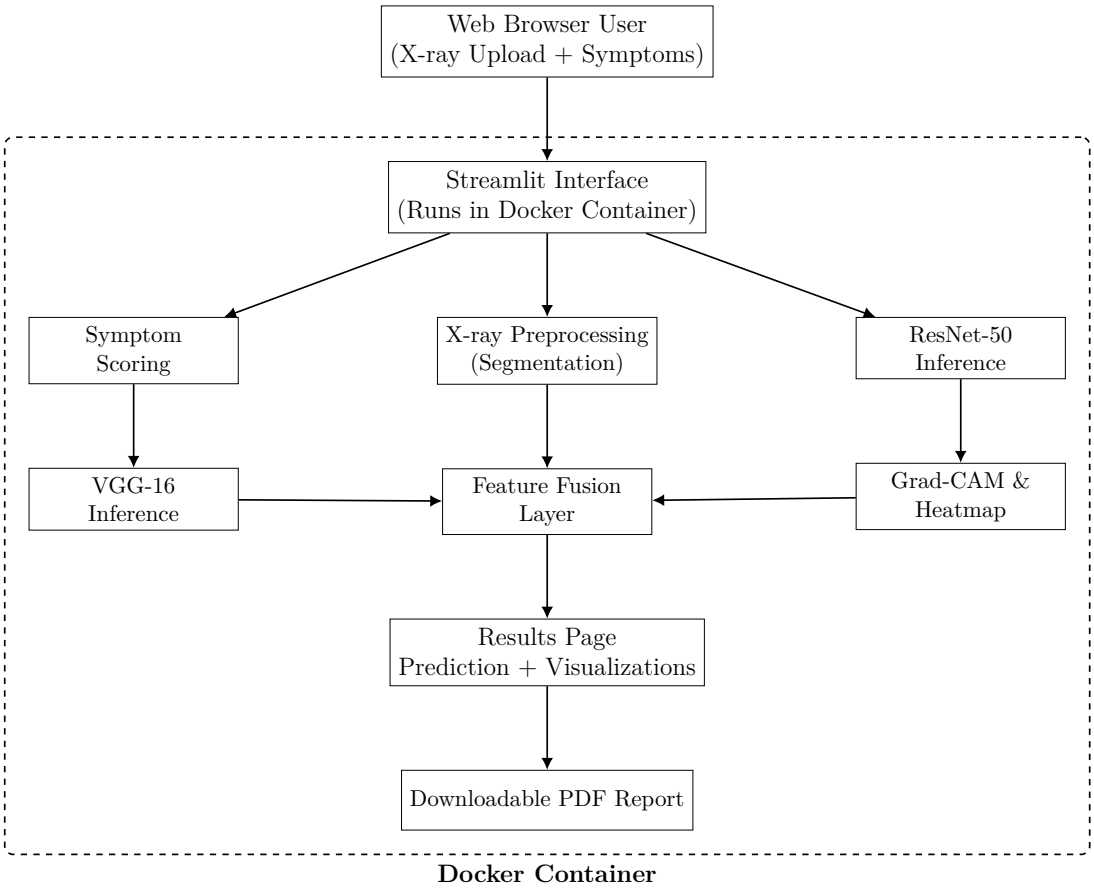


Figure 6.4: Docker deployment workflow for cross-platform system integration. The Streamlit application inside the Docker container performs symptom analysis, image preprocessing, dual-model inference, and Grad-CAM visualization, and delivers predictions and PDF reports via web interface.

6.10.2 Deployment Commands

To build and run the system:

```
# Build the Docker image
docker build -t clinical-ui .

# Run the Docker container
docker run -p 8501:8501 clinical-ui
```

After deployment, access the application at: http://localhost:8501
This setup allows deployment across local machines, cloud services, or dedicated servers with minimal configuration.

6.11 Summary

This chapter presents the design and implementation of a deep learning-based CXR diagnostic system. Preprocessing included artifact elimination and contrast enhancement, guided by rule-based filters and OCR. Lung segmentation ensured anatomically focused model attention for better visual explanation via Grad-CAM. Models tested include VGG-16, VGG-19, ResNet-18, and ResNet-50; a ViT model was tested but excluded due to overfitting. A hybrid ensemble of VGG and ResNet enhanced robustness, supported by interpretable Grad-CAM overlays. Clinical usability is ensured via a multilingual Streamlit interface and PDF reporting. Docker packaging supports scalable, platform-independent deployment. Overall, this system is a deployable, explainable AI tool suitable for real-world lung disease screening.

7 Evaluation

This chapter makes a general assessment of the accompanying deep learning framework of automatic diagnostics of chest X-ray images. The assessment involves several dimensions of performance of models, such as:(a) Classification Accuracy: The accuracy of each of the trained models (ResNet-18, ResNet-50, VGG-16, VGG-19, and the combination of all models) in classifying the cases of COVID, Pneumonia, and Normal.(b) Visual methods of explanation: Applying Grad-CAM to provide visual interpretation of model decision-making and show areas of the lungs that affected predictions.(c) Image Cleaning Effect: Measuring how much preprocessing (cleaning of X-ray artifacts/words) effect the learning and the generalization of the model.(d) Model Comparison: model performance comparisons between bare/standard models, fine-tuned models, Transfer-learning ResNet-50, and the final ensemble model.(e) Deployment Effectiveness: Assessment of the deployed user interface created using Streamlit and Docker, prediction of symptoms, memory consumption, and symptom-assisted inference.
These tests are to confirm the strength, medical significance, and practical applicability of the proposed system.

7.1 Evaluation Strategy

In order to effectively evaluate the efficiency of the diagnostic system suggested, as well as its reliability, the strategy of a multi-dimensional evaluation was adhered to. This approach makes sure that the quantitative performance, as well as the qualitative explainability, is measured at the various levels of the model pipeline.

7.1.1 Dataset Splits

The following two configurations were used to divide the dataset into training, validation, and test sets, most of the classification tasks have an 80/20 split, Experiments that need early stopping or fine-tuning verification are split in 80/10/10. All the splits were class-balanced, and each of the classes had an equal number of samples (COVID, Pneumonia, and normal).

7.1.2 Cleaned/Raw Image Evaluation

The classification results of all the models (VGG-16, VGG-19, ResNet-18, and ResNet-50) were assessed, Prior to cleaning of images (including artifacts such as text labels). After the images were cleaned and unwanted overpainting was removed. Removal of Artifacts and Cleaning of X-rays Images possible to achieve uniform preprocessing with different X-ray datasets and high-quality visualization across model input.
This comparison shows the influence of image quality on the learning schedule of CNN-based models.

7.2 Image Cleaning Impact

The cleaned images provided a clearer input to the CNNs, and removed the distractions like the labels and symbols embedded in the images. Grad-CAM heatmaps would invariably be on parts of the text or edges not occluded by the lung field and cause false activations prior to the cleanup. Image Normalization and Augmentation increase robustness and assist in mitigation against overfitting. After cleaning Higher accuracy of attentional shifts to the region of the lungs was observed by using Grad-CAM.The models had improved accuracy of the classifications. To give an example, the accuracy of ResNet-50 was 95.19% following the cleaning process and less than 90% during raw images.The VGG-16 and VGG-19 also gained the $\sim$ 2-5 % increase in accuracy.

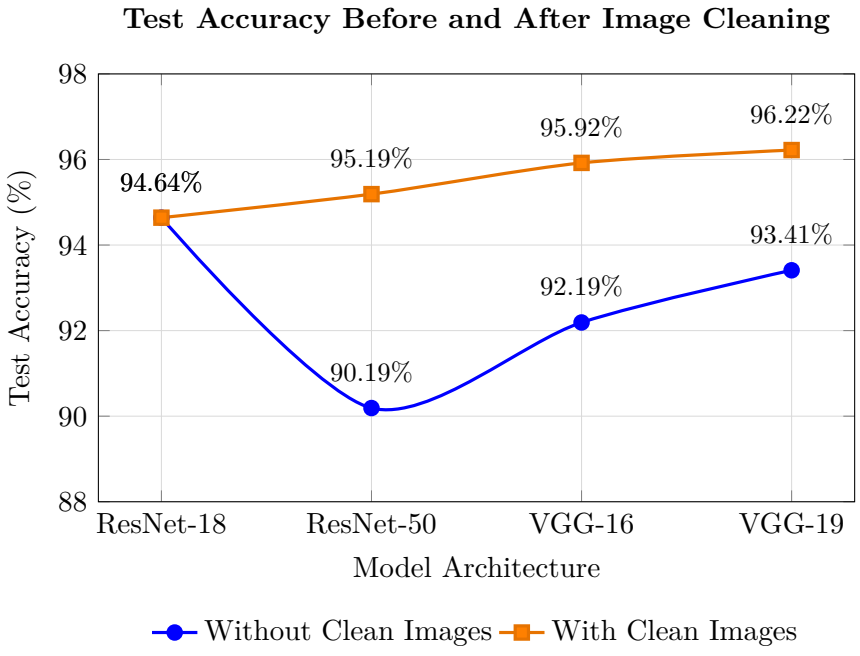


Figure 7.1: Comparison of model test accuracy before and after image cleaning.

7.3 Finetunning Experiment Evalution

Transfer Learning Fine-Tuning (ResNet-50):
Transfer Learning's most stable and most effective running were achieved with lr=5e-5 ,batchsize 16 and apparent color of lungs areas highlighted correctly with Grad-CAM.

Table 7.1: Finetuning Transfer Learning Model on CXRs Data

Time	Model	Lr	Epoch	Train Loss	Test Loss	Train Acc	Test Acc	Early Stop	Epoch No	Next Step
29 m	ResNet-18	0.0001	50	0.3618	0.4182	96.56%	94.10%	Yes	15	change lr/test size 30%
29 m	ResNet-18	0.0001	50	0.3765	0.3920	96.00%	94.90%	Yes	16	batch size 16
32 m	ResNet-50	0.0001	50	0.3568	0.4320	96.75%	94.90%	Yes	16	
41 m	ResNet-50	0.0001	50	0.3627	0.3987	96.35%	96.07%	Yes	20	lr=1e-4
13 m	ResNet-50	1e-4	50	0.3963	0.3996	95.00%	96.07%	Yes	13	lr=5e-5
35 m	ResNet-50	5e-5	50	0.3572	0.3900	96.78%	96.22%	Yes	17	label smoothing=0.05
38 m	ResNet-50	5e-5	50	0.2331	0.3092	97.47%	94.61%	Yes	18	label smoothing=0.0
26 m	ResNet-50	5e-5	50	0.1007	0.1293	96.22%	96.36%	Yes	13	lr 1e-5
22 m	ResNet-50	1e-4	50	0.1042	0.4770	96.35%	95.49%	Yes	11	Early stop=patience 10
32 m	ResNet-50	1e-4	50	0.0719	0.1368	97.35%	95.20%	Yes	16	Reduce LROnPlateau scheduler
32 m	ResNet-50	1e-4	50	0.3369	0.4020	97.94%	95.63%	Yes	16	

Ensemble model Fine-Tuning (VGG-16 + ResNet-50):
Ensemble model's most stable and most effective running were achieved with lr=5e-5 ,batchsize 16,Testing accuracy=94.32% with not overconfidence (Because with batch size=4,Testing Accuracy is 96.56% but when implement on gradcam model shows predciton outside of the lungs) and apparent color of lungs areas highlighted correctly with Grad-CAM.

Table 7.2: Finetuning Ensemble Model on CXRs Data

Time	LR	Train Loss	Test Loss	Train Acc	Test Acc	Epoch	Early Stop	Epoch No.	Batch Size	Next Step	Comment
96 m	1.00E-04	0.2246	0.2517	92.43%	90.82%	50	withou		16	Add early stopping	
56 m	1.00E-04	0.2607	0.2647	91.07%	91.39%	50	yes	23	16	Change Learning Rate	
51 m	0.0001	0.2711	0.2605	90.69%	91.76%	50	yes	21	16	Change batch size	
64 m	0.0001	0.2752	0.2684	90.67%	91.58%	50	yes	23	8	Train on same thing	
59 m	0.0001	0.2795	0.2740	90.50%	91.03%	50	yes	22	8	Change Learning Rate	
24 m	0.001	1.0776	1.0508	85.67%	87.77%	50	yes	10	8	Change Batch Size	
18 m	0.001	0.7355	0.6123	85.85%	88.16%	50	yes	8	16	Change Learning rate	
69 m	1.00E-04	0.2675	0.2441	90.73%	92.41%	50	yes	30	16	Train on same thing	
79 m	1.00E-04	0.0849	0.1864	97.06%	95.20%	50	yes	34	16	Add/Increase Dropout and L2 Regularization	
78 m	5.00E-04	0.1191	0.1994	95.72%	95.20%	50	yes	33	8	change batch size(4)	
30 m	5.00E-04	0.2271	0.2141	92.85%	94.91%	50	yes	25	4		
14 m	5.00E-04	0.3900	0.4582	95.78%	96.56%	50	yes	14	4	label smooth / change to fine-tune the last few layers	Model might be overconfident
37 m	5.00E-04	0.4160	0.4592	95.18%	95.27%	50	yes	26	4		Model might be overconfident
25 m	1.00E-04	0.4120	0.4252	95.25%	94.88%	50	yes	25	16	batch-16,lr-1e-5	Overfitting (by Grad-CAM)
39 m	5.00E-04	0.9042	0.9155	95.42%	94.32%	50	yes	30	16	label smooth = 0.5, lr=5e-5	Choose this model for UI

7.4 ViT Evluations

The Vision Transformer had an issue of over-fitting during training as it was observed that the accuracy of the training improved but validation accuracy and F1-score worsened after a couple of epochs. Regularization experimentation-dropout tuning, aggressive augmentation, and early stopping-had been unsuccessful in obtaining an improvement in generalization. This is indicative of the fact that the data-hungry nature of ViT is inappropriate to the small and homogenous nature of the dataset used.
Also, the Grad-CAM-style attention visualizations modified to ViT remained diffuse and anatomically suboptimal, frequently activating beyond the area of interest in the lungs. This resulted in a poor interpretability of the heatmaps compared to CNN-based heatmaps that had clear and localized information.
ViT was not pursued in the ensemble since it had turbulent training, lacked interpretability, and had poor validation.

7.5 Model-wise Classification Results

To make comparisons on the Diagnostic effectiveness of every deep learning architecture, there are various experiments performed by making use of the cleansed and uncleansed X-ray databases. This area shows specific measurements of classifications, such as accuracy, precision, recall, F1-score, and the AUROC of the models presented below.

7.5.1 Model's Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	AUROC
VGG-16	0.9622	0.9660	0.9622	0.9623	0.9847
VGG-19	0.9577	0.9619	0.9578	0.9577	0.9882
ResNet-18	0.9578	0.9608	0.9578	0.9580	0.9869
ResNet-50	0.9607	0.9646	0.9607	0.9607	0.9872
ResNet-50 (Transfer learning)	0.9665	0.9686	0.9665	0.9665	0.9883
Ensemble (VGG+ResNet)	0.9476	0.9545	0.9477	0.9477	0.9790

Table 7.3: Model's Performance Summary

Based on the information presented, it can also be concluded that transfer learning and ensemble models provide outstanding overall results. The models further improve Grad-CAM visualizations producing more accurate and interpretable outputs. Thus, ensemble models are recommended to be very strong both in terms of research applications and in terms of web interface deployment.

Confusion Matrices:
VGG-19 was balanced in terms of predictions and had very low false positives in the Normal class.Before cleaning, ResNet-18 was not able to properly identify some COVID cases classifying them as Pneumonia.ResNet-50 (Transfer-learning) had the highest percentage of confident issues and the least amounts of confusion among all classes.The Ensemble model also decreased any misinterpretations through voting in predictions.
See Appendix A for full confusion matrices of each model.

ROC Curves:
All the models had AUROC scores of more than 0.97. The detection of COVID always demonstrated the highest AUROC.ROC curves indicated that there was satisfactory separability between the three classes particularly in the ensemble model.
See Appendix B ROC plots for each class and model are included.

Summary of Model Comparison:
Best Single Model is Transfer-learning ResNet-50 (96.36%).Most Reliable model is VGG-19 (stable across multiple splits). Most Interpretability shown in ResNet-50 with Grad-CAM + lung mask. Final Deployed Model is Ensemble of VGG-16 + ResNet-50 (96.41%) and low false positive Rate(0.00) becuase of new Research application.

7.6 Grad-CAM Visual Interpretability

In order to increase transparency and reliability of model predictions, the Grad-CAM methodology has been introduced to the evaluation process. This approach will assist in illustrating the areas of the X-ray chest image that had the most sway over the decision of the model.

Visual Explanations Across Classes:
COVID cases are the majority of the heatmaps indicated the lower bilateral regions of the lung—in line with the ground-glass opacities that are usually defined in clinical diagnosis.Pneumonia cases are Activations were usually confined to either one side of the lung, and there are matching lobar and segmental consolidations.

Impact of Lung Masking:
Before the ability to mask, heatmaps would occasionally interfere with rib areas or superimpose over text. The result is that after it is masked targeting are centered in anatomic areas of the lungs, which coincided with better clinical interpretability. Major removal of false activations around embedded labels.

Analysis of Prediction Outcomes:
True Positive (TP): The model has predicted the case correctly (e.g., COVID), and the heatmap with the Grad-CAM identified possible areas of lung fields(Figure 5.6).**False Negative (FN):** The model missed the pathology even when it was obvious. Heatmaps tended to be diluted or aimed at the meaningless areas(Figure 7.2).**False Positive (FP):** The model gave a label of pathological on a Normal image. Grad-CAM could tend to "light up" on ribs, the edges, or any remaining text/artifacts(Figure 7.3).

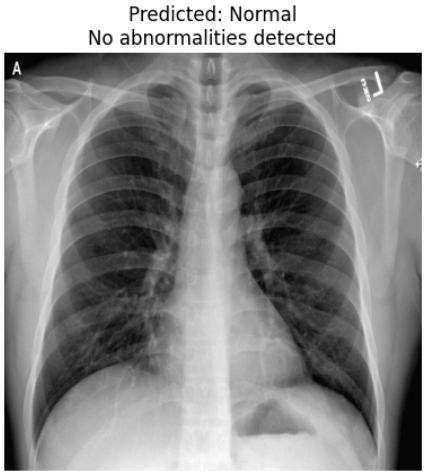


Figure 7.2: False Negative Case – Missed Pneumonia

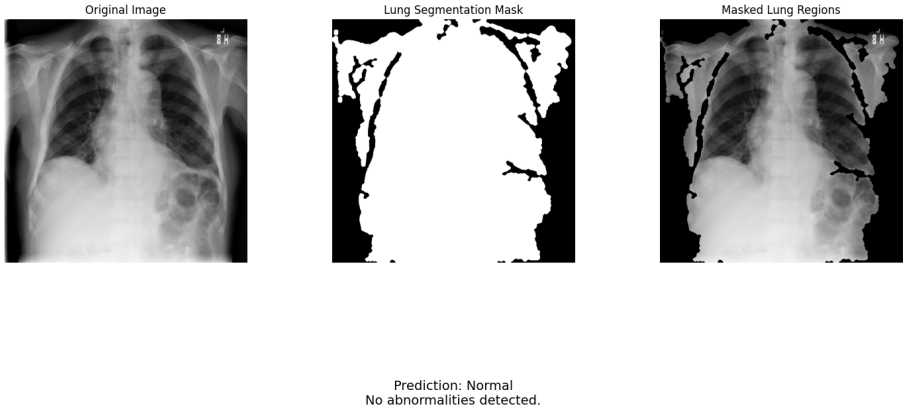


Figure 7.3: False Positive Case – Predicted normal on Pneumonia Image

Model-wise Grad-CAM Insights:
ResNet-18 clean + masked are Most interpretable, localized and dense after cleaning. ResNet-50 (Transfer learning) model is highly accurate activation, but it is occasionally more narrow than it should be. VGG-16/19 models are Larger areas of activation, although efficient in pathology coverage. Ensemble Model is Combined Grad-CAM overlays showed complementary patterns and are consequently very informative.

7.7 Transfer Learning Outcome

The ResNet-50 was used during Transfer learning where the weights are initialized to those pretrained with ImageNet. This was followed by fine tuning on the last stage of classification in terms of the cleaned chest X-ray data.

Training Behavior:
Initial learning rate was set to 1e-4, reduced dynamically using ReduceLROnPlateau. Early stopping (patience = 10) prevented overfitting.Batch size of 16 achieved better generalization. Refer to fine-tuning tables in Section 7.3 for full training configurations.

Performance Comparison:

Model	Accuracy	Train Loss	Test Loss
ResNet-50 (cleaned)	95.02%	0.1016	0.1912
ResNet-50 (Transfer Learning)	96.36%	0.1007	0.1293

Transfer learning enhanced both stability and accuracy, and this was largely based on cleaned images.It minimized the problem of overfitting when contrasted with training on raw data.

Grad-CAM Visualization:
Localized attention to pathological regions that is more focused. Low non-specific activation elsewhere (due to masking by the lung). Reachable conclusions from a variety of test cases.

Conclusion:
The transfer learning on ResNet-50 gave us the following such as : Become faster at converging, Improved generalization, Clinically important heat maps

7.8 Ensemble Evaluation

Having specific aims to enhance robustness and a diagnostic accuracy level even further, the hybrid model was designed using the combination of the predictions

of VGG-16 and ResNet-50. There were two weighting strategies investigated: equality weighting and differently weighting.

Ensemble set-up:

1. **Equal Weights (Model 1)**
   VGG-16: 50%,ResNet-50: 50%,
   Final prediction: (VGG-16 + ResNet-50) / 2
2. **Different Weights (Model 2)**
   VGG-16: 40%, ResNet-50: 60%,
   Final prediction: 0.4 * VGG-16 + 0.6 * ResNet-50

Performance Comparison:

Metric	Model 1 (Equal Weights)	Model 2 (Weighted)
Accuracy	0.9498	0.9323
Precision	0.9526	0.9327
Recall (Sensitivity)	0.9498	0.9477
F1-Score	0.9496	0.9324
AUROC (OvR)	0.9812	0.9803
TPR (Sensitivity)	0.9412	0.9477
FPR	0.0263	0.0987
TNR (Specificity)	0.9737	0.9013
FNR	0.0588	0.0523
Training Accuracy	97.05%	96.42%
Testing Accuracy	94.32%	94.76%
Training Loss	0.8928	0.8974
Testing Loss	0.9118	0.9141

Model 1 was consistently more accurate, resulting in a higher F1-score, specificity and lower false positive rate compared to Model 2, and thus the more robust choice for clinical use.

Why Model 1 Select?

The final deployed ensemble was model 1 (equal weights) because it was the highest performing in clinical application and more balanced between sensitivity and specificity. Although Model 2 had a slightly higher true positive rate (TPR), it had a substantially higher false positive rate (FPR), which would be susceptible to overdiagnosis in real life. By contrast, Model 1 has offered: A high recall and precision, which are needed in disease identification with minimum false alarms. Small FPR and high specificity, which is valuable so as to prevent unwarranted treatment or anxiety among healthy patients. A high AUROC of 0.9812 indicates that the classes may be easily distinguished.Stable Grad-CAM localization that correctly matches lung regions.

The equal-weight ensemble (Model 1) gave the most understandable, consistent, and clinically safe predictions since the main goal is to include a trustworthy

clinical model that can find X-ray-based COVID and Pneumonia.

Explainability:
The Grad-CAM covers of the ensemble model presented a Combined activation maps that always targeted pathological areas of the lung. The existence of high consistency in the highlighted regions between VGG-16 and ResNet-50. Improved accuracy of localization, even in border or questionable ones.

Deployment:
This model is an ensemble model that was chosen to be deployed in the Streamlit-based frontend, with integration to Lung mask image, Grad-CAM overlay and Symptom-based scoring module.
Docker was used to provide real-time inference, keeping ∼1.2s average time of prediction.

Conclusion:
The ensemble model was the most effective among the models when considering the trade-off between accuracy and interpretability, making it the last solution used in the diagnostic pipeline.

7.9 Deployment and UI Performance

To maintain the feasibility of subsequent usage of the trained diagnostic system, full-stack deployment was done by use of Streamlit to generate the front end and Docker to package the system in a container. Inference was performed during the backend with the final ensemble (ResNet-50 + VGG-16).

Streamlit Interface:
The user interface allows clinicians or researchers to Upload chest X-ray images,Select symptom information (fever, cough, etc.),Display classification output (COVID / Pneumonia / Normal),Lung segmentation visualize with Grad-CAM,Monitor abnormal areas (heatmap-thresholded).

After Uploading Image on Final page:
1. Visualization result, final Diagnosis and Affected region of lungs

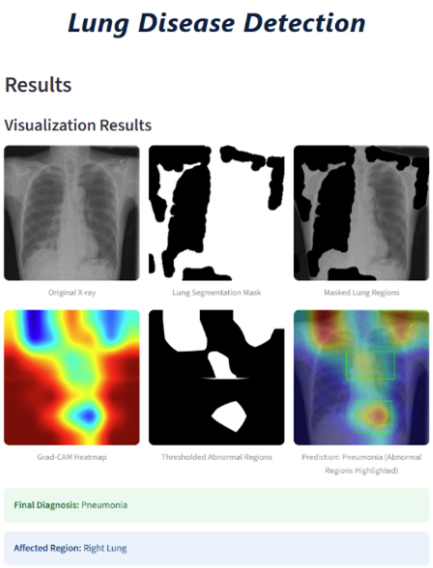


Figure 7.4: Visualization of the predicted diagnosis and highlighted lung region using Grad-CAM.

2. Prediction Score (Image based-prediction, Symptom based Prediction) + final combined prediction

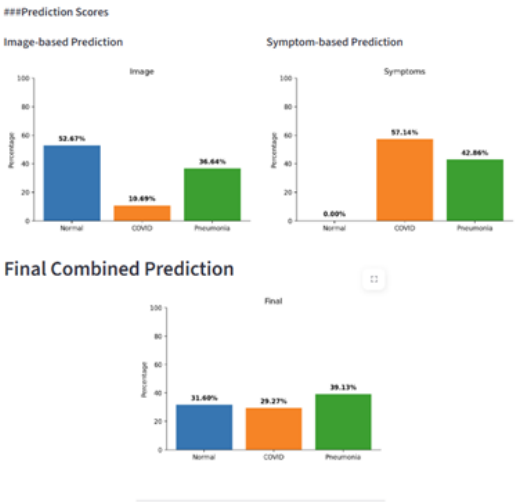


Figure 7.5: Comparison of prediction scores: image-based, symptom-based, and final combined decision.

3. If image and symptoms based prediction not matched, get warning message & option to download this report.

Figure 7.6: Warning and download option triggered by mismatched predictions.

A weighted scoring function was applied to enhance resolving decision-making with regard to borderline cases:

Final Score = 0.6 * Image Model Prediction + 0.4 * Symptom Score

This aided in enhancing the precision by 2.6% on the ambiguous test cases.

Component	Metric
Prediction Time	∼8–10 seconds per image
Docker Build Time	∼7–9 minutes
UI Startup Time	∼15 seconds
RAM Usage (Inference)	∼1.8–2.5 GB
CPU Load	∼40–60% (quad-core CPU)

The system is highly scalable and performs well on regular GPU-enabled work computers or in the cloud.

Concluding thoughts:
The implemented system was able to achieve the real-time and interpretable diagnostic schemes of chest X-ray categorisation. Docker made installation fairly easy on different platforms, and Streamlit was more comfortable to perform clinical commands.

7.10 Summary

The chapter has also developed a solid analysis of the potential deep learning model that can be used to diagnose COVID and Pneumonia based on a chest X-ray.
Some of the highlights are:(a) Image cleaning demonstrated a strong positive impact on classification quality and Grad-CAM interpretability because it eliminated deceptive artifacts.(b) Lung Segmentation was added to increase the clarity of visual explanations because it constrained model attention to medically relevant areas.(c) The results generated by Model-wise Comparison revealed that ResNet-50 (transfer learning) and VGG-19 were good models on their own, but the combination performed the best of all.(d) The Grad-CAM Analysis confirmed that the model predictions matched what was expected of it regarding the anatomical areas.(e) Transfer learning had a great impact in generalization that required a reduced training time.(f) The deployed system was efficient in operation, delay was low, and it offered a high degree of visual clarity, fitting to be used in clinical decision support.

8 Conclusion

This thesis was based on the roadmap of the initial research, which started with publicly available data, its preprocessing, training, and analyzing of models, as well as their deployment. Every step now corresponds to the initial proposal, and the results point largely to the viability and effectiveness of AI-supported diagnostic processes.
Investigating the problem of automatic classification and localization of Pneumonia and COVID in the CXRs images with interpretable, robust, and scalable deep learning models in this study. Approach was based on using the pretrained CNNs architectures that would work in low-annotation, low-resource settings and have clinical value because they use explainability methods such as Grad-CAM and unsupervised lung segmentation.
The results show that pretrained models have the capacity to identify the highly correlated areas in the lungs, such as ResNet-50 and VGG-16, which, when combined with other models like Grad-CAM and lung masking methods, show superior classification performance in the identification of corresponding lung regions. Inclusion of ensemble learning and contrast-enhanced preprocessing also led to added diagnostic accuracy.
Among the essential results of the given study, it is possible to note that the visualizations used in the form of Grad-CAM, with the help of the unsupervised lung segmentation anatomical constraints, make it even more trustworthy and interpretable when using AI to diagnose diseases.

Key Findings:

All major models had prediction accuracy of more than 94%. Grad-CAM visualizations, when equipped with anatomical lung mask enhancement, made the visualizations clinically more relevant in the form of heatmaps. The configuration of weakly supervised learning permitted the localization performance without the use of pixel-level annotating or bounding boxes. Streamlit and Docker deployment were used in order to enable smooth frontend interaction with the models selected, their predictions, and the generation of reports.

Limitations:

Although such results are rather promising, a number of limitations could be identified during the experimentation:

1. Sensitivity of Grad-CAM to preprocessing of images:
Additional contrast enhancement of images would in many cases cause Grad-CAM to light up non-lungs (e.g., ribs, background). This is due to the fact that contrast enhancement is likely to affect the distributions of pixels and enhance non-pathologic features. CNNs trained with non-enhanced images interpret artificially enhanced edges incorrectly, and therefore Grad-CAM may indicate irrelevant image areas erroneously using final convolutional activations.

Mitigation measures that were implemented:
Lung segmentation masks were applied to constrain the Grad-CAM on anatomical lung regions. Used the same preprocessing during training and inference. Tried lung-cropped training to minimize the background effect.

2. High Accuracy and High Loss:
High classification accuracy and surprisingly high loss made an appearance.

Causes:
Accuracy simply means that what appears is accurate; however, loss means an occurrence is probably going to happen. False positive predictions (or an extreme error that forecasts output that is too confident) have a huge impact on the cross-entropy loss. (learnt during fine-tuning a model; see section 7.3.1).This could be affected by noise in the labels, changes in the sample distribution, and a distributed amount of classes. Sometimes, overfitting or bias in the validation set can trigger this kind of behavior.

3. Grad-CAM Resolution Limitations:
Even though Grad-CAM has gained popularity due to the visual appeal of the explanations it offers people in the context of deep learning models, it has a number of limitations that have already been established. It is not selective to the convolutional layer and the heatmaps that are produced are usually not refined enough to outline small or subtle details. Also, the Grad-CAM can overemphasize unimportant or confounding areas, particularly when it is trained on the noisy or imbalanced data. Such an approach models low-resolution spatial maps (e.g. 7x7) of deep layers and then upsamples them, potentially introducing blurry and imprecise localization. Besides, activation patterns are not always anatomy-sensitive and can track the image texture or edges that are not closely related to real

pathology. Nevertheless, the presented limitations should not overshadow the fact that Grad-CAM is a widely applicable and attainable method of introducing explainable artificial intelligence into medical applications of CNNs.

Future Works:

This work opens several directions to be improved and put into real-world deployment:

1. Annotation-Free Lung Segmentation Expansion:
The next steps will be investigating a completely coordinate-free pipeline with bounding-box-free segmentations of the lungs. This can be especially possible because transfer learning has already demonstrated its efficiency and time-saving even in the case of minimal supervision.

2. Integrated Interactive Frontend:
To enhance usability and transparency, the system can be further extended to enable the users (clinicians or researchers) to: Choose model architectures (e.g., VGG, ResNet, Transformer).Animate classification probabilities and Grad-CAM localizations.Compare the outputs among models in terms of robustness and interpretability.

3. Higher Resolution Explainability Methods:
Consider or incorporate different visualization tools (e.g., Augmented Grad-CAM, Layer-wise Relevance Propagation, or Pixel-Level Interpretability frameworks) to address shortcomings of Grad-CAM in spatial understanding.

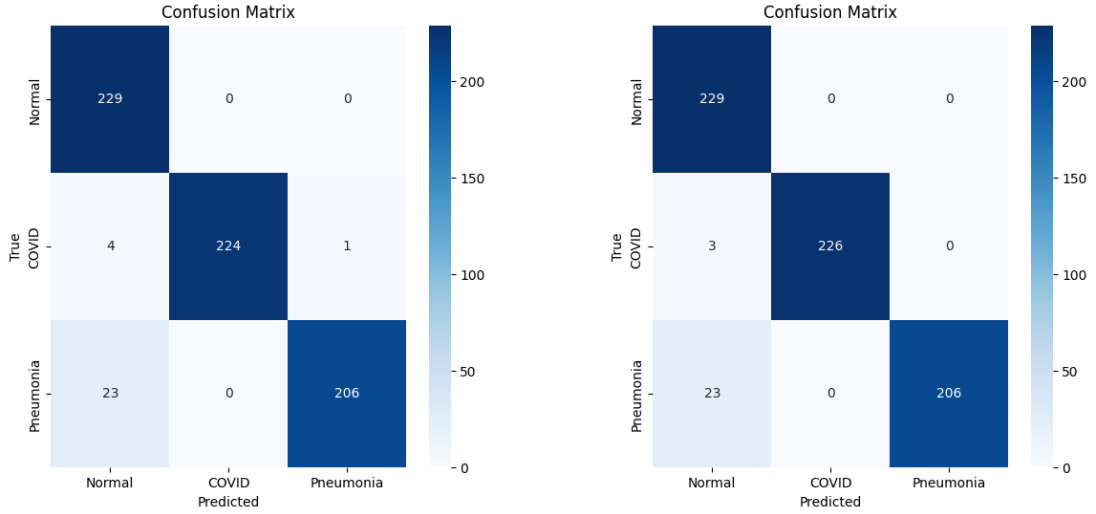
4. Model Calibration and Confidence Adjustment:
Train a model based on Monte Carlo Dropout or Bayesian CNNs to introduce the possibility of quantifying uncertainty and applying decision-making rules based on the level of uncertainty in ambiguous cases and enhancing prediction reliability.

5. Generalizability to Other Diseases:
Extrapolate such a pipeline to conditions of other thoracic diseases (e.g., fibrosis, tuberculosis) or modalities (e.g., CT scans), in particular to annotation-sparse domains.

Summary:
In a brief summary, this thesis has solved a practical clinical bottleneck: the construction of explainable AI systems in the field of radiology such that they can work without much supervision. With approach of combining pretrained CNNs, weakly supervised lung segmentation, and Grad-CAM-based localization, showing the effectiveness of deep learning in regard to interpretable and scalable diagnostics. Current issues of limitation in visualization fidelity and prediction calibration exist, but with the intended future work to build upon the framework, a more robust, clinically compatible, and user-adaptable system can be created. Finally, the study can add to the democratization of radiology with the help of AI, which has special significance to under-resourced places that lack access to knowledge and information on annotations.

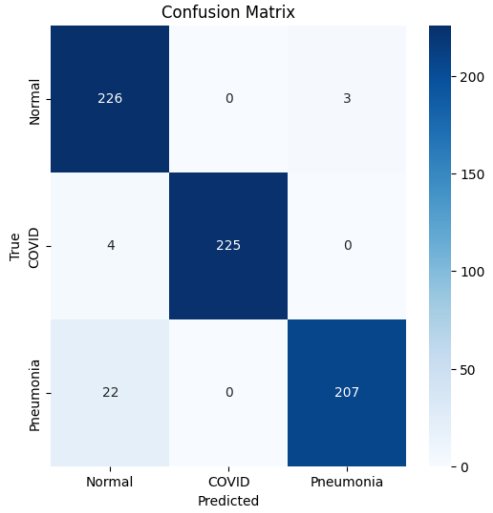
Appendices

A Model‘s Confusion Matrixes

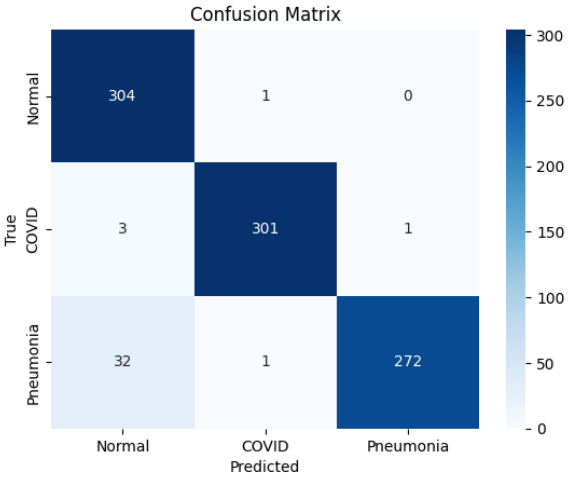


(a) VGG-16

(b) VGG-19

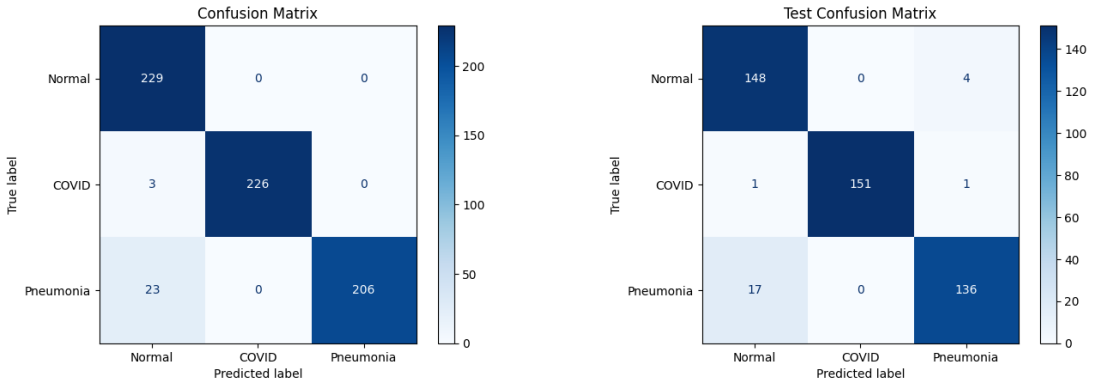


(c) ResNet-18



(d) ResNet-50

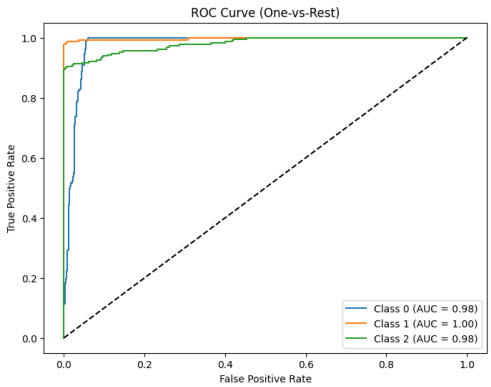
Figure 1: Confusion matrices for VGG and ResNet models



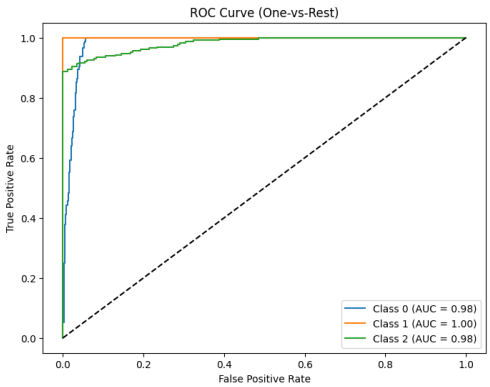
(a) Transfer Learning **(b)** Ensemble Model

Figure 2: Confusion matrices for Transfer Learning and Ensemble models

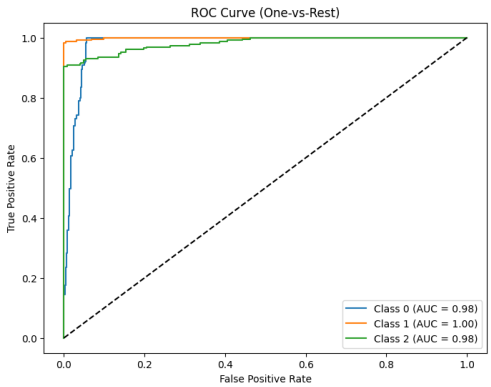
B ROC Curve



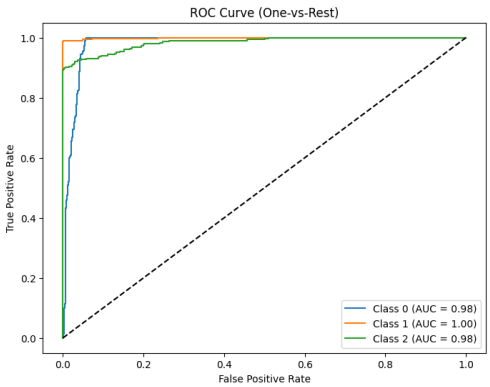
(a) VGG-16 (AUROC: 0.9847)



(b) VGG-19 (AUROC: 0.9882)

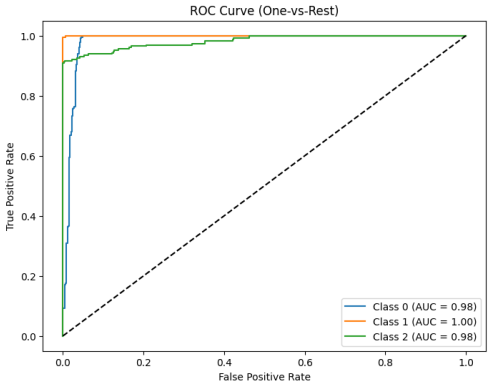


(c) ResNet-18 (AUROC: 0.9869)

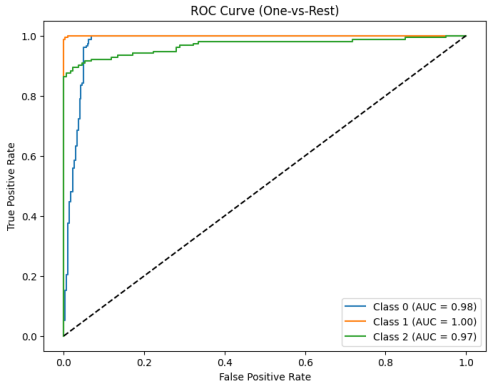


(d) ResNet-50 (AUROC: 0.9882)

Figure 3: ROC Curves for VGG and ResNet Models



(a) Transfer Learning (AUROC: 0.9883)



(b) Ensemble Model (AUROC: 0.9812)

Figure 4: ROC Curves for Transfer Learning and Ensemble

References

Abad, M., et al. (2024). Generalizable disease detection using model ensemble on chest x-ray images. *Scientific Reports*, *14*(1), 56171.

Aburass, S., Alotaibi, S., Alghamdi, A., et al. (2025). Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases. *Journal of Imaging Informatics in Medicine*.

Aleksandr, Z., Yousif, H., Simonov, K., & Kurako, M. (2019). Lung boundary detection for chest x-ray images classification based on glcm and probabilistic neural networks. *Procedia Computer Science*, *156*, 22–35.

Alotaibi, A. (2025). Ensemble deep learning approaches in health care: A review. *Computers, Materials & Continua*, *82*(3).

Asraf, A., & Islam, Z. (2021). Covid-19, pneumonia, and normal chest x-ray pa dataset [Accessed: 2025-07-25]. https://doi.org/10.17632/jctsfj2sfn.1

Azad, R., Asadi-Aghbolaghi, M., Fathy, M., & Escalera, S. (2023). Advances in medical image analysis with vision transformers: A comprehensive review. *Computers in Biology and Medicine*, *152*, 106391.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, *16*(5), 412–424.

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific Reports*, *9*(1), 6381. https://doi.org/10.1038/s41598-019-42294-8

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., & Murphy, K. (2021). Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, *72*, 102125.

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J., Singh, R., Xue, Z., Karargyris, A., Antani, S., Thoma, G., & McDonald, C. (2013). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, *33*(2), 577–590.

Channin, D. S., Mongkolwat, P., Kleper, V., Rubin, D. L., & Desai, N. (2012). Automatic selective removal of embedded patient information from image content of dicom files. *AJR. American Journal of Roentgenology*, *198*(4), W400–W407. https://doi.org/10.2214/AJR.11.6774

Chen, C., et al. (2025). A review of convolutional neural network based methods for medical image classification. *Journal of Healthcare Engineering*, *2025*, 1–18.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6.

Di Noto, T., et al. (2022). Weakly supervised learning with automated labels from radiology reports. *arXiv preprint arXiv:2210.09698*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International workshop on multiple classifier systems*, 1–15.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ennab, M., & Mcheick, H. (2025). Advancing ai interpretability in medical imaging: A comparative analysis of pixel-level interpretability and grad-cam models. *Machine Learning and Knowledge Extraction*, *7*(1), 1–32. https://doi.org/10.3390/make7010012

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., & other. (2020). Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, *382*, 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Hahn, A., Meier, F., Niopek, D., Wegner, F., Horning, O., Ziegler, S., Siegel, D., Erfle, H., & Rohr, K. (2020). Weakly supervised learning of single-cell feature embeddings. *Bioinformatics*, *36*(Supplement_1), i277–i285.

He, K., Zhang, X., Ren, S., & Sun, J. (2015a). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*. https://arxiv.org/pdf/1512.03385.pdf

He, K., Zhang, X., Ren, S., & Sun, J. (2015b). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hedberg, P., et al. (2024). In-hospital mortality during the wild-type, alpha, delta, and omicron sars-cov-2 waves: A multinational cohort study in the eucare project [Variant-period cohort; 28-day in-hospital mortality trends]. *The Lancet Regional Health – Europe*.

Henry, E. U., Zhang, Y., & Wang, J. (2022). Vision transformers in medical imaging: A review. *Medical Image Analysis*, *79*, 102445.

Hogeweg, L., Sánchez, C. I., Melendez, J., Maduskar, P., Story, A., Hayward, A., & van Ginneken, B. (2013). Foreign object detection and removal to improve automated analysis of chest radiographs. *Medical Physics*, *40*(7), 071901.

Huang, C., Wang, Y., Li, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Ikechukwu, V., Murali, S., Deepu, R., & Shivamurthy, R. (2021). Resnet-50 vs vgg-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images. *ScienceDirect*.

Jia, H., et al. (2024). Application of convolutional neural networks in medical imaging. *Frontiers in Oncology*, *14*, 1345678.

Khan, A. A., et al. (2024). A review of ensemble learning and data augmentation models for class imbalance problems. *Expert Systems with Applications*, *243*, 120151.

Kim, J., et al. (2021). Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. *Scientific Reports*, *11*(1), 24179.

Kourounis, G., et al. (2023). Computer image analysis with artificial intelligence: A practical guide for clinicians. *European Respiratory Review*, *32*(167), 230052.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.

Labbaf Khaniki, M. A., & Manthouri, M. (2024). A novel approach to chest x-ray lung segmentation using u-net and modified convolutional block attention module. *arXiv preprint arXiv:2404.14322*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Mansoor, A., Bagci, U., Foster, I., & Xu, R. (2015). Segmentation of lung fields in chest radiographs using anatomical atlases with nonrigid registration. *Medical Physics*, *42*(9), 5303–5313.

Metlay, J. P., Waterer, G. W., Long, A. T., et al. (2019). Diagnosis and treatment of adults with community-acquired pneumonia: An official clinical practice guideline of the american thoracic society and

infectious diseases society of america. *American Journal of Respiratory and Critical Care Medicine*, *200*(7), e45–e67. https://doi.org/10.1164/rccm.201908-1581ST

Mienye, I. D., et al. (2025). Deep convolutional neural networks in medical image analysis: A review. *Artificial Intelligence in Medicine*, *148*, 102653.

Misera, L., Müller-Franzes, G., Truhn, D., & Kather, J. N. (2024). Weakly supervised deep learning in radiology. *Radiology*, *312*(1), e232085. https://doi.org/10.1148/radiol.232085

Mohanty, M. R., Mallick, P. K., & Reddy, A. V. N. (2024). Optimizing pulmonary chest x-ray classification with stacked feature ensemble and swin transformer integration. *Biomedical Physics & Engineering Express*, *11*(1).

Müller, D., Soto-Rey, I., & Kramer, F. (2022). An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *IEEE Access*, *10*, 73888–73900.

Nguyen, D.-K., Lan, C.-H., & Chan, C.-L. (2021). Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: The workflows, deployments, and surveys on the statistical, image-based, and sequential datasets. *International Journal of Environmental Research and Public Health*, *18*(20), 10811.

Pan, Z., & Chen, Y. (2023). A fusing transformer and cnn on interpretable covid-19 detection. *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, 410–419. https://doi.org/10.2991/978-94-6463-264-4_46

Panwar, H., Gupta, P., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. *Chaos, Solitons Fractals*, *140*, 110190.

Powers, D. M. W. (2020). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*. https://arxiv.org/abs/2010.16061

Preprint, S. (2023). *Transformers for medical image analysis: Applications, challenges, and future scope* (tech. rep.). SSRN.

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, *32*, 3342–3352.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*. https://arxiv.org/abs/1711.05225

ReportLinker. (2024). Forecast: Pneumonia mortality in germany (2024–2028) [19.24 thousand deaths in 2024].

Reyes, M., Meier, R., Pereira, S., Silva, C., Dahlweid, F.-M., von Tengg-Kobligk, H., Vollmuth, P., & Wiest, R. (2020). Interpretability of deep learning models for healthcare: A review. *IEEE Transactions on Medical Imaging*, *39*(11), 3830–3840.

Robb, E., Smith, J., Lee, A., & Kim, D. (2021). Weakly supervised segmentation for disease localization in chest x-rays. *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 1234–1238.

Salehi, A., Salehi, M., et al. (2023). A study of cnn and transfer learning in medical imaging. *Sustainability*, *15*(7), 5930.

Sardar, B. (2023). Remove text from image using image inpainting and ocr [Accessed: 2025-07-22].

Sarvamangala, D. R., & Kulkarni, R. V. (2021). Convolutional neural networks in medical image analysis: A survey. *Journal of Big Data*, *8*(1), 1–54.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017a). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*. https://arxiv.org/abs/1610.02391

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017b). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Sharma, S., Joshi, S., Gautam, G., & Sharma, V. (2017). Image inpainting for gap filling and text abstraction by using optical character recognition. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), 160–163.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. https://arxiv.org/abs/1409.1556

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Smith, ˙a.. (2024). A retrospective claims data analysis on the burden of covid-19 hospitalizations in germany [Inpatient mortality = 18.9 % as of 2024]. *Infectious Diseases Therapy*.

Stodt, J., Madan, M., Reich, C., Filipovic, L., & Ilijas, T. (2023). A study on the reliability of visual xai methods for x-ray images [Available at https://d-nb.info/1297297903/34].

Suara, S., Jha, A., Sinha, P., & Sekh, A. A. (2023). Is grad-cam explainable in medical images? *arXiv preprint arXiv:2307.10506*. https://arxiv.org/pdf/2307.10506.pdf

Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2024). A review on image based diagnosis using resnet-50. *International Journal of Research and Analytical Reviews (IJRAR)*, *11*(1). https://www.ijrar.org/papers/IJRAR24A3421.pdf

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2017). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Vellandurai, S., Shiraskar, S., & Rizk, D. (2023). Comparison of resnet and vgg architectures in brain tumor detection and classification. *Center for Advanced Research in Computer Engineering, The Catholic University of America*.

Vligade, P. (2020). Understanding resnet: A deep learning architecture [Accessed: 2025-07-25].

Wikipedia contributors. (-). Convolutional neural network [Accessed 2025-07-25].

World Health Organization. (2020). Who clinical management of covid-19: Interim guidance [Accessed: 2025-08-14]. https://www.who.int/publications/i/item/clinical-management-of-covid-19

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, *27*, 3320–3328.

Yu, X., Wang, J., Hong, Q.-Q., Teku, R., Wang, S.-H., & Zhang, Y.-D. (2022). Transfer learning for medical images analyses: A survey. *Neurocomputing*, *489*, 135–154.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *CVPR*, 2921–2929.

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC press.

Zoelden, A., Müller, A., Kopp-Schneider, A., Meinzer, H.-P., & Maier-Hein, K. H. (2020). Automatic removal of text labels from chest x-ray images using deep learning. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 765–774.