BABEŞ-BOLYAI UNIVERSITY CLUJ-NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
SPECIALIZATION COMPUTER SCIENCE IN ENGLISH

DIPLOMA THESIS

Interpretable Deep Learning for Chest X-ray Disease Classification and Visualization

Supervisor
Prof., Dr. Darabant Sergiu Adrian

Author
Pop Cristian-Andrei

2025

UNIVERSITATEA BABEŞ-BOLYAI CLUJ-NAPOCA
FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ
SPECIALIZAREA INFORMATICĂ ÎN LIMBA ENGLEZĂ

LUCRARE DE LICENŢĂ

Învățare Profundă Interpretabilă pentru Clasificarea și Vizualizarea Afecțiunilor din Radiografiile Toracice

Conducător științific
Prof., Dr. Darabant Sergiu Adrian

Absolvent
Pop Cristian-Andrei

2025

ABSTRACT

The field of medical imaging is witnessing a notable increase in attention towards AI-enhanced diagnostics. This trend is largely driven by the core challenges linked to analyzing intricate medical images. A primary example is chest X-rays (CXRs), which, although widely used, pose challenges in attaining both precision and efficacy. This is due to the substantial case volume and the frequently subtle manifestations of pathologies.

This thesis proposes addressing the issue of diagnostic variability through the development of an interpretable deep learning model for thoracic pathology analysis. Starting from fine-tuning established single-task models' architecture, this work explores a multi-task learning approach, where the neural network is trained to both classify pathologies and predict their location. The core hypothesis is that this dual objective forces the model's shared backbone to learn more clinically relevant features. This method, guided by categorization and localization cues, ensures that the model develops a robust understanding of illness manifestation. The multi-task model attained a mean classification AUC of 0.789 across 14 diseases in the NIH ChestX-ray14 dataset, demonstrating commendable performance relative to single-task baseline models.

To demonstrate the practical application of developed AI, it has been integrated into a comprehensive, complete web application. The application is built with React and Firebase and provides a comprehensive workflow for the management of patients and cases. Clinicians can upload CXR images, review AI-generated multi-label classifications, and directly visualize model reasoning through Grad-CAM heat maps in the interactive case management. The application's main innovation is advanced annotation functionality, which allows clinicians to add their own diagnostic notes and draw accurate borders on images.

The accuracy and interpretation of AI models combined with web application capabilities can improve diagnostic workflow efficiency and improve clinician confidence in AI-driven tools. The application's privacy-preserving architecture, which uses TensorFlow.js to perform all client-side AI inferences, addresses key data security issues. Furthermore, the integration of annotation and data export capabilities transforms the system from a simple diagnostic tool to a dynamic research tool, allowing the creation of new high-quality data sets, which create a powerful feedback cycle for improving future models.

Contents

Chapter 1

Introduction

1.1 Context and Motivation

The field of medical imaging is seeing a substantial transformation accelerated by advances in Artificial Intelligence (AI), especially deep learning methodologies. These technologies possess the capability to augment diagnostic precision, optimize clinical processes, and ultimately enhance patient outcomes. One of the most widely used diagnostic procedures in the world is Chest X-rays (CXRs) because of the vital information about thoracic conditions they can provide. The interpretation of chest X-rays can be complex, requiring significant effort and often leading to differences in interpretation among observers, necessitating the development of automated assistance systems to aid physicians.

The motivation for this study arises from a significant issue in medical AI known as the ”black box” dilemma. Although deep learning models have demonstrated significant effectiveness in classification tasks, their internal decision-making processes remain obscure. For an AI tool to be genuinely helpful and reliable in a clinical environment, it must deliver precise forecasts and provide transparent, comprehensible justifications for its reasoning. This thesis addresses this need by focusing on the development of a system that is both diagnostically accurate and highly interpretable.

1.2 Objectives

The main objective of this thesis is to perform an in-depth exploration of creating an interpretable AI system for analyzing chest X-rays, encompassing everything from model testing to the deployment of a functional software application.

This investigation leads to the design and comparative evaluation of several AI models tailored for multi-label thoracic pathology classification. The research ex-

plores various popular Convolutional Neural Network (CNN) architectures, fine-tuned on a large-scale dataset. This leads to a key objective, precisely testing the hypothesis that a multi-task learning model, which learns to categorize X-rays based on disease while also localizing them, can outperform single-task models.

Furthermore, the thesis aims to implement and validate robust explainability techniques. A main goal is to use Gradient-weighted Class Activation Mapping (Grad-CAM) to generate visual heatmaps that provide insight into the model's decision-making processes and to evaluate and assess the faithfulness of these explanations against ground-truth localization data.

A web application will additionally be developed to demonstrate the practical applicability of the developed AI models. This application enables clinicians to manage patient records, upload X-ray images for analysis, and review AI-generated classifications and visual explanations. A novel objective of this system is to incorporate tools that enable clinicians to add their own annotations, including bounding boxes, thereby creating a framework for curating new, high-quality datasets for future research.

By addressing these objectives, the thesis aims to make important progress in the development of trustworthy artificial intelligence in medicine, not only through an effective diagnostic model but also through a practical and confidential software tool that supports clinical workflows and facilitates ongoing research.

Our main contributions include a rigorous comparative analysis of multiple CNN architectures (DenseNet121, EfficientNetB0, ResNet50, and MobileNetV2) and the demonstration of a substantial performance increase by moving from a single-task to a multi-task learning paradigm for chest X-ray classification. To our knowledge, the application of this specific multi-task approach, combining classification with a Faster R-CNN detection head on this dataset, represents a significant experimental contribution. Another significant contribution is the successful development of a comprehensive, privacy-preserving clinical support application featuring client-side AI insights. This architecture ensures that sensitive patient images are never sent to the server for analysis, directly addressing key concerns about data privacy. Finally, this thesis is a solution to the problem of data scarcity in medical AI by directly integrating data collection tools into the application. The data can be exported into data sets to improve the model in the future.

1.3 Thesis Structure

The thesis consists of multiple chapters, each concentrating on a distinct aspect of the research. The second chapter presents a comprehensive summary of relevant studies. It provides a comprehensive overview of deep learning in medical imag-

ing, highlights recent research, and opens the way for contributions to this work. The theoretical basis of the methods used is described in Chapter 3, "Theoretical Foundations".

Following these sections, the thesis delves into the practical development. It covers the methodology for the AI experiments, including the design of the single-task and multi-task models, and describes the architecture and implementation of the developed web application.

The experiments conducted in this thesis are then presented. This chapter provides a comprehensive analysis of the performance of the artificial intelligence models, including a comparison with state-of-the-art models and an overview of the explainability techniques used.

The final section of the thesis offers a thorough conclusion. This part outlines the main findings and contributions of the study, admits its shortcomings, and explores several avenues for its further development and expansion. This systematic approach ensures that the thesis delivers an in-depth analysis of the subject, yielding a thorough comprehension of the research undertaken and setting the stage for future advancements in this area.

1.4 Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Gemini in order to improve the clarity, structure, and academic tone of the text. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the thesis.

Chapter 2

Background and Literature Review

This chapter examines key developments in medical imaging and computer vision, with a particular focus on deep learning applications for chest X-ray analysis. Through a thorough review of the existing literature, we present the most important concepts, methodologies, and significant findings related to automated disease classification and localization in chest radiographs. We analyze the characteristics and importance of widely utilized datasets in this domain, such as ChestX-ray14. In addition, we examine current state-of-the-art models and methods for multi-label classification and interpretable localization methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM). Finally, the pertinent issues and ongoing challenges within this rapidly evolving field, particularly concerning dataset limitations, model generalizability, and clinical integration, are discussed.

2.1 Deep Learning in Medical Imaging

Artificial Intelligence (AI) has advanced rapidly in recent years, with machine learning (ML) gaining popularity and finding widespread use. Key developments include the evolution of neural networks (NNs) and, subsequently, deep neural networks (DNNs). They have been implemented in various areas, including medical imaging for a range of applications like image classification, disease recognition, and segmentation.

Medical imaging has been transformed by deep learning, which has automated procedures that formerly required human skill. The hierarchical structure of the DNNs allows them to extract details in the medical images, such as edges and textures, in order to find patterns that describe pathological conditions. This capability makes DNNs very useful in the identification of tumors, the detection of fractures, the segmentation of organs, and the evaluation of the course of diseases [MSS+23].

In medical imaging, DNN-based solutions have generated a great deal of interest

worldwide because of their promising outcomes. Deep learning development in the medical field follows a slower but similar path to computer vision. However, because medical and natural images differ, applying conventional computer vision techniques could not yield satisfactory results. To achieve positive outcomes, issues unique to medical pictures must be resolved [MSS+23].

Figure 2.1 demonstrates the application of trained deep neural networks (DNNs) across various stages of medical analysis, beginning with raw images and progressing to more specialized functions.

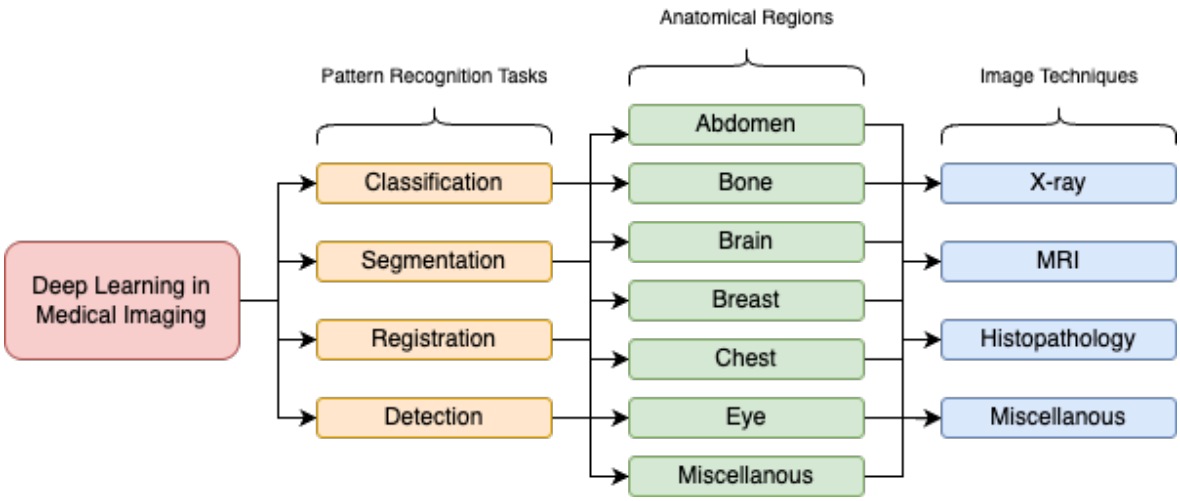

Figure 2.1: Diagram illustrating the application of DNNs in medical imaging.

2.1.1 Overview of Common Deep Learning Architectures

Convolutional Neural Networks (CNNs) with foundational concepts formalized in the late 1990s (e.g., LeCun et al., 1998), form the bedrock of many deep learning approaches in image analysis [LBBH98]. The key components of CNN-based frameworks typically include:

Convolution: Convolutional layers are responsible for the essential task of feature extraction by applying filters (or kernels) to the input image. These filters demonstrate a high degree of proficiency in identifying patterns such as edges, corners, and textures. This capability is achieved through the calculation of dot products between the filter's weights and the localized receptive fields of the input. This process culminates in a collection of feature maps that illustrate how each filter responds to the input image, thereby capturing localized spatial features to be used in classification tasks [GBC16].

Pooling: Pooling layers, such as max-pooling, reduce the feature maps' spatial dimensions. For each feature map, they operate on small rectangular areas (typically 2x2 or 3x3) and retain the most significant feature within each pooling window.

Max-pooling maintains the most critical aspects by selecting the highest activation. It also boosts translational invariance, making the network more robust to minor spatial variations [GBC16].

Fully Connected Layers (FC): The integration of information gathered from the convolutional and pooling layers enables the identification of global relationships in the data. These layers play an essential role in decision-making by merging relevant details to predict higher-level representations [MSS+23].

Activation: This layer's role is to transform the previous layer's outputs into a probability distribution over the target classes. This process enables the network to assign a probability to each class and subsequently determine the predicted label [MSS+23]. Softmax is a widely adopted activation function for mutually exclusive multi-class problems; however, in multi-label classification tasks where multiple categories may be assigned to a single input, alternative activation methods such as per-class Sigmoid functions are typically employed at the output layer.

In order to classify the image, the softmax activation function generates a probability distribution across different classes, convolutional layers find important features, pooling layers downsample while preserving important information, and fully connected layers capture general associations.

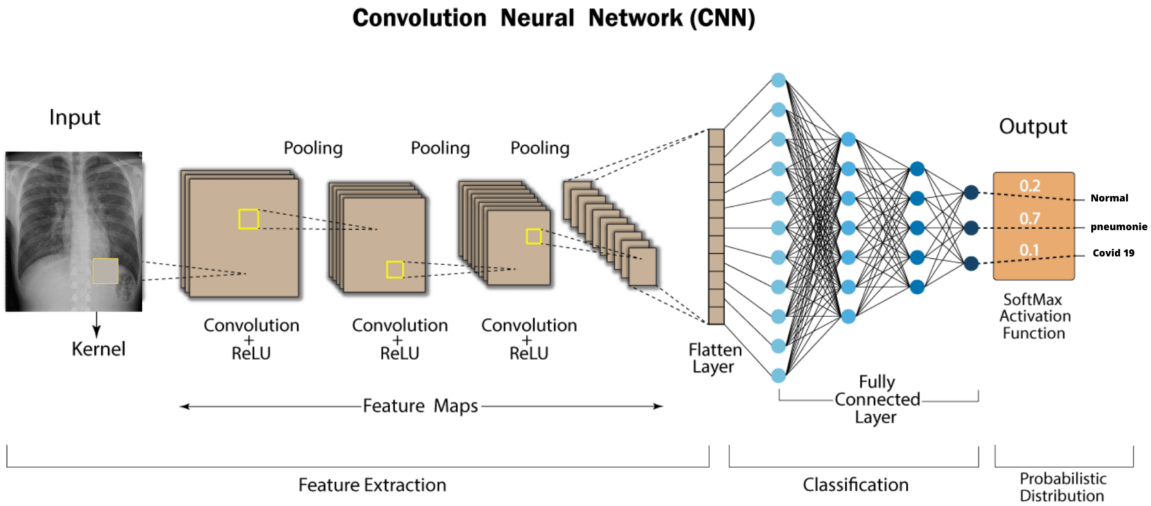

Figure 2.2: Convolutional Neural Network for disease classification [RHZ+22].

Recurrent Neural Networks (RNNs): In contrast with Convolutional Neural Networks, RNNs are designed for handling sequential data. In the domain of medical imaging, recurrent neural networks (RNNs) have demonstrated significant value in the analysis of time-series data, including dynamic contrast-enhanced magnetic resonance imaging (MRI) and video fluoroscopy. Their capacity to sustain an internal state, or memory, enables them to assess input sequences and discern temporal relationships, a capability that is indispensable. By addressing the vanishing gradient problem, variations such as Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks facilitate the network's ability to learn dependencies over extended periods. [GBC16].

Transformers: Initially created for natural language processing, transformers have lately shown exceptional capabilities in various computer vision applications, particularly in analyzing medical images. Transformers operate using a self-attention mechanism that enables the model to assess the significance of different segments within the input sequence during prediction. This feature allows Transformers to effectively capture both long-range relationships and the overall context of images, often surpassing traditional CNNs that mainly focus on local receptive fields [SKZ+23]. Within the Vision Transformer (ViT) framework [Dos20], the input image is generally segmented into a grid of fixed-size, non-overlapping patches (e.g., 16×16 pixels). Each patch is then flattened and linearly projected into an embedding, forming a sequence of visual tokens. These tokens are enhanced with positional embeddings to maintain spatial information, after which they are processed by a standard Transformer encoder. As a robust alternative to CNN-based structures, transformers are increasingly being investigated for analyzing chest X-rays. Recent studies emphasize their ability to model global context, which is advantageous for tasks like pathology classification and localization of chest radiographs. Some research shows that their performance is comparable to or exceeds that of CNN-based methods. [SKZ+23].

Hybrid Architectures: Often, the best performance is achieved by combining different architectures. Beyond coupling different network types, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for tasks like generating automated radiology reports from chest X-rays [YQW+19], or CNN-Transformer models like TransUNet [CLY+21] for improved medical image segmentation, hybrid approaches also include the integration of specialized modules within a primary architecture. One significant example is the integration of attention mechanisms within CNNs. Components such as the Convolutional Block Attention Module (CBAM) [WPLK18] enhance CNNs by enabling them to focus on more informative spatial and channel features. CBAM achieves this by sequentially deriving one-dimensional (1D) channel attention maps, followed by two-dimensional (2D)

spatial attention maps, which subsequently recalibrate the input feature maps. This not only enhances the representational ability of the model but also provides a level of inherent interpretability, as these attention maps can be visualized to identify which regions or channels the model focuses on during predictions. These hybrid approaches leverage the strengths of different architectural elements to address the unique challenges of chest radiograph analysis.

2.2 Applications in Chest X-ray Anomaly Detection

Chest X-ray imaging is one of the most extensively available diagnostic tools and has significantly benefited from deep learning approaches for anomaly detection. The efficacy of deep learning in this domain stems from its capacity to process large datasets, identify subtle features, and provide automated and scalable diagnostic support. This section delves into specific examples of how various deep learning architectures have been applied to chest X-ray anomaly detection.

CNNs for Disease Detection: Due to their ability to automatically learn spatial hierarchies of features, CNNs have become the standard for image-based anomaly detection in chest X-rays. There are several studies that demonstrate their effectiveness.

- **Pneumonia Detection:** Rajpurkar et al. [Raj17] developed CheXNet, a 121-layer DenseNet CNN trained on the ChestX-ray14 dataset, which showed performance comparable to that of radiologists in the detection of pneumonia. This study showed how deep learning could be used to automate the diagnosis of pneumonia.

- **Tuberculosis Screening:** Lakhani and Sundaram [LS17] used a CNN-based approach for automated classification of pulmonary tuberculosis from chest X-rays. Their model achieved high accuracy in distinguishing between normal and tuberculosis-affected images, highlighting the potential for large-scale screening in resource-limited settings.

- **COVID-19 Detection:** Numerous studies explored CNNs for COVID-19 detection using chest X-rays. For instance, Wang et al. [WLW20] proposed a modified ResNet architecture called COVID-Net, specifically designed for COVID-19 detection. This network incorporated a projection-expansion-projection design pattern to improve feature representation and achieved promising results in differentiating COVID-19 from other lung diseases.

Transformers and Vision Transformers (ViTs): Transformers, particularly Vision Transformers (ViTs), are gaining traction in medical image analysis, offering

advantages in capturing long-range dependencies and global context:

- **Anomaly Localization:** While CNNs excel at local feature extraction, Transformers can better capture relationships between distant image regions. This is crucial for localizing subtle anomalies. Although direct comparison studies focusing solely on localization with chest X-rays are still emerging, the general advantage of Transformers in capturing global context suggests their potential in this area. Some studies have applied transformers for segmentation tasks in chest X-rays, which inherently provides localization information [CLY+21].
- **Superior Performance in Specific Scenarios:** Chen et al. [CLY+21] introduced TransUNet, a hybrid CNN-Transformer architecture for medical image segmentation, including chest X-rays. This model replaced the CNN encoder in a traditional U-Net with a Transformer, demonstrating improved segmentation accuracy by leveraging the Transformer's ability to model long-range dependencies. This suggests that Transformers can outperform CNNs in tasks requiring global context understanding.

Hybrid Models: Combining different architectures can leverage their complementary strengths:

- **CNN-RNN for Report Generation:** Some studies combine CNNs with RNNs for automated report generation from chest X-rays. The CNN extracts visual features from the image, while the RNN generates a textual description of the findings. This approach can help radiologists with report writing and improve consistency. [YQW+19].
- **CNN-Transformer for Enhanced Feature Representation:** As mentioned earlier, TransUNet [CLY+21] is a prime example of a CNN-Transformer hybrid. Replacing the CNN encoder with a Transformer allows the model to benefit from the local feature extraction capabilities of CNNs and the global context modeling of Transformers. This approach has shown promising results in segmentation tasks, which are closely related to anomaly detection.

2.3 Datasets

Datasets are fundamental for developing and evaluating deep learning methods in anomaly detection for chest X-rays. The following table provides a summary of commonly used datasets in this domain, highlighting their key characteristics.

Dataset	No. Images	Class Labels	Annotations	Key Highlights
ChestX-ray14	112,120	14 thoracic diseases	Bounding boxes for some	Large-scale, imbalanced dataset.
CheXpert	224,316	14 observations	Uncertainty labels	High-quality labels; designed for AI tasks.
MIMIC-CXR	377,110	13 observations	Free-text reports	Rich metadata; publicly available.
PadChest	160,868	174 labels	Multi-label annotations	Includes lung segmentation masks.
VinDr-CXR	18,000	6 abnormalities	Localization boxes	Focus on localization and pathology.
COVIDx	13,975	COVID-19, pneumonia	Binary or multi-class	Specifically for COVID-19 research.

Table 2.1: Summary of commonly used datasets in anomaly detection for chest X-rays.

2.3.1 ChestX-ray14

ChestX-ray14 is a medical imaging dataset that comprises 112,120 frontal-view X-ray images belonging to 30,805 unique patients that were gathered between 1992 and 2015, featuring fourteen common disease labels extracted from radiological reports through natural language processing (NLP) methodologies. The incorporation of six additional thoracic diseasesnamely, edema, emphysema, fibrosis, pleural thickening, and herniasignificantly expands the scope of ChestX-ray8.

2.3.2 CheXpert

CheXpert is a large-scale chest radiograph dataset containing 224,316 radiographs from 65,240 patients, labeled for 14 common observations. The labels are extracted from free-text radiology reports using a rule-based labeler that classifies observations as positive, negative, or uncertain. The dataset includes a validation set that was annotated by three board-certified radiologists and a test set with consensus labels from five radiologists, enabling rigorous benchmarking. [IRK+19].

2.3.3 MIMIC-CXR

MIMIC-CXR [WJJ+19] is another extensive chest radiograph dataset that contains 227,835 imaging studies of 65,379 patients made between 2011 and 2016. Each imaging study may contain one or more images, typically in the form of front and lateral

views. The dataset under consideration contains a total of 377,110 images. To create MIMIC-CXR, three distinct data modalities had to be managed: images (chest radiography), natural language (free-text reports), and data extracted from electronic health records.

2.3.4 PadChest

PadChest [BPSDLIV20] is a significant, high-resolution dataset of chest x-rays labeled for the purpose of automated exploration of medical images and associated reports. It consists of over 160,000 images collected from 67,000 patients, covering six different positional views along with additional information on image acquisition and patient demographics. These images were analyzed and reported by radiologists at Hospital San Juan in Spain from 2009 to 2017. The reports encompass 104 anatomical regions, 19 differential diagnoses, and 174 unique radiographic abnormalities, which have been mapped to the standard nomenclature of the Unified Medical Language System (UMLS) and organized in a hierarchical structure. Out of the reports, 27% were annotated manually by certified doctors, whereas the rest were labeled through a supervised method utilizing a recurrent neural network with attention mechanisms.

2.3.5 VinDr-CXR

The VinDr-CXR dataset [NLL+20] is a meticulously curated collection of postero-anterior chest X-ray images, designed to advance research to better detect and identify thoracic pathologies. The dataset, obtained from two important Asian hospitals, includes localization for 22 significant findings (marked with bounding boxes) as well as classification for six prevalent thoracic conditions. The dataset is divided into two segments: a training subset containing 15,000 images and a testing subset with 3,000 images.

2.3.6 COVIDx

The COVIDx dataset, employed in creating the COVID-Net model [LQA20], is an extensive compilation aimed at identifying COVID-19 cases through chest X-ray images. It is categorized into three classes for healthy individuals, pneumonia not related to COVID-19, and cases confirmed positive for COVID-19. This dataset comprises 13,975 images from 13,870 unique patients. This dataset aggregates data from different sources, such as Covid-ChestXray and COVID-19 CXR datasets, improving both the quantity and diversity of cases. Using the COVIDx dataset, the authors

of COVID-Net were able to train their deep convolutional neural network effectively, improving its performance in identifying COVID-19 cases. This emphasizes the dataset's importance as a useful tool for developing automated medical imaging diagnostics.

2.4 Techniques for Anomaly Detection

Deep learning presents various methods specifically designed for detecting anomalies in chest X-rays, which can be mainly classified into supervised, unsupervised, and self-supervised learning. Each of these categories tackles unique challenges and utilizes distinct features of deep learning models.

2.4.1 Supervised Learning

Supervised methods require labeled data, where each X-ray is labeled with the presence or absence of specific anomalies (e.g., pneumonia, nodules, effusions). These methods frame anomaly detection as a classification problem. During training, the model learns the underlying patterns and characteristics that differentiate normal and abnormal images. Subsequent to the training phase, the model can be utilized to predict the presence or absence of anomalies in new, unseen chest X-rays. These models necessitate substantial data to achieve high accuracy in detecting abnormalities, as they depend on labeled data, a process that can be costly and time-consuming, particularly for rare anomalies.

A notable example of supervised learning in medical imaging is CheXNet, a Dense Convolutional Neural Network (DenseNet) with 121 layers that was trained on the ChestX-ray14 dataset [Raj17]. CheXNet makes use of labeled data for pneumonia detection, achieving performance comparable to that of radiologists in a classification task involving two categories. Initially, the network was pretrained using ImageNet before it underwent fine-tuning on the ChestX-ray14 dataset, which features more than 100,000 labeled chest X-ray images covering 14 distinct pathologies. In addition to excelling in pneumonia detection, CheXNet exhibits state-of-the-art performance across all 14 conditions represented in the dataset.

The supervised learning technique used by CheXNet underscores the significance of extensive, well-annotated datasets for training deep learning models. The ChestX-ray14 dataset was generated by automating the extraction of labels from radiology reports, facilitating the creation of a highly accurate model without the need for manual labeling for each disease [WPL+17]. However, despite these advancements, supervised approaches are limited by the quality and variety of the labeled data available. Rare diseases, for example, pose challenges due to the scarcity of

positive examples for training.

Even though supervised learning techniques such as CheXNet have produced remarkable outcomes, their dependence on labeled datasets highlights the necessity for innovative approaches to address labeling limitations and ensure scalability, especially for uncommon conditions.

2.4.2 Unsupervised Learning

In contrast to supervised learning, unsupervised learning does not necessitate the utilization of labeled datasets. Instead, the focus is on identifying patterns present within the data to recognize anomalies. In the domain of chest X-ray (CXR) anomaly detection, unsupervised techniques usually frame the task as a one-class classification problem, where the model is trained solely on normal CXRs. This approach is particularly advantageous for CXR analysis, given that complete annotation of all potential pathologies is often unfeasible due to the infrequency of certain diseases and the prohibitive cost of expert labeling. By training exclusively on normal CXRs, unsupervised techniques strive to identify anomalies as variations from the established normal distribution.

Several deep learning techniques are employed in unsupervised CXR anomaly detection. Reconstruction-based methods use autoencoders (AEs) or Variational Autoencoders (VAEs) to learn a condensed representation of CXRs and then rebuild the input. The reconstructed error is employed as a score for anomalies, a high score denoting potential anomalies. However, these methods can sometimes reconstruct out-of-distribution samples if the normal data is too diverse [MKRBD23]. Distribution learning methods, such as Support Vector Data Description (SVDD) and Deep SVDD, model the probability distribution of normal data, defining a boundary in feature space to identify anomalies lying outside this boundary. Generative models like VQ-VAEs and autoregressive models are also used for this purpose, though they often require careful calibration [MKRBD23].

Self-supervised learning (SSL), as detailed in [MKRBD23], provides a powerful approach. By defining pretext tasks on unlabeled data, such as predicting patch positions or solving jigsaw puzzles, SSL methods learn valuable feature representations. Contrastive learning techniques like SimCLR enhance these representations by maximizing agreement between different augmented views of the same image. A key advancement is the use of multi-resolution patch-based SSL [MKRBD23]. This approach addresses the limitations of fixed-size patches by capturing anomalies of varying scales and shapes through aggregating anomaly scores from different patch resolutions.

Unsupervised learnings primary advantage is its independence from labeled

data, enabling the detection of potentially novel anomalies not seen during training. However, it can have lower accuracy than supervised methods with sufficient labeled data, particularly for subtle anomalies, and is sensitive to the quality and representativeness of the normal data. Its evaluation is also more challenging due to the lack of ground truth for anomalies. Despite these limitations, unsupervised learning remains a crucial tool in medical image analysis, especially with the promising results from self-supervised approaches like multi-resolution patch-based learning [MKRBD23].

2.4.3 Emerging Techniques

Recent studies have explored innovative methodologies such as CLIP-based approaches that leverage pre-trained models on large datasets to enhance anomaly detection capabilities. These methods adapt pre-trained models through position-guided prompt learning techniques that focus on specific lung regions during diagnosis [SGL+24]. Such advancements highlight a trend towards integrating transfer learning with traditional deep learning frameworks to improve performance across varied clinical settings.

In summary, while supervised methods excel in environments with abundant labeled data, unsupervised approaches offer flexibility and adaptability in scenarios where labeling is impractical or impossible. The integration of self-supervised techniques further enhances these capabilities by allowing models to learn robust features from unlabeled datasets effectively. As research progresses, combining these methodologies may yield even more powerful tools for detecting anomalies in chest X-rays efficiently and accurately.

2.5 Issues and Challenges

The integration of deep learning in the domain of medical image analysis has demonstrated considerable potential. However, numerous challenges must be addressed to ensure its widespread adoption and efficacy.

A pressing challenge in medical image analysis is represented by the insufficiency of large datasets that are annotated by professionals. Deep learning algorithms need a significant quantity of labeled data to perform optimally; however, acquiring such datasets within the medical field is typically a laborious, lengthy, and expensive process. The consistency and quality of annotations can greatly differ between datasets, which may introduce biases and inaccuracies during the training of models. This problem is especially evident for uncommon diseases, where there are typically few positive cases available for training purposes [KAS+24] [DDBS23].

Additionally, the process of annotation can be subjective, shaped by the knowledge and experience of the annotators (such as radiologists). Differences in interpretation can result in inconsistencies in labeling, which may negatively impact the performance of the model. Tackling these issues requires the creation of effective annotation strategies and the investigation of semi-supervised or unsupervised learning methods that can utilize unlabeled data [KAS+24].

Deep learning models that are developed using particular datasets might face difficulties in adapting to unfamiliar populations or diverse imaging protocols. Medical imaging datasets frequently demonstrate intrinsic diversity resulting from discrepancies in patient demographics, imaging techniques, data acquisition protocols, and equipment configurations. For instance, a model that has been trained on images from a specific hospital might not perform well when tested on images from a different facility because of variations in imaging technology or patient demographics [DDBS23].

The "black-box" nature of deep learning models poses significant challenges in medical applications where understanding the rationale behind predictions is crucial for clinician trust and patient safety. Many healthcare professionals are hesitant to rely on automated systems without clear explanations for their decisions. This lack of interpretability can hinder the adoption of deep learning solutions in clinical practice [KAS+24], [QKH+24].

Initiatives to improve interpretability encompass the creation of Explainable AI (XAI) methodologies, including attention mechanisms, saliency maps, and gradient-based visualization techniques. These approaches aim to elucidate how models arrive at specific predictions, thereby fostering greater transparency and trust among clinicians [QKH+24]. While achieving a perfect balance between model performance and comprehensive interpretability remains an active area of research, these XAI methods represent crucial steps towards more clinically acceptable AI.

The vulnerability of deep learning algorithms to adversarial attacks, wherein minor alterations to input images can result in substantial misclassifications, is a matter of concern. In the domain of medical imaging, such deficiencies give rise to significant concerns, as erroneous predictions have the potential to exert a substantial adverse influence on the quality of patient care.[KAS+24, DDBS23].

It is crucial to develop robust training strategies that increase model resilience against adversarial examples. Approaches like adversarial training, which include adding adversarial samples to the training data, have been suggested as possible solutions [KAS+24]. However, ensuring that models remain dependable across various conditions continues to represent a significant challenge.

The successful deployment of deep learning models in clinical settings requires seamless integration into existing workflows. Radiologists often face heavy work-

loads and time constraints; thus, any new system must not only provide accurate results but also be user-friendly and efficient [LXD+20].

Moreover, regulatory hurdles and concerns regarding data privacy further complicate the integration process. Compliance with healthcare regulations (e.g., HIPAA) is essential when handling sensitive patient data. Therefore, developing frameworks that ensure both compliance and usability is necessary for promoting widespread adoption of deep learning technologies in medical imaging [LXD+20].

Tackling challenges related to data scarcity, generalization across populations, interpretability, adversarial vulnerabilities, and integration into workflows will be crucial for enhancing the reliability and acceptance of deep learning solutions in healthcare settings. Ongoing research efforts aimed at developing innovative methodologies and frameworks will play a pivotal role in overcoming these obstacles and unlocking the transformative power of deep learning in medical imaging.

Chapter 3

Theoretical Foundations

3.1 Fundamentals of Deep Learning

Deep Learning, a significant domain of machine learning, utilizes various neural network designs to identify complex patterns and facilitate automated decision-making. Common training methodologies include Supervised Learning, which relies on labeled datasets, and Unsupervised Learning, which extracts insights from unlabeled data. The research presented in this thesis adopts a Supervised Learning framework specifically for multi-label classification, where each chest X-ray image is associated with a set of target labels indicating various pathologies.

Neural Networks are composed of neurons or nodes that are coupled to one another. These are grouped into layers and connected by layers. The weights of a matrix are learned during a linear transformation that occurs in each layer. The information received by different layers is subject to alteration by various factors. It is important to note that signals may traverse a series of intermediary levels, which are occasionally designated as ”hidden layers.” These layers are situated between the initial and final layers. The neurons within each layer utilize a set of weights and biases to generate an output signal subsequent to receiving input signals from the preceding layer. This signal is then transmitted to the subsequent layer. A network is frequently referred to as a ”deep neural network” if it has more than one hidden layer.

Deep learning models use a robust mathematical framework to extract intricate patterns from data. Vectors, matrices, and tensors are essential building blocks of linear algebra, which they utilize to describe data and transformations. Calculus is essential for optimization during training, especially when it comes to derivatives and gradients. Furthermore, a variety of probabilistic models and loss functions are supported by probability theory.

The optimization of machine learning models typically involves the implemen-

tation of gradient-based methods, such as stochastic gradient descent (SGD), and variations thereof, including Adam. These approaches seek to reduce the loss function by iteratively adjusting the parameters of the network. The calculation of the gradients necessary for these updates depends significantly on backpropagation.

Loss functions measure the discrepancy between the target and the model's prediction. A common choice for such a function in the context of multi-label classification would be Multi-Label Soft Margin Loss or Binary Cross-Entropy.

3.2 Transfer Learning and Fine-Tuning

The paucity of annotated datasets constitutes a significant challenge in the domain of medical machine learning. Transfer learning is a method that addresses this issue by leveraging models that have been pretrained on large-scale, general-purpose datasets, such as ImageNet, and adapting them to domain-specific tasks. In this research, multiple architectures, including DenseNet, MobileNetV2, EfficientNet, and ResNet, were fine-tuned on the ChestX-ray14 dataset for multi-label classification. DenseNet, in particular, has garnered recognition [Raj17] for its densely connected structure, a feature that has been demonstrated to enhance feature reuse and mitigate the number of parameters in comparison to other architectures.

3.2.1 Fine-Tuning Strategies for Pre-trained Models

Fine-tuning a pre-trained model involves adapting its learned parameters to the new dataset and task. The extent and manner of this adaptation can vary, leading to different strategies:

- **Feature Extraction (Frozen Base Model):** In this approach, the convolutional base of the model that was pre-trained (e.g., DenseNet121) is "frozen", meaning its weights are not updated during training on the new task. Only the weights of a newly added classifier head, which is usually one or more connected layers placed on top of the frozen base, are trained. The pre-trained base operates in the capacity of a static feature extractor, leveraging its knowledge of general visual patterns. This strategy is often a good starting point, especially when the target dataset is scarce and considerably divergent from the pre-training dataset, as it prevents the pre-trained weights from being corrupted by potentially large error signals from a randomly initialized head.
- **Fine-Tuning Specific Layers (Partial Unfreezing):** A common and effective strategy, and the one adopted in this thesis, involves a two-stage process:

1. Initially, the layers of the base model are frozen and only the new classifier head is trained (as in the feature extraction approach). This allows the head to learn to interpret the features provided by the pre-trained base.
2. Subsequently, some of the top layers of the pre-trained base model are "unfrozen" (their weights are made trainable), and the model (both the unfrozen base layers and the head) is trained further, typically with a lower learning rate. This allows the model to gently adjust the more specialized features learned by these top layers to better suit the nuances of the target dataset. Earlier layers, which learn more generic features, often remain frozen. In this research, the top layers of the DenseNet121 base were unfrozen after an initial head-training phase.

- **Full Fine-Tuning (Training the Entire Network):** The entirety of the pre-trained model's layers, encompassing the base and the classifier head, are rendered trainable and undergo update procedures during the training phase on the novel task. The efficacy of this approach is contingent upon the size of the target dataset, which must be comparable to the pre-training dataset, or alternatively, an initial phase of partial fine-tuning may be necessary. This approach confers the greatest flexibility to the model, yet it carries the highest risk of overfitting if the target dataset is limited and necessitates meticulous learning rate management.

The strategy to be chosen depends on different factors such as the size of the target dataset, its imilarity ot the dataset on which the model was pre-trained, and the available computational resources.

3.2.2 Advantages of Transfer Learning in Medical Imaging

The use of transfer learning, particularly with pre-trained CNNs, offers several compelling advantages in the domain of medical imaging, primarily centered around efficiency in development and enhanced model performance.

Transfer learning significantly addresses the challenge of data scarcity, common in medical imaging. Pre-trained models, having learned general visual representations from large datasets like ImageNet, reduce the need for extensive task-specific medical data. This, coupled with starting from well-initialized weights that already capture useful features, leads to faster convergence during training. Consequently, both computational resources and development time can be substantially saved.

Additionally, the use of pre-trained models often translates to improved predictive capabilities and robustness. They provide a superior initialization point in the weight space, frequently resulting in higher accuracy and better generalization on

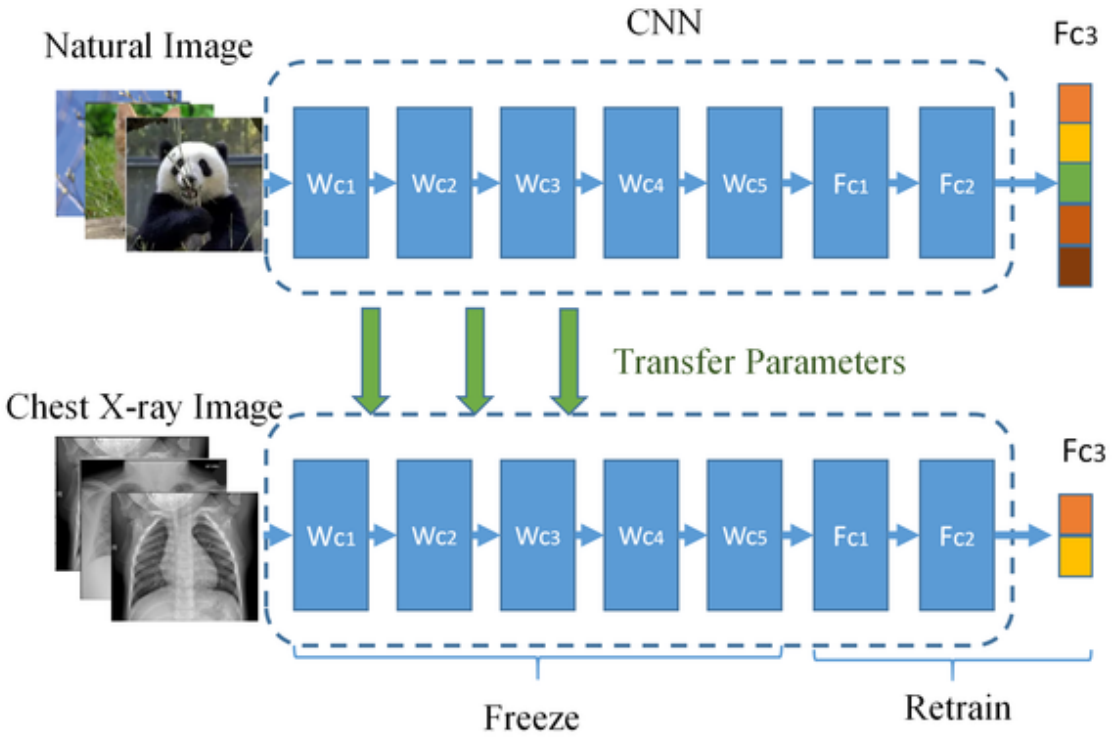

Figure 3.1: Flowchart of deep learning models for transfer learning [WMZ+20]

the target medical task, especially when dealing with limited datasets. The generalized features learned from diverse source datasets like ImageNet contribute to more reliable predictions on unseen medical images. Furthermore, by leveraging features learned from a large source dataset, transfer learning acts as a form of regularization, mitigating the risk of overfitting, which is a critical concern when training deep networks on smaller, specialized medical datasets.

By applying transfer learning to several contemporary CNN architectures, including DenseNet121, EfficientNetB0, ResNet50, and MobileNetV2, this research aims to leverage these advantages for the multi-label classification of pathologies on the ChestX-ray14 dataset. The process of fine-tuning refines the general visual features learned from ImageNet, adapting them to effectively identify and distinguish between the 14 thoracic pathologies. Furthermore, this thesis demonstrates how a pre-trained backbone serves as a powerful foundation not only for single-task classification but also for more advanced, multi-task learning models that simultaneously perform classification and localization, showcasing the versatility and practical utility of transfer learning in developing effective and interpretable diagnostic support tools.

3.3 Explainable AI (XAI) in Medical Imaging

While deep learning models, especially Convolutional Neural Networks, have exhibited remarkable performance in medical image analysis, their inherent complexity often leads to them being characterized as ”black boxes.” The decision-making process within these networks can be opaque, which can make it difficult for clinicians to understand why a model arrived at a particular prediction. In critical domains like healthcare, where diagnostic decisions have direct patient impact, this lack of transparency can hinder trust and adoption. Explainable AI, or XAI, is a set of techniques and methods that make it easier to understand how AI systems make predictions and the internal workings of AI systems, especially complex ones like DNNs, more understandable to humans.

Explanations derived from XAI build trust and confidence among clinicians by allowing them to verify that an AI model's decisions are grounded in medically relevant image features. This increased confidence is essential for facilitating the clinical integration of AI systems, transforming them from ”black boxes” into understandable decision support tools. Furthermore, XAI is invaluable for debugging and model improvement, as it can reveal whether errors stem from focusing on artifacts or overlooking critical pathological signs, thus guiding further development. In some instances, by highlighting which features a model deems important, XAI may even lead to the discovery of new medical insights or biomarkers previously unrecognized by human experts. Lastly, in an era of increasing AI prevalence, XAI helps meet crucial ethical and regulatory compliance requirements for accountability and transparency.

An important category of CNN's XAI technology is a visualization technique that highlights the areas of the input image that most significantly impact the decision-making of a particular class of a model. Grad-CAM is a widely used method in this category and is employed in this thesis. It employs the target class score gradients to flow to the CNN's last convolutional layer, thereby generating a coarse localization map. The map illustrates the areas of the input image that are considered salient for predicting the class. The calculation entails the weighted sum of the feature map in the selected convolutional layer, with the weight being derived from the gradient. The resulting heat map can be overlaid on the input image to provide a visual explanation of where the model ”sees” when predicting.

In the context of this research, Grad-CAM is utilized to visualize the regions in chest X-rays that the fine-tuned classification models deem indicative of specific pathologies. This aids in interpreting the models' predictions and provides a mechanism for qualitatively assessing whether the models have learned clinically relevant features. Furthermore, as will be explored in the evaluation chapter, these visual

explanations can be compared against ground-truth localization data to more rigorously assess the faithfulness of the model's attention. This thesis also explores a more advanced form of interpretability by developing a multi-task model that is explicitly trained to perform object detection. In this case, explainability goes beyond Grad-CAM's implicit localization to the explicit localization of predicted bounding boxes, offering a direct, quantitative way to assess where the model identifies pathologies. Understanding and comparing these forms of explainability is crucial to developing reliable, trustworthy AI diagnostic tools.

Chapter 4

Research Approach and System Implementation

4.1 AI Research Approach and Model Implementation

4.1.1 Backbone Architectures and Transfer Learning

Developing a deep neural network from scratch for medical image analysis is a task that demands vast datasets and significant computational resources. To overcome these challenges, we adopted a Transfer Learning approach. Transfer learning leverages pre-trained models that have already learned rich, hierarchical feature representations, usually from larger datasets with a more general purpose, such as ImageNet. These learned features, which include simple edges and textures as well as more complex details, provide an excellent initialization for domain-specific tasks, significantly reducing training time and improving performance, especially when medical datasets are limited in size.

The core of the AI models developed in this thesis is the backbone, which functions as the primary feature extractor. To identify a robust and effective backbone for chest X-ray analysis, a comparative study was conducted using several prominent CNN architectures, all pre-trained on ImageNet:

- **DenseNet121:** Chosen for its parameter efficiency and feature reuse through dense connectivity, which has shown strong performance on this dataset in seminal works like CheXNet [Raj17].
- **EfficientNetB0:** The baseline model from the EfficientNet family employs compound scaling to achieve a balanced compromise between depth, width, and resolution. This approach has been demonstrated to yield strong accuracy while maintaining computational efficiency and a relatively small number of parameters.

- **ResNet50:** A classic and powerful architecture that introduced residual connections to effectively train very deep networks by mitigating the vanishing gradient problem.
- **MobileNetV2:** A lightweight architecture designed for efficiency, particularly on mobile or resource-constrained devices, which uses depthwise separable convolutions to reduce computational cost.

These architectures serve as the foundation for the two main experimental paths explored in this research: single-task multi-label classification and a more advanced multi-task model for simultaneous classification and localization.

4.1.2 Single-Task Multi-Label Classification

The first experiment aimed to establish a strong performance baseline by implementing and comparing the previously mentioned architectures on the single task of multi-label classification. The implementation was done using TensorFlow/Keras.

Model Architecture and Fine-Tuning Strategy

A consistent architecture was used for each backbone. The pre-trained convolution base was followed by a `GlobalAveragePooling2D` layer that condensed the spatial feature maps into a single feature vector per channel. The vector was then inserted into a final dense (fully connected) layer of 14 output units, corresponding to 14 diseases. A Sigmoid activation applies to each output unit, allowing the model to independently predict the probability of each disease, which is appropriate for multi-label tasks where multiple diseases coexist.

- **Global Average Pooling (GAP):** The final convolutional layer of the backbone produces an output of a set of feature maps with spatial dimensions (e.g., $7 \times 7 \times 1024$. The GlobalAveragePooling2D layer reduces the dimensionality of each feature map to a single number by taking the arithmetic mean of all values. This dramatically reduces the number of parameters, helps prevent overfitting, and makes the model more robust to spatial translations of features in the input image.
- **Intermediate Fully Connected (Dense) Layer:** Then, the feature vector from the GAP layer is transmitted to a Dense layer comprising 512 units, along with a ReLU activation function. This layer functions as a bottleneck, processing increasingly complex, non-linear combinations of the features provided by the backbone.

- **Dropout:** This layer randomly sets a fraction of its input units to zero at each update step, which is a regularization technique to prevent co-adaptation of neurons and mitigate overfitting.
- **Output Layer:** The final Dense layer has 14 output units, one for each pathology. Each unit has a Sigmoid activation function applied. Unlike Softmax, which creates a probability distribution that sums to one, which is suitable for single-label classification, the Sigmoid function outputs a value between 0 and 1 for each class independently. This is ideal for multi-label classification, where an image can have zero, one, or multiple pathologies simultaneously.

Fine-Tuning Strategy

1. **Initial Head Training:** The weights of the pre-trained backbone were initially frozen. Only the newly added classification head (GAP and Dense layers) was trained. This allows the new, randomly initialized layers to learn to interpret the powerful features from the backbone without corrupting the pre-trained weights with large, erratic gradients.
2. **Full Model Fine-Tuning:** Once the head converged, the upper 20 layers of the backbone were unfrozen. Subsequently, the entire model underwent additional training at a significantly reduced learning rate. This approach enables the model to gradually refine the more specialized, higher-level features of the backbone to align more closely with the particular context of chest X-rays.

Weighted Loss Function

Because of ChestX-ray14's substantial class imbalance, illustrated in 5.1 we had to circumvent the model's tendency to prioritize predictions of pathologies and overlook less common ones. For this task, we integrated a weighted binary cross-entropy loss function. Every class out of the existing 14 was assigned a weight as the ratio of negative samples to positive samples in the training set.

$$weight = \frac{samples_{negative}}{samples_{positive}} \tag{4.1}$$

This results in a significantly higher penalty for misclassifying a rare positive case, thereby forcing the model to focus more on these underrepresented pathologies during training.

4.1.3 Multi-Task Model for Classification and Localization

To explore methods that could further improve classification performance and provide more direct explainability, a more complex multi-task learning model was developed in PyTorch. The hypothesis was that forcing the model to learn not only what pathology is present but also where it is located would lead to more robust feature learning in the backbone, thereby benefiting the primary classification task.

Multi-Task Architecture

This model uses a single, shared DenseNet121 backbone for feature extraction. The output features from this backbone are then fed into two separate heads, as illustrated in Figure 4.1.

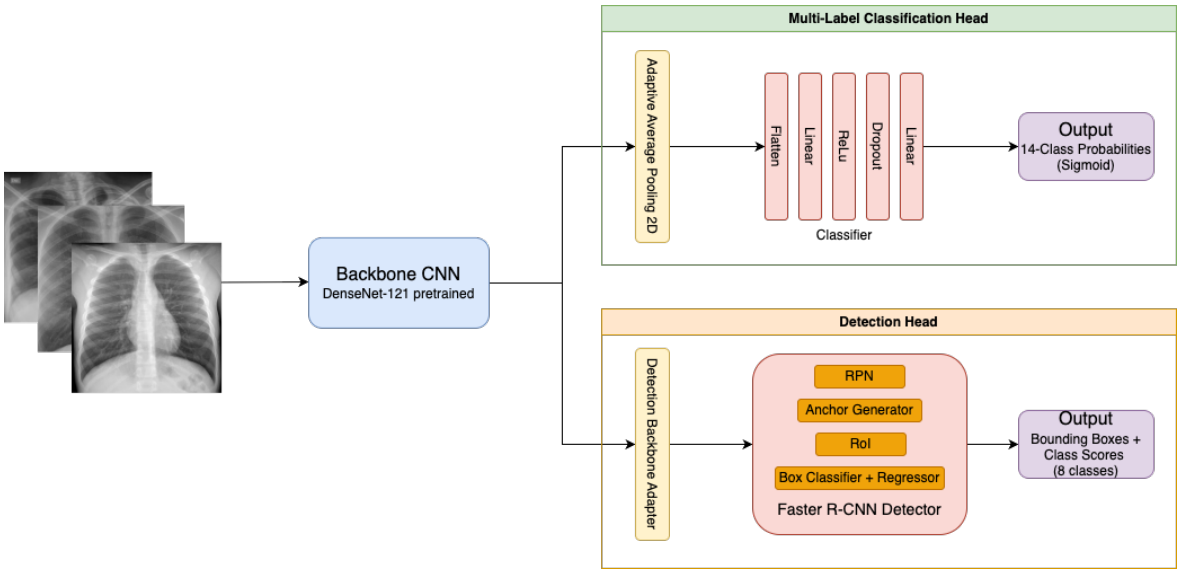

Figure 4.1: High-level architecture of the multi-task model, showing the shared DenseNet121 backbone feeding into both a classification head and a Faster R-CNN detection head.

1. **Classification Head:** This head is designed to process the high-level features from the shared backbone and produce the final pathology classifications. It begins with the average pool layer reducing the spatial dimensions of each characteristic map, effectively creating a characteristic vector. This vector is then flattened and transmitted through a sequential block containing:
    - A Linear (fully connected) layer that reduces the feature dimension from 1024 (the output of the DenseNet121 backbone) to an intermediate size of 512.
    - A ReLU (Rectified Linear Unit) activation function, that is used to add non-linearity by outputting the input if it is positive and zero otherwise.

- A Dropout layer, which reduces overfitting by randomly dropping units during the training phase to ensure that the model is well generalized to unseen data. All models use a dropout probability of 0.3.
- A final Linear output layer that maps the 512 intermediate features to the 14 target classes, producing the raw logits for each pathology.

2. **Localization Head:** This head is a complete Faster R-CNN detector [RHGS16], a foundational and widely recognized two-stage object detection architecture. It includes a Region Proposal Network (RPN) to identify potential object locations and a box head (using MultiScaleRoIAlign) to refine these proposals and predict a final bounding box and class label for each detected pathology. This head was trained on the subset of data with available bounding box annotations.

Composite Loss Function

The training of this multi-task model is supervised by a composite loss function, which is a sum of the losses from both heads:

$$L_{total} = L_{classification} + \lambda L_{detection} \tag{4.2}$$

where $L_{\text{classification}}$ is the BCEWithLogitsLoss for the classification task, and $L_{\text{detection}}$ is the combined loss from the Faster R-CNN head, which includes losses for region proposal, objectness, class prediction, and box regression. The hyperparameter λ, implicitly set to 1 in this implementation, balances the contribution of each task. By optimizing this combined loss, the shared backbone learns to produce features that are useful for both tasks simultaneously.

4.1.4 Implementation for Application Integration

For integration into the final web application, the trained models required a specific deployment pipeline.

- **Model Conversion:** Using the TensorFlow.js command-line tool, the Keras model (`.h5`) was converted to TensorFlow.js graphic model format. This generates a topology file of model.json and a set of binary weight files optimized for delivery to the browser.
- **Client-Side Explainability:** Grad-CAM algorithm has been reimplemented in JavaScript using TensorFlow.js. This allows the creation of visual heat maps directly in the client browser without the need to calculate on the server side, which further enhances data privacy.

4.2 Web Application System Design and Implementation

4.2.1 Application Goals and Use Cases

The main goal of the suggested system is to provide clinicians with an intelligent tool to help them diagnose thoracic pathologies from chest X-ray images (CXRs), putting emphasis on the model's interpretability and offering a robust platform for patient and case management. The system is designed to enhance diagnostic workflows and support potential data curation for future research.

The primary actors and their interactions with the system are

- **Doctor (Clinician):** The primary user of the system responsible for patient management, uploading CXRs for AI analysis, reviewing AI-generated predictions, annotating images with text and bounding boxes, generating reports, and managing case histories.
- **Patient (Indirect Actor):** Receives diagnostic information and care facilitated by the system but does not directly interact with it in the current scope.

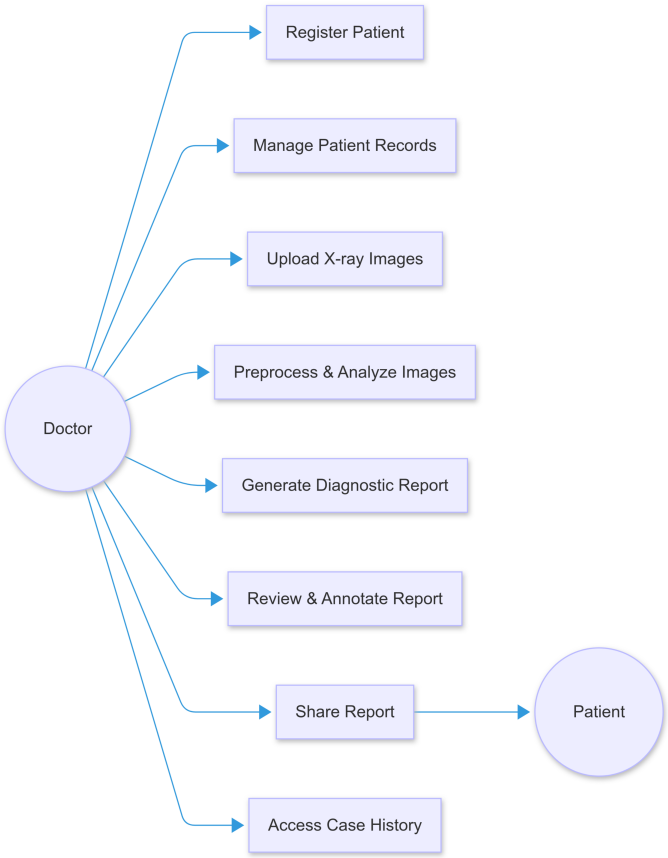

Figure 4.2: Use case diagram illustrating interactions between actors and system functionalities

Key System Functionalities: The system is conceptualized with two primary functional pillars: a core AI engine and a web application that integrates it.

AI-Driven Diagnostic Engine

The core of the system is an AI diagnostic engine responsible for automated analysis of chest X-ray images. This functionality includes the preprocessing of uploaded images and multi-label classification using fine-tuned CNN models, including DenseNet121, ResNet50, EfficientNet, and MobileNetV2 to predict 14 common thoracic pathologies. An important component of this functionality is the generation of Grad-CAM heatmaps to provide better visual explanations of the model's decisions, essential for clinical interpretability.

User Authentication and Management

Secure access to the application is managed through a robust user authentication system. Doctors must be able to register and log in to the application. This functionality is essential for ensuring data privacy, system security, and implementing role-based access control, where users have permissions appropriate to their clinical or administrative roles.

Patient and Case Management

Clinicians need to have the capability to oversee an extensive digital record

for every patient. This means performing thorough CRUD (Create, Read, Update, Delete) operations on patient profiles that encompass personal information and medical history. The system also lets you make and maintain separate diagnostic cases for each patient. This lets you organize various X-ray tests over time and gives you access to a full case history. This is a core feature for maintaining continuity of care and effective patient management.

AI-Assisted Diagnosis and Interactive Annotation The program prominently features an integrated case dashboard that facilitates physician interaction with patient data and AI insights. Doctors can upload a CXR image within a patient's case, which triggers the AI engine for analysis. Then, the system shows the AI-generated predictions right next to the X-ray. Clinicians can use the dashboard to add their own text comments and diagnostic summaries, as well as create precise bounding boxes on the image to mark areas of interest. This feature is very important for checking the results of AI and adding expert input to the diagnostic record.

Reporting and Dataset Curation Support The system allows clinicians to create and download detailed diagnostic reports that encapsulate a case by combining the CXR image, AI predictions, and their own notes. A notable aspect is the system's ability to facilitate dataset curation. Users can export case information, including anonymized identifiers, diagnoses, and bounding box coordinates, in a well-structured CSV format. This capability is essential for generating new, high-quality datasets for upcoming AI model training and research.

Home Dashboard and Statistics

Upon login, clinicians are presented with a home dashboard that provides an overview of clinical activity. This dashboard displays statistics such as the most frequently recorded diagnoses, lists of recent cases, and recently accessed patients. This feature aims to improve workflow efficiency and provide clinicians with immediate insights into their recent workload and diagnostic trends.

4.3 System Design and Technical Specifications

The architecture of the system utilizes a client-server model, yet it features a notable difference: AI model inference takes place on the client side, right in the user's web browser. This method improves data privacy and lightens the computational burden on the server. The complete architecture consists of a client-side front-end application that handles both the user interface and AI processing, along with a server-side back-end focused on data storage and user authentication.

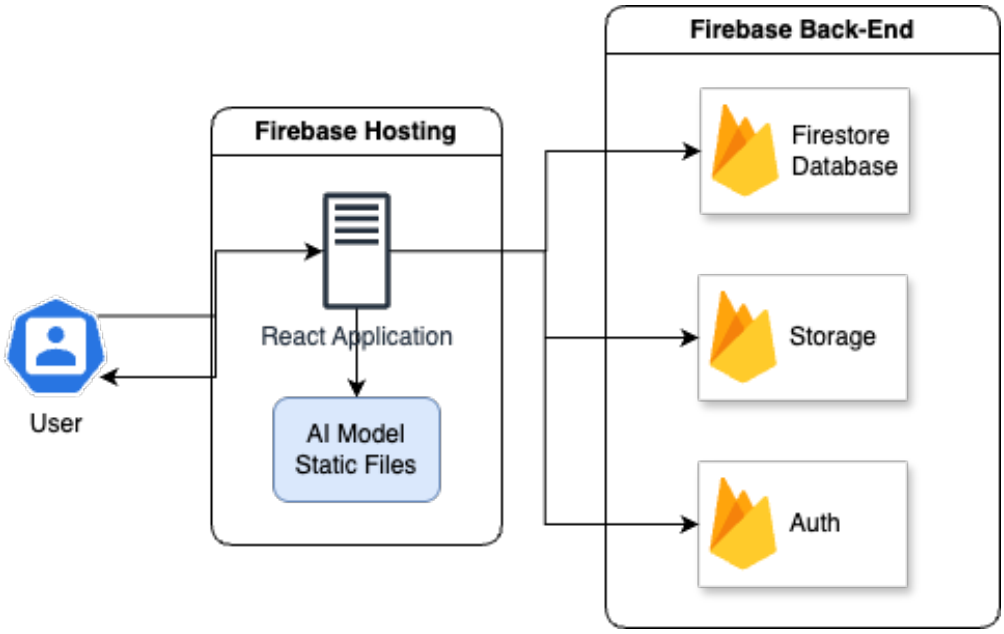

Figure 4.3: High-level system architecture illustrating the client-side AI model within the front-end, and the back-end for data services.

4.3.1 Front-End (View, Controller, and AI Inference)

The Front-End application is providing an intuitive user interface and handling all AI-related computations. It was developed using React, a JavaScript library popular for its component-based architecture and efficient rendering. The application manages user authentication views, a main dashboard with statistics, patient management forms, and a detailed case dashboard. This central view integrates an interactive X-ray viewer with tools for textual and bounding box annotations, report downloads, and data export. The AI model inference is executed directly within this front-end, making it a more cost-efficient solution. The overall user experience design prioritizes a clear, navigable workflow to efficiently present complex clinical and AI-generated information.

4.3.2 Back-End (Data Persistence and Authentication)

The back-end framework is established on the Firebase platform, serving exclusively as a back-end as a service (BaaS) for handling data and authentication, without engaging in any AI processing. Firebase Authentication is utilized to provide secure user registration, logging in, and session management. Data retention is accomplished through Cloud Firestore, a NoSQL database designed for storing structured data such as patient profiles, case information, annotations, and metadata related to AI outcomes. Large media files, including chest X-ray images, are stored in Firebase Cloud Storage. Interaction with these services is conducted directly from

the React client using the Firebase SDKs, all secured by Firebase's rule-based security framework.

4.3.3 AI Diagnostic Service

The AI Diagnostic Service is operating entirely on the client-side using TensorFlow.js. Several CNN architectures (DenseNet121, EfficientNet, ResNet50, MobileNetV2) were trained in Python and subsequently converted to the TensorFlow.js Graph Model format (`model.json` and sharded weight files). The application integrates one of these converted models. It takes an HTML image element as input, which is processed in the browser into a 224×224 tensor. The resulting output provides probability scores for the 14 specified pathologies. To enhance interpretability, Grad-CAM visualizations are also produced on the client side. This approach to client-side deployment offers notable benefits regarding data privacy, as sensitive patient X-rays are not required to be routed to a server for processing, and it reduces computational demands on the back-end. However, it depends on the performance of the user's device for speed during inference and requires an initial download of the model files.

4.4 Implementation

4.4.1 AI Diagnostic Module Implementation

- **Dataset and Preprocessing:** To maintain the reliability and reproducibility of the outcomes, we adhered strictly to the patient-wise data divisions supplied by the National Institutes of Health (NIH). We utilized the train set to distinguish between the training and validation sets, and the designated held-out test set was included in the official test set. Data preprocessing procedures were unified across all the experiments. We established an effective data loading pipeline utilizing TensorFlow's `tf.data` API, which conducted image resizing to 224×224 pixels, normalized pixel values to the $[0, 1]$ range, and implemented batching. To improve the generalization ability of the model over the training data, we put the images through a series of augmentations that consisted of a variety of random transformations, such as rotations, horizontal flips, shifts, and zooms, all executed within the `tf.data` pipeline to ensure optimal performance.

- **Model Training:** A transfer learning approach was employed across multiple pre-trained models, such as DenseNet121, EfficientNet, ResNet50, and MobileNetV2. The implementation adhered to a two-stage fine-tuning procedure.

Initially, the convolutional base from the ImageNet pre-trained model was kept static, while a new classification layer comprising a `GlobalAveragePooling2D` layer followed by a Dense layer coupled with a `Sigmoid` activation function was added. This stage utilized the Adam optimizer. Subsequently, the upper 20 layers of the base model were made trainable, allowing the model to be fine-tuned, thereby adapting pre-learned features to the specificities of the CXR field.

To tackle the notable class imbalance present in the ChestX-ray14 dataset, it was vital to use a weighted loss function in the training methodology. Each of the 14 pathologies had a class weight determined based on the inverse frequency of that class within the training data. This approach imparts a higher penalty for incorrectly classifying less common pathologies. This was realized through a weighted loss function, applied during both stages of training.

- **Model Conversion for Client-Side Deployment:** For seamless integration into the web application, the trained Keras model was transformed into the TensorFlow.js Graph Model format by employing the TensorFlow.js converter tool. This transformation results in a `model.json` file and binary weight shards, optimized for fast loading and execution in the browser. Lastly, Gradient-weighted Class Activation Mapping (Grad-CAM) was adopted both in Python for model assessment and in TypeScript with TensorFlow.js for the client-side application. This method targets the final convolutional block of the respective model architecture to generate heatmaps that visually elucidate the model's forecasts.

4.4.2 Implementation of the Patient Management and AI-Assisted Diagnostic System

- **Front-End (React & TensorFlow.js):**
  - **Component Structure:** Developed using functional components and hooks. Key views include Login/Register, Main Dashboard, Patient List, and the central Case Dashboard.
  - **AI Model Service:** A TypeScript service class was created specifically to oversee the lifecycle of the AI model on the client side. It manages the asynchronous loading of the TensorFlow.js graph model from its specified location, ensuring the model is loaded just once. The service provides methods to transform an HTML image element into a tensor and to execute inference. Additionally, it processes the prediction tensor to return a

ranked list of the most likely pathologies with probabilities exceeding a certain threshold (e.g., 0.5).

- **X-ray Viewer and Annotation:** Developed an interactive X-ray viewer within the Case Dashboard that enables the display of the X-ray, overlays AI-generated Grad-CAM heatmaps, allows users to draw bounding boxes on a canvas, and facilitates the input of textual notes.

- **Back-End (Firebase):**
  - **Data Modeling:** The application's back-end leverages Firebase purely for data persistence and authentication. The data model is structured within Cloud Firestore, a NoSQL database, using several key collections: `users` for clinician profiles, `patients` for demographic and medical history, and `cases` to link everything together. Each `case` document stores case-specific information, a reference to the corresponding image file in Cloud Storage, and all clinician-provided annotations. These annotations include an array of `Note` objects for textual entries and an array of `Diagnosis` objects, where each diagnosis can contain a bounding box object with its coordinates. The relationship between these entities is illustrated in Figure 4.4.
  - **Security Rules:** Firestore and Cloud Storage security rules are critical in this architecture. They are configured to ensure that authenticated users can only access and modify the data, ensuring data integrity and privacy at the database level.

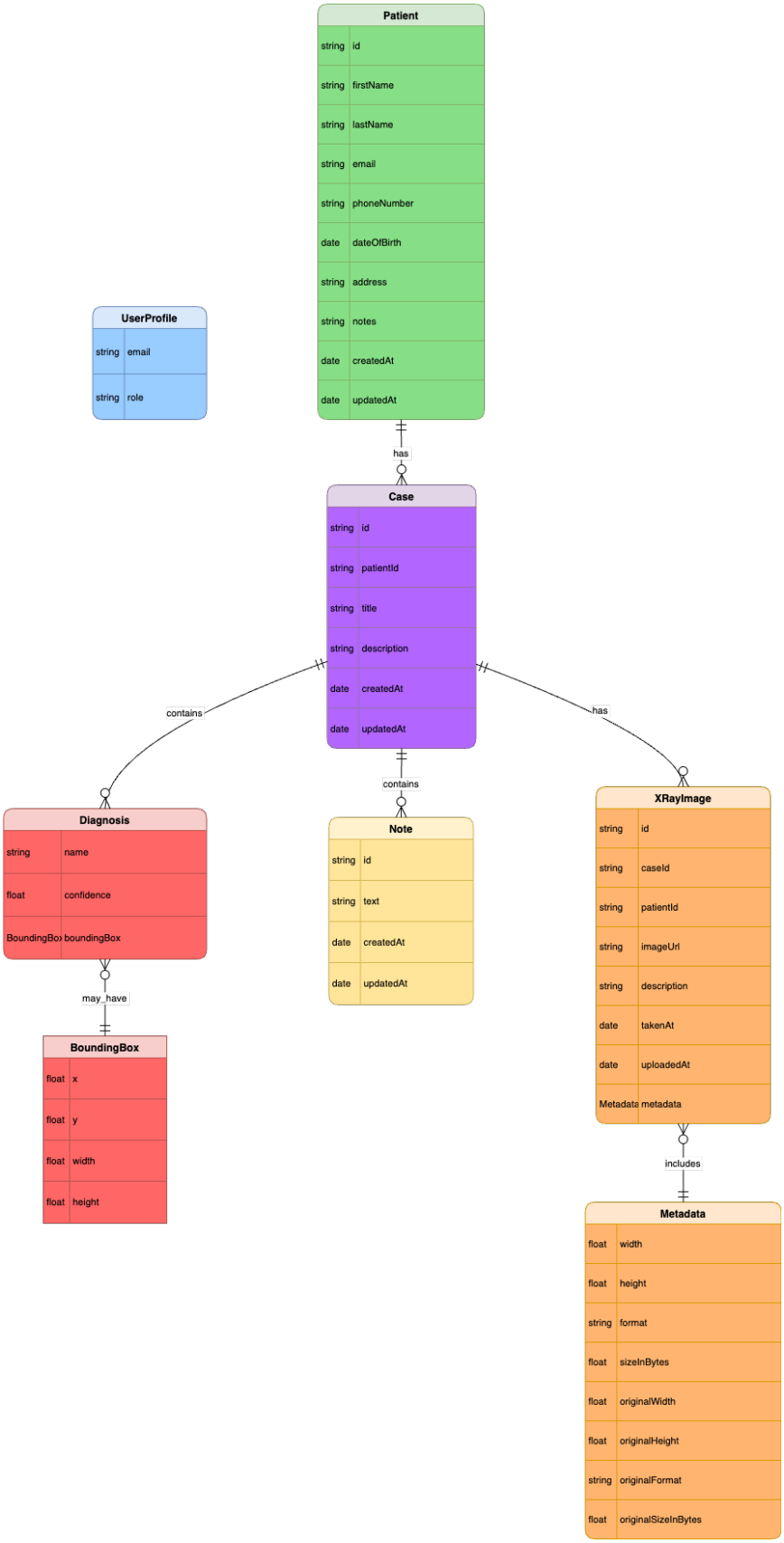

Figure 4.4: Entity-Relationship Diagram (ERD) of the application's data model in Cloud Firestore.

- **Client-Side AI Integration and Data Curation:**

  The client-side AI model integration follows a workflow designed for privacy and efficiency. When a doctor uploads a CXR, the image file is sent directly from the client to Firebase Cloud Storage, and a reference is saved in a new `case` document in Firestore. The browser then loads this image and passes it to the `ModelService` for local preprocessing and inference using TensorFlow.js. The resulting predictions and Grad-CAM visualizations are displayed directly in the React UI without the patient's image ever being processed on a server. The clinician's subsequent annotations are saved back to the `case` document in Firestore.

  This system is specifically designed to facilitate future research by curating datasets. A user-friendly tool was created for the client side that enables authorized users to query Firestore for a selection of cases. This tool formats the relevant clinician-verified diagnoses and bounding box annotations into an organized CSV format, which can be downloaded directly via the browser. This functionality offers an effective way to generate new, high-quality annotated datasets stemming from actual clinical interactions, establishing a valuable feedback loop for the development of future AI models.

4.5 Testing Strategy

4.5.1 Standalone AI Module Validation with a Gradio Prototype

The standalone prototype of the core AI diagnostic pipeline was validated with Gradio, a Python library designed to streamline the creation of interactive web interfaces for machine learning models, before it was integrated into the comprehensive web application. This quick prototyping tool facilitated concentrated testing of the AI model's complete inference and explainability pipeline within a controlled setting.

The primary objectives of this phase were to verify that the Python-based Keras models could correctly load, preprocess images, generate multi-label predictions, and produce plausible Grad-CAM visualizations. Key test cases included:

- **TC1.1: Prediction Generation:** Verifying that the trained models produce probability vectors of length 14 for any valid input X-ray.
- **TC1.2: Grad-CAM Visualization:** Ensuring Grad-CAM heatmaps are generated for specified classes and can be correctly overlaid on input images.

- **TC1.3: Grad-CAM Plausibility (Qualitative):** Visually inspecting Grad-CAM outputs for a subset of test images to assess if highlighted regions correspond to expected pathological areas, especially for high-confidence predictions. This initial check provided confidence before proceeding to more complex implementations.

This prototyping step was crucial for de-risking the project, confirming the core AI logic worked as expected before its conversion to TensorFlow.js and integration into the larger application.

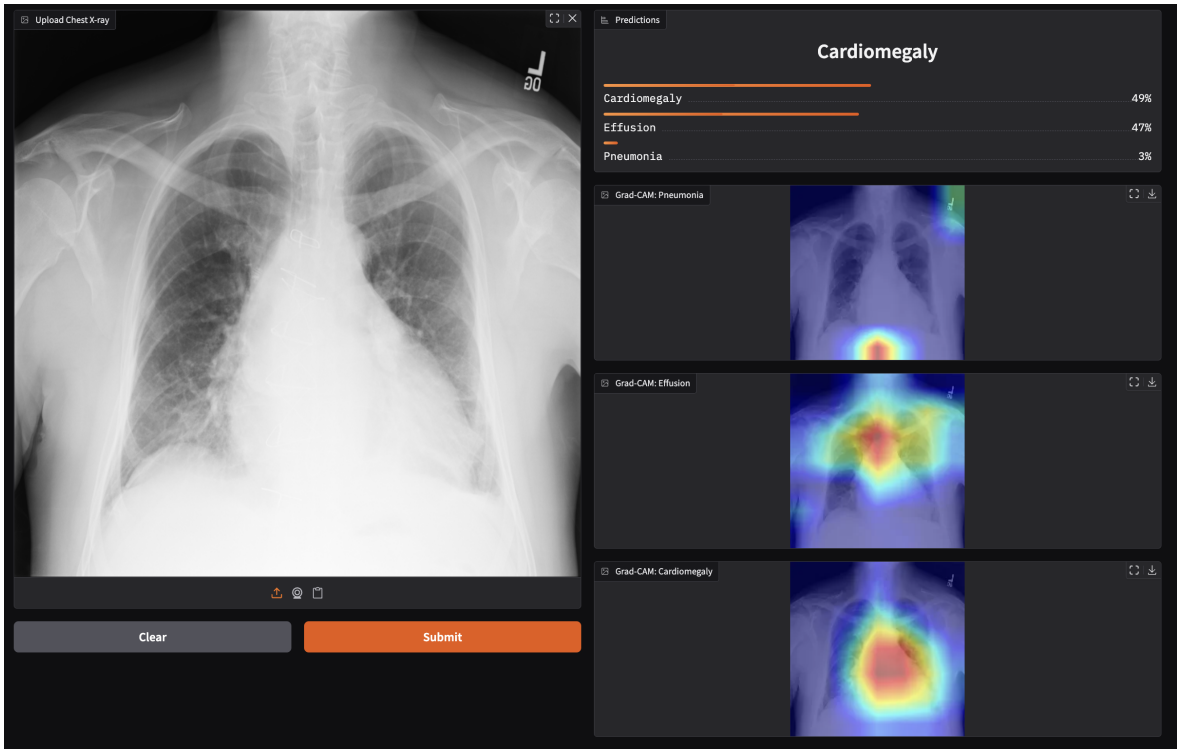

Figure 4.5: Example screenshot of the Gradio application interface displaying an uploaded X-ray, AI predictions, and Grad-CAM visualizations.

4.5.2 Testing the Integrated Patient Management and Diagnostic System

- **TC2.1: User Authentication and Security:** Testing user registration, login, logout, and verifying that Firestore security rules correctly prevent unauthorized data access.

- **TC2.2: Patient and Case CRUD Operations:** Testing the creation, viewing, updating, and deletion of patient profiles and their associated cases.

- **TC2.3: AI Analysis Integration:** Verifying that after X-ray upload, the client-side AI analysis is triggered correctly and the predictions and Grad-CAM visualizations are displayed in the UI.
- **TC2.4: Annotation Functionality:** Testing the drawing, saving, and retrieving of both textual notes and bounding box annotations, ensuring coordinates are stored and rendered correctly.
- **TC2.5: Report and CSV Export:** Testing the generation and download of case reports and the CSV export feature, verifying the correctness of the output format and data.
- **TC2.6: Error Handling:** Testing system behavior when the AI model fails to load, for invalid image uploads, or when there are network issues connecting to Firebase, ensuring error handling.

4.6 Application Functionality and Usage

4.6.1 Mini User Manual for the Application

The web application provides a comprehensive interface for managing patients and leveraging client-side AI for chest X-ray analysis.

1. **Login:** Clinicians access the system using their registered credentials. On first load, the AI model begins downloading in the background.
2. **Dashboard Navigation:** The main dashboard presents an overview of recent activity. Clinicians can navigate to the ”Patients” section.
3. **Managing Patients:** Clinicians can view patient lists, add new patients, or select an existing patient to view their case history.
4. **Creating a Case and AI Analysis:** Within a patient’s profile, a new case is created by uploading a chest X-ray. The image is displayed, and once the client-side AI model is ready, analysis is performed automatically in the browser.
5. **Reviewing Case and Annotating:** The ”Case Dashboard” displays the X-ray, AI-generated probabilities, and Grad-CAM overlays. The clinician can then add textual notes and draw bounding boxes to annotate findings.
6. **Generating Report & Exporting Data:** The clinician can download a report for the current case or use the data export feature to download a CSV of annotated cases for research.

4.6.2 Application Flows

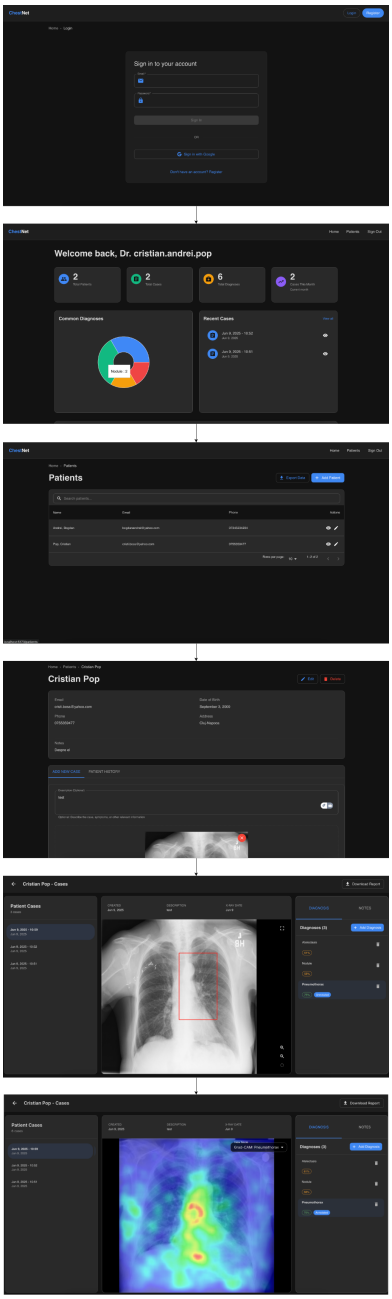

Figure 4.6: User flow for creating and uploading a CXR and creating a new case.

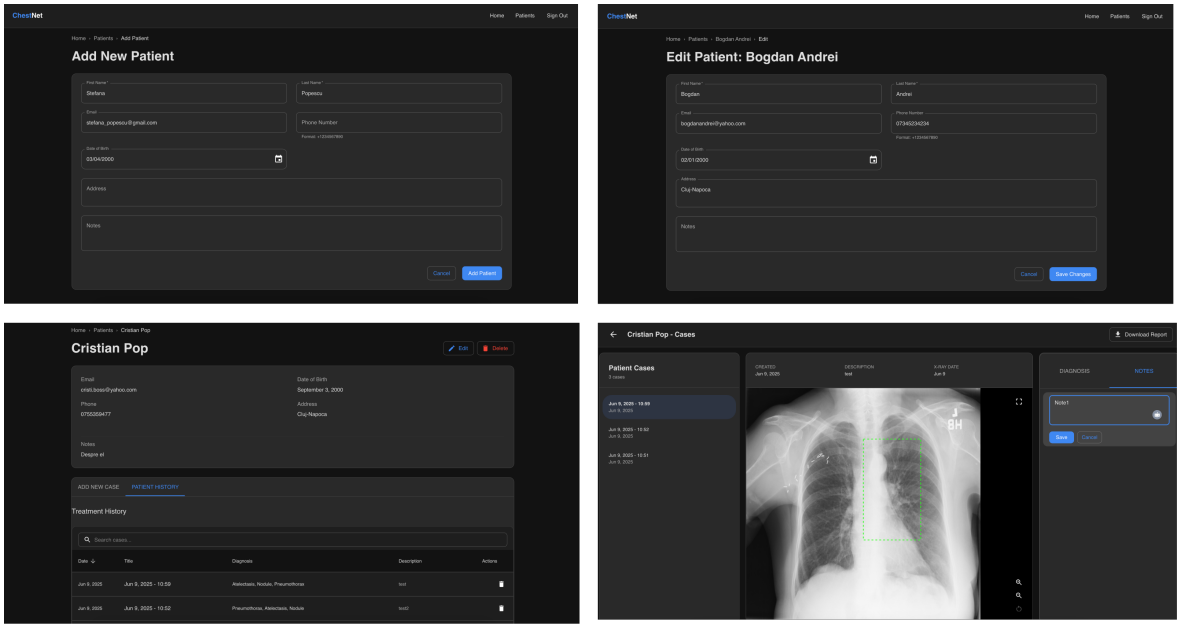

Figure 4.7: Other user operations.

Chapter 5

Experimental Results and Discussion

In this chapter, we detail the results of the experiments conducted for this thesis. It presents the performance of the fine-tuned CNN architectures on the task of multi-label thoracic disease classification. First, the experimental setup, including the dataset and training parameters, is outlined. The core of the chapter focuses on a comparative analysis of several classification-only models, benchmarked against established literature. Following this, results from a more advanced multi-task learning model, designed to perform simultaneous classification and localization, are presented. The model's explainability is critically examined through Gradient-weighted Class Activation Mapping (Grad-CAM) and, where applicable, by comparing model outputs to ground-truth localization data.

5.1 Experimental Setup

5.1.1 Dataset and Data Splits

The primary dataset for all experiments was the NIH ChestX-ray14 dataset [WPL+17]. For consistency and reproducibility, the official patient-wise data splits provided by the NIH were used. The official splits file was used to source images for the training and validation sets. After processing and cleaning (as described in Chapter 4), the final dataset was partitioned into:

- **Training Set:** 28,775 images.
- **Validation Set:** 7,194 images.
- **Test Set:** 15,704 images.

All 14 common thoracic pathologies were used as target labels for the multi-label classification task. For localization validation, the official CSV file for bounding boxes was used, which provides ground-truth bounding box annotations for 8 of

these 14 pathologies on a subset of the images. To better understand the challenge of class imbalance within the dataset, the distribution of the 14 pathologies was analyzed for the training and validation sets, as shown in Figure 5.1 and Figure 5.2

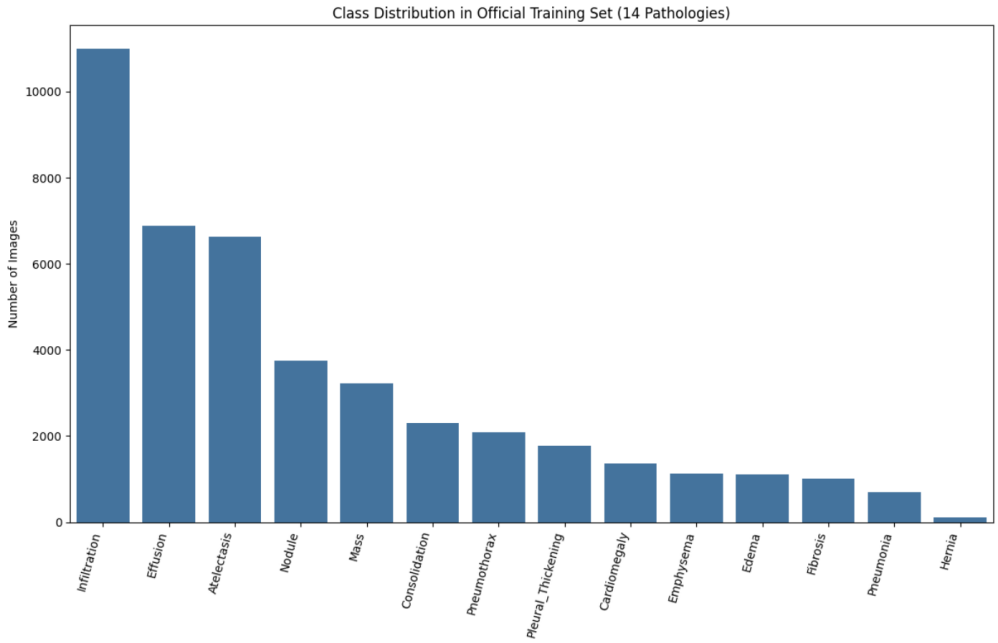

Figure 5.1: Class distribution of the 14 pathologies in the training set (28,775 images).

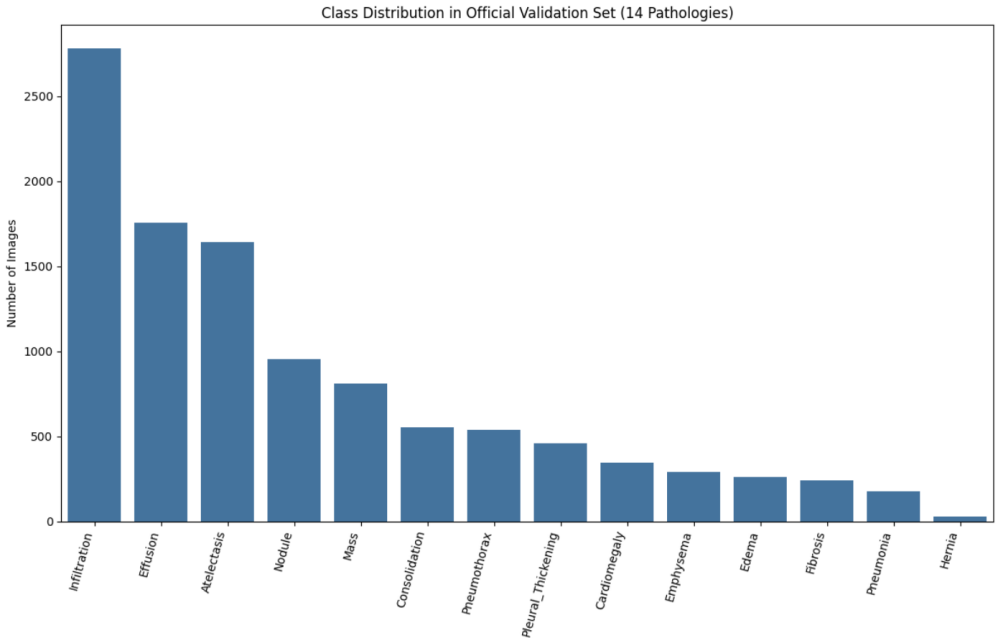

Figure 5.2: Class distribution of the 14 pathologies in the validation set (7,194 images).

5.1.2 Model Configurations and Training

A comparative analysis was conducted on several architectures. The first set of experiments focused on fine-tuning four single-task multi-label classification models trained on ImageNet: DenseNet121, EfficientNetB0, ResNet50, and MobileNetV2. These models were trained using a two-stage fine-tuning strategy with a weighted binary cross-entropy loss function to address class imbalance.

A second, more advanced experiment involved developing a multi-task learning model. This model was designed to perform simultaneous classification and localization. It utilizes a shared DenseNet121 backbone, a classification head for predicting the 14 pathologies, and a detection head that uses the Faster R-CNN architecture. This model was trained with a composite loss function, combining binary cross-entropy for the classification task with the standard RPN and ROI losses for the detection task.

5.1.3 Evaluation Metrics

Model performance was evaluated using the following metrics:

- **Area Under the Receiver Operating Characteristic Curve (AUC):** This fundamental metric evaluates how well the classification model differentiates positive from negative instances at any given decision threshold. The outcomes are presented for each class and as a macro-average.
- **Precision:** This metric was derived from classification reports to evaluate model performance at a specific decision threshold. It is important to note that the commonly used default threshold of 0.5 is often suboptimal, and precision values become most informative after appropriate threshold optimization.
- **Intersection over Union (IoU):** For localization validation, IoU was used to measure the overlap between bounding boxes generated from Grad-CAM heatmaps and the ground-truth bounding boxes.

5.2 Comparative Analysis of Single-Task Classification Models

The test set was used to assess the four classification-only models. A summary of their overall performance, based on the macro-average AUC across all 14 pathologies, is presented in Table 5.1.

Model Architecture	Average Test AUC (14 Classes)
DenseNet121	0.694
EfficientNetB0	0.604
ResNet50	0.638
MobileNetV2	**0.723**

Table 5.1: Comparison of average AUC scores on the test set for the evaluated models.

A more detailed breakdown of the per-pathology AUC scores is provided in Table 5.2. This table highlights the variability in performance across different conditions and models.

Pathology	DenseNet121	EfficientNet	ResNet50	MobileNetV2
Atelectasis	**0.679**	0.583	0.589	0.678
Cardiomegaly	0.767	0.631	0.673	**0.774**
Consolidation	0.642	0.581	0.617	**0.683**
Edema	**0.805**	0.708	0.736	0.780
Effusion	**0.735**	0.607	0.628	0.732
Emphysema	0.751	0.599	0.621	**0.768**
Fibrosis	0.707	0.679	0.733	**0.752**
Hernia	**0.879**	0.724	0.853	0.843
Infiltration	0.627	0.488	0.566	**0.633**
Mass	0.647	0.471	0.591	**0.675**
Nodule	0.666	0.595	0.621	**0.688**
Pleural Thickening	0.633	0.545	0.603	**0.659**
Pneumonia	0.582	0.554	0.545	**0.641**
Pneumothorax	0.733	0.589	0.654	**0.747**
Average	0.694	0.604	0.638	**0.723**

Table 5.2: Per-pathology AUC scores on the test set for each model architecture.

All single-task models were trained until convergence or until the early stopping callback was triggered, showing stable learning patterns. For clarity, the training history for the best-performing model in this category, MobileNetV2, is presented in Figure 5.3. The plots show a consistent decrease in loss and an increase in AUC throughout the fine-tuning phase.

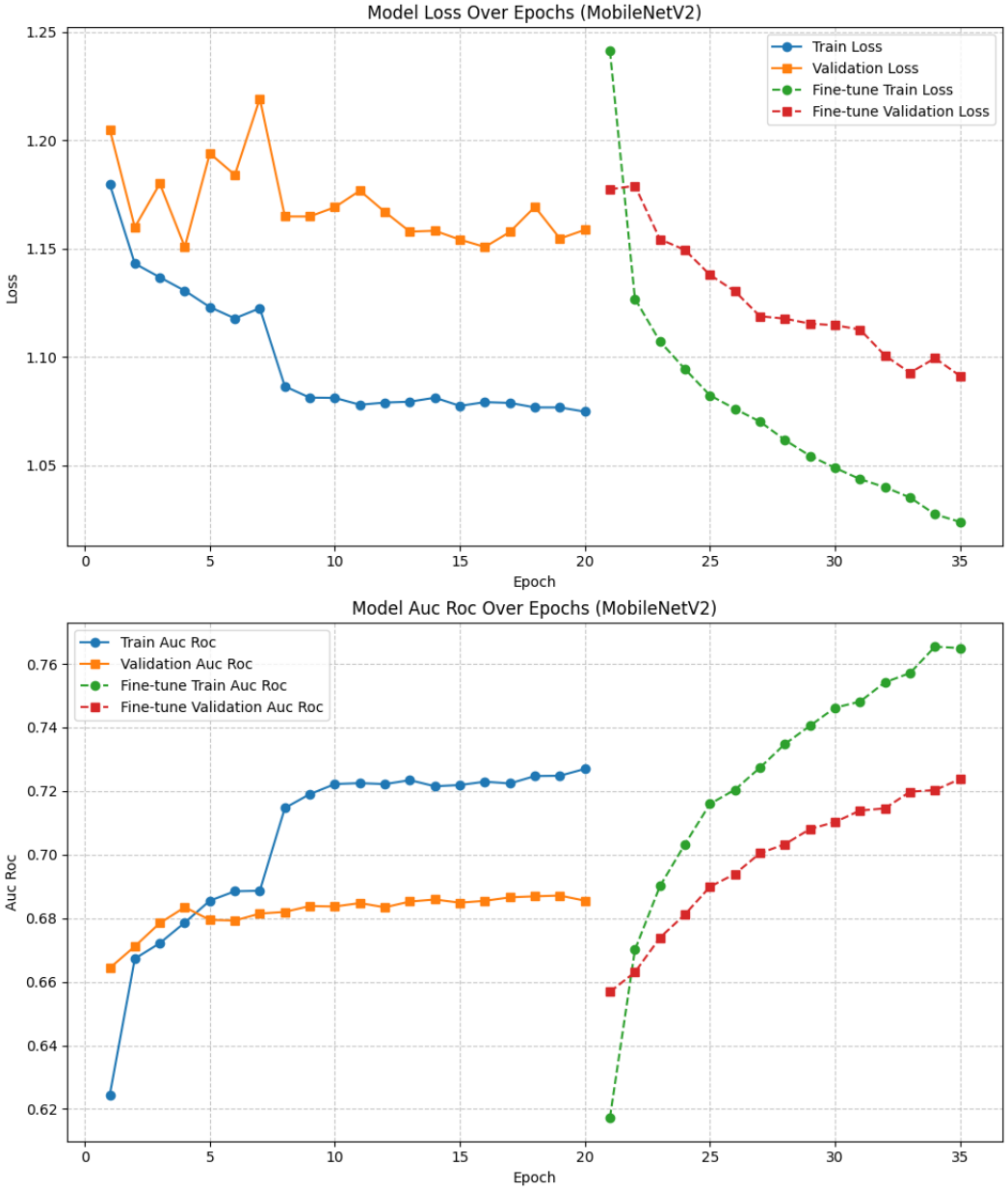

Figure 5.3: Training and validation AUC and loss curves over epochs for the MobileNetV2 model.

5.3 Multi-Task Model for Classification and Localization

To explore a more advanced approach that moves from implicit localization (via Grad-CAM) to explicit localization, a multi-task learning model was developed in PyTorch. This model was designed to simultaneously perform 14-class pathology classification and predict bounding boxes for 8 of those pathologies, leveraging a shared DenseNet121 backbone.

5.3.1 Multi-Task Model Performance

This model was trained for 15 epochs, balancing the classification and detection losses. The evaluation on the test set yielded significantly improved classification results when compared to the single-task, obtaining an average AUC of 0.789. This suggests that the inclusion of the localization task acted as a powerful regularizer, compelling the shared backbone to acquire more robust and generalizable features that enhanced the classification task.

For the detection task, the model achieved a mean Average Precision (mAP) of 0.213. While this score is modest, it indicates a non-trivial localization capability, especially given the limited number of ground-truth bounding boxes available for training. The per-class results, detailed in Table 5.3, show that the model performed particularly well in localizing larger, more distinct findings like Cardiomegaly (Det AP: 0.423).

Pathology	**Classification AUC**	**Classification AP**	**Detection AP (IoU@0.5)**
Atelectasis	0.778	0.526	0.137
Cardiomegaly	0.871	0.421	0.423
Consolidation	0.670	0.162	-
Edema	0.854	0.233	-
Effusion	0.819	0.617	0.254
Emphysema	0.891	0.487	-
Fibrosis	0.778	0.135	-
Hernia	0.880	0.520	-
Infiltration	0.702	0.585	0.221
Mass	0.793	0.461	0.121
Nodule	0.704	0.381	0.000
Pleural_Thickening	0.734	0.174	-
Pneumonia	0.687	0.057	0.317
Pneumothorax	0.823	0.407	0.268
Average	**0.789**	**0.369**	**0.213**

Table 5.3: Per-pathology performance for the PyTorch Multi-Task Model on the test set. Dashes (-) indicate pathologies for which bounding box data was not used for detection training.

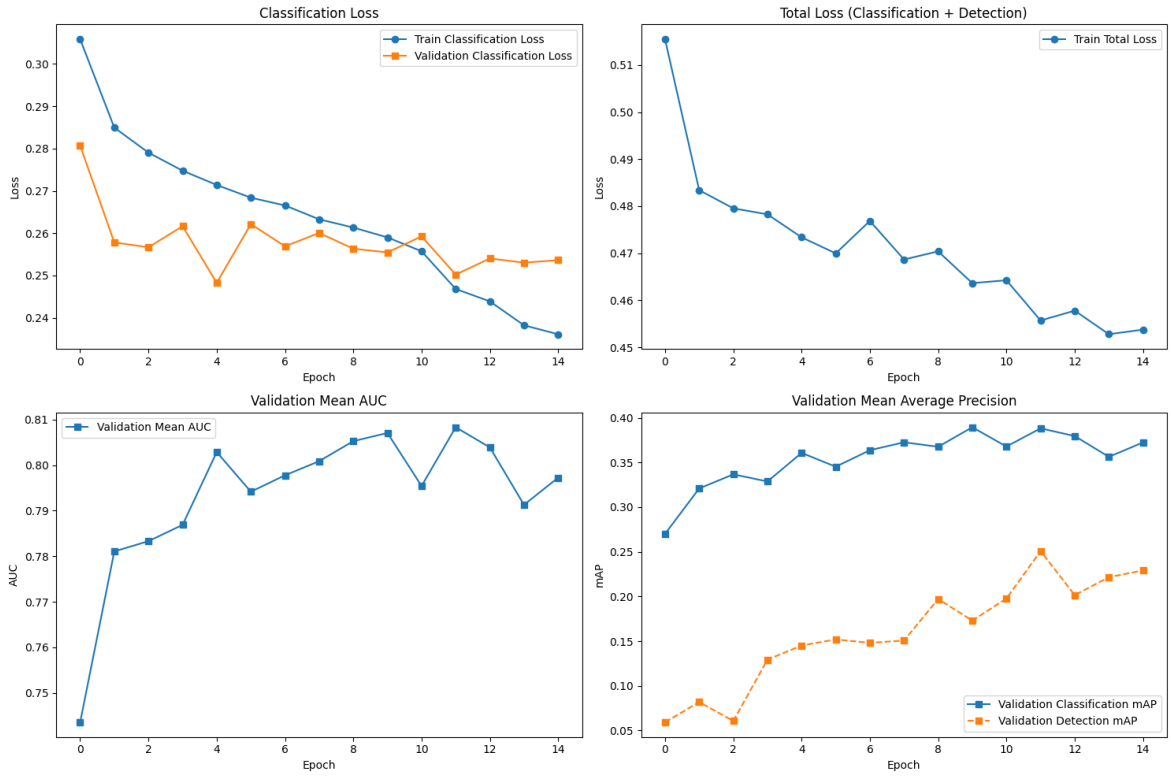

Figure 5.4: Loss and AUC for the Multi-Task model over 15 epochs.

5.4 Explainability and Localization Analysis

A key objective of this thesis was to ensure model interpretability. For the multi-task model, explainability is provided both implicitly by its learned features, which can be visualized with Grad-CAM, and explicitly by its predicted bounding boxes.

To validate the implicit localization capabilities, heatmaps from the multi-task model were compared against ground-truth bounding boxes. Figure 5.5 illustrates this comparison for several pathologies. The algorithmically generated bounding box from the Grad-CAM heatmap (green) is shown alongside the ground-truth box (red). The examples demonstrate a varying but often reasonable degree of alignment, providing qualitative evidence that the actual location of the pathology corresponds to where the model's attention is focused. This alignment supports the use of Grad-CAM as a reliable, albeit coarse, tool for understanding the model's focus.

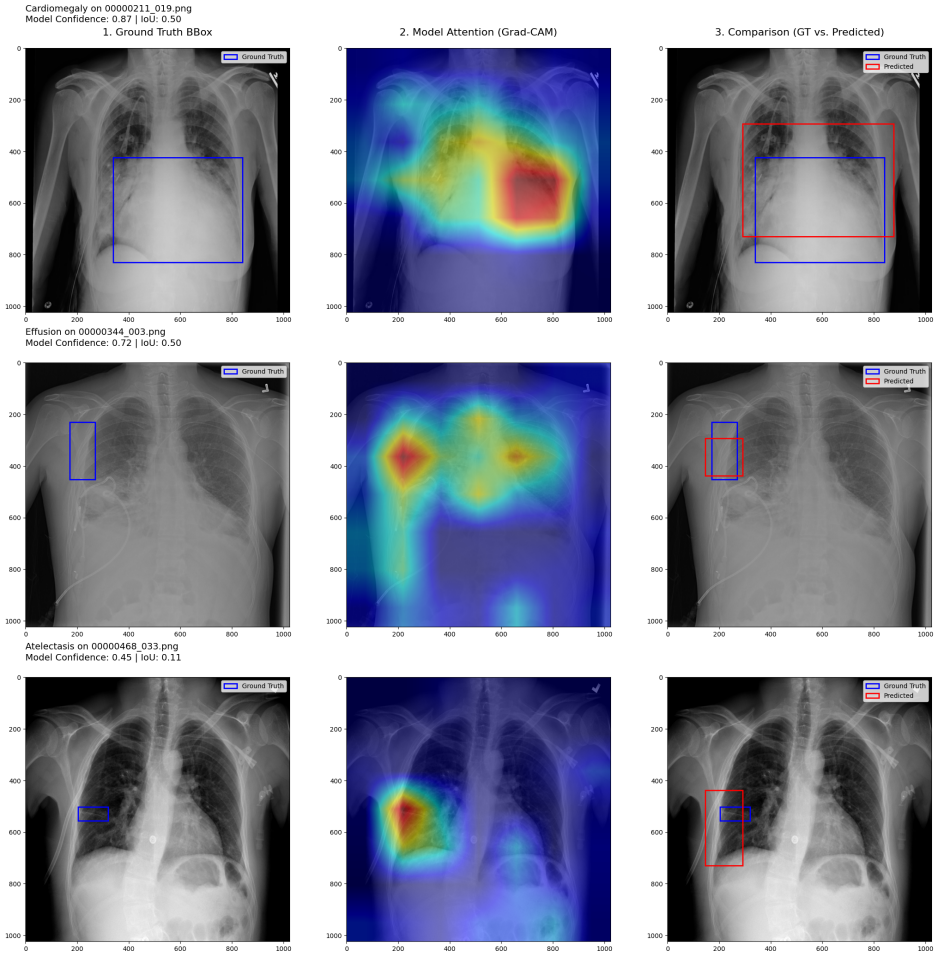

Figure 5.5: Comparison of ground-truth bounding boxes (blue) with generated bounding boxes from Grad-CAM heatmaps (red) for the multi-task model.

5.5 Discussion of Results and Comparison with State-of-the-Art

Several important insights are revealed by the experimental data. The most significant finding is the performance uplift achieved by the multi-task learning model. With a mean AUC of 0.789, it substantially outperformed the best single-task classification model (MobileNetV2, AUC 0.723). This suggests that forcing the network to learn both ”what” (classification) and ”where” (localization) acts as a powerful regularization technique, encouraging the development of more robust feature representations in the shared backbone, which ultimately benefits the classification task.

To contextualize these results, the performance of this study is compared against two notable benchmarks: CheXNet [Raj17], the foundational paper using DenseNet121, and SynthEnsemble [AMAA23], a recent state-of-the-art model, which also utilized a DenseNet121 architecture. It is important to note that direct comparisons can be influenced by minor differences in preprocessing, framework versions, and specific

training hyperparameters.

Pathology	CheXNet	SynthEnsemble	DenseNet Multi-Task	MobileNetV2
Atelectasis	0.810	0.809	0.778	0.678
Cardiomegaly	0.925	0.919	0.871	0.774
Consolidation	0.790	0.815	0.670	0.683
Edema	0.835	0.910	0.854	0.780
Effusion	0.864	0.890	0.819	0.732
Emphysema	0.829	0.929	0.891	0.768
Fibrosis	0.785	0.833	0.778	0.752
Hernia	0.916	0.917	0.880	0.843
Infiltration	0.735	0.741	0.702	0.633
Mass	0.867	0.873	0.793	0.675
Nodule	0.780	0.806	0.770	0.688
Pleural Thickening	0.806	0.813	0.734	0.659
Pneumonia	0.768	0.776	0.687	0.641
Pneumothorax	0.888	0.902	0.824	0.747
Average	0.829	0.854	0.789	0.723

Table 5.4: Comparison of AUC scores for various models on the ChestX-ray14 dataset.

In both experimental setups, the weighted loss function proved crucial. The classification reports show that this strategy successfully prompted the models to recognize and predict underrepresented pathologies, indicated by the high recall scores across most classes. However, this came with the expected trade-off of lower precision. This reinforces the need for per-pathology threshold optimization as a critical post-processing step before any clinical application to find a suitable balance between sensitivity and specificity.

The explainability analysis also yielded different but complementary insights for each approach. For the classification-only models, Grad-CAM provided valuable qualitative evidence that the models were learning clinically relevant spatial features. For the multi-task model, the explainability is more direct: the model's explicit bounding box predictions provide clear, interpretable localization information, moving beyond the coarse heatmaps of Grad-CAM. The modest detection mAP of 0.213 indicates that while the model has learned basic localization, it struggles with precision, which is a common symptom when having a small amount of annotated boxes.

Recognizing the limitations of the developed models in this thesis is important, as the performance of the model is shaped by multiple factors. The primary limitation is the inherent label noise within the ChestX-ray14 dataset, which was derived from radiology reports via NLP and sets a ceiling on achievable performance. Secondly, while several architectures were evaluated, hyperparameter tuning was not exhaustive, and more extensive tuning could yield further improvements. Finally, the detection performance of the multi-task model was constrained by the small

number of available ground-truth bounding boxes, limiting its ability to generalize for the localization task. Despite these limitations, the results demonstrate a successful and methodical approach to developing and evaluating a complex diagnostic AI system.

Chapter 6

Conclusions

This thesis addresses a crucial need for accuracy and transparency in AI-driven medical diagnostics by presenting a thorough method of explainable deep learning for the visualization of diseases from chest X-rays. The research has successfully navigated the complexities of multi-label classification and explainability, culminating in the development of an advanced multi-task learning architecture and a feature-rich web application. The findings and developed tools hold substantial implications for enhancing diagnostic workflows, fostering clinician trust through model transparency, and creating a novel feedback loop for medical dataset curation.

The experimental assessment of the AI models has shown the effectiveness of the suggested approach. The multi-task model, which integrates classification and localization, has improved model interpretability and robustness, providing an enhanced framework for concurrent disease prediction and specific identification of anomalies. This dual-function capability represents a significant advancement in the development of clinically applicable diagnostic tools.

Parallel to the AI research, this thesis has successfully showcased the practical deployment of the trained models within a user-oriented diagnostic support system. Through the implementation of a web application featuring a React-based front-end and Firebase back-end, the system allows comprehensive patient and case management, AI-assisted diagnostic predictions, and image annotations. Client-side inference using TensorFlow.js ensures privacy-preserving diagnosis, aligning the system with the real-world requirements of healthcare data.

Furthermore, by enabling clinicians to export labeled case information, such as diagnostic notes and bounding boxes, into standardized formats appropriate for future model refinement and research, the system facilitates structured data curation. A useful feedback loop for the iterative development of AI models in medical imaging is established by this combination of clinical interaction and dataset generation. The diagnostic support system highlights the potential of the suggested AI pipeline

to improve actual radiological workflows in addition to highlighting its usefulness.

This research highlights its relevance and utility in modern medical practice by providing clinicians with an empowering system that includes interpretable AI tools. By addressing both accuracy and explainability, which are essential for clinical adoption, the proposed solution contributes meaningfully to the field of AI in healthcare.

6.1 Future Improvements

One key limitation of this study lies in the presence of label noise in the publicly available ChestX-ray14 dataset. The labels introduce errors that naturally limit the maximum performance attainable by any model trained on this dataset because they were automatically generated using natural language processing techniques. For this reason, future work could explore fine-tuning with additional datasets such as CheXpert or MIMIC-CXR to improve generalizability across diverse clinical environments. Secondly, while several architectures were evaluated, the hyperparameter tuning was not exhaustive, and further optimization could potentially yield additional performance gains. Finally, incorporating transformer-based or hybrid CNN-transformer architectures may further enhance both classification accuracy and spatial awareness.

In terms of the web application, enhancements could include integrating a reporting system that supports structured outputs in standardized formats (e.g., DICOM SR), improving UI/UX design based on clinician feedback, and implementing a complete administrator dashboard for user management, role control, and system-wide analytics. By deploying the system in a controlled research or clinical environment, the annotation and data export features can be used to curate a new, high-quality dataset with clinician-verified labels and bounding boxes. This new dataset would be invaluable for retraining the multi-task model to improve both its classification and detection accuracy significantly.

This research lays the groundwork for developing a stronger, more scalable, and efficient medical AI system with future enhancements. Continued advancements on the platform hold the promise of significantly aiding radiologists in their diagnostic processes, fostering quicker and more precise patient care.

Bibliography

[AMAA23] S.M. Nabil Ashraf, Md. Adyelullahil Mamun, Hasnat Md. Abdullah, and Md. Golam Rabiul Alam. Synthensemble: A fusion of cnn, vision transformer, and hybrid models for multi-label chest x-ray classification. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, page 16. IEEE, December 2023.

[BPSDLIV20] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[CLY+21] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[DDBS23] Tribikram Dhar, Nilanjan Dey, Surekha Borra, and Robert Sherratt. Challenges of deep learning in medical image analysis -improving explainability and trust. *IEEE Transactions on Technology and Society*, PP:1–1, 03 2023.

[Dos20] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[IRK+19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[KAS+24] Thakur Gopal Kumar, Thakur Abhishek, Kulkarni Shridhar, Khan Naseebia, and Khan Shahnawaz. Deep learning approaches for medical image analysis and diagnosis. *Cureus*, 2024.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LQA20] Wang Linda, Lin Zhong Qiu, and Wong Alexander. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 2020.

[LS17] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.

[LXD+20] Zhang Ling, Wang Xiaosong, Yang Dong, Sanford Thomas, Harmon Stephanie, Turkbey Baris, Wood Bradford J, Roth Holger, Myronenko Adriy, Xu Daguang, and Xu Ziyue. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging. 2020 Jul*, 2020.

[MKRBD23] Kim Minki, Moon Ki-Ryum, and Lee Byoung-Dai. Unsupervised anomaly detection for posteroanterior chest x-rays using multiresolution patch-based self-supervised learning. *Scientific Reports*, 13, 2023.

[MSS+23] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, 4:100216, 2023.

[NLL+20] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020.

[QKH+24] Wang Alan Q, Karaman Batuhan K, Kim Heejong, Rosenthal Jacob, Saluja Rachit, Young Sean I, and Sabuncu Mert R. A framework for interpretability in machine learning for medical imaging. *IEEE*, 2024.

[Raj17] P Rajpurkar. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv abs/1711*, 5225, 2017.

[RHGS16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[RHZ+22] Zakaria Rguibi, Abdelmajid Hajami, Dya Zitouni, Amine Elqaraoui, and Anas Bedraoui. Cxai: Explaining convolutional neural networks for medical imaging diagnostic. *Electronics*, 11(11), 2022.

[SGL+24] Zhichao Sun, Yuliang Gu, Yepeng Liu, Zerui Zhang, Zhou Zhao, and Yongchao Xu. Position-guided prompt learning for anomaly detection in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–577. Springer, 2024.

[SKZ+23] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, 2023.

[WJJ+19] Johnson Alistair E. W., Pollard Tom J., Berkowitz Seth J., Greenbaum Nathaniel R., Lungren Matthew P., Deng Chih-ying, Mark Roger G., and Horng Steven. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 2019.

[WLW20] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020.

[WMZ+20] Dingding Wang, Jiaqing Mo, Gang Zhou, Liang Xu, and Yajun Liu. An efficient mixture of deep and machine learning models for covid-19 diagnosis in chest x-ray images. *PLOS ONE*, 15(11):1–15, 11 2020.

[WPL+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[WPLK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.

[YQW+19] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE international conference on data mining (ICDM)*, pages 728–737. IEEE, 2019.